Text Mining Based on the Prototype Matching Method

Antonina Kloptchenko

Turku Centre for Computer Science

TUCS Dissertations

No 47, December 2003

Text Mining Based on the Prototype Matching Method

Antonina Kloptchenko

Doctoral Dissertation
to be presented, with the permission of the Faculty of Economics and Social Sciences of Åbo Akademi University, for public discussion in Auditorium 3120, Datacity, Åbo, on the seventeen of December, two thousand three, at twelve o'clock

Åbo Akademi University
Institute for Advanced Management Systems Research
Turku Centre for Computer Science
FIN-20520 Åbo, Finland

Åbo, 2003

SUPERVISED BY

Professor Barbro Back
Institute for Advanced Management Systems Research
Department of Information Systems
Åbo Akademi University
Åbo, Finland

REVIEWED BY

Professor Erkki Sutinen
Department of Computer Science
University of Joensuu
Joensuu, Finland

Professor Heikki Topi
Department of Computer Information Systems
Bentley College
Waltham, MA, USA

ISBN 952-12-1257-8

To my parents
Моим родителям

ACKNOWLEDGEMENTS

It has been said that in the pursuit of scientific achievement, especially in the rapidly changing world of IT and IS, PhD students will be constantly buffeted with new technologies and methods such that they would never be able to sit down and write a thesis that is on the cutting edge of their field. Before the ink dries on this dissertation, there will be newer and better methods exploding into this field. PhD students must have the courage, assiduousness and confidence to put all their research together in one place at some point and summarize their achievements. The person who gave me this courage is my supervisor, Barbro Back. She has been instrumental in assisting me with the pursuit for an interesting research question, one that is exceptionally applicable in today's Internet society. She has introduced me to the scientific community of the western world, which is particularly important for me, coming from the former Soviet Union. I sincerely thank her for keeping me motivated and focused.

Sincere gratitude goes out to my reviewers, Professor Heikki Topi and Professor Erki Sutinen, who had the patience and fortitude to read earlier versions of my thesis and provided constructive criticism to help me defend it. Professor Heikki Topi summarized and evaluated my scientific achievements. Professor Erki Sutinen helped me revise my work and gave me a fresh perspective from a computer science point of view. Their guidance not only improved my dissertation but also will benefit my future work. I would like to thank my opponent, Professor Henry Tirri, who accepted the invitation to act as a public challenger and reviewer of my dissertation.

I owe special thanks to the GILTA research group; Professor Ari Visa, Professor Hannu Vanharanta, Jarmo Toivonen, Antti Arpe, Camilla Magnusson, Jonas Karlsson, Tomas Eklund and Tomi Vesanen. Professor Ari Visa undoubtedly deserves the greatest recognition for the completion of my thesis. He headed the research group and offered new perspectives with which to examine and explain text mining problems. Professor Vanharanta gave me endless encouragement. Tomas Eklund offered me tremendous support as a co-author, friend and critic of my research. Jarmo Toivinen was always willing to answer any mathematical questions I had about the prototype matching method. Thank you Camilla Magnusson for exploring the same problem with me from a linguistic viewpoint.

My thanks go to the professors and administrative staff at Turku Center for Computer Science (TUCS) and Institute for Advanced Management Systems Research (IAMSR) at The Abo Akademi. Professor Christer Carlsson deserves many words of gratitude for his valuable input and advice. He provided me with an inspiring work environment and excellent working conditions. Professor Pirkko Walden who generously shared her knowledge experience with me.

Thank you TUCS for the research and travel grants that enabled me to get my papers recognized at many international conferences. The administrative staff helped me resolve many day-to-day difficulties with being in a foreign country. Lastly, my teachers. Professor Leonid Berstein and Professor Irina Tsaturova, at the Taganrog State University of Radio Engineering, in Taganrog, Russia, provided me with a good solid education enabling me to further my studies.

Heartfelt appreciation goes to my friends and fellow colleagues at IAMSR. Their interest in my research and endless discussions helped me to improve my work. Piia Hirkman's unquestionable loyalty, support and friendship made it possible for me to finish my PhD while being overseas. Nikko, Vladimir, Adrian and Minna encouraged me during times of frustration and disappointment. The warmth and friendship of Viktor, Natasha, Katya and Bea made possible to withstand the cold Turku winters. Whenever I had a seemingly unsolvable problem, Viktor always pointed me in the right direction. He was always there for me.

Finally, my closest friends and family all played important roles in my development as a person and scholar. Vika, Oksana, Alexander, Marina, Andrey and Nata encouraged me to continue my education and were always happy for my successes. My husband's parents, Richard and Angela Durfee, supported me while I studied in the United States. Angela helped me stay focused on my work when I became a new mom. My parents made all this possible. My father, Vladimir Klopchenko, being a Physics professor, set high goals for me and drove me to become an intellectual equal. Natalia Klopchenko, my mother, gave me endless valuable advice and help. Her unconditional love, wisdom and encouragement made it easier for me to achieve my goals. My sister Anna always kept my spirits high. My husband Kevin was my in-house spell checker and was my best critic. Last of all, the most joyful addition to my life, my son Aivan, has brought a smile to my face every day.

13 October 2003
Narragansett, RI, USA

TABLE OF CONTENTS:
Part I: Research Summary

TABLE OF CONTENTS:
Part II: Original Research Papers

TABLE OF FIGURES

LIST OF ABBREVIATIONS

KDD - Knowledge Discovery in Databases
DM - Data Mining
TM - Text Mining
IR - Information Retrieval
IT - Information Technology
IS - Information Systems
AI - Artificial Intelligence
GILTA - manaGIng Large Text MAsses
NN - Neural Networks
SOM - Self Organizing Map
MFNN - Multilayerd feedforwards neural network
HICSS - The Hawaiian Conference on System Science
NL - Natural language
TPS - Transaction Processing system
ERP - Enterprise Resource Planning system
CRM - Customer Relationship Management system
MIS - Management Information System
DSS - Decision Support System
WAP - Wireless Application Protocol
GPRS - General Packet Radio Switch
PDA - Personal Digital Assistant
OAS - Office Automation System
KMS - Knowledge Management System
VSM - Vector Space Model
TF - term frequency
IDF - inverse document frequency
MI - Mutual Information

PART I
Research Summary

1. INTRODUCTION

The proliferation of computers and the World Wide Web make access to various data sources very convenient. The availability and accessibility of many up-to-date data sources offers an extensive support for managers in the dynamic, complex, and demanding business environment of today. Modern managers, decision makers and knowledge workers have access to larger amounts of facts than the managers of earlier times. It would seem possible to make optimal business decisions with the support of a variety of data sources. However, the amount of data, with its obscure structure, has swamped managers resulting in loss of time and productivity. Modern information technology (IT) has outpaced human capability to process, utilize, and explore stored data. While computing power doubles every 18 months according to Moore's Law, computers' capacity of digital data storage has doubled twice as fast every nine months (Fayyad and Uthurusamy 2002). Kahle's Internet Archive (www.archive.org) has collected 100 terabytes of data since October 2001, which is about four times greater than the United States Library of Congress, the world's largest library with about 20 to 50 terabytes of data. Analysts estimate that unstructured textual information is doubling in quantity every three months. With technological advances, our technological abilities to collect, generate, distribute, and store data has outgrown our ability to process and understand it. Business units, collecting and storing information by the click of a mouse button, now want to understand the trends lying behind this information quickly. Understanding those trends allows enhancing decision-making within the company and reaching customers more effectively. The rapid rise of information available in electronic format has turned a dream of creating an information-rich society into a nightmare of information overload.

The situation around information and technology proliferation is threefold. First, the technological realm offers managers a wide variety of available data sources, with various data types, about almost everything they want to know. Second, the business realm puts strong pressure on managers for instant and efficient decision making. Third, the capacity of human short term memory to hold and digest information is very limited (Miller 1956; LeCompte 2000). Therefore the utilization of intelligent information technological solutions is required. Although wireless technologies such as WAP-mobile phones, laptops, and PDA devices can deliver vital data for decision-making purposes anywhere, anytime, the utilization of these advances is mediocre. These technological devices do not offer additional value and understanding of data to their users but rather contribute to data multiplication, leading to information overload. Information overload creates data tombs resulting in unavoidable losses and missed opportunities. This environment dictates the strong need for intelligent solutions for data analysis and exploration. The times of "data culture" in business, when accuracy of data collection, consolidation, storage and access were worshiped, are slowly evolving into the time of "knowledge culture" which worships the wise utilization of experiences decoded from data.

1.1 Knowledge Discovery in Databases and Data Mining

Knowledge discovery is a relatively new discipline gaining visibility due to the exponential growth of data collections, and the need to understand and explore them. According to Fayyad (1996), *Knowledge Discovery in Databases* (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. A parsimonious summary of a subset of data is considered to be a pattern (Fayyad, Piatetsky-Shapiro et al. 1996, Fayyad and Uthurusamy 2002). KDD is drawn from various scientific fields: cognitive psychology in addition to database and system science, statistics, artificial intelligence (AI), management information systems, and information retrieval. *Data mining* (DM) is the essential and arduous step in the process of KDD with the goal of extracting high-level knowledge from low-level data. Although some specialists use the terms DM and KDD interchangeably, there are many steps preceding the DM step in the KDD process, which is illustrated in Figure 1-1. These steps include data integration and retrieval from data sources, selecting the appropriate data subset to work with, cleaning the data and dealing with missing data, data transformation, and dimensionality reduction. After the DM step, one should decide whether the extracted information is valuable knowledge by evaluating discovered patterns, representing them via advanced visualization methods, and consolidating them with existing domain knowledge. The evaluation of the value of the discovered pattern, and its applicability to the investigated problem is, however, more of an art combined with common sense than a science.

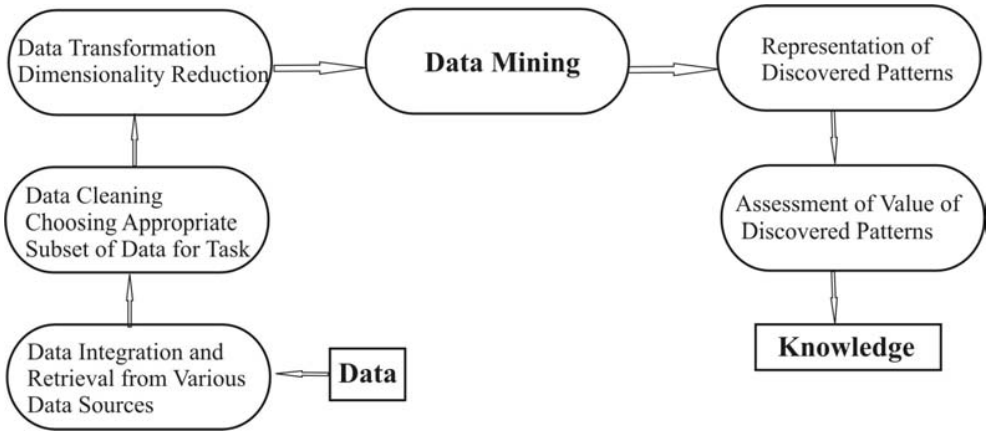


Figure 1-1. The KDD process (adapted from (Fayyad, Piatetsky-Shapiro et al. 1996; Han and Kamber 2001))

DM, which has its roots in various scientific disciplines, from exploratory data analysis in statistics to machine learning in AI, has recently started to play a dominant role in data processing. Statisticians had used the term DM in the 1970s with a negative connotation of "sloppy exploratory analysis" with no prior hypothesis to verify. Lately, however, it has evolved into the most promising

solution for business intelligence, as well as technological, biological, medical, military, and other purposes. DM applications are sets of problems with similar characteristics across different domains, therefore, same algorithms and models that are used for developing fraud-detection capability for credit card transactions can be used to develop health insurance fraud detection applications. One of the most promising new topics in IT in 1996, according to the online journal of Database Management System, DM has now moved beyond the early adopter stage. DM solutions contribute to human understanding of huge masses of data in various representations, including tabular data domains, spatial data domains, text-based domains, and image-based domains. DM in business and business analytics can be used interchangeably to give business users strategic insights from collected massive databases (Kohavi, Rothleder et al. 2002). Also, DM can be used for secondary analysis of data that have been collected for some other purpose (Laurikkala 2001). For example, using term domain distribution analysis for the medical records collected for monitoring the quality of doctors' work, the surprising discovery was made that the primary thoracic lung cancer tumor appears in the right lung more often than in the left lung (with a ratio of 3:2) (Goldman, Chu et al. 1998).

A significant distinction between DM tools and other analytical tools is how they are utilized for exploring data interrelationships (Moxon 1996). While analytical tools support verification-based approaches of intuitively posing queries to the database, DM tools use discovery-based approaches to determine the key relationships in the data by employing different algorithms of pattern matching. For instance, the Knuth-Morris-Pratt matching algorithm finds copies of a given pattern (short string of symbols) as a contiguous subsequence of a larger text (Knuth et al. 1977). The approaches used in DM depend on the goals of the person who is analysing the data and on the types of pattern to be extracted. DM has borrowed a number of algorithms and techniques from its parental fields to accomplish the main tasks of data analysis, classification, clustering, categorization, and prediction. The algorithms differ from each other in the type of data mined, starting from the most trivial numeric data, moving toward transaction data, data streams, graphical and scientific data. Below is the categorization of DM tasks adopted from (Han and Kamber 2001). It not unique, and allows further division into finer tasks:

Association analysis aims at discovering any association relationships or correlations that occur frequently together among a set of items. An example of association is the extraction of association rules from textual databases. *Classification* is the process of finding a set of models or patterns in the training data that describe and distinguish data cases or concepts. Classification constructs a model to predict the class of objects whose class type is known. The derived models may be presented as a set of if-then rules, decision trees, mathematical formulae, or neural networks. *Prediction* is used for predicting possible values of missing data or missing class attributes of data objects. Prediction also encompasses the identification of distribution trends based on the available data. *Clustering analysis* identifies clusters embedded in the data, where a cluster is a collection of data objects that are "similar" to one another. The objects are grouped based on the principle of maximizing intraclass similarity and minimizing interclass similarity.

Clustering can be used to generate labels for classes of data objects, and for taxonomy formation, which is the organization of observations into a hierarchy of classes that group similar events. *Outlier analysis* refers to the analysis of outlier data objects that do not comply with the general behaviour or model of the data. In some applications, such as fraud detection, the rare events can be more interesting than the more regularly occurring ones, and, thus, should not be treated as noise. *Evolution analysis* describes and models regularities or trends for objects whose behavior changes over time. This may include classification, clustering, association, and discrimination of time-related data.

1.2 Text Mining as Data Mining from Textual Databases

Although most of the previous research in DM has been focused on structured data such as relational, tabular or transactional data, in reality, a substantial portion of the available information is stored in text or document databases. This information resides in the company's internal and external documents, technical and financial reports, customer feedback on products, market analyses and overviews, electronic mail and notice board messages, advertisements, managerial notes, business related publications, business plans, correspondence with partners and creditors, competitor releases, news articles, research papers, books, digital libraries, and various company-related web pages. *Text mining* (TM) or *data mining from textual databases*[1] has a definition that almost repeats the most popular definition of DM: TM is an essential part of discovering previously unknown patterns useful for particular purposes from textual databases (Hearst 1999), (Dörre, Gerstl et al. 1999). While the bulk of textual data is stored in file systems, some organizations have begun to store and manage it in relational databases in text-based columns (MicrosoftCorporation 1999). However, there is little similarity between data and text mining in KDD due to the principal differences in the types of data explored. In general, whereas numeric data quantitatively describe some measurements that are related to phenomena interpretation, textual data describe the phenomena qualitatively. While users of numeric data can explore a new, previously unknown pattern in numeric data stored in a database, the users of natural language (NL) text can only associate the discovery of "previously unknown" patterns in textual data with rediscovery, or with a new interpretation of what the author of a text had already written. It is very argumentative to state that an author of a text might not know him what he has stated in a text. A new interpretation of all the facts stated in the text can only appear because different readers understand the same text differently based on their

[1] A database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. The most prevalent type of database is the relational database, a tabular database in which data is defined so that it can be reorganized and accessed in a number of different ways. A distributed database is one that can be dispersed or replicated among different points in a network. An object-oriented programming database is one that is congruent with the data defined in object classes and subclasses. (Glossary for enterprise resource professionals from http://www.searchwebservices.com) A textual database deals with retrieval and manipulation of documents or segments of text. It allows a user to search on-line complete documents or parts of documents rather than attributes of documents (Tseng, Yang et al.1990).

backgrounds. Witten, Bray et al. (1998) argue that TM has potential because one does not have to understand the text in order to extract useful information from it.

1.2.1 Text Mining as a Confluence of Multiple Disciplines

TM has its roots in computational linguistics, NL processing, text analysis, cognitive psychology, information retrieval, machine learning, statistics, and information and library sciences. The confluence of multiple disciplines in the area of TM is presented in Figure 1-2. Some of the parental disciplines, for instance, statistics, AI, and information science, are the same for TM and DM because TM can be seen as a subpart of DM that deals with one specific format of data, namely, text. Moreover, some disciplines are interrelated and, thus, some of them intersect in Figure 1.2. TM borrows methods from AI, statistics, and computational linguistics that are inspired by linguistics, dictated by the needs of information science and limited by constraints of database technology. In this dissertation I use methods from several disciplines: linguistics (collocational networks in Chapter seven), NL processing (a word coding method used in the prototype matching method in Chapter five), AI (quantization approach used in the prototype matching method in Chapter five and throughout the research papers), and visualization (in the form of collocational networks in Chapter seven and proximity tables in Chapters six and eight, and research papers 2, 3, 6, and 7).

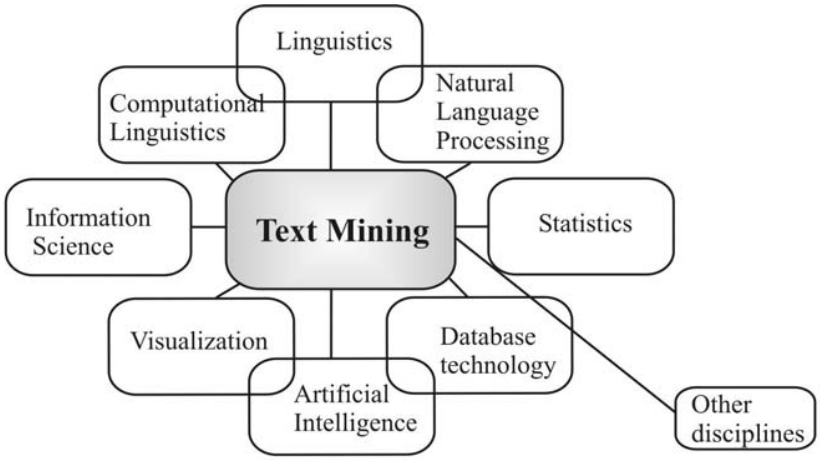


Figure 1 -2. TM as a confluence of multiple disciplines

1.2.2 Text Mining Tasks

The fulfilment of TM tasks, such as clustering, categorization, feature extraction, thematic indexing, and information retrieval by content, is associated with satisfying the primary information needs of text users, such as *searching*, *browsing*, and *visualization*.

In searching, the user specifies an information request in terms of a finely defined query, and asks the system to locate individual documents that correspond to that query. Search engines, such as www.google.com, www.altavista.com, and www.overture.com are the most successful web applications that try to satisfy the needs of searching and information access.

In browsing, the user navigates the text collection with the help of links between individual documents that are provided by the system. Hypertext technology can link web pages creating a hierarchical structure of a document collection such as the one of www.yahoo.com. Summaries and semantic maps can be created for navigating and browsing large textual databases. A summary represents the content of the document in a more compact form than originally, such as keywords or key sentences (Neto, Santos et al. 2000). A semantic map graphically represents concepts and relations that compose a concept. It assumes multiple relations between a concept and the knowledge that is associated with the concept. Searching is used to obtain a suitable starting point of browsing (Lagus 2000). The map displays categories or associations, learning and memorizing of which improve users' searching and browsing abilities (Lin 1995).

Searching and browsing systems require an explicit description of the information needed by the user in the form of queries. However, while searching and browsing text collections for relevant information, users face problems in constructing smart queries. It is easy to imagine situations where the user might not be fully acquainted with established terminology in a field, or not fully sure about the content of the documents he needs to retrieve. Most users, as was noticed in (Anick and Vaithyanathan 1997), prefer to answer questions about the relevance of information already presented to them by the system, rather than to describe explicitly what they are looking for. Additionally, the vocabulary problem that is studied in human-system communication (described in Section 1.2.4) negatively influences users' ability to construct efficient queries.

For visualization of any type of information something familiar, such as a hierarchy or map, is used as a means of illustrating something more complex or unfamiliar. Under text visualization, researchers consider illustration of similarities, differences, overlaps, and other relationships existing in documents and document collections. Text visualization contributes to faster and more intuitive understanding of the entire document collection by filtering out uninteresting items. Graphical interpretation of the relationships in mailing lists and semantic maps is an example of text visualization, e.g. WebSOM (Kohonen 1999).

The most general and common task of textual data analysis is exploring patterns in text. This general task can be divided into finer, more specific tasks, some of which are similar to the *clustering* task described earlier under DM tasks, some of which are different: *categorization*, *feature extraction*, *thematic indexing*, and, according to Hand, Mannila et al.(2001), *information retrieval by content* tasks are employed for TM.

Clustering in TM corresponds to its counterpart in conventional DM, described in Section 1.1. Clustering in TM is the process of partitioning a given collection into a number of previously unknown groups of documents with similar content. Clustering allows for the discovery of unknown or previously unnoticed

links in the subset of documents or terms in any particular document collection. Categorization assigns documents to pre-existing categories, called "topics" or "themes". Applications of text categorization include indexing text to support document retrieval and extracting data from the text (Lewis 1992). Feature extraction refers to the extraction of linguistic items from the documents to provide a representative sample of their content. Distinctive vocabulary items found in a document are assigned to the different categories by measuring the importance of those items to the document content. Thematic indexing refers to the identification of the significant terms for a particular document collection. Indexing identifies a given document or a query text by a set of weighted or unweighted terms obtained from a document or a query text. Those terms are often referred to as index terms or keywords. According to (van Rijsbergen 1979), information retrieval (IR) is the process of locating the subset of the documents that are deemed to be relevant to a posed query. IR by content is the process of inference of documents that are semantically similar to a query pattern.

1.2.3 Distinction between TM and DM

Table 1-3 contains a brief description of similarities and differences between the data and text mining fields as I view them in this dissertation. TM can be seen as a subpart of the more general field of DM. In general, DM consists of methods for exploration of various types of data ranging on the ladder of difficulty from numeric (numbers), textual (free-text, documents) to multimedia (images, video and audio streams). Although general DM algorithms can process all of these types of data, DM is also considered to be a subpart of general DM. A certain distinction can be made between TM and DM as subparts of general DM based on the specific data types and algorithms involved. Throughout this dissertation, I refer to DM as a field that analyzes quantitative data (numbers) and to TM as a field that analyzes qualitative data (text)[2].

Table 1-3. Comparison between Data and Text Mining

	Data Mining	**Text Mining**
Object of investigation	Numeric data (numbers)	Texts (documents, web pages, etc.)
Object structure	Relational databases, numbers	Free form texts
Goal	Predict outcome of future situations, analyze the reasons that affect the desired outcome, visualize data interrelations	Retrieve relevant information, distil the meaning, categorize content, compare and evaluate texts
Methods	Machine learning: decision trees, genetic algorithms, neural networks; statistics: multinomial regression,	Indexing, special neural networks, clustering and categorization algorithms,

[2] Here and later, quantitative or structured data refers to numeric data, while qualitative or unstructured data refers to text

	regression analysis	linguistics, ontologies
Estimated[3] *current market size*	100,000 analysts at large and medium size companies	100,000,000 corporate workers and individual users
Maturity	Broad implementation since 1994	Broad implementation from 2000

1.2.4 Working with Textual Data

Since text is the most popular and convenient way of transferring meaning from authors to readers, the amount of digitally available text is mounting. The professional website for the knowledge discovery community (http://www.kdnuggets.com) reports that about 17% of data mined worldwide during 2001 was text. Other popular data sources were web content that is partly textual, web clickstreams, and time series, with 14%, 14% and 16% respectively. A recent study indicates that 80% of a company's information is contained in text documents (Tan 1999). Managers and knowledge workers spend a lot of time dealing with textual information overload looking for useful points in it. For instance, a Gartner group survey reveals that 75% of managers spend more than an hour per day sorting out and answering their e-mails (Marino 2001), because the amount of incoming e-mail messages is overwhelming. The dynamic business environment does not allow managers the luxury to devote enough time to read and analyze all available documents that might contain information that might impact managerial decisions.

Effective DM methods enable computers to analyze tens of dimensions of numeric data. Unlike numeric data, textual data is highly multidimensional and consists of tens of thousands of dimensions (Fayyad and Uthurusamy 2002). Business intelligence systems operating only with numeric data warehouses succeed in telling managers what happened and when, but they are not very good at answering why. Being the most common vehicle for written communication, text has a complicated and ambiguous multilevel structure. Structural principles exist in the formation of words (morphology of language), in the creation of grammatical sentences (syntax), and representation of meaning (semantics). The three components of text: word usage, grammatical construction, and content vary very much within every individual language. Furthermore, the authors and readers of the text often represent the same semantics using different words (synonymy) or describe different meanings using words that have various meanings (polysemy). For instance, in human-system communication (Furnas, Landauer et al. 1987) discovered that two people favoured the same term with a probability of less than 0.20, which consequently resulted in 80-90% failure rates in communication. The probability that two typists would use the same main verb while describing an editing operation is less than one in fourteen; that two cooks would use the same first keyword for a recipe is less that one in five. This feature of NL words as the

[3] According to personal communication with S.Ananyan, executive manager at Megaputer Intelligence, Inc, 1998. Although 1998 was the year of the high-tech hype, the source is cited to illustrate the relative difference in the potential sizes of market with numeric vs. textual information.

basic units of text confuses TM technologies such as document management systems, automatic thesauruses, and search engines. These technologies are based primary on keywords (indexes) or text properties (such as author, subject, type, word count, printed page count, and time last written) approaches. Those approaches are less effective in working with NL text because of the ambiguity, polesymy, synonymy, complexity of syntactic construction, and multi-variance of interpretations discussed above. TM technologies aim to increase the productivity of managers, decision makers and knowledge workers and reduce textual information overload by retrieving "golden nuggets" or insights from textual databases. The user of TM technology should be able to categorize, prioritize, compare documents, and understand and utilize the meaning of any particular document without browsing, reading and analyzing an entire document collection.

Because of the convenience for humans of delivering information via text, and despite the difficulties for computers in operating with text, there are a number of basic operations that computer-based applications are accomplishing. Figure 1-4 presents routine operations associated with text collections.

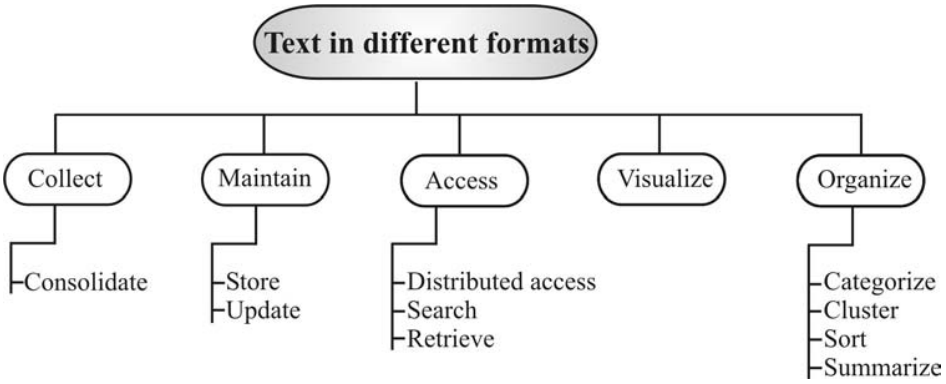


Figure 1-4. The routine operations performed with text collections

If text sources are distributed in various databases and online repositories, then computer-based technologies should *consolidate* those sources to *collect* necessary information. This information should be *updated* and conveniently *stored* to provide further *access* to users, who might *search* the consolidated databases in order to *retrieve* a particular document and represent it in a suitable *visual* form. The efficient *organization* of text collection is a challenging process that is achieved by *categorization, clustering, sorting,* and *summarizing* documents. Although all of the routine operations outlined in Figure 1-4 can be described in a couple of sentences, the number of processes is large and the priority of their efficient accomplishment is not obvious.

1.2.5 Difficulties in Working with Text as a Written Form of Natural Language

Text as a written form of spoken NL provides the most effective communication bridge between people of the same and different generations. Unlike programming or formal languages, NL is ambiguous if not understood in terms of all its parts (Hearst 1999). While modern science tends towards mathematics, only words lend coherence, build understandable narratives, and explain what science is all about (Turney 2001). Even though science and technology should be precise, unambiguous, logical and universal, NL is none of these things.

Popular view, a language is merely a fixed stock of words. Therefore, the field of modern IR that deals with computer-based text processing conveniently relies on language representation as a bag of words. Words interact in many ways: some words co-occur near certain words with higher probability than others. The product of the frequency of use of words and their rank (the order of importance) is, according to Zipf's law, approximately constant (Zipf 1972). However, the extraction of important keywords or indexes from text does not guarantee the extraction of meaning from text. The first danger of the keyword approach is in the slim chance of using the same keywords by different individuals to describe the same concept because of the words' synonymy. Therefore, a part of a document that does not include a query-matching keyword is ignored by conventional IR systems. A more fundamental reason for a language not being a word stock is that NL expressions have syntactic structure. As a result we understand "Mohamed will come to the mountain" and "The mountain will come to Mohammed" completely differently. Moreover, NL does not comply with "unique readability" – there is no one-to-one correspondence between sound strings and syntactic structure, or between syntactic structure and meanings. According to Pullum (2001), other features of NLs that confuse automated systems are the unlimited complexity of NL expressions, and NL syntax variability that allows us to understand even ill-structured expressions. In other words, the sentence below illustrates another aspect of human cognition that was lately discovered by British researches and quoted at www.techdirt.com/articles. "You can slitl raed this sentence, even thugoh the midlde ltteers are mxied up". Randomizing letters in the middle of words has little or no effect on the ability of skilled readers to understand the text. We are able to read and understand the sentence as long as the first and last letters are correct and are placed in the beginning and at the end of the word. This would quickly confuse automated systems. Moreover, we can use a verb phrase such as "look after" with numerous meanings embedded in variety of different clauses and complex sentences. Even by omitting some words from the context, the reader or listener is usually able to infer the meaning from ill structured but still delivered text information.

Furthermore, a speaker or writer who constructs NL expressions to deliver some information to a listener or reader, allows personal preferences and background to determine the structure of textual expressions. Humans can create an

astronomic variety of sentences from a limited number of discrete units – words (Ferrer i Cancho 2001). A language that conveyed all information unambiguously would consist of a separate word for every thing, concept, or action it referred to. Although such a language would be formidably complicated for the speaker, computer analysis would be much more possible. The ideal language for a speaker would be a language of only few simple words, which serve many purposes; the teenager minimal-effort slang is an example of this type of language. In light of this reasoning, Ferrer i Cancho and Solé constructed a new mathematical model of language, which was described by Ball (2003). In this model the cost of using a language depends on the balance between the conflicting preferences of speakers who want to use few words and listeners who want to use many. The notable feature of this model is the variation of language interpretability. In other words, a listener interprets the meaning of a NL expression to the best of his/her knowledge of vocabulary and context, despite the content and lexicon implied by speaker. This model is better for understanding the phenomena of NL, and explains the obstacles that prevent computers from effectively "understanding" text.

In summary, NL, and text as a more tangible representative of it, exhibit a unique combination of characteristics: semantic word-to-word relations which are analogous to signaling among primates and can be articulated in text analysis systems; syntactic structures as complex and exact as in formal languages; and openness, flexibility and interpretability depending on the content that formal languages do not allow. Combinations of NL characteristics embedded in the written form of it restrict text analysis and prevent computers from "understanding" text. However, TM can be performed without thorough understanding of the textual content of documents in a collection because TM aims more at extracting useful patterns and indications from text for supporting decision-making in various fields and reducing textual information overload, than at "understanding" textual content (Witten, Bray et al. 1998).

1.3 Aim of the Research and Research Questions

This research has grown from the context of TM as an intersection of multiple disciplines. In the dissertation I investigate the applicability and characteristics of one particular method of prototype matching to accomplishing several TM tasks.

The systems that suggestively can be built based on the prototype matching methodology use prototypes or examples as inputs. Instead of spending time on constructing a smart query based on well-thought out keywords, a potential user presents a text as a prototype to a system. The prototype (or example) is a text segment or entire document that contains information of specific interest to a particular user. A prototype holds a semantic pattern – a set of syntactic features that occur in a text segment (Navarro 1999). Documents satisfying pattern specifications that are encoded in a prototype are said to "match" the pattern and the prototype. The located matching documents are retrieved as semantically similar in some specific way to the document presented as a prototype to the system. The

methodology gets its name because it utilizes prototype pattern matching against the document collection to obtain clusters of documents that are semantically similar or different from a document-prototype.

I create and apply several possible systems that incorporate prototype matching method for text clustering and IR by content on two illustrious text collections: financial reports and scientific publications. The intended users of the systems are knowledge workers who need to analyze rough textual data, or decision makers who need to utilize available new data and turn these data into knowledge to make crucial decisions quickly. The decision makers in the conference setting who can benefit from the proposed methodology are the conference organizers who try to build a workshop-like setting at the conference and schedule all the sessions carefully to create a high degree of interaction and discussion among the conference participants. The decision makers who can benefit from automatic mining of financial reports instead of manually reading them are the managers of the companies' partners and competitors, creditors and auditors, financial brokers, market analysts, investors, and other stakeholders.

The goals of the research evolve around the phenomenon of working and being overloaded by the masses of digital text in today's business setting. This research work elaborates on the potential use of the prototype matching method for allocating semantically similar scientific papers from a collection of scientific papers and for comparing the financial performance of companies based on a collection of their financial reports. These goals contribute to the reduction of textual information overload by diminishing the amount of textual information presented to the user while retrieving relevant papers in the conference context; and while benchmarking and forecasting companies' performance in the financial context. Information overload is characterized in two ways: first, when individuals are given more information than they can absorb; second, when the demand on an individual's time for performing interactions exceeds the capacity of time available for such processing (Farhoomand and Drury 2002).

Within the research frame I intend to achieve the following objectives:

a) to explain the nature of the relationship between textual information overload and the technologies that contribute to its occurrence and reduction
b) to explore the extent of use of prototype matching to knowledge discovery from financial textual collection
c) to demonstrate how knowledge extraction can be performed from low-level textual (qualitative) and numeric (quantitative) data with the help of prototype matching
d) to suggest ways to exploit the prototype matching in combination with other DM methods to deliver additional insights into phenomenon described in text to potential users
e) to determine the suitability of the prototype matching method to IR by content from scientific text collection

There are two main questions to be answered in the dissertation: does the prototype matching method, which hypothetically defines semantic similarities among documents in the collections, really discover some relationships among the documents; and, what can the user of the system built on the prototype matching method learn from the discovered relationships among the documents (i.e. scientific papers and financial reports).

The experience and findings needed for the elaboration of this dissertation have been collected in the GILTA (manaGIng Large Text MAsses) research project run in collaboration with Turku Centre for Computer Science, Åbo Akademi University, Pori School of Technology and Economics, and Tampere University of Technology between years 2000-2002. The wide scope of the projects was to build the prototype matching method and evaluate it form different perspectives. The project was conducted from a system development perspective, where the research team has proposed and investigated advantages and drawbacks of the new methodology in various contexts.

1.4 Related Work

Information systems (IS), such as transactions processing systems, management and decision support systems, office automation systems, multi-agent and knowledge management systems, enabled by modern network communication technology, intelligent technology, database and wireless technologies, contribute to both creation and reduction of textual information overload. In Chapter three I describe the impact of different types of IS on textual information overload according to the classification of IS by the TM tasks they perform, and give examples of existing IS systems available in the academic sphere and commercially. Although IS systems are available for TM tasks, they are not perfect because text representation and distillation, knowledge discovery algorithms, and extracted knowledge representation algorithms are computationally heavy and require further development. Automatic text representation is a challenging task because of word ambiguity in various contents. The ongoing research in AI, information systems, computer science and computational linguistic communities offer various algorithms for performing information retrieval by content, clustering and categorization. The Vector space model, the Boolean retrieval model, and probabilistic retrieval model described in Chapter four are among the most popular models for IR based on term weighting and introduction of indexes. Text clustering and categorization are performed by either traditional or modified statistical algorithms, i.e. k-means or hierarchical, or by heuristic algorithms, i.e. neural networks, self-organizing maps (SOM), and genetic algorithms. The representation of patterns extracted from text collection after performing IR, text clustering or categorization is an attractive field of research because text brings multidimensional information, which is difficult to depict graphically to human cognition. A number of attempts in text visualization have been made by introducing 3D trees, WebSOM, semantic maps, and by constructing summaries. Section 4.3 discusses some particular examples of the visualization of TM results.

1.5 Overview of the Dissertation

Table 1-5 shows the outline of the dissertation. It consists of two major parts: the first one is a comprehensive summary and outline of the research; the second one is the combination of seven research papers.

Table 1-5. Dissertation Outline

Theme	Part I	Part II
Introduction	Chapter 1	
Research Approaches	Chapter 2	
Information Systems and Technologies to Support Managerial Work in Dealing with Textual Information Overload	Chapter 3	*Papers* 1, 5
State-of-the-art in TM	Chapter 4	*Paper* 3
Research Methodology of the Prototype Matching Method	Chapter 5	*Papers* 1, 2, 3, 4
Combination of Data and Text Mining Methods	Chapter 6	*Papers* 2, 5, 7
Mining Content of Financial Reports	Chapter 7	*Papers* 2, 6
Information Retrieval by Content of Scientific Publications	Chapter 8	*Papers* 3, 4
Conclusion	Chapter 9	

In Chapter two, I present the pluralistic research methodology that is used in the research process. Here, mainly, the combination of *interpretive*, *explorative* and *constructive* approaches constitutes the pluralist methodology of the research. The interpretive research approach produces understanding of the context of IS and the processes that influence and are influenced by the context of IS. The explorative approach leads to insights and to increased familiarity with the problem area. The constructive approach unites the processes of new artifact building based on existing research knowledge, most likely gained from the exploratory part of the research. I discuss how these two approaches are combined to form my research methodology, which contributes to solving the problem of textual overload by extracting "golden nuggets" from financial and scientific text collections. The retrieved "nuggets" should provide decision makers with concise and lucrative insights into the information coded in text collections, aiming to enhance the decision maker's understanding of the underlying phenomenon. This chapter intends to present the methodological elements that I have used in my research process.

In Chapter three, I summarize the recent advances in IS and IT that contribute to the problem of textual information overload. IS are seen as the process of applying IT in social context for solving various business related problems. IS and IT can be used to support managers, knowledge workers, and decision makers in their work by helping to digest the masses of text in order to find useful knowledge in them. I provide a brief overview of today's TM systems with respect to the common TM framework that unites text distillation, knowledge sophistication, and knowledge representation processes. This chapter does not aim to scientifically contribute to the content of the dissertation, instead it sets the

technological context in which textual overload occurs, and suggests technological means for dealing with it.

In Chapter four, I describe the state-of-the-art scientific domain for text processing that assists in reducing textual information overload. I discuss the following aspects: What are the common processes that constitute TM? How should text from documents be represented in order to become digestible for computers? What are the different approaches and methods used to discover valuable knowledge from text collections? What are the methods for representing discovered TM results?

In Chapter five, I present the prototype matching method that has been used as my primary mathematical approach for dealing with textual information overload in two different text domains: financial reporting and scientific publications. Chapter fives provides the key definitions and a description of the key steps of the method. This content-based method, which was invented in the GILTA project, came up during my research as a "common denominator", and thus is used as a base methodology in *Papers* 1-4, and *Papers* 6-7. For my part, I studied its suitability for performing IR by content on a collection of scientific articles and for performing financial analysis on a collection of quarterly reports. I explore and analyze the applicability of the prototype matching method for various TM-related tasks in distinct and well-defined text domains. As a starting point of view, *Paper* 1 reports various applications of the method, some of which I researched more thoroughly and describe in the subsequent chapters.

In Chapter six, I illustrate how a combination of various data and text mining methods can be used to uncover some unexpected trends in financial data. I explain how findings from TM can become a source for DM in retrieving further insights from data collections with both qualitative and quantitative data. I propose a conceptual model for a knowledge-building system based on the integration of DM techniques for both numeric and textual data. The summarized experiences and findings gathered in the research project together with the review of the theoretical concepts from the state-of-the-art chapter have evolved into the model suggested in *Paper* 5. *Paper* 5 describes a conceptual model of a knowledge-building system for decision support. The model is based on a society of software agents, and a combination of data and text mining methods. The details of using a combination of data and text mining methods for discovery of complex patterns to be used in financial forecasting are provided in *Paper* 7. Additionally, I briefly present another related stream of research on combining data and text mining methods, which I discuss in more detail in the subsequent chapter and in *Paper* 2.

In Chapter seven, I introduce additional applications of our content-based TM method for discovering indications of companies' future financial performance from their quarterly reports. This chapter describes the practical aspects of existing methods and tools for financial report processing and competitor financial benchmarking. While *Paper* 2 serves as a more detailed illustration of how future financial performance can be semi-automatically inferred from the textual parts of the reports, *Paper* 6 provides linguistic validation of the TM results. *Paper* 7 shows how those indications of companies' future financial performance can be used in prediction.

In Chapter eight, I present another application of the prototype matching method. This content-based method is used for IR of relevant scientific articles from the domain of scientific publications. I regard the allocation of scientific papers submitted to conferences as an integrated task of TM and IR, namely IR by content. The chapter starts by providing the motivation for performing information retrieval by content for scientific publication allocation in a multidisciplinary conference setting. *Paper*s 3 and 4 offer a detailed description of the applicability of the prototype matching method to a collection of scientific publications, based on abstract and full-paper levels respectively.

In Chapter nine, I summarize the research results and the contributions of the dissertation. I also state the limitations of my study and propose directions for further research.

1.6 Contributions and Publications

The novelty of this dissertation is in the utilization of a new content-based method – the prototype matching method – developed in the GILTA project by Prof. Ari Visa (Visa, Toivonen et al. 2000) from Tampere University of Technology, along with or in combination with other DM techniques for TM purposes. The GILTA project has been led by Prof. Ari Visa and I have been a member of the group. My work makes three primary contributions. First, it has been shown that the prototype matching method provides a platform on top of which various data exploration tools can be constructed. I combined data and text mining methods for processing quantitative and qualitative financial data with the aim of obtaining additional knowledge about the financial performance of the analyzed companies. Secondly, I constructed frameworks for content-based data exploration of free text – both from the scientific and financial domains. Thirdly, I illustrated how to ease textual information overload for decision makers, knowledge workers and managers using the proposed frameworks. The later objective is achieved by retrieving documents semantically similar to the search criteria, without an obligatory reading of the entire document collection. Although it could be difficult and time consuming to identify the patterns that characterize the document-prototype and those similar to it, these can be very valuable to managers, decision makers, or knowledge workers. During the course of this research, my main findings were reported in a number of scientific papers.

Paper 1, *"Prototype-matching Methodology Applications in Text Mining"*, is an introduction to the problems associated with textual information overload in various real-life settings in different textual domains. The paper discusses the applicability of the proposed content-based methodology of prototype matching to help users of textual information to satisfy their information needs in a diverse range of tasks: organization of scientific conferences, news clustering, analyses of financial information, or authorship attribution. The paper has a descriptive character and aims to review the potential applicability and usability of the novel methodology. I was the main author of the paper, for which I had collected the

experience that was accumulated by my research group in the early stage of development of the prototype matching.

Paper 2, *"Combining Data and Text Mining Techniques for Analyzing Financial Reports"* explores the possibility of discovering knowledge from two types of data (quantitative and qualitative) describing the same facts about companies' financial performance. Pattern discovery is achieved through a combination of data and text mining methods. The paper introduces a research model for separately analyzing the quantitative and qualitative parts of financial reports using, for instance, the SOM and the prototype matching method. This paper can be viewed as an extension of one particular application mentioned in *Paper 1*, namely analysis of financial information in the form of annual/quarterly reports. The novelty with respect to earlier applications of the SOM for numeric data clustering is its combination with the prototype matching method for textual data clustering, which allows the identification of complex patterns in quarterly reports that highlight the strategic focus and benchmark positions of the companies. My part of the research consisted of gathering the reports from Nokia, Ericsson, and Motorola for the qualitative data set, obtaining the results from the prototype matching of every report in the entire data set, and analyzing the results. The obtained results were explained after thorough reading of the reports and comparing the current study to the previous study of (Back, Toivonen et al. 2001).

Paper 3, *"Toward Content based Retrieval from Scientific Text Corpora"* can be considered a shorter version of the research on the applicability of the prototype matching method to the domain of scientific publications. This methodology was applied for information retrieval by content of scientific papers based on the processing of their abstracts. The results contribute to multidisciplinary conference organization by helping conference organizers in establishing semantic similarities among submitted papers. Although the paper was joint work, I was in charge of investigating the applicability of the novel content-based methodology of the prototype matching method for scientific text corpora's case study, and describing, explaining and validating the obtained results with domain knowledge.

Paper 4, *"Prototype-Matching System for Allocating Conference Papers"* suggests the use of the prototype matching method to allocate scientific papers into tracks, minitracks, and themes according to their content in a multidisciplinary conference setting. On the one hand, the proposed system assists the conference organizers to automatically establish semantic similarities among submitted papers and allocate them into common themes. On the other hand, the system aims to assist the attendees in retrieving papers from the conference proceedings based on the similarities of their contents. The allocation is performed on full-text versions of the submitted papers and, thus, can be regarded as an extension of *Paper 3* with more thorough validation of the linguistic peculiarities of the mined textual data. The center of my research attention was to collect and preprocess articles accepted to the Hawaiian International Conference on System Science–34 as a sample qualitative data set. I used every article as a prototype for retrieving articles that are similar by their content. I performed interpretation of the similarities among articles together with validation of those similarities using domain knowledge.

Paper 5, *"A Conceptual Model for a Multiagent Knowledge Building System"*, introduces a conceptual model of a knowledge building system based on a society of software agents, and the combination of data and text mining methods. The proposed system can be used to monitor new financial updates from a variety of sources, and calculate financial ratios for different companies. This model, based on qualitative and quantitative data, could be used for accomplishing various tasks, for example, financial benchmarking and assessing creditworthiness of different companies. In this paper I paid more attention to binding various technologies to create a construct of a knowledge building system. The paper proposes a general idea and some enhancements of using multiagent technology for discovering complex patterns in quantitative (numeric) and qualitative (text) data that describe the same phenomena. My contribution was in proposing the initial idea of implementing a combination of data and text mining methods based on a society of intelligent software agents for knowledge building from financial related data. I suggested the initial architecture of the conceptual model of the "knowledge-building system" using an agent metaphor.

Paper 6, "*Mining Textual Contents of Quarterly Reports"* studies more carefully the language, messages, and contents of quarterly reports from linguistic and TM points of view. I compare the results obtained from linguistic analysis of quarterly reports by means of collocational networks and the results obtained from automatic TM by means of prototype matching. Although the results are somewhat controversial, this paper justifies the TM approach used in *Paper 2* by linguistic validation of the obtained results. I cross-justified the results received by the methods from two different sciences, namely TM and linguistics. In this part of the research I was responsible for the coordination of applied methods, implementation of elementary training, and analysis and comparison routines. I was mainly responsible for consolidating the results from TM and linguistic analysis, drawing the conclusions from the consolidation, and writing the paper.

Paper 7, *"Determining Companies' Future Financial Performance From Their Historic Quarterly Reports"* elaborates on the possibility of using neural network technology to predict the future financial performance of the companies from the discovered patterns embedded in the numeric and textual parts of their quarterly reports. This research stream continues and builds on findings from *Paper 2*, in which the combination of data and text mining methods was proposed for the discovery of complex patterns in the qualitative and quantitative parts of financial reports. I was responsible for creating the idea of constructing backpropagation neural networks to predict future financial performance for companies, whose historic financial performances were analyzed using the SOM and the prototype matching method. I proposed the framework that I have applied on the data set, which I have collected from the quarterly reports of Nokia, Ericsson and Motorola.

All of my papers use the same method of prototype matching for accomplishing different TM tasks. The entire dissertation utilizes a pluralist research paradigm (Mingers 2001) allowing for benefits from several methodologies for an effective solution to the textual overload problem, rather than following one specific methodology. As a result, *Paper 1* is a descriptive research

paper that provides an overview of the methodology applications to different domains and generalization of possible outcomes of various text-related tasks. *Paper 2* exploits the TM ability of this methodology to discover hidden useful patterns or nuggets in the financial quarterly reports that contribute to creating insights for decision makers about companies' future financial performance. *Paper 3* uses the IR by content ability of the method to establish semantic relationships among scientific publications. *Papers 2* and *3* belong to exploratory research because they aim to find out more about the method and its applicability for different problems. *Paper*s *4*, *6*, and *7* develop the ideas and lessons learnt from both *Paper*s *2* and *3*, and offer constructions (conceptual models, instantiations, or methods) for dealing with different subtasks (financial performance prediction, TM evaluation, and conference organization).

Moreover, *Paper 1* describes the applicability of the prototype matching method for information reduction in authorship attribution, financial analyzis, and clustering of scientific collections tasks. *Papers 2* and *7* illustrate the use of the prototype matching method for TM in combination with other DM tools on financial text collections (research objectives *b*, *c*, and *d* from Section 1.3). *Paper 6* suggests a conceptual model that can be applied for knowledge extraction from low-level textual data (research question objective *c* from Section 1.3). *Paper*s *3* and *4* elaborate on the applicability of the prototype matching method to IR by context from scientific text collections, which is stated as one of my research questions (research q objective *e* from Section 1.3).

Papers 1, *2*, *3*, *4*, *6* and *7* have been published in the proceedings of International Information Systems conferences and were presented at these conferences. They have been reviewed through a blind peer-review process. The revised version of *Paper 2* was accepted for publication in the Journal of Information Systems in Accounting, Finance and Management (IJISAFM). The findings from *Papers 6* were presented at the XXVI Annual Congress European Accounting Association and published in the research reports series of the Turku Centre for Computer Science. The revised version of it was accepted for publication in The International Journal of Digital Accounting Research. *Paper 7* has been accepted by the committee of the First Annual Pre-ICIS Workshop on Decision Support Systems in Seattle, Washington, USA, December 14, 2003.

2. RESEARCH FRAMEWORK

Information systems (IS) as an emergent interdisciplinary field (Lee, Gosain et al. 1999) combines computer, organization, and management sciences to study the application of IT in organizations and society, with the aim of directly or indirectly improving some aspects of social, technical, and organizational activity (Keen 1980). In an applied field, such as IS, it is common to depend on theories prepared by "reference disciplines" (Clarke 2000). Because of its interdisciplinarity, IS encompasses fundamentally different schools of thoughts based upon different perceptions of its core concerns, definitions, as well as different approaches to investigate them (Checkland and Holwell 1998). The multiplicity of the approaches has influenced the choice of appropriate research methodology in my case.

In this chapter I will discuss the theoretical foundations and beliefs used in the research framework of the dissertation. I outline the definitions of *data*, *information,* and *knowledge* because a unique understanding of the terms is important for KDD and TM. I believe that when proposing novel solutions for practical problems the choice of research methods depends mostly on the nature of the problem at hand and that different research objectives require the application of different research approaches. Therefore, various research approaches from dissimilar schools of philosophy of science are to be used to accomplish the tasks posed in this dissertation. As Moody and Buist (1999) stated: The real question is not whether the research method is appropriate *per se*, but whether it is appropriate to answer the question being asked. In my case to answer a question of how to extract valuable patterns from textual data and reduce textual information overload with the help of the prototype matching method I found appropriate to rely on a combination of research approaches. My standpoints on the methodologies and philosophical issues on definitions, and concepts, and the combination of research methods used in the dissertations are clarified in the subsequent sections.

2.1 Definitions: Data-Information-Knowledge with Respect to Text

The concepts of data, information, and knowledge concepts are closely related to KDD processes, such as DM and TM, which aim to transfer data into knowledge. The understanding of the concepts is an important step in both performing the KDD process and creating modern knowledge-based organizations. Those organizations apply knowledge management practices for achieving and sustaining competitive advantage. There are a number of different definitions of *data, information, and knowledge* and varying interpretations of their relationship.

Data. In this dissertation I adopt an understanding of the term data as a set of discrete, objective facts about events (Davenport and Prusak 1998). Data can be interpreted as a raw statement of the facts in the form of symbols that merely exist and have no significance beyond that existence. Data by itself, unlike information, has little relevance or purpose. Modern organizations store data concerning their

transactions in technology systems by fulfilling the record-keeping aim. Managers argue that if data were accurately gathered then objectively correct decisions would automatically suggest themselves. Efficiently recording millions of transactions is a very difficult but achievable task using IT, unlike making sense of these data to assist in efficient decision-making. Here a contradiction resides – more data is not always better than less data. Although data does not provide any judgment or interpretation, both information and knowledge can be communicated through data.

Information is data organized into meaningful relationships and structures that are endowed with relevance and purpose. According to many researchers, information can be described as a message that has a sender and a receiver. Information, usually as a text document or audible communication, is meant to change the way a receiver perceives the reality. The Latin word "inform" means "to give a shape to". Only the receiver of information can decide whether or not the message that has traveled from the sender is truly satisfying the receiver's information needs. For instance, a report full of scientific ramblings may be considered information by the writer but regarded as noise by the reader. The solitary message it may imply unintentionally to the receiver is about the quality of sender's intelligence. Information can convey an explicitly or implicitly stated message that can at the same time be different to sender and receiver. Because information has meaning, it has a shape when it is organized for some purpose (Davenport and Prusak 1998). To transform data into information one should conceptualize, categorize, calculate, correct, and condense. Although computers can help in some aspects of transforming data into information, they require human assistance with adding context, categorizing, and condensing of irrelevant raw statements. In other words, computers require human supervision and guidance in operations that form meaning from the data. Information is an important brick in the building of coherence, because it can be viewed as a collection of symbols, which has the potential to alter the cognitive state of a decision maker.

Knowledge. There is no consensus in literature on what knowledge is. Over the years, the dominant philosophies of each age have added their own definition of knowledge. All the definitions of knowledge suggest that knowledge is a broader, deeper, and richer concept than data and information. Scientific knowledge is viewed as a construct and defined as "understandings that the cognitive system possesses and uses to take effective action to achieve the cognitive system's goal." Nonaka (1995) defines knowledge as justified true belief. I adopt the modern working definition of knowledge as information organized into meaningful patterns that are applicable for particular purposes. Knowledge consists of the 'mental models' inherent in humans derived from education, experience, culture, genetics, character, and mood. Knowledge usually emerges when people use (share, discuss, evaluate) context-relevant and validated information during their interaction. Knowledge derives from information as information derives from data. This transformation happens through knowledge-creating activities that according to Davenport and Prusak (1991) are comparison, consequences, connections, and conversation. Knowledge is close to action and must be validated and improved through the course of actions.

Textual messages and documents with audio interchange of thoughts are the best examples of information in the context of human communication. In this dissertation I analyze mainly two types of data: quantitative (numeric) and qualitative (text). The numerical representation is more convenient over other representations because it allows manipulation of the numbers easily and relatively effortlessly (Berthold and Hand 1999). While the objects of numeric data analysis are the things, which have given, rise to the numbers, such as measurements, the objects of text data analysis are the blocks of text themselves. In text data, the basic symbols are words rather than numbers. Words can be combined in more complex and loaded ways than numbers can. Text data in the form of documents, which I explore in this dissertation, can be regarded as information because documents have syntactic, semantic, and morphological structure, as well as purpose, receiver, and sender. Due to the duality of text, as data and information, it usually requires additional human expertise to automatically decode its meaning. The complicated nature of text allows the interchangeable use of the terms qualitative or "unstructured" data, text, and textual information.

2.2 Pluralist Research Methodology

Research is a process of inquiring and investigating activities that are focused on mapping theory to practice and developing theory based on practice for obtaining fuller scientific knowledge of a studied subject (Järvinen 2001). Theory can be created, refined, and improved by the application of the research to any particular studied area. Research, in general, is constituted from its objectives and methods (methodologies or approaches). There is a common confusion between the terms *method* and *methodology* in the IS literature and research. Methodology can either mean a study of meaning (Checkland 1981), or in the more general view, methodology refers to the actual research methods that are used in a certain piece of research (Mingers 2001). In the current research I use the term methodology according to Mingers as the combination of processes, methods and tools for conducting research. Furthermore, research methods and methodologies make explicit assumptions about the nature of the world and of knowledge, relying on a certain *paradigm*. In social sciences Burrell and Morgan (1979) constructed a set of antithetical paradigms that specifies a general set of philosophical assumptions that could exist simultaneously. A paradigm covers ontology (what is assumed to exist), epistemology (the nature of valid knowledge), axiology (what is considered valued or right), and methodology (Fitzgerald and Howcroft 1998). According to Nunamaker, Minder et al. (1991), the *research process* involves understanding the research domains, asking meaningful research questions, and applying valid research methodologies, to address these questions and contribute to expanding the body of knowledge in a given domain.

The existence of cross-categorization of research and research methods is possible because of the *pluralistic view* on methodologies that penetrates IS research. The pluralist methodology uses the idea of grouping different paradigms to achieve a desirable combination of methods that have substantially different assumptions but can complement and benefit from each other without leading to

anarchism (Landry and Banville 1992). IS, according to Lee, Gosain et al. (1999), lent theoretical paradigms from its three parental fields - computer science, management science, and organizational science – in order to constitute the necessary foundation for IS research. Those fields encompass very different research traditions for achieving their goals, which puts IS in a position characterized by a plurality of research paradigms similarly to other applied[4] disciplines (Moody and Buist 1999; Mingers 2001). Mingers (2001) states two main arguments for promoting pluralism. On the one hand, it is the ontologically stratified and differentiated real world that consists of a plurality of structures constituting any particular event. On the other hand, the nature of research is not usually a single discrete event but a process consisting of a number of phases with individual tasks and problems. Following Moody and Buist (1999), all research methods may be appropriate depending on the research problem addressed. The choice of a research method is not a question of paradigms, but rather a function of the research problem (Decrop 1999). Therefore I adopt the pluralist methodology, varying research methods to accomplish different objectives in the different parts of the dissertation. For instance, the first objective of the dissertation (objective *a* from Section 1.3 – the explanation of the relationship between textual information overload and IT)– offered in Chapter three is achieved following the interpretive research approach. The second objective of the dissertation – the exploration of the prototype matching usage for knowledge discovery from financial textual collection (Sections 7.1 and 7.2) is achieved following the exploratory research approach. The third and forth objectives – demonstrating knowledge extraction from low-level qualitative and quantitative data using a combination of the prototype matching method and other DM methods by delivering additional insights into phenomenon described in text (Sections 6.2, 6.3, and 7.3) – are carried out following the constructive research approach. The fifth objective – determination of the prototype matching applicability to IR by content from scientific text collections (Chapter eight) –is accomplished in the exploratory and constructive manners. Section 8.2 uses the traditions of exploratory research to gain insight into using the prototype matching method for IR by content of abstract versions of scientific articles. Section 8.3 uses the traditions of constructive research to determine the extent of the applicability of the prototype matching method for retrieving full versions of scientific articles.

According to Nunamaker, Minder et al. (1991), research in IS can be classified in various ways based on its objectives and methods:

1. *Basic* and *applied research*. Basic (*fundamental* or *pure*) research is driven by a scientist's curiosity or interest in a scientific question. The main motivation is to expand humans' knowledge, not to create or invent something commercially valuable. Most scientists believe that a basic, fundamental understanding of all branches of science is needed in order for

[4] While natural science is descriptive and explanatory in its intent, it is inclined toward basic research in the sense that it aims at understanding reality. Design or applied science is creating and evaluative in intent. It attempts to create things that serve human purposes by offering prescriptions and artifacts (March and Smith (1995)..

progress to take place because basic research lays the foundation for the applied science. Applied (*design*) research aims at improving human level of living and, thus, is designed to solve practical problems of the modern world, rather than to acquire knowledge for knowledge's sake (Spengler, Pinkas et al. 1998).

2. *Scientific* and *engineering researches* have no logical distinctions in methods, but are distinct in the objectives that engineers and pure scientists attempt to achieve. According to Davies (1973) cited in (Nunamaker, Minder et al. 1991) in an engineering approach, the artistry of design and the spirit of "making something work" are more essential than in scientific research.
3. *Evaluative* or *developmental (constructive) research* is directed toward solving problems. The developmental type of research "involves the search for (and perhaps construction or synthesis of) instructions" that yield a better course of action (Ackoff, Gupta et al. 1962). Although developmental research has largely been disregarded by some researchers, it is clear that without research efforts directed toward developing new solutions and systems, there would be little opportunity for evaluative research.
4. *Research* and *development* give rise to exploratory, advanced, engineering, and operational studies. Development itself is the systematic use of scientific knowledge directed toward the production of useful devices, methods. It includes design and development of prototypes and processes.
5. *Formulative* (*exploratory*) and *verification research*. The goal of formulative research is to identify problems for a more precise investigation, to develop hypotheses, as well as to gain insights[5] and to increase familiarity with the problem area. The goal of verification research is to obtain evidence to support or refute formulated scientific guesses and hypotheses.

The classification proposed by Nunamaker, Minder et al. (1991) is given to illustrate the multiplicity of IS research approaches available. The classification of the IS research categories is not mutually exclusive and one particular research effort can belong to several categories. The research presented in this dissertation fits comfortably into the category of applied science and belongs to the engineering, developmental (constructive), and formulative (exploratory) types of research.

Text as qualitative data, which identify things that exist rather than how many of them exist, was the predominant subject of my study. On the methodological level, however, I mainly follow the *quantitative approach* of using mathematical techniques to identify relationships, build models, and derive conclusions. Historically, quantitative methods have dominated in IS research because they deliver results in a conclusive and structured fashion (March and Smith 1995). However, I operate with the meaning and context by attaching

[5] Insights are fresh and unexpected perspectives that can be gained from data or information. According to *The Collins English Dictionary*, one of the definition of insight as a noun, (psychology) is the immediate understanding of the significance of an event or action, or more generally, the ability to perceive clearly or deeply, a penetrating or often a sudden understanding, as of a complex situation or problem.

different measures to different meanings. Different research dichotomies were considered for the two basic tasks that I tackled in the course of my research – discovering complex patterns in financial reports (TM in Financial Reporting, see Chapters six and seven, research objective *b* in Section 1.3) and retrieving scientific papers according to their contents (IR in Scientific Proceedings, see Chapter eight, research objective *e* in Section 1.3). Table 2-1 presents a summary of the dichotomies used for research objectives *b* and *e* from Section 1.3. The definitions adapted from (Fitzgerald and Howcroft 1998) initially described "soft" vs. "hard" research dichotomies, so that interpretivist, relativist, subjectivist research, qualitative method, and relevance attributes relate to the standpoints of "soft" research dichotomy; while positivist, realists, objectivist research, quantitative method, and rigor attribute relate to the standpoints of "hard" research dichotomy. This dissertation unites both "hard" and "soft" standpoints by utilizing the advantages of pluralism in research methodology. The categories' definitions of ontological, epistemological, methodological levels are borrowed from (Iivari, Hirscheim et al. 1998).

Rigor and relevance determine the value of the obtained scientific knowledge when doing scientific research. Rigor is the strength of inference made possible by the given research (Staw 1985). The relevance of research is the external validity of the research question to the practice. The tradeoff between relevance and rigor in research is an argumentative issue in the IS community. The theories explaining the tradeoff and interrelationship between rigor and relevance are presented in (Fallman and Gronlund 2000). The research conducted in this dissertation is rigorous because it aims at testing out a new method, the prototype matching method, and its combination with established mathematical methods for extracting knowledge from low level data found in the mist of information overload (research objectives *a* and *c* from Section 1.3). The internal validity of the research is determined by the soundness of the mathematical methods that were used in it. The research is relevant to the practice because it has direct bearing to the matter of textual information overload and automatic knowledge discovery, and it savings in the manual efforts of knowledge workers, decision makers, and managers.

Table 2-1. Summary of research dichotomies used in the research papers

	TM in Financial Reporting	**IR in Scientific Proceedings**
Paradigm level (concerns the general agreement of belief of how the world works)	**Positivist** – belief that the world conforms to the fixed law of causation. Emphasis on objectivity, measurement and repeatability. Concerns with *what is.* (We do not need to understand text to mine the important text dimensions that help to reduce complexity from textual information overload in order to narrow down an answer of what is the that performs better.)	**Interpretivist** – understanding the situation offered by social members from the researcher's own frame of references. (In the evaluating part of the research we interpret the meaning of a retrieved text according to our background and research interests, keeping in mind our frame of reference.)
Ontological	**Realist -** belief that the external	**Relativist** – multiple realities

level (concerns the structure and properties of "what is assumed to exist", i.e. building blocks)	world consists of pre-existing tangible structures independent of an individual's cognition. (Assuming that financial performance and text builds a one-to-one relationship that can be detected automatically.)	exists as subjective constructions of the mind. (Assuming that any retrieved result has multiple interpretations and meanings.)
Epistemological level (concerns the nature of knowledge and the proper methods of inquiry)	The **Objectivist** researcher remains detached from the research situation. Neutral observation of reality takes place in the absence of any contaminating values or biases on the part of the researcher. (all of the experiments in applying prototype matching to scientific or financial text collections were performed objectively using computers, thus results did not depend on the researcher's opinion)	
Methodological level (refers to procedures used to acquire knowledge)	**Quantitative** methods use mathematical and statistical techniques to identify facts and causal relationships. Results are generalized to larger populations within known limits and errors. (the prototype matching used for establishing semantic similarities between text documents, along with SOM or neural networks is a mathematical technique) **Exploratory** methods concern discovering patterns in research data, and understanding and explaining them. They lay out the basic descriptive foundation. (the study of the applicability of the prototype matching method for TM tasks from scientific and financial text collections has a predominantly exploratory nature)	
Axiological level (judges and refers to values of acquired knowledge)	**Rigor** is characterized by the testing of paradigms with emphasis on internal validity through tight experimental control and quantitative techniques (predominant). **Relevance** is the external validity of the research question. It assumes that research relevance to practice is vital (minor).	

2.3 Methods Used in the Research Papers

March and Smith (1995) considered IT research as research that studies artificial as opposed to natural phenomena and deals with human creations, i.e. organizations or IS. Referring to Simon (1981), natural science is concerned with explaining how and why things are the way they are, while design science is concerned with "devising artifacts to attain goals." Moreover, natural science is descriptive and explanatory in intent. Design science offers prescriptions and creates artifacts that embody those prescriptions. March and Smith (1995) argued that IT as artificial phenomena can be both created and studied under a broad notion of science that includes natural and design science. In this dissertation this duality has impacted the combination and choice of research methods, which resulted in the

adoption of the interpretive research approach for original research *Paper 1*, the exploratory research approach for *Papers 2* and *3*, and the constructive research approach for *Paper*s *4*, *5*, *6* and *7*.

Figure 2-2 schematically outlines which research methods were used in the different research papers, what topics were discussed in those papers, and in which chapters of the dissertation the conclusions on those topics are given.

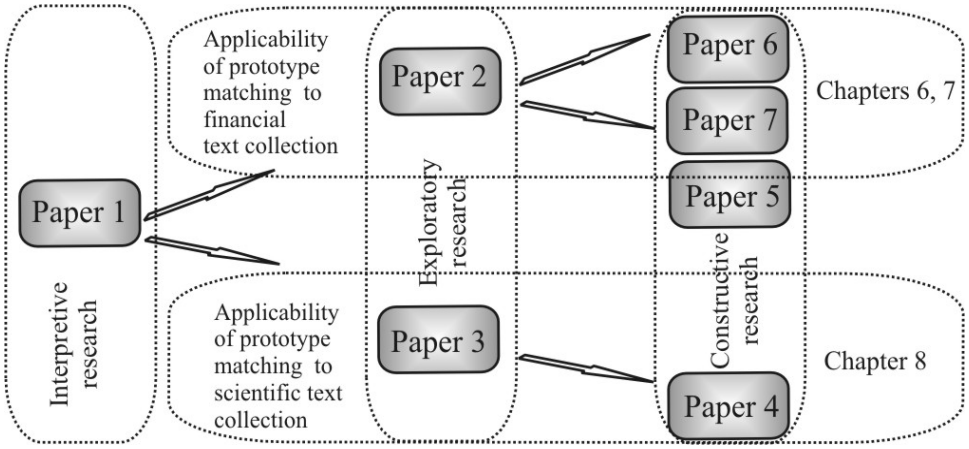


Figure 2-3. Interrelationship between papers and research methods used in them

In an interpretive manner *Paper 1* provides a state-of-the art depiction of the applicability of the prototype matching method for comparing the books of the Bible in different languages, organizing scientific conferences, clustering news, analyzing financial information, and attributing authorship. Exploratory in nature, *Paper*s *2* and *3* investigate particular problems associated with mining financial and scientific textual data, and provide insights for constructing artifacts for solving those problems. Consequently, *Paper*s *4*, *5*, and *7* follow the artifact-building branch of the constructive research approach, while P*aper 6* follows the artifact-evaluative branch of the constructive research approach.

2.3.1 Interpretive Research: Paper 1

Interpretive research generally attempts to understand phenomena through the meanings that people assign to them in any particular context, and to processes whereby the IS is influenced by the context (Myers and Walsham 1998). Interpretive research assumes that our knowledge of reality can be gained through social constructions, such as language, tools, documents, and other artifacts (Klein and Myers 1999). *Paper* 1 consists of a description of the case studies that were performed using the prototype matching method for executing various text related tasks for different text collections.

Paper 1 explains how the prototype matching method can contribute to solving various TM tasks to ease the burden of textual information overload

(research objective *a* from Section 1.3). *Papers 2*, *5*, *6*, and *7*, while following different research methods, accomplish a number of the main objectives of the dissertation. Exploration of the applicability of the prototype matching method to knowledge discovery from financial textual collection, and demonstration of knowledge extraction from low-level textual and numeric financial data with the help of the prototype matching method in a combination with other DM are performed (research objectives *b*, *c* and *d* from Section 1.3). In *Paper*s *3* and *4*, while following different research methods, I determine the applicability of the prototype matching method to IR by content from a scientific text collection, accomplishing the last objective of the dissertation (research objective *e* from Section 1.3).

2.3.2 Exploratory Research: Papers 2 and 3

Exploratory studies, while learning new phenomenon, are isolated from other, more specific studies. Exploratory research pertain to all strategies for answering research questions of how, why, what, how many, how much and where ((Yin 1989) p.13 in (Järvinen 2001)). According to Tull and Hawkins (1987), exploratory research discovers and classifies the general nature of the problem and the variables or hypothesis that relate to it. Exploration can be seen as a preliminary step that helps to ensure that future rigorous study does not begin with an inadequate understanding of the nature of the problem according to Zikmund (2000) cited in (Anckar 2002).

As the first step in the series of research tasks with the objective of investigating the applicability of the prototype matching method to reduce financial textual information overload, an exploratory study was conducted in *Paper 2*. It shows that application of the proposed methodology to analysis of quarterly financial reports can detect some indications of the future financial performance of the researched companies. The study was conducted on quarterly reports from three leading companies from the telecommunication sector gathered from the Internet. By using a clustering technique for mining quantitative data in the form of financial ratios, and clustering qualitative data in free-form text from quarterly reports, I detected the discrepancies in the qualitative and quantitative parts of the reports. Those discrepancies, after additional investigation, led to the discovery of indications or insights about companies' future financial performance. The initial study, conducted in January 2002, was later followed by a further constructive study that aimed at building a methodology for predicting future financial performance from companies' financial reports using the prototype matching method and neural networks. The prototype matching method was used to compare text in financial reports, and this is how computers were utilized for mining text to form meaning from data and produce knowledge.

Another exploratory study was undertaken in *Paper 3* on the applicability of the prototype matching method to retrieve scientific publications according to their content. The first step in this branch of the research focused on a smaller data collection – abstracts of the papers submitted to the Hawaiian Conference on System Science 2001. Because of the limited vocabulary and similar sentence

constructions used in the abstracts, I decided to expand the researched text collection and perform a study on the full-paper versions by proposing a prototype-system for IR by content from scientific publication domain. In the scientific text, I compared papers to connect similar papers to each other, in order to help a conference organizer allocate similar papers to the same session. In this sense we again give meaning or form to data.

By using small data samples, much external validity of the research was, without doubt, sacrificed. Nevertheless, considering the exploratory nature of the studies, the use of limited textual data collections seemed justifiable, especially since quarterly reports and scientific abstracts are real-life textual data collections obtained from the Internet and easily available in the same format to the potential users of prototype-matching based systems.

2.3.3 Constructive Research: Papers 4, 5, 6 and 7

Drawing on work by March and Smith (1995), Iivari, Hirscheim et al. (1998), and Järvinen (2001), design science can be pursued by constructive (or developmental (Nunamaker, Minder et al. 1991) or engineering (Clarke 2000)) research, that is defined by Kasanen, Lukka et al. (1993) as "problem-solving through the construction of organizational procedures or models." According to Järvinen (2001), it is typical for constructive research that a new artifact is built in a process based on existing research knowledge or advancements, and that the utility of the artifact is – sooner or later – evaluated. Moreover, Järvinen distinguishes artifact-building and artifact-evaluating approaches under the heading of constructive research. Constructive research may be conducted by using both quantitative and qualitative methods, or a combination of those and has a normative nature, as opposed to positivist, because it is an "inherently goal-directed problem solving activity" (Kasanen, Lukka et al. 1993).

Järvinen (2001) supports the vision on constructive research with a research framework in IT proposed by March and Smith (1995), which is driven by a distinction between design science research outputs or artifacts (*construct, model, method and instantiation*) and research activities (*build, evaluate*, *theorize* and *justify*) (see Table 2-3). The research framework, in their opinion, should be concerned both with utility as a design science, and with theory explaining how and why IT works as natural science. According to their definitions, constructs or concepts form the vocabulary of a domain to constitute a conceptualization for describing problems and specifying solutions. While achieving the objectives of the dissertation I follow the artifact-building approach of constructive research by building models, methods, and instantiations.

A model is a set of propositions or statements expressing relationships among constructs. A method is a set of steps (an algorithm or guideline) used to perform a task. An instantiation is the realization of an artifact in its environment, or in other words, operationalization of a construct, model, or method. Research activities in design science are *build* and *evaluate*, where build refers to the construction of an artifact and the demonstration that it *can* be constructed, while evaluate refers to the development of a criteria for assessing an artifact's

performance. Research activities in natural science are *discover* and *justify*. Discover (or theorize) refers to the construction of theories for explaining how and why something happens, i.e. explaining how and why an artifact works within its environment. Justify refers to theory-proving by gathering scientific evidence to support it. March and Smith claim that "the research contribution lies in the novelty of the artifact and in the persuasiveness of the claims that it is effective. Actual performance evaluation is not required at this stage." However, March and Smith (1995) state that research in the evaluate activity requires the development of suitable metrics and propose the following evaluation criteria for evaluating models: *their fidelity with real world* phenomena, *completeness*, *level of detail*, *robustness*, and *internal consistency*. Järvinen adds the following evaluation criteria: *form and content*, *richness of knowledge*, and *user experiences*. In the constructive approach, the value and the actual working of the construct have to be shown (Kasanen et al. 1993; Järvinen 2001). In this dissertation, I propose the employment of the prototype matching methodology as a part of a system for financial benchmarking (see Section 7.3.1) as well as a part of a system for financial forecasting (see Section 6.2.2). I evaluate the applicability of the prototype matching method based on the fidelity and adequacy of the instantiations to real world phenomena (how well the results from the proposed instantiations reflect reality), and richness of knowledge provided (how much information about reality do they provide). Existing information from text as well as industry specific literature is used to verify the findings of the tool and framework. The completeness, level of detail, robustness and internal consistency of the prototype matching method should be judged from mathematical, algorithmic standpoints, which are not my primary focus. The value of actual functioning and robustness of the methodology is established by the richness of knowledge extracted from the textual collections.

Table 2-5. A research framework in design and natural science (adapted from (March and Smith 1995))

Research activities/ outputs	Design Science		Natural Science	
	Build	***Evaluate***	***Discover***	***Justify***
Constructs				
Model	*Paper 5*		*Paper 5*	
Method	*Papers 6,7*	*Papers 6,7*		
Instantiation	*Paper 4*	*Paper 4*		*Paper 4*

Paper 4 draws on the findings from the series of exploratory experiments conducted in *Paper 3* to investigate the applicability of the prototype matching method for retrieving semantically similar papers from a collection of scientific publications originating in the HICSS-34 conference proceedings. I discovered in *Paper 3* that IR by content from the abstracts of scientific publications gives fairly low precision results for allocating abstracts of the papers from the same tracks and slightly higher precision results for allocating abstracts of the papers from the same cross-track themes (see Section 8.2). Although the nature of the relatively low number of the retrieved publications from the tracks or themes remained unclear,

the proposition was put forward to use the prototype matching method for IR by content from the full-paper versions from the scientific conference proceedings. Following constructive research manner, I built a prototype system (instantiation) that offers the potential user the opportunity to retrieve semantically similar papers by inputting into the system the paper from a conference collection that he has an interest in. *Paper 4* suggests a prototype version of the system that can be used for retrieved scientific papers from conference databases based on semantic similarities between submitted papers. For evaluation of the developed instantiation, the retrieving results were compared with the real-world conference organization, assuming that conference tracks or themes unite the most semantically similar scientific papers.

Based on the promising findings from the exploratory research *Paper* 2, a number of instantiations were created (see Figures 6-3, 7-1). A framework for justifying the fact that TM results contain indications upon which the future financial performance of a company can be determined was suggested in *Paper 7*. Another framework for validating TM results linguistically was proposed in *Paper 6*. *Paper 6* proposes a way to validate the semantic similarities in clusters of financial reports obtained from prototype-matching clustering of qualitative parts from companies' quarterly financial reports (summarized in Section 7.3.3). Collocational networks, as a linguistic method for outlining the central concepts in a text, were used for cross-validation of TM results for every report investigated in *Paper 2* in an artifact-evaluating manner. The level of resemblance among the collocational networks of a prototype-report and its closest matching reports determined by the prototype matching method has established the criteria for TM results evaluation. The level of resemblance among the collocational networks is established by the similarities in network outline and the number of similar collocates in the networks. *Paper 7* proposes a method for predicting future financial performance from the qualitative and quantitative parts of financial reports based on the prototype-matching method and neural networks in the form of the SOM and the backpropagation neural network. The hidden indications of future financial performance coded in co-occurrences and financial positions of the report's closest matches from the exploratory study (*Paper 2*) were justified by the use of the more objective neural network technique, rather than human judgments. I used the prototype matching to transform a lower level data representation – numeric and textual parts of quarterly reports – to another higher level knowledge representation – prediction of future financial performance of an analyzed company. The evaluation of the proposed methods was measured in terms of the prediction accuracy of the neural networks, by comparing actual financial performance of a company with the predicted one.

In research *Paper 5* a conceptual model of a knowledge building system for processing numeric and textual data was proposed. A model, from the design science point of view (March and Smith 1995), is a description of how things are. Models represent situations as problem and solution statements. Kasanen, Lukka et al. (1993) emphasize that all problem-solving exercises do not pass as constructive research, arguing that the novelty and the actual working of the solution have to be demonstrated as well to differentiate the approach from analytic model building.

With reference to this, it is important to stress the fact that the original research *Paper 4*, which presents the structure and possible usage (see Section 6.2.3) of the created artifact, does not include empirical evidence concerning the functioning of the agent application, because the provision for such an experiment was beyond the scope of the research. The research neglected the constructing of artifact as a whole following artifact building constructive approach. Nevertheless, parts of the artifact, such as data classifying, and data and text mining clustering agents were tried out and compared to other clustering methods in a number of research papers (Costea, Kloptchenko et al. 2001; Costea, Eklund et al. 2002; Costea and Eklund 2003).

Following the procedures of a research framework, such as choice of problem area, creation of research design, and interpretation of the obtained results, the original research papers used different research approaches for executing different procedures. For instance, the exploratory study in *Paper 3* concerning discovering hidden patterns in financial reports, served the purpose of choosing appropriate methods for clustering numeric data and data coding. They were further used in the constructive research in *Paper 7* for creating a method for predicting the future financial performance of companies based on the qualitative and quantitative data from their financial reports. The additional value of creating insights, or a fresh unexpected perspective about the level of future financial performance, was achieved by combining descriptive, exploratory and constructive approaches in a pluralist manner. The important information deducted from low level data (text and numbers from financial reports) can be delivered to the potential decision makers, managers, or knowledge workers.

3. INFORMATION SYSTEMS AND TECHNOLOGIES TO SUPPORT MANAGERIAL WORK IN DEALING WITH TEXTUAL INFORMATION OVERLOAD

One of the objectives of this dissertation is to offer a new technological method for reducing textual information overload. With the use of the prototype matching method, I aimed at saving managerial and knowledge workers' efforts in detecting important patterns in textual data that can bring additional insights to the facts, which are stated in text collections. Although text is intuitively easier to interpret and understand than numbers, textual information overload causes frustration and productivity slowdown for managers, decision makers, and knowledge workers. IT significantly alters the traditional concept of writing, causing textual data duplication and overload. Here I make an account of the latest available and relevant IT that both contribute to a problem of textual data overload and to the solution of it. These technologies, when properly combined, provide IS solutions that managers or decision makers can use to extract valuable nuggets from the piles of digital documents and written texts. Some particular IS solutions offered by leading software companies and research teams for the purposes of TM and reducing textual data overload are reviewed in this chapter.

3.1 Nature of Information Overload

The problem of information overload is rather serious because it results in wasted time and lost opportunities. Information overload is characterized in two ways: first, when individuals are given more information than they can absorb; second, when demands on an individual's time for performing interactions exceed the supply of time available for such processing (Farhoomand and Drury 2002). As McGovern (2002) notes, IT has become the Trojan Horse of information overload. As IT is being introduced into the organization as some magical gift to bring greater efficiency and reduced cost, it feeds on resources and spews out unimaginable quantities of low quality data. As E. Sachs, VP of research and development at Wolf Communications in Houston, a provider of network services for Notes users, remarks, on a typical day, he answers 20 telephone calls, replies to 60 e-mail messages, monitors six news services and filters 12 others, participates in several online discussion forums (three on an hourly basis), and gets two reports from a news retrieval service (Foley 1995). It was found that two thirds of managers suffer from increased tension and one third from poor health because of information overload. The negative effects of too much information include anxiety, poor decision-making, difficulties in memorizing and remembering, and reduced attention span (Heylighen 1999). Primarily, knowledge management systems and office automation systems are designed to get the right information into the right hands, and block out unnecessary. However, they accumulate duplicated data and result in data tombs.

Instinctively, overload can be fought either by abstracting knowledge workers or managers from the situation that causes overload by delegating

information activities, or by training other individuals to handle their personal information flows (Hall 1998). However, hiring another expert-librarian to sort relevant document into categories can be a good solution for senior managers and a nonexistent option for a regular knowledge worker. Another option is to train knowledge workers to hold files of gathered information and pointers to save time and effort of retrieval in the future. To avoid data paralysis they should change their information need attitude from "maximisers" to information "satisfiers" (Driver 1993). Whereas maximisers aim to uncover every detail from data prior to action, satisfiers use just enough information to make a decision. Beneficial for information overload reduction will be a refinement of communication skills by the information publisher or editor, who should pre-analyze, filter, and distribute only valuable details. Conversely, such careful text preprocessing by the authors can potentially limit information richness for reader. IT bounded in IS is called upon to ease information overload by carrying out various processes cheaply, quickly, and effectively. The body of literature applicable to the topic spans computer and information science, human-factors engineering, management science, and social sciences.

Since we live in an information society where the economy is driven largely by information, companies that are able to first acquire and apply information efficiently are likely to be more successful. Checkland and Holwell (1998) pointed out that the key element in information system development is a rich initial study of meanings and purposes that should be essential in order to arrive at the served system, which the data processing will support. Moreover, once the action to be supported has been established, it should be decided whether computer support should take the form of either one or both of the following: automating action which is currently being carried out by people, or by providing 'informational' support to people as they carry out tasks. In the latter case, two kinds of informational needs are to be considered: information that will help people who take the desired action, i.e. decision makers, and/or information that will help them to monitor and control. IS that perform the described actions belong to decision support IS and can be realized by means of expert systems (Back, Toivonen et al 2001), computing with words-systems (Zadeh 2000), etc. The systems that aim at helping to decide what action should be taken and what process should be supported can be classified as knowledge discovering systems, which offer data exploration and TM capabilities. The emergent combination of decision support and knowledge-discovery systems can benefit decision makers in the form of decision advisory systems that can provide hints and insights into what action or process should be explored, why, and how.

3.2 Information Technologies

IT includes all the devices that use electronics to gather, communicate, process, display, and store information. IT not only changes but also transforms our lives, the way we communicate, work, learn, play, and move around. Technology is fundamentally changing the way and the speed of doing business by companies. Consolidation, globalization, and deregulation available through the advent of

network and database solutions put pressure on managers to better understand their business and shorten decision-making cycles. Technological solutions change radically over a short time: the technology, which is available at the end of a doctoral dissertation research is often more advanced than it was in the beginning. In this section, I make an account of the year 2003, although the research work was done with alternative technology available earlier. In terms of technology contribution (creating and reducing) to textual information overload I identify four recent basic trends: (1) networking communication technology, (2) intelligent technology, (3) database technology, and (4) wireless technology. This categorization is not hierarchical or exclusive; some technologies, such as software agents can belong to networking database technology and be intelligent at the same time.

Network communication technology. Many companies gained a major competitive advantage by utilizing the Internet and networked solutions in their business processes throughout their entire value chains. Software systems that used to be relatively autonomous entities, such as accounting systems and order-entry systems, are now interlinked in large networks comprising extensive information infrastructures. Besides the undelivered hope of electronic commerce, the worldwide advent of the Internet provided a number of possibilities to support internal organizational processes with new types of IS. Integration is the big word in network technology. Examples include collaboration technologies, such as Lotus Notes from IBM, and messaging technologies, such as e-mail and file-sharing software. HTML, XML, IP, and LAN standards enable and accelerate the growth of networks. Network solutions present distributed access to common knowledge repositories accessible via Extranets or Intranets. Networking communication technologies: the Internet, e-mail, and faxes have accelerated information generation and duplication. In order to combat information overload, organizations design and implement search engines (such as www.overture.com for companies or www.yahoo.com for individuals) into retrieval tools.

Intelligent technology. With the lower cost of increasing computing speed, intelligent or smart technologies are becoming widespread. These technologies acquired the most advanced algorithms from AI: expert systems, neural and belief networks, genetic algorithms, fuzzy logic, probabilistic reasoning, NL processing, hybrid systems, and static and mobile intelligent agents. These algorithms are implemented into diverse decision support systems for resolving ill-structured problems that are impossible or costly to solve using, for instance, operation research methods. Intelligent agents or software agents, or personal information agents, brought a new perspective into the AI and machine learning fields. For instance, Nardi (1998) describes the collaborative, programmable intelligent agents implemented by Apple Data Detectors that are able to extract semantics from everyday documents. However, the discussion about the rightness of the definition of intelligent agents along the lines with the discussion about intelligence of any technology is ongoing. Various intelligent technologies can be bound together for designing information-customization software that helps to deal with information

overload. An example of this is software can be found behind personalized news sites. Intelligent technologies have empowered search engines with the ability to construct more flexible queries run on inference engines with built in fuzzy logic (www.google.com).

Database technology. The preconception of developing the database technology was to provide a single data source for data processing in transaction-oriented systems. Since the dawn of database in the 70s, database technology has undergone a lot of changes. The goal of database technology is to reliably, efficiently, and consistently store large amounts of data in a certain location. Today's data warehouses are integrated databases that provide current and historic data, as well as detailed and summary information (Neary 1999). Summary data is a great way for managers to see the bigger picture of their business quickly and use detailed data for further detailed analysis. Oracle database solutions and OLAP (On-line Analytical Processing technology) are acquired by many organizations worldwide. Intelligent and network technologies can be used to empower the access, retrieval, and presentation of information from the database. Poorly maintained and updated databases cause duplication of data in various formats and information overload turning warehouses into data tombs (Fayyad and Uthurusamy 2002). KDD as an individual research stream has grown out of the database engineering community (Piatetsky-Shapiro 2000). The algorithms that provide assistance in managing and understanding data in the databases belong to DM and TM.

Wireless technology. Mobile devices, such as personal digital assistants, communicators, laptops and mobile phones have shrouded the planet with new mobile networks since the end of the 1990s. Information availability anywhere, anytime, provided by such devices, promises to bring more pressure of fast business decisions making. New technological standards (Wireless Application Protocol (WAP), Bluetooth, MexE, Wireless Local Area Network etc.) are accelerating mobile network technology, mobile service technology and other technologies to support cooperation of mobile devices with their computing environment, the Internet or organizational databases. In terms of state of the art, WAP and GPRS (General Packet Radio Switch) represent the most innovative technologies widely available for real-time mobile users. Problems of data security have arisen in the discussion of new mobile networks. Because the adoption and reliability of such light-weight devices are not strong yet, managers and knowledge workers tend to copy information from their mobile devices on to their desktop ones or server, creating even bigger information overload and confusion concerning old and updated information.

3.3 Information systems

IT alone cannot deliver their full potential unless they are entered in a context of reality to create IS. The implementation of IS has implications on the work of managers and knowledge workers who deal with textual information overload on a daily basis and need to process this overload to perform their tasks.

Various applications of IT in IS by supporting different text-related operations (see Figure 1.4) can impact greatly on information overload reduction. I have investigated the following IS that deal with information overload: transaction processing systems, management and decision support information systems, office automation systems, and knowledge management systems.

Transaction processing systems (TPS) are recognized as the operational IS of the organizations by (Beynon-Davies 2002). Transaction processing systems function on operating data, including order entry, accounts payable, stock control reports, customer and purchase orders, etc. The advent of TPS came as a result of the "data-culture" that has prevailed in organizations since computers achieved the ability to perform hundreds of operations a second to decode, deliver and store data. TPS using networked and web-based solutions aims to integrate all parts of the value chain. TPS are sometimes referred to as the life-blood of the organization because they hold the data about all critical and effective organizational activities. Lately, Enterprise resource planning systems (ERP) united various separate TPS systems, and became popular for middle and large-sized companies, e.g. SAP software. One of the main ideas for ERP systems is data integration from separated TPS throughout the supply chain. This integration results in huge data masses with some duplication of the same data in various formats (O'Leary 2000). An example of textual TPS could be CRM (customer relationship management systems) that tend to accumulate mostly textual data in form of customer reports and complaints.

Management and decision support information systems (MIS/DSS) enable effective short-term tactical and long-term strategic decisions concerning operations and a strategic development of organization. DSS generally utilize management data generated by MIS to model short-term scenarios of company performance. A properly designed DSS is an interactive software-based system, intended to help decision makers compile useful information from raw data, documents, personal knowledge, news, and business models, in order to identify and solve problems and make decisions. The number and quality of technologies that can be used to support managerial decision-making has expanded thanks to increased speed of computation, processing and storage capacity, cost effectiveness, and quality support (Tetard 2002). The technologies of intelligent agents, neural networks, fuzzy logic, DM and data warehousing allow the creation of up-to-date and effective expert systems, DSS, and group support systems. Designing useful DSS systems that facilitate intelligent interaction between a user and the system is a cumbersome task. The majority of DSS acquired operations on NL expressions and allows the user to visualize analyzed data in the form of graphs and time sequences.

Office automation systems (OAS) were introduced at the end of the 1950s as a result of developments in telecommunications and computer systems (Wainwright and Francis 1984). In the 1980s the concept of OAS started to evolve from the "paperless office" towards integrated office automation systems as soon as the first desktop computer allowed the performing of multiple operations, such as concurrent use of word and spreadsheet applications. Modern OAS benefit from the integration

of communication channels, access to several data sources, information space sharing, synchronization of various information processing devices, and applicability of software agents. OAS started to support project management, collaborative authoring, and large-scale knowledge based development along with office work (Mahling and Craven 1995). The Microsoft Office products and Lotus Notes from IBM are good examples of integration of OAS in different office-related processes.

Knowledge management systems (KMS) support processes that include knowledge creation, knowledge codification and knowledge transfer (Davenport and Prusak 1998). Contemporary KMS heavily depend on the development of data warehousing and data and text mining technologies; the first one is for storing and organizing data, the last one is for knowledge creation from the available data. KMS utilize networked technological solutions, wired or wireless, for collaboration and knowledge transfer. An example of KMS can be a system proposed by NovoSolutions Inc. that supports employee training, document management for governmental or consulting organizations, and online customer service. KMS are coming into play in the information society with a new "knowledge culture" in business. P. Drucker once said "in the post-capitalism, power comes from transmitting information to make it productive." Knowledge is a more complicated phenomenon than data, and is consequently difficult to capture and/or discover, create, store, transfer, and enter into effective business processes. DSS and KMS can benefit from each other and offer decision advisory systems to managers. These systems will run on various processed data sources and choose the optimal scenario for company to follow.

All of the information systems described above can be empowered by web-based (network) or wireless technologies to assist managers or knowledge workers in delivering information anytime, anywhere. According to (Nyberg 2001), modern knowledge management system reduce information overload by helping to find the most relevant, most useful data. I argue in the thesis that IS built on the prototype matching method fight information overload not only by retrieving and delivering the most relevant crude data by performing IR by content, but also by producing sophisticated precise information from the crude data and delivering it to the user.

Multi-agent systems are concerned with coordinating intelligent behavior among a collection of constructed autonomous software agents. Those technologies can be used in KMS, DSS, and TMS to help users filter unnecessary information by exploring personalized intelligent agents. For example, AOL, Yahoo, BBC, and Dow Jones websites offer the opportunity to personalize their new portals, and www.msn.hotmail.com allows the setting up of spam filters to refine incoming e-mail messages on user's behalf. Document brokering with agents proposed by the "Persona" approach provides an appropriate web document presentation for different users or user groups on the fly (Suzuki 1998). Intelligent agents can provide active executive support to the managers by accomplishing business environment scanning, which often consists of business reviews, incident reports

and new streams (Liu 1998). Mladenic (1999) analyzed commercially available intelligent agents that perform text learning.

3.4 Technologies for Text Mining

TM solutions tend to automatically provide an overview of the documents, in order to grant a user overall understanding of what the text documents are about, without the need to read them. Visually, the TM framework consists of three parts. Very generally the TM practice can be represented as a sequence of the following processes:

1. *Text representation and distillation* transforms and represents free-form text in a chosen format and/or consolidates documents from various sources. Text from a string of symbols has to be encoded in some numeric format within a document or through document collections, i.e. words are represented as numeric vectors or ranges. Encoded text from every individual document is transformed further into lower dimensional and more appropriate for computer formats. This transformation is achievable via different procedures, such as word stemming, i.e. to only use the canonical form of a word (i.e. analyzed, analysis, analyzing – "analy"); word disambiguation, to determine which of the senses of an ambiguous word is invoked in a particular use of the word in order to fight both word polysemy and synonymy by constructing a dictionary of word senses (i.e. the word bank in a financial document means a depository or financial institution, but not a sloping land); and stop word removal, to exclude words that do not contribute to document meaning (i.e. the, it, a)
2. *Knowledge sophistication* deduces concepts (tokens[6]) and patterns from the distilled text by utilization of knowledge discovery algorithms. A document or entire document collection can be clustered, categorized, or visualized to reveal interdocument or interterm relationships. The extracted features from a document or collection can be summarized to present new knowledge to a user.
3. *Knowledge (relationship) representation* that delivers and presents the deduced knowledge to a user. The discovered relationships from the previous part are presented in some graphical or other visual form that a user can easily interpret (i.e. lists and tags, hierarchies, hypertext diagrams, semantic maps, tables or matrixes).

A general TM framework is presented in Figure 3-1, where the text distillation step is concerns consolidating documents into a sub-base, filtering, and converting them from various formats into one acceptable by the TM method, such as conceptual graph representations or vectors. The knowledge sophistication step identifies and extracts concepts and patterns from the distilled documents, which supposedly capture the main meaning of text in the form of tokens and consolidating discovered patterns into evocative knowledge. The former can be achieved via clustering, categorization, or associative

[6] I adopt M. Hearst's term of tokens as meaningful concepts or terms (Hearst, M. A. 1997).

discovery methods. In most of the available TM programs tokens are picked out by calculating word frequencies and co-occurrences. Tokens are chosen according to Zipf's law and collocations. Collocations are frequently used phrases such as fixed expressions accompanied by certain connotations (Hearst 1997). The knowledge representation step concerns retrieving the documents, which have notable tokens or patterns in them. This part visualizes documents with tokens and patterns and/or interrelationships among them by building semantic maps, hierarchies, summaries, or hypertexts with highlighted links.

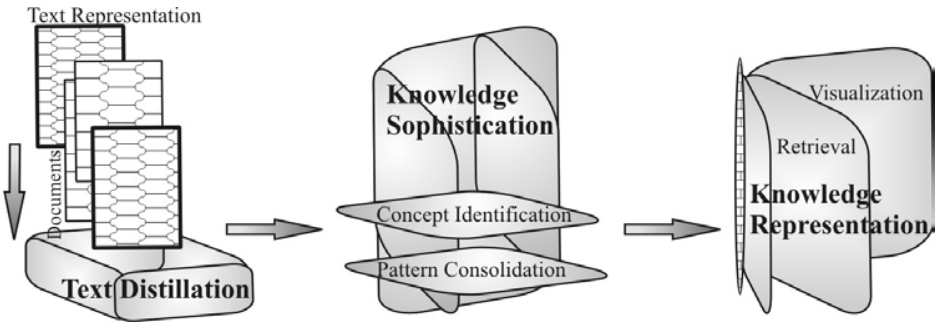


Figure 3-1. General TM framework

Table 3-2 has been compiled to enumerate the best-known TM systems at the end of 2002. It shows the main research endeavors from the corporate and academic worlds. The first four systems in Table 3-2 were developed and implemented at different universities. Companies specialized in data and text mining solutions for business intelligence, such as Megaputer Intelligence Inc. and Temis, as well as large corporations, such as SAS Institute and IBM Corp., designed the rest of the systems presented in Table 3-2. The descriptions of the software systems are borrowed from the websites and white papers of their developers using their descriptions. The technologies from universities came from academic publications, while the commercial technologies came mostly from promotional websites, news portals, and business journals. The absolute majority of the presented TM applications described in Table 3-2 use stemming, synonym list[7], stopword[8] removal, text parsing[9], dimension reduction[10], and clustering, or categorization procedures of the coded data. Some of them are query-based

[8] Synonym table composition assigns one meaning to every term used in document vocabulary (Riloff, and Hollaar, 1996).
[8] *Stopwords* are frequently appeared words among the documents, i.e. articles, prepositions, and conjunctions. Elimination of stopwords reduces the size of the indexing structure. (Baeza-Yates. and Ribeiro-Neto 1999)
[9] *Text parsing* algorithms convert text into word, phrases or clauses and put them into short memory. Text parsing decomposes text and generates a quantitative representation suitable for DM (Mayes,. Drewes et al., 2002).
[10] Dimension reduction algorithms treat every document as a vector where each dimension is a count of occurrences of a different word. It results into tens or hundreds of dimensions of every document (Isbell, 1998). . Independent Component Analysis and restricting weighting to the words specified in the query only are applied for dimension reduction (Kolenda and Hansen 2002). . The importance of executing word sense disambiguation based on a thesaurus or on-line dictionaries for increasing performance of IR systems was discussed by Sanderson (1994).

methods, which rely heavily on the use of term (keywords, items, indexes) extraction, i.e. SONIA, TextMiner, Sapere. The limits of performing TM tasks with query-based methods are given in Chapter 4.

Table 3-2 depicts an illustrative list of TM products and applications based on knowledge sophistication and knowledge representation, as well as the status and domain of the development.

Table 3-2. TM Systems

System Name (reference)	Company Organiza-tion	Status/ Domain	Knowledge Sophistica-tion	Knowledge Representa-tion	Approac h
SONIA (Service for Organizing Networked Information Autonomously) (Sahami, Yusufali et al. 1998)	Stanford Digital Libraries, Stanford University	Working on testbed/ Digital Library	Dynamic hierarchical document categorization based on full-text articles	Navigation in topical information space	Multi-tier extracting of terms, statistical clustering, Bayesian classificati on
TextMiner 1.0 (http://www.cs.sfu.ca/~ddm/dmsoft/Clustering/tm_index.html)	DDM Lab, Simon Fraser University	Working Prototype	Hierarchical clustering by measuring the similarity using large items		Text clustering on notion of frequent items
WebSOM (Kohonen 1999)	Helsinki University of Technology	Working prototype	SOM Clustering	Visualizati on, Navigation	(Online) Text clustering, mapping
Sapere (http://www.ai.mit.edu/research/abstracts/abstracts2002/natural-language/04katz.pdf)	Artificial Intelligence laboratory, MIT	Prototype/ Store Knowledge and Answer Queries	Representatio n of linguistic relationship as ternary expressions	Retrieval	Using ternary expression to facilitate easier indexing
Text Analyst 2.0 (http://www.megaputer.com/products)	Megaputer Intelligence Inc.	Commercial software package for Engineering, Educational, Customer Documents	Summarizatio n, Clustering, Semantic Information Retrieval. Topic structure explication	Navigation Visualizati on	Creating concept-based Semantic Hierarchica l Neural Network
CINDOR (Conceptual INterlingua DOcument Retrieval) (http://www.textwise.com/government/)	TextWise Inc.	Commercial software	Query-term expansion.	Query-based retrieval	NL retrieval in several languages

VisualText 1.5 (http://textanalysis.com/body_index.html)	Text Analysis International Inc.	Commercial software /Text mining, Knowledge Engineering	Information Extraction, Categorization, Summarization, Parsing	Information Extraction, Visualization	Text Analyzer, NL query, converting text to XML, SQL, text from speech
Docu/Master XML (http://www.dsisolutions/home3f.htm)	Document Systems, Inc. and Hostbridge Technology	Commercial mainframe software/ Knowledge Management	Online Retrieval and Integration	Retrieval	Creating Document Metafile, Complete full-text search, Synonymy Tables, Indexing
DolphinSearch (http://www.dolphinesearch.com)	Dolphin Search Inc.	Commercial/ Management System for Law Department Intranets and e-mails	Extraction of keywords to form semantic profile of every individual document	Retrieval	Assigning keywords to the stored Intranet documents and e-mails. Represent texts by semantic profile
Insight Discoverer: Categorizer, Clusterer, Extractor (http://www.capital-k.com/text_mining.php)	Autentica/ Temis	Commercial software/ Search Result Organization, Document Mapping	Categorization, Clustering, Information Extraction	Visualization	Using trained classification model of document collection
Online Miner (http://www.temis-group.com)	Autentica/ Temis	Working Intranet application Knowledge/ Management, Financial Market Analysis, Competition Intelligence	Clustering, On-line Categorization from Insight Discoverer: Clusterer and Categorizer	Visualization, Retrieval	Stores and analysis of online large document collections
TextQuest 1.5 (http://www.textquest.de/tqe.htm)	Textquest	Working/ Content Analysis	Categorizing by key-word-in-context	Retrieval	Text Analysis

Autonomy (http://www.autonomy.com/Content/Products/IDOL))	Autonomy Inc.	Commercial Software Solutions for Unstructured Information/Business Intelligence, Email Routing, E-commerce, ERP and Knowledge Management, Content Publishing	Bayesian Inference on pattern-matching	Retrieval	Conceptual Search, Language Independent
LexiQuest Mine, LexiQuest Categorize, LexiQuest Guide (http://www.spss.com/home_page/wp130.htm)	SPSS Inc.	Commercial software package/ Integration into Legacy Systems, CRM, Investment Research, e-mail filtering, Combining Data and Text Mining for Business Intelligence	Categorization, Term Extraction, Statistical Proximity Matching	Retrieval	Proprietary Language Recognition based on 600000 word dictionary instead of keyword queries, NLP
digiMine Customer Analytics, Predictive Recommen-dations, Retail Advisors (http://www.digimine.com/solutions/howdigimineworks.asp)	digiMine Inc.	Commercial software package/ Customer Analytics and Customer Interaction Optimization from emails, databases and web logs	Collects Data from various sources, parse and clean the data, produce analytical reports based on business profiles,	Analytical Assertion, Results Viewer, Visualization	Provides cross-sell recommend-dations, runs on analytics of web activity
Bullfighter (http://www.dc.com/insights/bullfighter/index.asp)	Deloitte Consulting	Commercial plug-in into Microsoft Word and Power Point	Establishes linkage between readiness of financial reports and business performance	Issues Bull Composite Index to measure readability of a report, recommends improvements	Based on a dictionary of 350 "bull" overused words in business
SemioMap (Semio_Corporation 2001)	Semio Corp.	Commercial Software/ Graphical Knowledge Extraction by concept	Categorization	Map, 3D Visualization, Navigation	Creating multi-layer concept maps for searching

IBM Intelligent Miner for Text (Tkatch 1997)	IBM Corp.	Commercial software package/Knowledge Discovery, Information Mining Solutions	Hierarchical, binary relational clustering, Centroid neighbor classification, Feature Extraction	Retrieval, Browsing, Visualization	Language Identification, Finding Similarities based on Lexical Affinities
SAS Text Miner as a module in Complete Data Mining Solution (http://www.sas.com/technologies/textminer)	SAS Institute	Commercial software package/ e-mail filtering, Routing news, Document Routing, Predicting of stock prices, etc.	Categorization, Hierarchical and Expectation-Maximization Clustering, Feature Extraction	Result-viewer Visualization	Concept-based analysis of document collections

One group of products described in Table 3-2 focuses on document organization, visualization, and navigation-retrieval in a document database. Another group focuses on text analysis from documents, notably information retrieval and extraction, categorization, clustering, and summarization, to provide text understanding to a user. The leaders of the software revolution, such as IBM, dictate the current trend in TM technologies. Recently, IBM research has offered WebFountain – an open platform for unique text access with "multidisciplinary text analytics to provide the power of analysis of vast data stores to produce valuable business insight" (www-1.ibm.com/mediumbusiness/ venture_ development/ emerging/wf.html). Integration of successful clustering and categorization algorithms for TM with easy to interpret representation of the results is a tendency of the described above commercial and academic systems. Another tendency is to enable systems for TM to work on-line in real time with different forms of textual data. The upcoming goal for data management software is "to find that bit of information no matter where it is, whether it's in spreadsheets, audio form, a Word document, plain old flat files, or a database engine " and to help computers understand human emotions" coded in information (Chase 2003).

Moreover, the big corporations with commercially available software packages, such as IBM, SAS, and SPSS offer the combination of data and text mining solutions in the form of toolboxes or add-on modules of various applicable algorithms, i.e., hierarchical and k-means clustering (described in Section 4.2.3), etc. Notably, TM modules in those packages were introduced only recently as an individual exploration tool for different types of unstructured data. In other words, TM discovers the patterns that can serve as hints to unlock the knowledge contained in test data so that it can be combined with data from numeric databases to build better models. Consequently, binding sophisticated DM and TM algorithms requires very specific mathematical and domain expertise from those who wish to apply them for effective problem solving.

TM technologies can be used as parts of advisory or decision-support systems. TM technologies aim at enhancing the support provided by decision support systems that, according to Doukidis, Land et al. (1989), traditionally assist the users in problem structuring and exploring lines of analysis, without offering solutions. For instance, TM can be extensively used for spam filtering to reduce information overload. The spam classifiers and filters that pay somewhat more attention to structural than linguistic features are likely to fail as spam tries to look more legitimate.

One of the main open questions in all these applications is how to integrate domain knowledge with the results of TM tools. As Tan (1999) noticed, domain knowledge can be used in a process of knowledge discovery from text as early as in the text refining-distillation stage. The interpretation and evaluation of the discovered patterns are still cumbersome and include intensive human involvement. The requirements for well-trained users who can interact with TM systems are still obligatory. Managers, who are heavy consumers of textual information in their work of decision making, very rarely have the time or technical expertise to master complicated TM applications and to gain the experience to recognize valuable discovered patterns. In the current work, I have tried to create systems using a novel TM method, namely the prototype matching method, in conjunction with other DM algorithms, i.e. the SOM (see Chapter six and *Paper* 2) to simplify the steps of interpretation and knowledge consolidation from the discovered patterns, which are difficult to interpret for non specialists of domain knowledge.

The current chapter has aimed at resolving research objective *a* from Section 1.3 – explaining the nature of the relationship between textual information overload and the technologies that contribute to its occurrence and reduction. First, I gave the definitions of data, information, and knowledge that form the foundation for the dissertation. Second, I explained the nature of textual information overload, and IT that cause its occurrence and reduction. The general TM framework, which consists of text distillation, knowledge sophistication, and knowledge representation, was described. Third, I provided the list of modern TM systems based on the approaches used for knowledge sophistication and representation, and the common trends in the development of those systems.

4. STATE-OF-THE ART IN TEXT MINING

In Chapter three, I have presented the recent developments in the fields of IS and IT that contribute to both creating and solving the problem of textual information overload. In this chapter, I summarize the state-of-the-art of what is known concerning TM approaches and algorithms, and their classification according to the purposes they serve. The body of literature applicable to the topic spans computer and information sciences, human-factors engineering, statistics, NL processing, and management science.

Below, I describe TM algorithms that are used for knowledge sophistication. They differ on the basis of the users' information needs they attempt to satisfy (whether a user can formulate a query explicitly or not) and the dimensions of a textual collection to be mined (single document or a collection of documents). Later I discuss the issues associated with the representation and visualization of TM results.

4.1 Text Mining Approaches

TM approaches can be theoretically divided into the following categories depending on the scale of text to mine (one document or an entire collection) and the user's knowledge of desirable information to be discovered (if the user can explicitly formulate information need or not). Table 4.1 represents TM categorization according to the mentioned scale and scope, where, for instance, IR by content and categorization approaches belong to quadrant III, and clustering and summarization approaches belong to all the quadrants.

Table 4-1. Categorization of TM tasks according to scale and scope

User's request:	**when user knows what he wants to discover**	**when user does not know what is to be discovered**
within a document	I	II
within document collections	III	IV

Below I briefly present various knowledge discovery algorithms and studies where they were implemented. The order of the algorithms follows the categorization of the TM tasks and purposes described above. I devote special attention to the related studies in IR by content and text classification in the form of clustering, since those TM tasks are accomplished in my studies using prototype matching for TM of a collection of financial reports (see Chapter six) and a collection of scientific publications (see Chapter eight). Text categorization and clustering belong to text classification, which is defined as the problem of mapping a document to a set of topics (Scheffer and Wrobel 2002). I discuss concisely the differences between clustering and categorization.

4.1.1 Information Retrieval by Content

IR is the oldest and most established field in text processing that, according to Hand, Mannila et al. (2001), can be regarded as subtask of TM. IR deals with "the representation, storage, organization, and access of information items" (Baeza-Yates and Ribeiro-Neto 1999). To perform IR, the user is supposed to have an idea in mind about what question should be answered by retrieving what kind of information. The IR system does not directly provide an answer, but rather points a user to an appropriate document. It is assumed that the user has a classification system in mind that separates the relevant documents from nonrelevant ones. Successful IR systems aim to discover and characterize this dichotomy to assist users in achieving their information needs in a search process. Traditionally, IR systems are query-based, and it is assumed that users can describe their information needs explicitly and adequately in the form of a query. The IR system aims at satisfying the searching and browsing needs of a user that were described in Section 1.2.2 by performing text indexing and using a particular searching strategy.

The grounds for IR systems lie in Luhn's idea which states, "frequency data can be used to extract words and sentences to represent a document", as well as Zipf's law described earlier. As a result, a document representation can become a list of class names that are referred to as the document's index terms (items, keywords or indexes) (van Rijsbergen 1979). Indexes can be described manually or retrieved automatically by deriving from the text of documents. Indexing usually consists of identifying index terms (keywords) and a data structure called an index which points to the specific locations of index terms in the text. A typical search consists of formulating a query based on the information provided by the user, and finding and retrieving documents that are relevant to the query.

The commonly used approaches for modern IR by content are term-based. In term-based representation the weighted terms that represent a document or a query are derived directly from the document or indirectly through thesauri or domain maps (Strzalkowski, Perez-Carballo et al. 1996). The terms (features or items) that are either user-defined or automatically extracted can be in the form of keywords (Sparck-Jones 1971, Dewey 1876, Chien 1997, Jo 1999), or indexes (van Rijsbergen 1979, Lawrence, Bollacker et al. 1999, Karanikas 2000, Chiaramella, Defude et al. 1986).

Keyword-based approaches have been studied by Sparck-Jones (1971). Keywords from the Dewey Decimal Classification of books according to the keywords assigned to the documents by their authors were used to characterize text in the United States back in 1876 (Dewey 1876). Chien (1997) characterized the content of documents by creating Patricia trees of the keywords assigned by the author. (Jo 1999) consigned different categorical substantial weights for informative, functional, and alien keywords, to perform text categorization. Turney (1997) evaluated four available algorithms of keyword extraction, namely Microsoft Word 97, Eric Brill's Tagger, Verity's Search 97, and NRC's Extractor on email messages and web page collections. Extractor, which uses supervised learning to generate a list of key phrases of the input documents, performed at least as good in

terms of precision and recall, and was as fast as the best of the competing algorithms.

The interface of an index is quite familiar to most people, since indexes are extensively used as lists of references in the back of books, or as tables of contents of online software manuals. C. van Rijsbergen (1979) described classical indexing IR approaches. Lawrence, Bollacker et al. (1999) created a full-text index of scientific literature on the web aimed at dissemination, retrieval, and accessibility of scientific literature. The authors used the standard practice of indexing by building a hash-table of words (inverted index) that contained a compressed version of a word and a pointer to a block of record files corresponding to the positions in a matching document. Broccoli (2001) analyzed the problem of improving IR by using human indexing to make a search more intelligent. Karanikas (2000) proposed utilizing automatically extracted features, like event names, to label each document based on the meaningful feature discovered in it. Salton, Wong et al. (1975) described the applicability of VSM for establishing correlation between documents in a document space in order to pursue automatic indexing. Full-text indexes make the entire text of all documents available for retrieval, as opposed to the more restrictive list of keywords (Witten, Nevill-Manning et al. 1996). Chiaramella, Defude et al. (1986) emphasized the importance of the overall principle of query processing in automatic indexing of highly structured documents. Indexing terms for effective IR from document collections are used as knowledge representation and provide the possibility to construct thesauri for performing word disambiguation (Manning and Shutze 1999; Sparck-Jones 1988).

An overview of three classical retrieval models that are applicable to situations where users know what they are looking for, and can explicitly formulate their query (from quadrants III and I in Table 4.1), is briefly given below.

Vector retrieval model can be used for indexing as well as for document representation because it encodes documents in a way suitable for fast distance calculation between them (Salton and McGill 1983). Briefly, each document is represented as a vector in a multidimensional space, where the number of dimensions is equal to the number of terms used in a collection vocabulary. The document relevant to a query is the closest document vector for a query vector in terms of either the smallest distance or the angle between them. This model allows utilization of a number of general data processing methods and algorithms, and thus, variants of distance-based algorithms underlie research in many modern IR systems. The vector model was used in the SMART project that was developed for automatic document retrieval at the dawn of research in IR (Computation_Laboratory 1961; (Salton 1991; Buckley and Walz 1999). Vector Space Model (VSM) gained popularity because of the availability of standard algorithms for model selection, dimension reduction, and visualization of vector spaces, and the relatively high speed of operations with vectors. The main problem with this approach is high dimensionality because the number of words used in a document collection can easily rise to hundreds of thousands (Fayyad and Uthurusamy 2002). Compound words, spelling errors, and word variations add dimensionality to the document model. Obtaining accurate information of semantic relatedness from textual

information automatically is intricate because any two words are by definition considered unrelated in VSM.

In the **Boolean retrieval model** (or exact match approach) a document is represented by a set of index terms that appear in the document. A query is a combination of a set of index terms and Boolean operators (and, or) and, thus, is semantically well-defined. The frequency of terms has no effect on retrieval results. Each document either fulfills the Boolean condition or not. Because Boolean expressions inherit simplicity and neat formalism, this model is widely used in modern commercial search tools (Lagus 2000) and early bibliographic systems . However, as formulating a suitable query is not always possible for user, the size of the retrieved list of documents can be very long, since there is grading of the matching and further ordering of results is not possible.

The **Probabilistic retrieval model** explicitly relies on the Probability Ranking Principle. It states that for a given query, it is possible to estimate the probability that a document belongs to a set of relevant documents, and return the documents in order of decreasing probability of relevance. The key goal in this approach is to obtain estimates regarding which documents are relevant to a given query. In the original Binary Independence Retrieval model, index terms are independent and their occurrences are considered binary and can thus be utilized according to some weighting scheme (analogous to VSM). Other forms of probabilistic models, such as Bayesian Inference Networks, were used for IR by, for instance, (Cheeseman and Stutz 1996). Chen (1993) discussed the use of the Hopfield neural network, evolution-based genetic algorithms, and ID3/ID5R symbolic learning algorithms based on the probabilistic retrieval model to create knowledge-based systems in IR. Nie, Simard et al. (1999) used the probabilistic model to translate an IR query into a different language.

On the contrary, some valuable information hidden in the documents, which is not outlined by manually or automatically chosen keywords, indexes and markups, cannot be retrieved. The simplest term-based representation of content is relatively better understood but usually inadequate, because, as was noted earlier, single words are rarely specific enough for accurate discrimination. A better method is to identify groups of words to create a meaningful *concept* or *phrase* and IR based on those groups of terms (Strzalkowski, Perez-Carballo et al. 1996).

Although a lot of research has been done in the past 10 years to study the effects of different *term weighting strategies* on the performance of an IR system, no strategy could be found that consistently worked well across all of the queries and text collections. One of the most widely used statistical features in term weighting strategy is term frequency (TF), which measures how many times a term has appeared in the document or query. Another commonly used feature is the inverse document frequency (IDF), which is a logarithm of a fraction of the total number of documents in the text collection measured by the number of documents in which a term has appeared. In other words, IDF refers to the rarity of a term in a document collection (Lam, Ruiz et al. 1999). A term weighting system proportional to $TF \times IDF$ assigns the largest weight to those terms which appear with a high frequency in individual documents, but are at the same time relatively rare in the collection as a whole (Salton, Wong et al. 1975). The weight of an index term is

proportional to its relative frequency in a document and inversely proportional to the number of documents containing this term according to the following formula:

$$weight_{i,j} = tf_{i,j} \times \log_2 \frac{n}{df_i},$$

where i is an index term; j is a document; n is a total number of documents in a collection; tf_i is a number of occurrences of term i in document j; df_i is a number of documents in a collection containing the index term i (Tokunaga and Iwayama, 1994; Hoch 1994).

More statistical features used in term weighting can be found in Salton (1983). There are a number of studies and IR systems that adopted the $TF \times IDF$ weighting strategy as a combination of those statistical features. (Fan, Gordon et al. 2000; Han and Kamber 2001 2001) discuss usage of multiple term-weighting schemes to design a ranking function that orders documents in terms of their predicted relevance to a particular query to different people.

The problem of evaluating the performance of a particular IR by content system is complex and subtle because retrieval is a human-centered process (Hand, Mannila et al. 2001). The performance evaluation of retrieval is based on the notion of the relevance and usefulness of a document to the information need articulated in a query by the user. The basic evaluation measures are *precision* and *recall.* Assuming that there exists a set of binary classification labels for all documents in a collection indicating which are relevant and which are not. According to (Han and Kamber 2001):

$$\mathrm{Pr}ecision = \frac{Number\_of\_documents\_retrieved\_relevant\_to\_a\_query}{Total\_number\_of\_documents\_retrieved}$$

$$\mathrm{Re}call = \frac{Number\_of\_documents\_retrieved\_relevant\_to\_a\_query}{Nnumber\_of\_relevant\_documents\_in\_a\_collection}.$$

These measures treat relevance as an absolute notion in the sense that the relevance of any document for a given query is the same for any pair of users. Although the measures are a great simplification, they are the most widely used ones by the IR research community for evaluation of IR systems.

4.1.2 Text Categorization

According to Hidalgo (2002), *text categorization* is an automatic assignment of documents to a set of predefined classes, and, thus, can only be used for tasks from the I and III quadrants. The classes are usually content based, and can be labeled, for example, topics, keywords or subject headings, but can also reflect genres or authors. The dominant approach is to classify (label) a document (quadrant I) or a set of documents (quadrant III) according to terms used in them by means of machine learning and IR techniques to induce an automatic classifier or label for new documents. The initial classification that can be used for training is usually given. The intermediate steps in text categorization are feature selection and extraction, which supply a content space for categorization. Automatic document categorization for knowledge-sharing purposes, document indexing in libraries, web

page classification into Internet directories, and some other tasks can be accomplished by implementing categorization algorithms.

Chen (1993) described several systems that utilized neural networks and evolution-based genetic algorithms for creating classifiers of text. Lam and Lee (1999) proposed a way to implement feature reduction using neural network for text categorization. Ruiz and Srinivasan (1998) reviewed literature on the applicability of neural networks for automatic text categorization in general. Additionally they explored the use of a counterpropagation neural network on the MEDLINE dataset, consisting of medical journals. The MEDLINE dataset requires substantial human effort in categorizing articles according to Medical Subject Headings. Jo (1999) discussed text categorization considering categorical weights and substantial weights of information keywords. Lam, Ruiz et al. (1999) investigated the applicability of instance-based learning from queries and retrieval feedback algorithms for automatic text categorization. Hoch (1994) classified German business letters by deriving statistical hypotheses about the underlying document structure according to EDIFACT message types. Dörre, Gerstl et al. (1999) concluded with regards to IBM's Intelligent Miner for Text that a categorization schema based on a set of defined categories matches the subject matter of an incoming document. However, they discovered that the topic categorization tools agree with human categorizers to the same degree as human categorizers consent with one another. Lin, Shih et al. (1998) built the classifier based on extracted association rules for mining Internet documents. Mining association rules were applied to discover important associations among items (keywords) in a text collection, in order to automatically establish the categories.

4.1.3 Text Clustering

Document clustering has been extensively explored for TM, since researchers historically have perceived clustering techniques as discovery tools. Clustering does not require any predefined categories for grouping the documents (Jain 1999) as opposed to categorization, and thus, is considered to be exploratory in nature. Document clustering not only allows the classification of text domains, and improved document search and retrieval (Willett 1988), but also provides more information about text collections by finding similarities among documents. As was mentioned earlier in Section 1.2, clustering techniques partition a collection into subsets of documents (clusters), so that every individual cluster represents a group of documents having features that are similar, compared to the features of documents from other groups (Hand D. 2001). Clustering is applicable in situations where a user has vague ideas about his or her information needs that can be satisfied via document collection exploration. Therefore, the approach belongs to quadrants II and IV in Table 4-1.

The central assumption proposed by van Rijsbergen in 1979, and known as Cluster Hypothesis, has made document clustering a powerful method. Cluster Hypothesis states that the documents relevant to a query are more likely to be similar to one another than to nonrelevant documents. This hypothesis has received an experimental validation in the context of the Scatter/Gather system that used

document clustering as its primary operation (Cutting 1992). There is a major division of clustering algorithms between *agglomerative* and *divisive*. Agglomerative clustering proceeds by iteratively choosing two documents groups to agglomerate into a single document group (Cutting 1992). At the same time, any divisive clustering algorithm considers two sub-problems of how to select the cluster for splitting; and how to split the selected cluster (Savaresi, Boley et al. 2002). Hierarchical, k-means, and Probabilistic Clustering are the most popular and best-known text clustering methods (Karanikas 2000).

Hierarchical clustering in the form of agglomerative or divisive clustering is often portrayed as the better quality clustering approach, because it presents textual information in the intuitively understandable form of hierarchies. Hierarchical clustering assumes a similarity function for determining the similarity of instances in a cluster. Agglomerative hierarchical clustering based on the Generalizable Gaussian Probabilistic Mixture model was successfully tested for segmentation of e-mails by Szymkowiak, Larsen et al. (2001). A hierarchical document clustering method, based upon a series of nearest neighbor searches, was addressed in (El-Hamdouchi and Willett 1986). Cutting (1992) and Schutze and Silverstein (1997) studied ways to improve clustering algorithms to make them computationally feasible in order to implement them in real-time. Jain (1999) described the limitations of book-grouping according to the classification offered by the American Library of Congress in their review of data clustering methods. The scientists favor the method of hierarchical agglomerative clustering and proximity dendogram creation to depict the contents of the books more efficiently. IBM implemented hierarchical and binary relational clustering in Intelligent Miner for Text, so that vocabulary analysis and determination of important pairs of terms is archived using hierarchical clustering. Finding hidden topics in documents and establishing relationships between topics are achieved using binary relationship clustering.

Steinbach, Karypis et al. (2000) argued that the quadratic time complexity of hierarchical algorithms makes them less appealing than **k-means clustering**, which has a time complexity linear in the number of documents. K-means is a direct clustering method that uses a specified number *k* of clusters, *centroids* as attributes of the clusters' description, and an assigned clustering evaluation function. The author shows that the "bisecting" k-means technique produces results that are as good as, or better than, tested hierarchical approaches. Larsen (1999) clustered the features extracted from documents using unified k-means to discover topic hierarchies in gigabytes of documents.

Baker and McCallum (1998) used a probabilistic framework for semantic word clustering. Goldszmidt and Sahami (1998) used a probabilistic approach to full-text document clustering that allows documents to overlap. They follow the probabilistic considerations: similarity is scored according to the expectation of the same words appearing in two documents. Soft clustering, contrary to "hard clustering," assigns probabilities to an instance belonging to every subset of clusters, so that each document is assigned a probability distribution across a set of discovered categories, and the probabilities of all categories sum up to 1. There are

a number of algorithms for soft clustering such as fuzzy k-means and Gustafson Kessel clustering (Kaymak and Setnes 2000).

Text collection clustering can assist in organizing text collections for enabling retrieval by content (Anick and Vaithyanathan 1997; Merkl 1997; Lee 1999; Lin, Soergel et al. 1991) and searching (Cutting 1992). Cutting (1992) used a clustering technique that supports an iterative searching interface by dynamically scattering a document collection into smaller semantic clusters. A user navigates the document search space by selecting relevant documents among the clusters to regroup the results. Anick and Vaithyanathan (1997) exploited document clustering and paraphrasing of term occurrence for document retrieval by content. Merkl and Schweighofer (1997) used a different approach for detection of similarities between documents in organized legal text corpora to enable document retrieval by content. They combined the vector space model, cluster analysis, and the clustering ability of neural networks, in the form of the SOM to organize the legal text corpora as the hypertext and knowledge base of descriptors, probabilistic context-sensitive rules, and meta-rules of legal concepts. SOM is an unsupervised tool for clustering and visualization, also applicable for very large document collections (Kohonen 1998). SOM organizes high-dimensional input data so that similar inputs are mapped close to one another. Lee (1999) presented a SOM-based clustering approach based on word co-occurrences for IR on a Chinese corpus from the web. The WebSom system is built on SOM clustering abilities and allows browsing and retrieval of the resulting match list to perform multi-level search of text collections with an increasing navigating role of the user (Kohonen 1999; Lagus, Honkela et al. 1996). Lin, Soergel et al. (1991) explored the potential of a SOM semantic map as a retrieval interface for an online bibliographic system. Roussinov and Chen (1999) discovered that Ward's clustering, proceeded by a pair-wise comparison of texts is slightly more precise in detecting associations between documents than a SOM, when applied to the study of text messages obtained from a group brainstorming meeting. Jones, Robertson et al. (1995) tried out another AI technique – a genetic algorithm for nonhierarchic document clustering. They clustered the Cranfield standard test collection with documents and queries on the subject of aerodynamics, then searched the resulting sets of clusters and evaluated the effectiveness of those searches in comparison with nearest-neighbor clustering. The researchers concluded that while genetic algorithms provide an obvious approach to the generation of partitioning-type clustering of document databases, such clustering does not offer any obvious advantages over the currently available methods for document grouping.

In the majority of the algorithms for IR by content mentioned in Section 4.2.1, a user participates actively in the clustering and navigating processes, controlling how his/her information needs are met. Clustering can be performed either on the word or term-level within a document (see quadrant II in Table 4.1) or within an entire document collection (see quadrant IV in Table 4.1). Studies described in the previous paragraph are concentrated on document clustering within collections.

Document clustering helps to tackle the information overload problem in several ways. One way is exploration; the top level of a cluster hierarchy gives a

brief overview of the documents' contents in a collection, enabling users to selectively drill deeper to a more specific topic of interest (Lagus, Honkela et al. 1996; Larsen 1999). Another way is retrieval; clustering organizes search results by topic similarity to help find potentially useful documents more quickly (Sahami, Yusufali et al. 1998; Zamir 1998). The later approach is applicable in situations where users know what information they want to extract from text (see quadrants I and III in Table 4.1).

4.1.4 Text Clustering vs. Text Categorization

Both text clustering and text categorization are used for text classification. The word "classification" is ambiguous in a sense, since it can either mean assigning a new object to one of an existing set of possible classes, or finding the classes themselves from a given class of "unclassified" objects (Cheeseman and Stutz 1996). Although text categorization requires a set of predefined categories and, thus, a priori human knowledge about them, text clustering discovers the intrinsic structure of a document or collection without any a priori knowledge. Sometimes, clustering algorithms are used to find a set of classes, and categorization algorithms are used to classify new cases in the defined by clustering classes. For instance, Roussinov and Zhao (2003) used Ward's clustering to discover the initial categories in messages from a customer relationship management system, and then, categorized new incoming messages into those identified categories. Makoto and Takenobu (1995) described the simplest strategy to search documents: first, cluster training documents in *k*-nearest neighborhoods, then use those clusters as the categories assigned to training documents.

Although clustering techniques are mostly used in situations when users do not know what they want to discover in a text, some clustering approaches can be applied in situations when users know what information they want to extract from a text (quadrants I and III in Table 4-1). For example, Deogun and Raghavan (1986) discussed the user-oriented approach to document clustering based on the user's concept of closeness between documents. The user-defined query to a database, and the number of retrieved documents, determine the documents that are consistently seen by users as belonging together. Then, an attempt was made to characterize the resulting clusters by considering the power of various index terms. In other words, the developed classification of the documents became consistent with the behavior of past users.

Clustering and categorization algorithms compose the bulk of content-based methods as an alternative to query-based methods discussed in Chapter 4.2. The majority of content-based retrieval systems are based on approaches from computational linguistics and linguistic knowledge about the text collection. Hatzivassiloglou, Gravano et al. (2000), for instance, noticed that linguistically motivated features in conjunction with the full word vectors increases the overall clustering performance. Miike, Etsuo et al. (1994) developed a Japanese full-text retrieval system that analyzes text and enables the user to generate an abstract interactively. The system was based on linguistic knowledge and clues, such as idiomatic expressions, and was domain independent but required a dictionary of

60,000 entries for morphological analysis of sentences. However, heavy exploitation of linguistic knowledge makes the TM system content and language dependant, and often requires laborious manual preprocessing. At the same time, the ideal TM system should require as little human intervention as possible in order to enable processing of very large document collections (Honkela 1997).

4.2 Representing Text Mining Results

Research on data visualization and exploratory data analysis methods has flourished, providing methods and tools capable of illustrating properties and relationships of complex data sets graphically (Vesanto, Himberg et al. 1999). Although text is written in a linear fashion, text affects readers' cognition in complex way. The richness of information can be presented differently, for example, the beauty of a landscape can be described in text or using a drawing. Research in human cognition as a process of perceptions, has made possible the creation of some models that can be visually interpreted by computers and then presented to users.

Although graphically picturing knowledge that was derived by humans from text collections is very problematic, several attempts to map it or create reflections of it have been made in a number of studies. Wise (1999) proposed an "ecological approach" to text visualization, and introduced the SPIRE text visualization system that visualizes free text as natural terrains. (Vesanto 1999, Flexer 2001, Kohonen, Kaski et al. 2000, and (Himberg 2000) investigated the applicability of the SOM for text visualization and navigation of document collections. Gershon, Eick et al. (1998) discussed the use of hierarchies in the form of tree-like nodes, such as cone 3D trees and link diagrams, for document visualization. Kontkanen, Lahtinen et al. (2000) suggested Bayesian model-based visualization to represent multidimensional data by considering two vectors similar if they lead to similar predictions. Kim, Johnson et al. (2002) introduced the query fingerprinting text visualization approach to inform users simultaneously about the frequency of queries, distributional information about queries, and segmental structures of the document. Wong, Cowley et al. (2000) proposed visualization of sequential patterns for TM purposes in large text corpora. Dubin (1995) explored the visualization abilities of the VIBE tool for mapping word co-occurrences in document clusters for easier browsing. Girolami, Vinokourov et al. (2000) used a probabilistic hierarchical mixture method to develop models that visualize topographic relations between similar documents on a 2D grid.

Text summaries can be regarded as tools for visualization of discovered knowledge from a text. They offer new information for users in the form of an overview, presenting the meaning hidden in text in a different form. Text summarization also belongs to the methods described in quadrant IV in Table 4-1, since users do not have an a priori idea about the content of the mined documents. (Amini and Gallinari 2001) presented an automatic summarization system that is based on text-span extraction and self-supervised learning. Based on a query, the system extracts the most relevant sentences from a document. It does not rely on the availability of labels for learning to rank sentences to compose a summary of a text

collection. Hatzivassiloglou, Klavans et al. (2001) explored nonhierarchical clustering technique, the exchange method, for creating SimFinder – a flexible tool for summarization.

In summary, the ground work in TM's parental fields of IR and NL processing was done by(van Rijsbergen 1979, Salton and McGill 1983, Sparck-Jones 1971) and (Manning and Shutze 1999) while the most influential contributions to the field of TM were made by (Hearst 1997), and (Witten, Bray et al. 1998). The practical estimations of fundamentals and the applicability of the various proposed algorithms were achieved in the numerous studies described in Sections 4.2.1–4.2.4. Despite the extensive research in TM and its parental fields, there are still a number of unsolved or poorly solved problems that negatively affect the performance, applicability and diffusion of TM systems. One of those problems lies in the creation of an adequate text representation that reflects word ambiguity and the linguistic structure of a text. The computational heaviness of clustering and categorization algorithms complicates the development of TM methods. More effective and intuitively easily decodable representations of TM results would accelerate validation and justification of TM methods. More flexible text representation models that rely on representation of the meaning of text rather than stand-alone concepts introduced in the text, and models that reflect the interrelationships between concepts, can potentially burst the penetration of TM methods. More flexible TM systems should allow their users to retrieve meaningful patterns from a text without thorough composition out well-thought of and well-defined queries. Language independent TM methods that can be applied to different alphabets, and to languages with different syntactic structures, are needed for the development of real-time online mining applications. In general, while I explain the nature of the relationship between textual information overload and modern technologies that contribute to its occurrence and reduction (research objective *a* from Section 1.3) in Chapter four, in Chapter five I describe the methods and algorithms that form the scientific basis and are currently implemented in technologies for dealing with textual information overload. The majority of the methods used in TM is query-based, and require their user to know what is to be discovered. Those methods fit comfortably into quadrants I and III. The most challenging part in the development of TM methods comes in situations when a user does not know what is to be discovered from large document collection (quadrant IV in Table 4-1). It seems that modern TM systems require their users to presuppose intelligent answers to the analytical questions about the subject before even asking them. This resembles a situation in which a reader is required to learn a lesson from a book by looking at its cover. Content-based methods that can create queries independently based on learning patterns of interest from a user are a powerful way to go.

5. RESEARCH METHODOLOGY OF THE PROTOTYPE MATCHING METHOD

In this chapter, I introduce the prototype matching method, which I have explored and employed for accomplishing TM tasks from financial and scientific document collections. I explain thoroughly all the steps and show how the applied methodology transforms free-text into a feasible form for TM. The methodology fulfills the information needs of the user when he only has a very vague idea about a document collection, or does not even know what is to be discovered from it. According to the categorization of the TM approaches from Chapter four, the prototype matching method belongs to quadrant IV in Table 4.1. Moreover, it is a content-based method as opposed to query-based ones. It tries to establish semantic similarities between the prototype-document, which contains the description of the user's informational needs, and the rest of the documents in a collection. The prototype matching method allows for analysis of document collections on both word and sentence levels. The evolution of the method throughout the developmental cycle, a part of which the current research is, has earlier been discussed in several publications by (Toivonen, Visa et al. 2001), (Back, Toivonen et al. 2001), and (Visa, Toivonen et al. 2002). This dissertation uses the third developmental version of the system.

The system, GILTA, which is built on the prototype matching[11] method, is domain adjusting. The methodology eliminates the use of any language-dependent linguistic methods in a clustering process. It offers a user the opportunity to input into the system either a part of or the entire document (prototype or example) he has an interest in, and to locate and retrieve the documents that are semantically similar to it in some specific way. Although a prototype is a part of a document, primarily, it is an articulation of the user's information needs, not a reflection of the document. This is because the user might not even know what pattern of information in a document he is looking for. The concept of prototype (or template) matching exists as a separate theory in psychology to explain pattern recognition by bottom-up processing (Legge 2001). Finding the set of documents with the most relevant patterns is called matching. Matching problems rely on some distance measures, e.g. Euclidian distance. There is no guarantee that such measures really reflect useful similarities and dissimilarities between documents (Kontkanen, Myllymaki et al. 1997). Therefore I have validated the similarities and dissimilarities determined by the prototype matching method manually, by reading the documents.

The system built on the prototype matching method consists of three processes that were introduced in Figure 3-1 in Chapter three. Document collection preprocessing and encoding perform text representation and distillation. Knowledge sophistication is achieved by document processing and matching. Document retrieval and cluster representation conclude the knowledge visualization process. Figure 5-1 depicts all parts of the system; the gray blocks indicate the input and output through which the system interacts with its users.

[11] The newer working title of the methodology is an example-based text matching (Visa, Toivonen, et al. 2002).

The method executes clustering in a specific manner, since it considers only similarities between the document-prototype and the closest-matching documents, and does not consider dissimilarities between documents from different clusters. In other words, for every prototype, there are two clusters to be created: one cluster consists of the documents that are similar in some specific way; another cluster consists of the documents that are different from a prototype in some specific way. Similar documents share the same patterns with prototype. Because the user is not presented with those patterns explicitly, he might not even recognize the similarity. A cluster of similar documents is formed from documents that fire as the closest matches at the top of a ranked list. A ranked list is an ascending list of the distances between the prototype and the other documents in a collection. The cluster of documents different from the prototype consists of documents that share few or no patterns with the prototype. The documents that constitute this cluster fire at the farthest distance to a prototype on a ranked list. The method constructs a ranked list for every document-prototype, and creates clusters from the first hits on the ranked list. As was noted in Chapter four, clustering is the most challenging task, since there is no pre-existing set of categories created by human experts and thus, clustering results are difficult to evaluate. The patterns that combine or separate documents into those clusters are not obviously noticeable by a user of the information, thus the evaluation of the retrieved results becomes subjective and tricky.

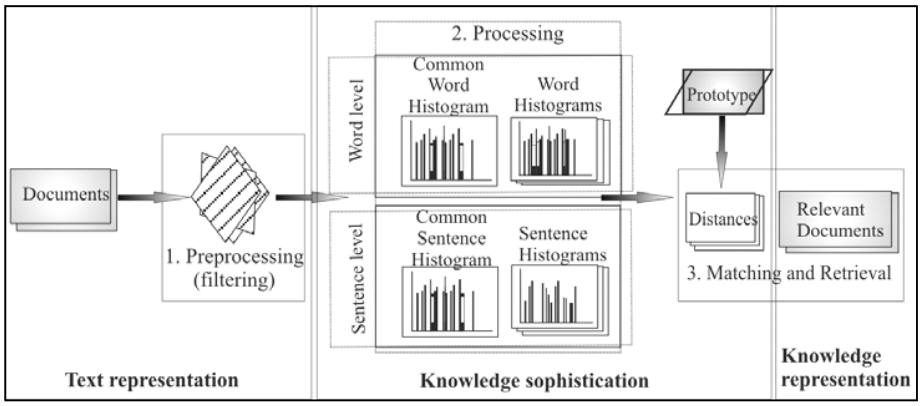


Figure 5-1. Parts and processes of a system with the prototype-matching clustering

Figure 5-2 schematically depicts the process of comparing documents (process 2 and part of process 3 from Figure 5-1) from a collection of documents. The comparison of the documents is achieved by calculating the distances between histograms that represent those documents. It consists of collecting various text documents into a single repository, creating word histograms based on the distribution of word frequencies in a collection, and creating sentence histograms based on the distribution of sentence vectors in a collection. The approach utilizes the idea that all the words with their frequencies, and rare words in the histograms,

distinguish documents (Manning and Shutze 1999). The same idea is expanded to sentence level analysis. Distance-based comparison of word (sentence) histograms corresponding to every single document distinguishes clusters of the documents from a collection on the word or sentence levels. Generally speaking, word level clustering shows the similarities in vocabulary between documents, and sentence level clustering shows the semantic similarities between documents.

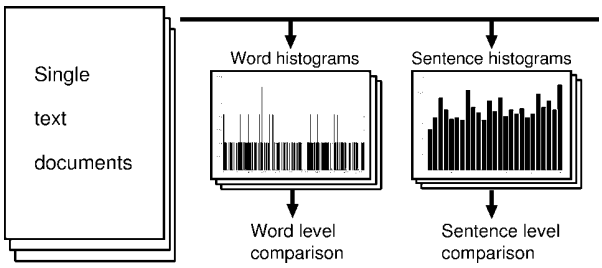


Figure 5-2. The process of comparing documents based on extracted histograms on word and sentence levels

Below I describe the clustering methodology as a combination of processes and sub-procedures, and provide their intuitive explanations. Document collection preprocessing and encoding can be termed as the process of text representation and distillation, if one follows the TM framework introduced in Chapter 3.4 and continued in Chapter four. Document processing is a knowledge sophistication process. Finally, document retrieval belongs to the knowledge representation process.

5.1 Document Collection Preprocessing and Encoding

a. Preprocessing takes place before text documents are processed for text clustering and comparison. A basic filtering devotes an own line to every sentence in a text collection, rounds the numbers, and separates punctuation marks with extra spaces. Extra carriage returns, mathematical signs, and dashes are excluded. Compiling abbreviation, synonym, and compound word files performs synonym and compound word filtering. The filtered text is data suitable for further encoding. I did not omit the stop words and did not perform word stemming in an attempt to keep as much initial information and structure in the preprocessed documents as possible because, as I noted in Chapter one, word order, their combinations, concurrences, and conjunctions can convey important insights or understanding to a reader. For example, the significant semantic difference between “Thanks, I have no objections” and “No thanks, I object” can be destroyed after ommiting *no*, *thanks,* word stemming, and disregarding word order.

b. After basic filtering of the documents in a text collection, every document is encoded. Among the variety of encoding approaches, some of which were described earlier in Chapter four, words were analyzed character by character, based on the key entry of the characters. Although this encoding approach is

accurate and sustainable for statistical analysis, it is sensitive to capital letters and conjugations. Every word *w* in a document is transformed into a unique number according to the following formula: $$y = \sum_{i=0}^{L} k^i \times c_{L-i} \quad (1),$$

where *L* is the length of a word as a string of characters; c_i is the ASCII value of every character within a word *w,* and *k* is a constant. Since 8-bit ACSII character set was used, *k*= 256 was chosen empirically. The encoding algorithm produces a unique number for each word disregarding word stems, capitalization and synonyms, so that only the same word attains equal number. The codes of every word and punctuation mark, from every document, form feature word vectors to represent documents in individual files.

5.2 Document Processing

The methodology is based on the frequency distribution of all words from a training set based on all the documents in a collection, contrary to indexing approaches that characterize documents based on a keyword, index, or term co-occurrences. Very often the entire text collection is used as a training set. The word histogram and, later, the sentence histogram creation processes, which allow the comparison of different documents to each other, rely heavily on the frequency distributions of words or sentences from an entire document collection. After each word has been converted to a code number, we consider the distribution of the code numbers of the words.

Word Quantization:

a. From a set of word codes resulting from equation (1) the minimal and maximal values are chosen for setting parameters (*a* and *b* in equation (2)) to characterize and estimate the word distribution from an entire document collection. The Weibull distribution is a versatile distribution that can take on the characteristics of other types of distributions based on the value of the shape parameters (ReliaSoft_Corporation 2002). The Weibull distribution was selected for estimation of the distirubution of the words' code numbers.

b. In the training phase, the range between the minimal and maximal values of words' code numbers is divided into N_w logarithmically equal bins, where N_w is the quantity of the words used in the text collection. A word's frequency in each bin is calculated and further normalized according to N_w. Using selected precision, the number of Weibull distributions are calculated using various possible values for *a* and *b,* which that are restricted by the established minimal and maximal values. Then the best fitting Weibull distribution corresponding to the textual data is determined by examining the cumulative distribution. Every estimated Weibull distribution is compared with the code distribution by calculating the Cumulative Distribution Function (CDF) according to:

$$CDF = 1 - e^{\left((-2.6 \times \log(y / y_{\max}))^b\right) \times a} \quad (2),$$

where *a* and *b* define the shape pf the Weibull distribution. The ration *y/ymax* tells the actual portion of the total density mass. The only way to compare distributions is to use cumulative functions that are the same as the integrated probability distributions. By doing the integration to the Weibull distribution in our case you will get 2.6 The comparison between estimated Weibull distributions and the Cumulative code number distribution is performed in the smallest square sum sense.

The training phase is where the quantization for the words is formed. In the testing phase the histogram for every single document is created. The best fitting Weibull distribution is divided into N_w bins of equal sizes. Every word is assigned to a bin that can be found by using the code number and the best fitting Weibull distribution. This way every word can be presented as the number of the bin of the distribution (quanta) to which it belongs. The quantization is best where the words are the most typical to a text (usually 2-5 symbol words - the most widespread length of English words). The distribution and thus quantization of longer words is sparser.

Sentence Quantization:

c. Similarly to the word level, on the sentence level every sentence has to be converted to a number. First, every word in a sentence is changed to a bin number (bn_i) in the same way as words on the word level were changed. The whole encoded sentence is considered as a sampled signal (vector). Since the sentences in the text contain different numbers of words, the sentence vector's lengths vary. To compensate this fact, Discrete Fourier Transformation (DFT) is applied for converting every sentence vector from a collection into an input signal. The input signal for DFT is a vector (bn_1, bn_2,…, bn_m), where *m* is a word's number in a sentence. The output signals from DFT are coefficients B_i (i= 0 .. n). The coefficient B_1 was selected to further represent the sentence signal, as it was observed during the experiments, that the coefficient B_0 is too heavily affected by the sentence length.

d. After every sentence has been converted into numbers, a cumulative distribution is created from the sentence data set using coefficients (B_1) in the same way as on the word level. The range between the minimal and maximal parameters of the sentence code distribution is divided into N_s equally sized bins, where N_s is the length of the histogram vector. The frequency of sentences belonging to each bin is calculated. Then the bins' counts are normalized in accordance with N_s. Finally, the best fitting Weibull distribution corresponding to the sentences' distribution is found. A graphical representation of a sentence quantization process is borrowed from (Visa, Toivonen et al. 2002) and provided in Figure 5-3.

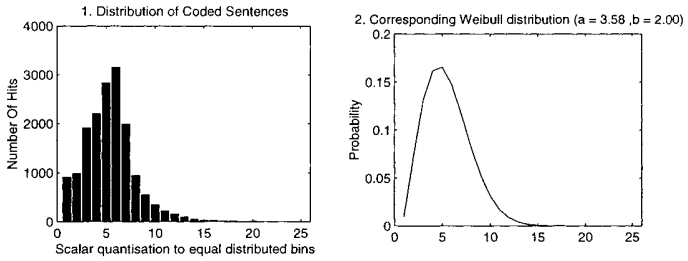


Figure 5-3. Example of a sentence distribution

Word and Sentence Individual Histograms:

Furthermore, every document in a collection is examined once again to create the individual word and sentence level histograms. The histograms are built for the documents' word and sentence code numbers (levels), according to the corresponding values of their quantization.

e. The filtered document from a collection is encoded word by word on the word level. Each word code number is quantisized using word quantization created using all the words from a collection. The right quantization value is determined, an accumulator corresponding to this value is increased, and a word histogram A_w is created. The histograms A_w are normalized according to the length of the histogram vectors.

f. Similarly, the histograms on the sentence level for every document in a collection are created. Every single document is encoded into sentence code numbers (B_1) and the hits according to the corresponding place in the quantization are collected into histogram A_s, that is, further normalized according to the total number of the sentences in a document.

5.3 Document Matching and Retrieval

Having individual word and sentence level histograms of all documents in a collection allows their comparison with a histogram corresponding to a document-prototype (or example). It is not necessary to have any prior knowledge about the structure or content of the actual text documents in order to use the described method. The comparison of a histogram corresponding to a document-prototype with the histograms corresponding to the rest of the documents in a collection is called matching. Calculating the simple Euclidian distance measure, although any distance measures can be used, establishes similarity among the histograms. The closest documents to a document-prototype in terms of the smallest Euclidian distance, form a cluster. Only the comparison of same-level histograms is robust, since closeness of word level histograms identifies the similarity in terms of vocabulary of the documents, and closeness of sentence level histograms identifies the semantic resemblance of the documents in a collection. In other words, matching the document-prototype with the rest of documents in a collection is performed by establishing the Euclidian distances between histograms of sentence

or word level distributions for the document-prototype and the other documents in a collection.

In the retrieval phase, the documents within the smallest distances to a document-prototype are chosen from the top of the ranked list. The system creates a proximity table of all distances among the documents in a collection. The documents from the top of the proximity table for a given document-prototype are retrieved and presented to a user within a specified recall window. The recall window is the quantity of closest-matching documents that a user wants to retrieve and consider for further analysis. The order of the retrieved documents within a recall window is not considered.

The same operations and analysis can be performed for another linguistic unit of text – the paragraph, if mined documents are very lengthy.

5.4 Applications and Validation of the Prototype Matching Method

The prototype matching method is designed to assist in solving a variety of real-world problems associated with documents and text processing. In *Paper* 1 good results have been reported with high-content, small and medium sized text collections. GILTA has been used to cluster a range of text materials, including full-text versions of books (Visa, Toivonen et al. 2001; Visa, Toivonen et al. 2000), full-text versions of scientific articles (Kloptchenko, Back et al. 2002), abstracts of scientific articles (Kloptchenko, Back et al. 2002), financial reports (Back, Toivonen et al. 2001; Kloptchenko, Eklund et al. 2002), and news items (Visa, Toivonen et al. 2000) with some degree of success. It was deemed that the GILTA tool can be implemented as a module either in document clustering decision support systems, financial analysis tools, or as an author attribution tool (Visa, Toivonen et al. 2001). Additionally, the methodology can also be used for text mining of documents written in languages whose linguistic structures are different from English, e.g. the Finnish language.

One larger procedure was undertaken to evaluate and validate the prototype-matching method. For validation purposes, the Bible written in Greek, Latin and two versions in Finnish from the years 1933 and 1938 were chosen as testing material. Bible translations are considered to be very accurate and bear the same meaning in different languages. The idea was to compare all the books in the Bible using them as the prototypes against the whole text of different versions of the Bible. Every book of the Bible was taken one by one as a document-prototype, and the ten closest matches were examined (Toivonen, Visa et al. 2001). The difference between the books from the Old and the New Testament was expected to be discovered by the methodology. In fact, the methodology found an average of six books out of ten that belong to the same category of books in Bible, i.e. for Greek version of the Bible, matching the Genesis (book number 1) retrieves eight books from the Old Testament on the word level and six books on the sentence level. Receiving similar results with the books in the different languages would serve as evidence that the methodology works (Visa, Toivonen et al. 2002). Word, sentence, and paragraph level histograms were created for the books of the Bible in Greek,

Finnish, German, and English. The co-occurrences of the books from the different languages, not the order within the examined recall window, were considered. In other words, for the different languages, i.e. Greek and Finnish, the same books were expected to be retrieved among the closest matches to the same book-prototype. On average, there were 4.52 of same books within the window in the English and Finnish versions based on the word map, 7.94 of same books based on the sentence map, and 5.56 of same books based on the paragraph maps. Based on a random sample there should have been only 2, i.e. the results can be considered statistically significant. The tests provided evidence that the method works.

As a potential future application of the methodology, routing (Visa, Toivonen et al. 2002) or search engines can be considered. The main emphasis is on extract valuable information hidden in the documents, which is not outlined by keywords, and thus, cannot be retrieved by user-defined information retrieval methods. As a result, instead of typing the keywords united with Boolean operators in the query line, the whole paragraph that is of interest to the user can be copy-and-pasted into the system that employs the prototype matching method. This helps in dealing with textual information overload, and delivers targeted textual information necessary for decision-making purposes.

In summary, this chapter presents the prototype-matching (example-based) method that is used in the consequent chapters to perform clustering and information retrieval by content tasks from collections of financial reports and scientific publications. The position of this method in the TM categorization was explained. The prototype matching method is a content-based, language-independent method that aims at establishing semantic similarities between the chosen document-prototype (example) and the rest of the documents in a collection. The analysis of documents and their semantic similarities is carried out on the word and sentence levels. The method is expected to capture the semantics of a particular document because it encodes the document word by word taking into account word frequencies on the word level, and by preserving word order on the sentence level. The GILTA system, which utilizes all the procedures from the method was described along with its application areas.

6. COMBINATION OF DATA AND TEXT MINING METHODS

Throughout the previous chapters, I have outlined the problems and difficulties that managers, knowledge workers, and decision makers face in dealing with textual information overload in organizational and business settings. Powerful TM and information management tools are needed for getting accurate, limited, updated information, and to extract meaningful insights from it ontime. TM systems represent a novel approach for knowledge creating and decision support, and have much to offer as building blocks of business intelligence and decision a dvisory systems. TM solutions are able to contribute to a more systematic, scalable, and faster knowledge extraction process than human process alone. DM solutions, on the other hand, based on well-established statistical techniques, constitute part of modern DSS systems. Managers, knowledge workers, and decision makers do not have enough time to hunt for insights and nuggets in the deep forests of available text and numeric information that can help them to give their companies a competitive advantage. In this chapter, I present a conceptual framework using a combination of data and text mining techniques for a constructing knowledge-building systems and discovering more complex patterns. The chapter is a summary and extension of *Paper*s 2, 5, and 7.

6.1 Text Mining as a Data Source for Data Mining

As I describe in Section 3.1, modern organizations and decision makers within them are not aware of all the content contained in their intranets, libraries, and knowledge management repositories. Manual attempts to monitor external information sources often reach only the peak of critical information useful for gaining a competitive advantage. DM solutions explore thousands of records with numeric data from well-maintained internal data warehouses. Additionally, DM algorithms embedded in software agents can scan business related sources, and bring to decision maker's or knowledge worker's attention the important trends from online numeric feeds. However, in some cases, it is difficult to determine from numeric data what particular data stream should be mined, and to establish what discovered data relationships can be beneficial for business processes. TM solutions can assist in those tasks by navigating decision makers to particular questions to be explored further, by bringing to their attention important concepts from various text collections. Questions to explore include how those concepts are used and occur together, what else they are linked to, what they indicate and predict, and how. In other words, DM and TM methods can be combined to discover more complicated patterns from either structured or unstructured data and to deliver greater financial returns.

Results from TM and DM analysis of the data that describe the same real-world phenomena, such as changes in stock market prices or the history of product sales, can not only enrich each other by providing more thorough and accurate discovery and specific explanations for the grounds for those discoveries, but can

also assist in predicting some important values. For instance, users of financial news, financial analysts, employees of securities companies, and investors have to deal with vast amounts of news and information, upon which the progress of the capital market depends daily. Investor expectations, the political situation, company performance, industrial trends, and market situation are coded in a mass of numeric and textual data, and published in various private and public sources that are costly to overview manually. The sophisticated combination of TM and DM solutions can contribute to discovering important points from those sources. For example, a Hong Kong research group exploited TM and DM techniques, such as rule-based, k-nearest neighbor algorithms, and neural networks for daily prediction of major stock indices concerning the data gathered from closing values of major stock market indices in Asia, Europe, and America as well as textual analytical and news articles about them from influential on-line financial newspapers, for example *Wall Street Journal*. Wuthrich, Permunetilleke et al. (1998) suggested a trading strategy based on analytical patterns discovered from news trends that were found to have an impact on price changes. The forecast system saves time and efforts spent on manual reading of numerous newspaper articles. Moreover, a research group from the University of Massachusetts developed a system for predicting trends in stock prices based on the content of news stories that precede the trend. Textual documents, such as news stories from *BizYahoo!,* and numeric historic all time series data were explored using TM and DM techniques (Lavrenko, Schmill et al. 2000). Major software companies, e.g. SPSS Inc. and SAS Institute, also provide tools for the combination of TM and DM techniques. Their tools have a restricted area of application due to the limitations of the mathematical algorithms used in them. I described the limitations and advantages of the most popular mathematical methods for TM in Chapter four. For instance, text categorization algorithms cannot discover a novel category or theme in a textual collection, unless it is either learned from the training data set or defined by a user. In a real business environment, obtaining an accurate and adequate training data set is a cumbersome task. Categorizing a text collection requires special expertise and time input from the decision makers and knowledge-workers who are potential users of these analytical tools.

The physical inability to analyze all data sources to determine the golden points upon which to perform further exploration can result in loss of effectiveness and profitability in a lucrative financial sector. In *Paper 5*, a conceptual model of a knowledge building system is based on a society of software agents. Each agent exhibits intelligence by using different data and text mining methods. The system aims at monitoring new financial updates from a variety of sources, and calculates financial ratios for different companies that can be used for various tasks: financial benchmarking, assessing creditworthiness of different companies, etc. In *Paper 2*, I suggested the utilization of DM techniques to identify the interesting trends in the comparative financial performance of companies from one industry, which can be further investigated using TM techniques. Clustering of quantitative data in the form of financial ratios was performed using the SOM, and clustering of qualitative data in the form of the textual parts of quarterly reports was performed using prototype matching. In *Paper* 7, I proposed a methodology for forecasting a company's

financial performance in the next analyzed period using feedforward multilayered neural networks (MFNN) with backpropagation learning trained on combined results from clustering of qualitative and quantitative parts of financial reports.

6.2 Combination of Data and Text Mining Techniques

6.2.1 Complex Patterns Discovery in Data

The majority of DM techniques are designed for extracting meaningful patterns from numeric, well-structured databases. At the same time, some valuable knowledge about phenomena hidden in unstructured data can help to describe these phenomena more accurately. More than ever, the ability to automatically combine quantitative and qualitative data processing to discover more complex patterns is needed. The majority of studies have investigated data mining opportunities from the financial data domain, and only a few attempts have been made to explore both types of data: Lavrenko, Schmill, et al. (2000), Back, Toivonen et al. (2001), and Kloptchenko, Eklund et al. (2002) have combined quantitative and qualitative financial data analysis using TM and DM techniques.

For benchmarking purposes, Back, Öström et al. (1998) compared 120 companies in the international pulp and paper industry. The study was based on standardized financial statements for the years 1985-89. The objective of the study was to investigate the potential of using the SOM for the process of analyzing large amounts of quantitative financial data. The results of the study indicate that the SOM is an adequate tool for processing vast amounts of numeric financial data. Additionally, the relative company financial performance is determined from the company's position on the map. Eklund, Back et al. (2002) collected the updated data set of financial performances from pulp and paper companies for 1994-2000 years for benchmarking purposes. They clustered seven financial ratios for about 80 companies using one SOM and determined the companies' financial position and movements over the years. Karlsson, Eklund et al. (2001) benchmarked telecommunication companies for the years 1995-2001 using the SOM. The relative position of companies, and sudden changes in financial position, were discovered and brought to the attention of a potential decision maker.

One of the first attempts to semi-automatically analyze company's performance by examining quantitative and qualitative information from annual reports, was made by Back et al. (2001). The authors compared numeric and textual information for 76 companies in the international pulp and paper industry for the period of 1985-1989 using several SOMs.

My research on the combination of the SOM and the prototype matching method for clustering numeric and textual data reveals that clustering results from both methods do not coincide because numeric and textual data relate to different time periods. It was discovered that annual/quarterly reports tend to state information about company's past performance, but also contain indications of its future performance. For example, the tables of financial numbers indicate how well a company has performed, while the linguistic structure and written style of the text

may indicate how well a company will do. The results that were obtained after analyzing the qualitative and quantitative information from quarterly reports have proven that some future changes in financial performance can be anticipated by analyzing the textual parts of reports. Before a dramatic change occurs in a company's financial performance, a change in the written style of a financial report can be seen. The tone tends to be closer to the company's future performance because in anticipation of drastic changes in the future, especially when performance worsens, management tends to cautiously choose vocabulary and sentence constructions similar to the previous performance in order to avoid revealing huge performance fluctuations to the public. If the company's position will be worse in quantitative terms during the next quarter, the report of the current quarter tends to become more pessimistic, even though the actual financial performance remains the same. Using TM or DM techniques alone could not disclose such a complicated pattern in the data and reveal dependencies between companies' strategies and real financial performances.

I aimed at revealing the above-mentioned observation about connections between changes in the financial performance of a company, reflected by the companies' financial ratios, and the tone of the financial report, for the following time period. The limitations of the study are carefully discussed in *Paper 2*. The biggest drawback is the small size of the data collection in text clustering, which limits the ability to generalize the results. The limited vocabulary (terms related to finance and the telecommunications sector), extensive use of proprietary names (such as Motorola, Nokia, and Ericsson), and indications of time period (quarter, year, annual), might have slightly influenced the clustering ability in our qualitative analysis, although much less than on the word level.

6.2.2 Forecasting Future Performance from Complex Patterns Discovered in Data

After revealing the complicated dependencies in qualitative and quantitative data clustering, the classification model for predicting future financial performance was designed. I attempted to create a methodology that can mine the financial reports of companies' competitors or partners for forecasting their future financial performance. The methodology consists of several procedures. First, quantitative data in the form of financial ratios from quarterly reports is clustered using the SOM. Second, the clustering of qualitative data from the textual part of quarterly reports is preformed using the prototype matching method. Third, backpropagation neural networks are trained and used to classify the results from quantitative and qualitative clustering. Aiming at making all the major steps in the methodology automatic, I used MFNN with backpropagation learning as the final step. The remaining manual steps are associated with the nature of neural networks, and require human expertise in choosing appropriate architectures of SOM and MFNN, and in the interpretation of quantitative data clustering using domain knowledge. A schematic overview of the methodology is provided in Figure 6-1, where the first two steps of data and text mining can be performed consequently. The forecasting of financial performance is a classification problem that can be performed using

neural networks, probabilistic models, fuzzy rule-base system, discriminant analysis, etc. Classification procedures predict the value of a single class variable of a new partially observed data vector, based on a model constructed from the sample (Tirri, Silander et al. 1997).

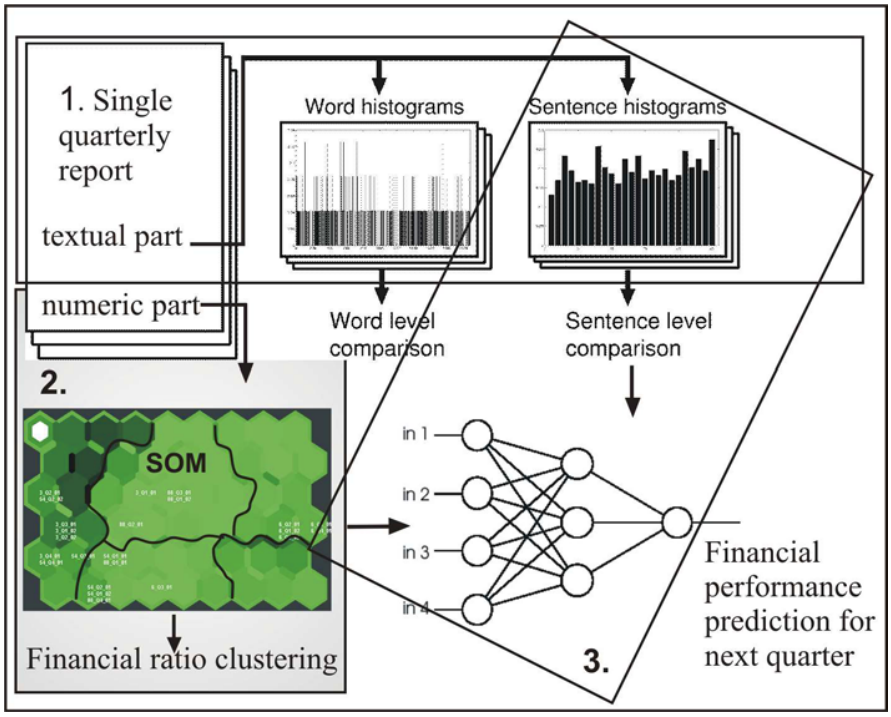


Figure 6-1. The three-step methodology for forecasting financial performances from quarterly reports

Following the strategy of collecting data from the Internet, the dataset was collected from quarterly reports for three telecommunications manufacturers: Nokia, Ericsson, and Motorola, for years 1999-2001. The dataset consists of both quantitative and qualitative data, so that the quantitative data consist of a number of calculated financial ratios, and the qualitative data of the textual description from each report.

Using the SOM, seven financial ratios describing the performance of telecommunications companies for years 1995-2001 were clustered. By carefully analyzing the output map, six major clusters of companies were identified. The two classes of best-performing companies showed good profitability ratios, with very high values in Operating Margin, Return on Total Assets, and Return on Equity. There was one class of well performing companies whose profitability is fairly good and, particularly, their ROE values are excellent, two classes of moderately performing companies who possess decent profitability, good liquidity, with good values in Equity to Capital. The last class, poorly performing companies, was

characterized by low profitability and solvency, average liquidity, and varying Receivables Turnover ratios.

The results from the prototype matching for Nokia, Ericsson, and Motorola were obtained by matching every quarterly report against the entire qualitative data collection. The clusters of the most semantically similar reports to every chosen prototype-report were discovered. The similarities in sentence construction and word choice, which constitutes the language structure and written style, determined the clusters. Word choice has a smaller impact on the clustering results than the sentence construction, since quarter names and proper names, e.g. Nokia, Motorola or Ericsson, did not influence cluster construction on the sentence level clustering.

By combining the patterns discovered by mining the quantitative and qualitative parts of quarterly reports, the summary table was constructed. It contains the prototype-report in the header and the four semantically closest matches in the consequent rows, with the indexes of the class that a particular report belongs to from the quantitative clustering. The information from the summary table served as input data to the MFNN in order to predict the future financial performance of the analyzed companies. An individual MFNN was constructed for every analyzed company because each company has its own trend, cycle of development, and speed in applying changes that affect its performance. Additionally, every company has a unique style to describe its evolution during the reported period in the quarterly report. For validation, the predictions were compared with the real performance of the companies in the following quarters. The limitations of the study, such as size of the data set, are discussed more fully in *Paper 7*. The prediction task can be performed via different soft classifiers. Finite mixtures were reported to be more successful as opposed to neural network and fuzzy rulebased classifiers by (Koskimaki, Gloos et al. 1998)

6.3 Integration of Quantitative and Qualitative Data Mining into a Knowledge Building System: the Conceptual Model

The financial sector forces decision makers, their partners, competitors, investors, analysts, and stakeholders to transform new data into valuable knowledge very quickly in order to react to rapidly changing conditions and make crucial decisions in a timely manner. Very often valuable knowledge resides in qualitative and quantitative data that are dispersed in a number of different storages and sites. A *knowledge building system* can be viewed as the next step in the evolution of IS and AI, because it transforms low-value data – a raw statement of the facts in the form of symbols that merely exist and have no significance beyond that, into high-value knowledge that is organized into meaningful patterns applicable for particular purposes (see Section 2.1). Eventually, the company's desire for deeper insights and actionability drives the need to shift the focus of the analysis to more complex data that provide greater contextual relevance. Any executive or group of decision makers can utilize but a small fraction of all data sources available. Strategic information from differently organized sources tends to overlap considerably or partially, or even contradict each other. Sometimes, managers and decision makers have to put significant effort into corroborating their information by using a number

of sources (Mintzberg 1973). A knowledge-building system that collects basic facts from distributed sources on the Internet and intranet, rationally, is based on the use of a multi-agent software system and aims at assisting managers in using a large range of information sources. A multi-agent software system consisting of a collection of individual software agents, each of which provides a certain task (Lesser 1995) and/or uses different DM techniques, is a possible solution for accomplishing the task of knowledge discovery from qualitative and quantitative data. These software agents are designed to be cooperative, mobile, learning, and manageable in a multidimensional data environment that has often been described as "data rich but information poor" (Liu 2000). A knowledge-building system tends to become a remedy for information ignorance, ambiguity, irrelevance, and overload.

The conceptual model of a knowledge-building system is based on a society of software agents, each of which exhibits intelligence by using different data and text mining methods. The software agents that attempt to execute tasks on behalf of a business process, computer application, or an individual, are well suited for dealing with collecting, processing, and compiling vast volumes of dynamic data from distributed sources. The system could monitor new financial updates from a variety of sources, and calculate and compare financial ratios for different companies. These data could be used for various tasks, for example, financial benchmarking (Bendell, Boulter et al. 1998), assessing creditworthiness of different companies (Tan, van den Berg et al. 2002), clarifying companies' strategies by analyzing the economic environment at the macro level (Lansiluoto, Back et al. 2002), and evaluating countries' relative economic performance (Costea, Kloptchenko et al. 2001). The model suggests the integration of several computing techniques, namely SOM for clustering quantitative information, decision trees and/or multinomial logistic regression for classifying new cases into previously obtained clusters, prototype-matching for semantic clustering of qualitative information, and techniques for text summarization.

Figure 6-2 depicts the proposed conceptual model of a knowledge-building system that consists of six agents, namely the *Data Collection Agent,* the *Generic Mining Agent,* the *User Interface Agent,* the *Clustering Agent,* the *Visualization Agent*, and the *Interpretation Agent*. Each agent carries out its own functions and uses information provided by other agents connected to it. These agents handle three main activities (that are provided by three autonomous agents): data collection and storage (*Data Collection Agent*), searching for hidden patterns (*Generic Mining Agent*), and user-interface design (*User Interface Agent*). The Knowledge Building System aims at creating new knowledge by consolidating the obtained new information from the Generic Mining Agent that can be executed by the *Data Mining* and *Text Mining Agents*. The Knowledge Building System will behave reactively to the goal of the system.

Figure 6-2. Architecture of a Knowledge Building System

The Data Collection Ag*ent* collects, assembles, and sorts the quantitative and qualitative data from various Internet resources, such as *Bloomberg*, *Reuters*, *Wall Street Journal*, *MSNBC*, and individual companies' web sites. These data consist of, for example, market updates, quotes, financial reports, market reports, etc. The data collection engine should run through a number of specified and predefined websites similarly to EDGARSCAN from PricewaterhouseCoopers (http://www.pwcglobal.com/gx/eng/ins-sol/online-sol/edgarscan/).

The User Interface Agent is responsible for providing the communication channel between the system and the human user who chooses the goal of the system. It establishes interaction with the users of a system by offering them to choose the future purpose for the system, i.e. it is guided by the users, and, thus, is shaded gray. It offers the choice of a number of tasks defined by a user in and is associated with the choice setup of the system. For example, two possible applications are financial benchmarking and credit rating. These tasks are defined by the data (numeric and textual) included, as well as by the importance placed on each piece of data (for example, the importance of a particular financial ratio) that are determined by user interface agent. The system also allows for drill-down capabilities, for example, viewing individual financial ratios or the actual financial values that contribute to a particular ratio. In short the agent presents the system options, takes the user input commands, and shows the final results after it has interacted with the other agents.

Depending on what mining techniques and data are used, there are two main instances of the Generic Mining Agent: DM Agent (Figure 6-3) and TM Agent (Figure 6-4). The Generic Mining Agent is seen as a generic class (in programming terms), which does not exist physically, but rather is an abstract class that is implemented via its instances. Whatever the instance of the Generic Mining Agent is, it includes at least three activities in data processing: clustering the data, visualizing the intermediary results of the previous process, and interpreting the final results. The first two steps could be applied cyclically if one, for example, applies clustering via visualization. The clustering techniques are instance dependent, in the sense that we can apply different clustering algorithms when performing data and text mining. One has three agents for the three distinct steps in

data processing: the *Clustering Agent*, the *Visualization Agent*, and the *Interpretation Agent*.

The distinction between the two instances of the Generic Mining Agent is based on the data type to be mined and the mining techniques that they use. The Data Mining Agent is used for processing numeric data; and Text Mining Agent is used for processing text data.

6.3.1 Instances of the Generic Mining Agent: Data Mining Agent

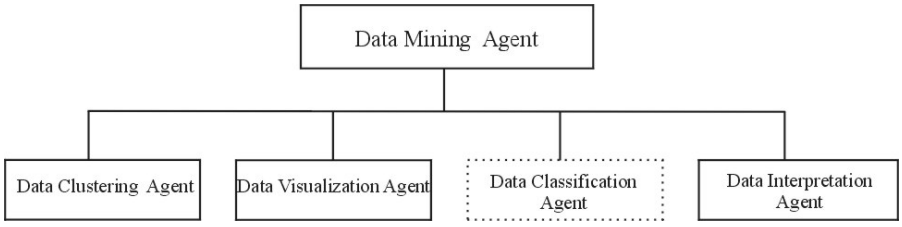


Figure 6-3. The DM Agent

The DM Agent, represented in Figure 6-3, is responsible for numeric data processing and pattern discovery. The DM Agent provides the Knowledge Building System with the cluster that a company (or other data, depending upon the intended goal) belongs to, as well as the characteristics of the clusters (high profitability, low solvency, etc.), i.e. the results of the entire clustering. The Data Clustering Agent calculates the chosen financial ratios for the chosen companies, standardizes the data, and clusters them using the clustering ability of the SOM. The Data Visualization Agent presents a visual U-matrix map on which chosen labels (data) are displayed, along with feature planes for every financial ratio used in the analysis. After the visualization of the clusters on the map provided by the Data Clustering Agent, the *Data Classification Agent* is used for creating classifiers, such as a decision tree and/or a multinomial logistic regression model for classifying new financial data (Costea et al. 2002). The classification models that achieve the highest accuracy in the training phase can be further used for performance prediction. Then, using all the information from the previous agents, combined with knowledge from other agents in the system, the Data Interpretation Agent would attempt to explain the findings. For example, in quantitative clustering, it is important to find explanations for a particular event, such as decreased profitability. This type of information is published in the textual parts of the annual reports and other financial statements.

6.3.2 Instances of the Generic Mining Agent: Text Mining Agent

The TM Agent, represented in Figure 6-4, is responsible for processing textual information, and identifying the essential indications and concepts in it that can potentially enrich the understanding of results obtained by Data Mining Agent. It could use the *Summarization Agent*, which deals with domain information, to

create news summaries for any chosen company, or for general market information for any chosen time period, and report them to the user. Then, the Text Clustering Agent performs financial statement clustering using the prototype matching method and reports which financial reports are close in meaning to each other, how accurately the textual description corresponds to the real financial performance of companies represented by the quantitative data, and draws decision maker's attention to the discrepancies in clustering results and misused concepts. The Text Visualization Agent presents a visual U-matrix map with a cluster representation and the labels of the companies, which are clustered according to the similarity of their financial statements. The Text Interpretation Agent has the same functionality as the Data Interpretation Agent, the difference being the type of data that is processing.

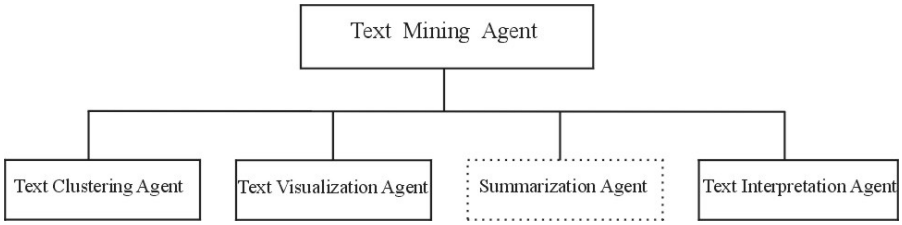


Figure 6-4. The TM Agent

6.3.3 Multiagent Knowledge-Based System at Work

Linking Data and Text Mining Agents, especially Data and Text Interpretation Agents, to support, deliver, and explain the patterns discovered in qualitative and quantitative data aims at building more accurate knowledge about the investigated phenomena, deeper understanding of the causes and forces effecting the phenomena, and explaining the possible outcomes. For example, when a multiagent system is set on a financial benchmarking task of a user-determined set of companies, the Data Collection Agent gathers quantitative data from individual companies' homepages, and from global financial sources such as Reuters and CNBC. The DM Agent starts and discovers benchmarks for chosen companies from the quantitative data, after the Data Clustering Agent has calculated the chosen or standard (recommended) set of financial ratios, grouped them, and ranked the companies according to their performance. The Data Visualization Agent represents the clustering results on a map that is easy to interpret, i.e. shows where the company is positioned and how the company's position and rank depend on each individual financial ratio. The Data Classification Agent classifies updates in financial data for chosen companies. The Data Interpretation Agent provides drill-down capabilities to explain why one specific company is ranked in some particular way. Additionally, the Data Interpretation Agent borrows some textual explanations from the Text Interpretation Agent to confirm or reject results from the quantitative DM. At the same time, the TM Agent processes qualitative data in form of free text, news stories, and companies' financial reports gathered from specified data sources.

The Text Clustering Agent groups these different text pieces according to their semantics and similarities in key concepts. The Text Visualization Agent depicts the similarities and differences between collected text documents in an intuitively understandable format, such as a semantic SOM, or collocational networks (see Section 7.3.2). The Summarization Agent provides condensed ideas from the specified text document in a one-sentence or one-paragraph format. The Text Interpretation Agent allows drill-down capabilities, and exchanges findings from the TM Agent and the DM Agent to explain why the financial position of a particular company chages dramatically. A dramatic change can be explained by one-time pay-off to settle a lawsuit, for example.

Overall, almost all traditional tasks of data and text mining that were described in Chapter one can be executed. Additionally, other activities for creating a knowledge building system, such as constructing classification models in the case of the DM Agent and information summarization for the TM Agent, are to be implemented. Two agents can perform these two different activities: the Data Classification Agent (see Figure 6-3, dotted-line rectangle) and the Summarization Agent (see Figure 6-4, dotted-line rectangle).

There are, of course, a number of problems associated with building a system of this complexity based on data that is freely presented on the Internet. The problems are discussed thoroughly in *Paper* 5. They can be divided in, at least, two categories: limitations that are specific for each individual agent and techniques embedded in agents, and limitations regarding the integration of different agents.

The combination of data and text mining techniques allows further exploration of multiformat data sources, and the extraction of more complex patterns in them. Those patterns contribute to a richer understanding of the underlying data phenomena and lead to creation of additional knowledge and insights for potentially quicker decision making. In Chapter six, I provided an example of complex pattern discovery from quarterly reports. I used the patterns discovered by combining the SOM and the prototype matching method for forecasting the future financial performance of leading telecommunications companies. Here, I suggested a way to exploit the prototype matching method in combination with the SOM and feedforward neural networks to deliver additional insights, such as forecasting future financial performances, into phenomenon described in quarterly financial reports (research objectives *d* from Section 1.3) and deliver this new strategic knowledge to the decision makers. An explanation of the possibilities for combination of the prototype matching method with other DM methods for financial benchmarking, performance analysis, etc. follows in Chapter seven. Moreover, I described the architecture of a knowledge-building system for complex pattern discovery and interpretation based on multiagent technology and a combination of mining methods. The conceptual illustration of how knowledge extraction can be performed from low-level textual and numeric data with the utilization of the prototype matching method (research objective *c* from Section 1.3) was given.

7. MINING THE CONTENTS OF FINANCIAL REPORTS

In Chapter one, I described text as the written phenomena of NL and explained how textual information has surpassed human and computer capacities of intelligent processing. Chapter three examines textual information overload, and IT and IS that help to deal with it in support of managerial work. I elaborated further on the framework of TM that should be implemented as a core in information systems that aim at discovering valuable knowledge from massive of text. In Chapter five, I presented the content-based prototype matching method that was applied in the course of my research to accomplish various TM tasks. In Chapter six, I showed how this particular method can be used in combination with DM methods to discover more complex patterns from data that can be utilized, for instance, for predicting companies' future financial positions.

In this chapter, I demonstrate how TM of financial reports through the use of the prototype matching method could save time and efforts on behalf of decision makers in accomplishing benchmarking tasks. Mining the contents of companies' financial reports can bring new insights in understanding the financial position and strategy of any particular company. This chapter is based on research carried out in two papers. In *Paper* 2, I conducted a series of experiments on using prototype matching for mining the quarterly reports of telecommunications companies. In *Paper* 6, I compared TM results obtained by prototype-matching clustering of the same quarterly reports to the results obtained by linguistic analysis of these.

In the following three sections (7.1-7.3), I present quarterly financial reports, as a subtype of annual financial reports, which are the most important mediums of companies' communication with the investing public. I describe how elements of the investment community read and process financial reports, what sort of information they focus their attention on, and what kind of indications they look for in it. I explain how the prototype matching and collocational networks can be used for extracting meaning from quarterly reports without thorough reading of the reports.

7.1 Financial Reports: Annual and Quarterly Reports

Modern corporate communication includes things such as fact books or fact sheets, news releases, websites, financial reports, and meetings or conference calls with analysts or investors. New communication and information technologies have simultaneously increased the types of media and decreased the companies' costs of direct communication with all elements of the investment community. Consequently, companies' choices of communications with the investment community involve not only matters of content disclosure but also of delivery media. Nonetheless, the annual report remains a centerpiece of corporate communication. Its historical and symbolic value, along with the breadth of its distribution, makes it able to convey the company's facts and message.

[12] The following chapter discusses annual and quarterly reports, their aims and language, as they are used in the modern Western civilization.

While annual reports have a long history, it was not until 1933 that U.S. securities regulations established a set of disclosure requirements for publicly traded firms (Bricker 2000). Annual reports are published for the benefit of shareholders. Financial accounts are required by law to be published. Special regulations specify how these must be published. Even small companies must report their accounts, which can be seen as annual reports. While specific content is not required for annual reports, they typically include extensive financial information. Annual reports have evolved and now contain several key elements, including a set of financial statements with related notes and auditor's letter, a CEO's letter, and Management's Discussion and Analysis. Publicly traded companies publish annual reports and submit filings to the authorities, for example, in the USA to the Securities and Exchange Commission (MIT-Libraries 2001). Regulations have specified that certain types of disclosures and discussions have to be included in a company's financial report. Disclosing certain discussions to the public aim at preventing companies from providing false or incomplete information to mislead investors and disturb the market. Research indicates that annual reports have multiple audiences, including stockholders and the financial community, and varying objects, ranging from questions of stewardship to outright promotions of the company (Hawkins and Hawkins 1986). Annual reports, while being important documents to stockholders and financial communities and firmly regulated, are still controversial documents. They generate disagreement regarding audience, objectives and credibility. Within the annual reports, the five types of the most interesting information to explore are recent development and outlook for the company's industry, annual company earnings, company's position in the marketplace, risks to the company, and recent significant events, according to SRI International, Investor Information Needs, and the Annual Reports cited by (Bricker 2000).

During the last decade there have been a number of studies attempting to resolve the controversial nature of company annual reports as a medium of corporate communication with investors. Thomas (1997) concentrated on transitivity, thematic structure, context, cohesion, and condensation in the language used in the reports, by studying the annual reports of a machine tool manufacturer during a period that began with prosperity and ended with severe losses. During the time of analysis, the structure of the language used in the reports had changed: the researchers saw an increase in the use of passive constructions, which present the actor (i.e. the company) as "being" rather than as "doing" as profits decrease. This indicates management's attempts to present itself as a victim of unfortunate circumstances to create an impression of objectivity to the reader. Nonetheless, when the company's profit increased, it presented itself as aggressive and forward moving through the use of the active voice and verbs with both an actor and a goal. A close look at the language structure in the letters to stockholders made by Thomas (1997) showed that the structure of the financial reports indirectly reveals some strategic things that the company may not wish to announce directly to its outside audience. A similar opposition between the actions of the company and circumstances created by nonhuman agents has been noted by (Kendal 1993). The researcher introduced the concept of drama, and classified the words and phrases

describing actors and objects in the drama into two groups, *god terms* and *devil terms*. Some god terms such as *growth*, *increased sales* and *competitive position* are words representing unquestionably good concepts in the eyes of the company. Devil terms are terms like *losses*, *decline in sales,* and *regulations*. In the annual reports analysed by Kendall, the company clearly plays the part of the hero, and the U.S government is presented as the demon of the drama, trying to obstruct the actions of the hero in the American economy setting.

Several other studies have been made with a focus on the relationship between the readability of the annual reports and the financial performance of a company (Subramanian, Isley et al. 1993). As research has shown, the annual reports of the companies that performed well were easier to read than those of companies that did not perform well. Moreover, writers of annual reports see the message they put in the report as their personal representation (Winsor 1993). The annual reports are not only the best possible description of a company, but are also a description of a company's managerial priorities. Thus, the communication strategies hidden in annual reports differ in terms of the subjects emphasized when the company's performance worsens (Kohut 1992). This body of research suggests that annual reports are valuable to investors because they contain rudimentary information that is relevant to forecasting the future performance of a company. Another important conclusion of the study by Thomas (1997) was the confirmation of the Pollyanna Hypothesis that states that regardless of the financial state of the company, the language in the annual letters is predominantly positive. While examining the content of annual reports in the best and poorest performing companies in the Fortune 500, (Kohut 1992) concluded that such technical characteristics as word count and number of sentences may serve as predictors of the future performance of a company. The changes in communication strategies are of great interest to human experts, and can potentially be detected by automatic TM methods because, as was noted in Chapter one according to Witten, Nevill-Manning et al. (1996), computers do not have to "understand" the text in order to extract useful information from it.

The language of quarterly reports has not been studied as extensively as annual reports, both within linguistics and business communication studies. However, for the subsequent studies, I used quarterly reports since they contain more up-to-date information and influence decision-making in the short-term. As a genre, quarterly reports closely resemble annual reports because the same writers produce quarterly and annual reports for the same readers within the same community. The reports have similar structures, conventions, and communicative purposes. The basic functions of a quarterly report are similar to those of an annual report, but the time spans are different. The study of the linguistic contents of quarterly reports has nevertheless been overlooked in favor of the study of the language of annual reports. From a short-term perspective, quarterly reports are important means for companies to appraise past performance and project future opportunities to the readers, who primarily consist of investors and analysts.

7.2 Exploring the Meaning of Annual/Quarterly Reports

Hyland (1998) reminded his audience that, although frequently criticized as "five pages of information and 40 pages of fluff", the production of annual reports is a major corporate endeavor that costs about 5 billion US dollars in the United States alone. Annual and quarterly reports are important tools for building credibility and imparting confidence, and convincing investors in the effectiveness and soundness of company's strategy. Even though research suggests that financial and investment decisions are based on financial data (Siegel 1994), analysts and decision makers widely read the annual and quarterly reports to validate major quantitative measures, predict the company's future, make intuitive judgments about the company's position, and evaluate strategic priorities that are often coded in the numeric information in the reports.

Financial analysts examine and actively seek nonfinancial information in the production of their reports, such as market share, competitive position, industry and economic conditions, the competitors' capabilities and products, business risks and uncertainties, value name, research and development expenditure and other intellectual property, and the company's principal strategy. Grant, Fogarty et al. (2000) examined the annual reports of 16 Fortune 500 companies in eight different industries. The analysis showed that the content of annual reports is contingent upon company culture, management, performance, regulatory requirements, and a variety of other factors. The study found that changes in the composition of the company had an impact on the content of the annual report, but not in easily predictable ways. The profitability of companies did not systematically affect the amount of their disclosures for the current year, but affected the composition of disclosures and can be used for sensing a company's future.

In order to assess the comparability of financial numeric and nonfinancial intangible information hidden in annual (quarterly) reports, the entire enormous quantity of reports must be read and analyzed. Industrial analysts have individual intuitive methods to uncover indications and hints about the future financial performance of the company by reading their financial reports and making "professional guesses" while they follow the development of the company for several years. Manual detection of important hints is a time-consuming process that requires a lot of training, background knowledge, and experience. The availability of computerized TM solutions for detecting companies' future financial intentions can be used in two ways: it can lighten the work load of analysts, saving them money, time, and efforts, or it can also conceivably help companies' officials to sway the public opinion by manipulating and faking positive writing style. The ready-made commercially available solutions, such as (FinGloge 2003), do not seem to provide solutions for the entire range of TM tasks introduced in Chapter one. FinGlobe, for instance, accomplishes only summarization and translation of annual reports, saving money and efforts for decision makers that would be expended for manual summarization, but does not discover "new" information in reports. Another attempt for computer based analysis of financial reports was undertaken by Osborn, Stubbart et al. (2001), to explore the strategic goals of the companies. The researchers examined themes in a president's letter to stakeholders

from annual reports in order to derive the mental models and strategic goals of the companies. After performing computer-aided content analysis of more than four hundred president's letter to shareholders, and examining the empirical linkages between themes in annual reports and the company's performances, they concluded that the text in annual reports reflects the strategic thinking of the management of a company

While the automatic analysis of financial ratios from annual reports is common (Eklund, Back et al. 2002; Martín-del-Brío and Serrano-Cinca 1993), the automatic analysis of their textual part and qualitative nonfinancial information is more rare. Attempts to semi-automatically analyze a company's performance by examining the quantitative and qualitative parts of annual reports have been made in a number of studies (Back, Toivonen et al. 2001, Kloptchenko, Eklund et al. 2002). Back, Toivonen et al. (2001) indicated that there are differences in qualitative and quantitative data clustering results due to a slight tendency to exaggerate the performance in the text. Kloptchenko, Eklund et al. (2002) attempted to explain this tendency using quantitative analysis by means of the SOM for financial ratio clustering, and qualitative analysis by means of the prototype matching method. In both studies the researchers noticed that combining two mining techniques for two different types of data describing the same phenomena could bring additional knowledge to a decision maker (see Chapter six). While the qualitative part of annual/quarterly reports explicitly states information about a company's past performance, they also contain some indications of its future performance, i.e. the tables with financial numbers indicate how well a company has performed, while the linguistic structure and written style of the text indicates what a company intends to do. The study has shown that sophisticated semi-automatic analysis of the style and content of the financial reports can help reveal insiders' moods and anticipations about the future performance of their company.

7.3 Mining Annual/Quarterly Reports

Annual and quarterly financial reports seem to offer excellent material for TM because they not only contain crucial and lucrative information for instant decision-making, but also they have limited, nonambiguous vocabulary. According to the linguistic studies described in Section 7.2, exploring nonfinancial and qualitative data from annual reports and presidents' letters to stakeholders can more carefully assist in the benchmarking of the companies, converging strategic goals of the companies, and predicting companies' future priorities and financial performance. As was noted earlier, there are no satisfactory TM tools for discovering hidden indications and hints upon which analysts base their predictions, decisions and professional "intuitive guesses." I have applied prototype matching for finding and retrieving those hints semi-automatically for accomplishing financial benchmarking (*Paper 2*) and future financial performance prediction (*Paper 7*) of the companies whose reports were analyzed. Moreover, using the linguistic method of building collocational networks, the contents of reports were visualized, and the TM results from prototype matching were validated in *Paper 6*. The research in *Paper*s *2*, *6,* and *7* is based on a data collection of quarterly reports

from the leaders of telecommunication companies: Nokia, Motorola, and Ericsson. The tasks in Figure 7-1, which were accomplished using the prototype matching method, are united by the dotted-line rectangles and highlighted using italic font. The central technology for accomplishing those tasks is content-based clustering using the prototype matching method, labeled number 1 in Figure 7-1. The techniques that are used in combination with the main one are shaded gray.

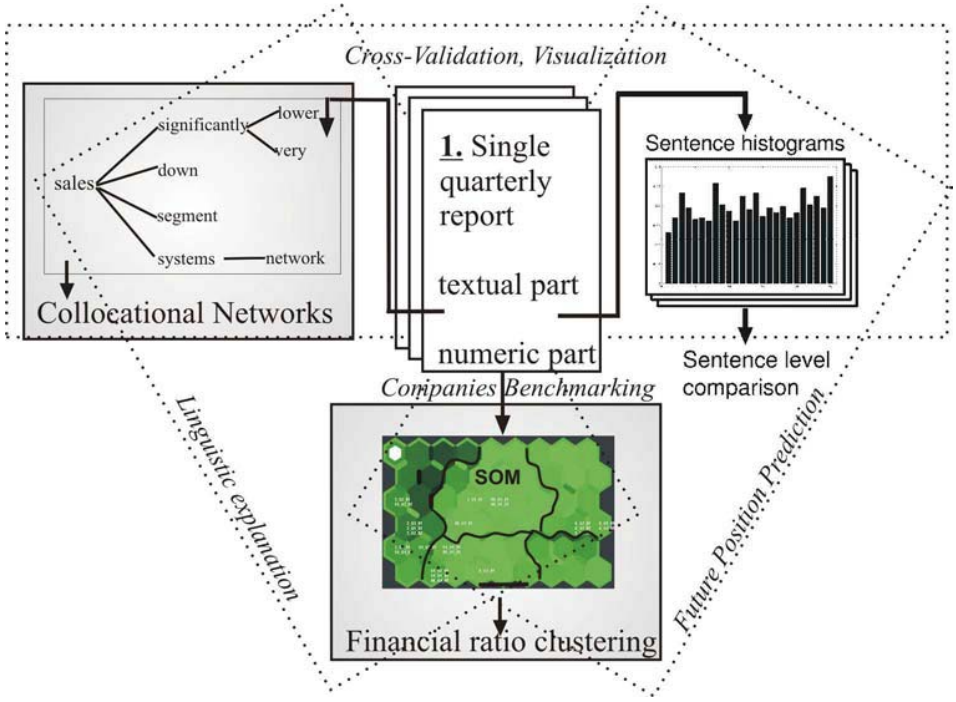


Figure 7-1. Mining Quarterly Reports: Tasks and Techniques

In order to apply the content based methodology of prototype matching, all the reports should be accumulated in the common repository, by either manually scanning the reports if they are paper-based, or by using software agents to collect them from the Internet if the reports are available online. Consequently all the reports were processed according to the procedures described in Chapter five. In other words, every word and sentence from the reports was encoded, and common and individual word and sentence level histograms were constructed. The intended user could match every quarterly report against the entire data collection and the system would compare all of the quarterly reports in the data collection by calculating the Euclidian distance between their sentence histograms. Within the chosen recall window, the user receives the list of the closest matches to the prototype-report that are the most semantically similar. All the reports in one cluster that are the closest-matches to the prototype would convey the same message and contain similar "hints." Therefore, instead of reading and comparing all the reports in a data collection, the user would pick only one report (prototype report) to

determine the main characteristics of the group constituted by the closest-matches reports. The obtained results could be presented to the user in form of a table with the closest matches, such as the three closest matches to Ericsson reports for four quarters of the year 2000 and the first two quarters of year 2001 in Table 7-2.

Table 7-2. Example of the three closest matches to Ericsson reports in the limited data collection (Sentence level)

Ericsson2000Q1	Ericsson2000Q2	Ericsson2000Q3	Ericsson2000Q4	Ericsson2001Q1	Ericsson2001Q2
Nokia2000Q1	Ericsson2000Q3	Ericsson2000Q4	Ericsson2000Q3	Ericsson2001Q2	Nokia2001Q3
Nokia2000Q3	Nokia2000Q2	Ericsson2000Q2	Motorola2001Q2	Ericsson2001Q3	Ericsson2001Q1
Motorola2001Q2	Ericsson2000Q1	Ericsson2000Q1	Nokia2000Q1	Nokia2001Q3	Ericsson2001Q3

For example, for the Ericsson report from the first quarter of 2000 the closest report by content on the sentence level is the report from Nokia the first quarter of 2000. The second closest is the report from Nokia, 2000, quarter three. This means that the Nokia reports from first and third quarters of 2000 and the Ericsson report from first quarter of 2000, have similarities in sentence construction and word choice, which constitutes the language structure and written style. In other words, the prototype-report from Ericsson quarter one, 2000, and its closest match, Nokia quarter one, 2000, share different sets of common patterns than with the closest matching report for Nokia quarter 3, 2000.

In this dissertation I present how the discovered closest matches combined with other techniques are applied to accomplish specific business-related tasks, such as competitive benchmarking which is discussed below, along with predictions for future financial performance, which was described in Chapter 6.2.

7.3.1 Competitor Financial Benchmarking Using the Prototype Matching Method

According to Merriam-Webster's Collegiate Dictionary, benchmarking is defined as "the study of a competitor's products or business practices in order to improve the performance of one's own company"[13]. Benchmarking is the process of comparing the activities of one company to those of another, using quantitative or qualitative measures, in order to discover ways to increase effectiveness (Eklund 2002). Bendell, Boulter et al. distinguished four types of benchmarking methods: *internal*, *competitor*, *functional,* and *generic* (Bendell, Boulter et al. 1998). Internal benchmarking is a method in which the performance of one part of an organization is compared to other parts. Competitor benchmarking is much more difficult to implement than internal benchmarking because competitors are used as benchmarking partners. Competitor benchmarking contributes to achieving the best practices within an industry. Functional benchmarking involves making comparisons with non-competitor organizations, which are good at a particular activity that a company is interested in, and can lead to novel practice in the industry. Generic benchmarking is the most extensive form of benchmarking, and

[13] http://www.m-w/dictionary.htm

involves benchmarking across several, not necessary related, industries. This particular study is an example of competitor benchmarking using quantitative and qualitative data from quarterly financial reports. Different companies, that is, competitors, within the same industry are benchmarked against each other using various financial ratios and nonfinancial qualitative data from their quarterly reports. The detailed process of using the SOM to cluster and analyze 89 companies within the telecommunications industry for years 1994-2001 is described in *Paper 2* and in (Karlsson, Back et al. 2001). Briefly, the SOM technique creates a two-dimensional map from *n*-dimensional input data. This map resembles a landscape in which it is possible to identify borders that define different clusters (Kohonen 1997). These clusters consist of input variables with similar characteristics and are illustrated below in Figure 7-3.

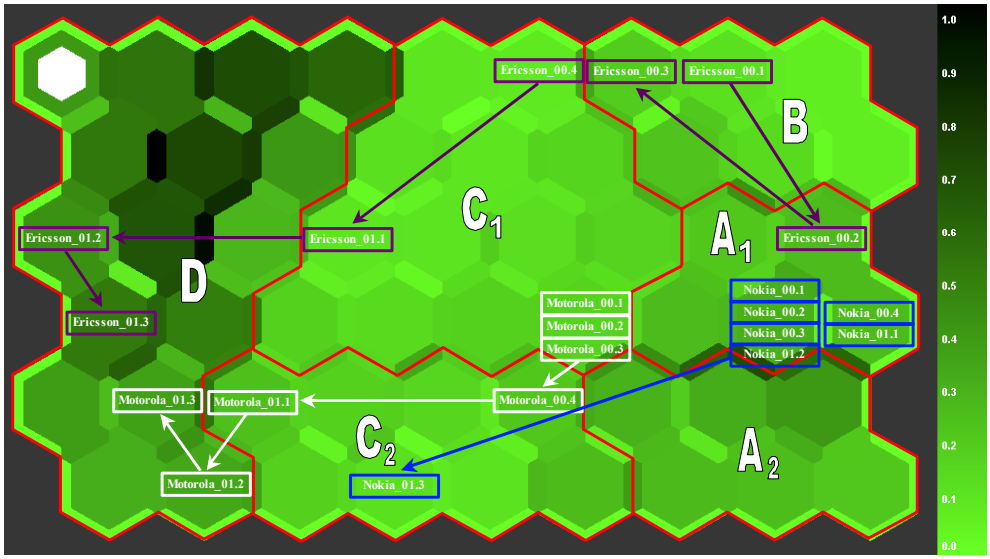


Figure 7-3. The Identified Clusters and the Quarterly Movements of Ericsson, Motorola, and Nokia

Summing up the results from the studies of Karlsson, Eklund et al. (2001) and Karlsson, Back et al. (2001), six major clusters of companies were identified ranging from best performing to the worst performing companies (in descending order: A_1, A_2, B, C_1, C_2, and D). The created SOM allows the determination of the competitive position of any particular company within the industry by visualizing the clustering results. The arrows on the SOM reflect the quarterly changes in the financial positions of the analyzed companies.

Here I provide an example of the benchmarking process and analysis of quantitative and qualitative data for Ericsson during 2000 and 2001. It was discovered by Karlsson (2001) that during the first and third quarters of 2000, Ericsson was situated in Group B among well performing companies, the same group as for the previous six years. The second quarter was particularly good for

Ericsson, and the company moved into Group A_1, consisting of the best performing companies. During this quarter Ericsson showed significantly increased values in its financial ratios. In the fourth quarter Ericsson began to experience difficulties that resulted in falling back into Group C_1, consisting of the moderately performing companies. This was mainly due to decreased profitability and solvency. In 2001, Ericsson experienced severe difficulties in the telecommunications market; almost all key ratios showed decreased values during the first quarter, and Ericsson dropped within Group C_1, ending up close to Group D – the cluster of the worst performing companies. In the second quarter of 2001, Ericsson showed very poor performance with low profitability and solvency and was positioned in the group of worst performing companies. The only category that improved during this period was liquidity. Even though the negative result was slightly smaller, in the third quarter of 2001 Ericsson managed to accomplish a minor improvement in the key ratios, the company remained in Group D. In other words, Ericsson traversed the whole range of financial performance in five quarters moving from a group of good performing into the best and, finally, falling to the group of worst performing companies.

Qualitative data from quarterly reports are according to linguistic studies able to explain and predict variations in financial performance. By obtaining TM results from content-based clustering of qualitative data from quarterly reports, it was discovered that the first indications of worsening performance occurred in the third analyzed quarter, when reports stating poor financial performance started to fire as the closest matches. Even though Ericsson performed at the average level, and was in Group C_1 in the first quarter of 2001, the majority of its closest matches were from the worst performing companies, such as from Ericsson 2001 quarters 2 and 3 from Group D. The consolidated mining results are represented in Table 7-4, where the header of each column contains the prototype-report followed by its three closest matches. The bold letters by the report codes denote the cluster from the quantitative clustering that a particular report belongs to.

Table 7-4. The closest matches to every report for Ericsson and their benchmarking position

Ericsson2000Q1 B	**Ericsson2000Q2 A_1**	**Ericsson2000Q3 B**	**Ericsson2000Q4 C_1**	**Ericsson2001Q1 C_1**
Nokia2000Q1 **A_1**	Ericsson2000Q3 **B**	Ericsson2000Q4 **C_1**	Ericsson2000Q3 **B**	Ericsson2001Q2 **D**
Nokia2000Q3 **A_1**	Nokia2000Q2 **A_1**	Ericsson2000Q1 **B**	Motorola2001Q2 **C_2**	Ericsson2001Q3 **D**
Motorola2001Q2 **C_2**	Ericsson2000Q1 **B**	Ericsson2000Q2 **A_1**	Nokia2000Q1 **A_1**	Nokia2001Q3 **C_1**

Some useful information from qualitative data in quarterly reports was captured in addition to determining the competitive benchmarking position of the analyzed companies. Before a dramatic change occurs in a company's financial performance, we see a change in the writing style contained in a financial report. The tone tends to be more similar to the company's future performance. If the company's position will be poorer in quantitative terms during the next quarter, the report of the current quarter tends to become more pessimistic, even though the actual financial performance remains the same. The further development and interpretation of this conclusion is presented below in the next section.

7.3.2 Interpretation of Collocational Networks of Quarterly Reports

In order to visualise the central concepts and their connections within a quarterly report, collocational networks were used (see *Paper 6*). *Collocation* was interpreted simply as "the occurrence of two or more words within a short space of each other in a text" following (Sinclair 1991), a central work within corpus linguistics. Collocational networks are visual constructions of collocations forming a unique frame of reference for any "word" within a given sublanguage (Furnas, Landauer et al. 1987). *Significant collocation* is an important factor that takes place when two or more words occur together more frequently than would be expected by coincidence. Following Williams (1998), significant collocation is measured using the *Mutual Information (*MI) score**,** which compares the frequency of the co-occurrence of a node and collocate with the frequency of their occurrence independently of each other. In these cases the MI score works "counter-intuitively": decreasing as the absolute number of collocates increases. This means that two words which always occur together get a higher MI score if they occur only once than if they occur more frequently.

The contents of each report were analyzed separately, so that pairs of words which are referred to as collocations are patterns which occur within a single text, and therefore cannot be considered to be typical for business English in general. Collocational networks give an opportunity to examine which concepts are emphasized by the company in a particular report and how these concepts are reflected through the words that constitute the nodes of a network. The decision maker can see which concepts are most frequently linked to each other, by revealing which words regularly appear within a close proximity to each other by looking at the presented collocation networks. The method does not always bring out combinations of words that are perceived by speakers of the language to belong together as phrases or compound words, such as *balance* and *sheet*, unless they occur very frequently in the text.

A brief overview of the collocational networks based on Ericsson's quarterly reports shows that the analyzed reports never exhibit the some similarity in their architecture, as the financial performance over the analyzed period of time would seem to indicate. Both the structure and the content of the networks vary considerably. This is also obvious when looking at the texts in the reports: during this period the reports undergo several structural changes that coincide with worsening financial performance. New headings are introduced and old ones are abandoned or reorganized. A particularly remarkable structural change in the networks occurs between the third and fourth reports of 2000. There is also a significant difference between the lexical items used and the number of lexical items in the networks. A full coverage is given in *Paper 6*. The collocational networks for Ericsson reports from the third and fourth quarters of year 2000 are presented in Figures 7-5 and 7-6 respectively. The network for the third quarter of 2000 starts with the most frequent word, *Ericsson*, which is linked to five collocates. One of these collocates, *increased*, is linked to *sales*, which has four other collocates of its own. One of these collocates, *systems*, is linked to *mobile*, which has five more collocates. These linkages mean that the main network for the

third quarter of 2000 consists of three parts, connected by collocational pairs. In addition, there are several separate collocational pairs and small networks outside the main network.

The structure of network for the fourth quarter of 2000 is very different from the structure of reports from the previous quarters of 2000. It consists of a main network attached to the most frequent word, *we*, and a smaller, separate network around *operating*. *We* is a new word in this network, and the most frequent word in the network, which substituted the main word *Ericsson*, which has disappeared. Starting from this network, the company now refers to itself using a pronoun instead of the name *Ericsson*. In addition to these two major networks, there is one separate collocational pair, consisting of two new words, *additional* and *restructuring*. The appearance of the passive voice can be the first indication of the upcoming worsening of financial performance that was discovered by clustering of qualitative data using the prototype matching method.

In the following network, the first quarter of 2001, the change continues. Figure 7-7 depicts the structure of the collocational network from this period of time. This network contains even fewer words than the previous one. Now there is only one word, *expect*, connected to *we,* as opposed to five collocates in the previous network. A new addition is the collocation *efficiency program*, a term bearing obvious negative connotations. Consequently, the closest matches to this report from content-based clustering are mostly reports from poorly-performing companies.

In the last three networks of 2001, more words start to appear and the structures become more complicated. Structurally these networks resemble the networks representing earlier quarters of year 2000. They include different words chosen by the method as the main concepts in the reports. These networks contain collocations such as *restructuring charges*, *increased borrowing* and *efficiency program*, all of these pointing to a negative development within the company.

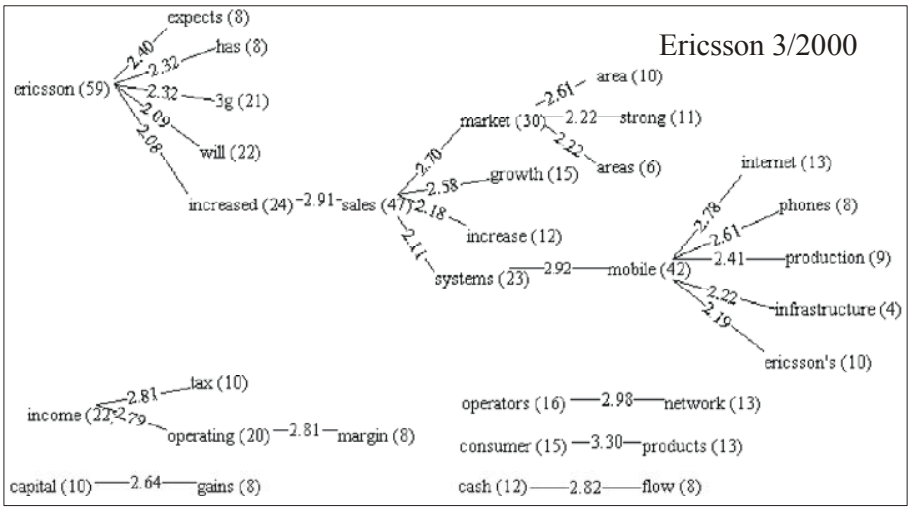


Figure 7-5. Collocational Network for Ericsson report from the third quarter of 2000

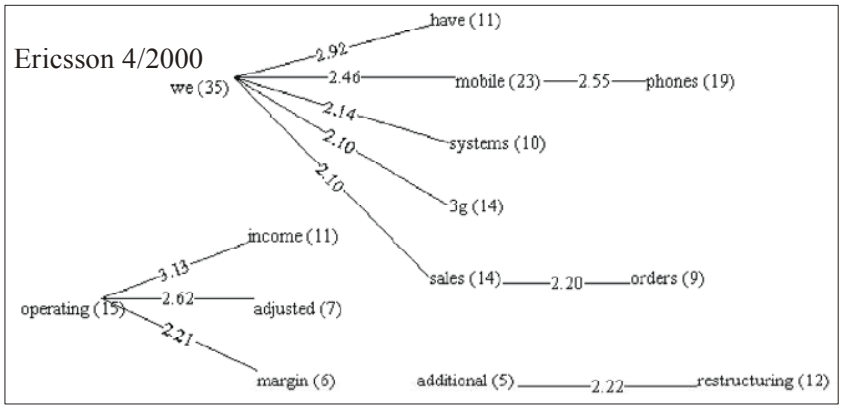


Figure 7-6. Collocational Network for Ericsson report from the fourth quarter of 2000

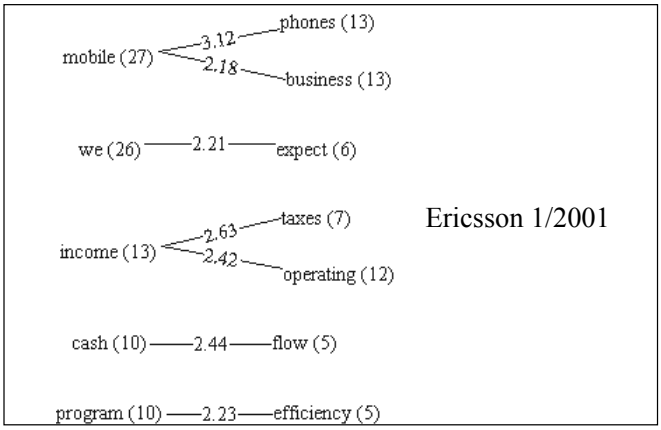


Figure 7-7. Collocational Network for Ericsson report from the first quarter of 2001

7.3.3 Consolidation of the Results from the Prototype Matching and Building Collocational Networks of Quarterly Reports

It is believed that text in a particular context bears more diverse information than numbers do. Annual/quarterly reports, for instance, tend to state information about a company's past performance, but also contain indications of its future performance, i.e. the tables with financial numbers indicate how well, and why any company has performed, while the linguistic structure and written style of the text may indicate how well a company will do. It was a desire to come up with a better scheme to provide ways of finding hidden indications about a company's future financial movements. After looking to benchmarking techniques based on SOM clustering for classification and visualization of the performance of the companies, and mining the textual parts of quarterly reports using the prototype matching

method for the same period of time, the collocational networks of the reports were built in order to validate the heuristic relationship between the writing style and facts stated by the numbers, and to visualize the content of the reports.

Summing up the results presented in Table 7.2 with the logic and interpretation of contents of the report offered in Section 7.3.2, I arrived at the following conclusions. If a company reports a good steady performance over a certain period of time (from groups A_1, A_2, or B in the quantitative data analysis) then the reports from companies with similar performance start to appear among the closest matches to the analyzed quarterly report, e.g. a report from Nokia from the first quarter of 2000. When a company performs well, and expects to continue doing so, the tone of the report is positive with extensive use of optimistic vocabulary consisting of good terms (such as *increase, share growth, higher, our profitability, new, strong demand*), active verbs (such as *doing, not being*), and short clause constructions (such as *operating margin, demand growth, increased share).*

If a company reports an abrupt worsening of its performance, more companies with poor performance (from Groups C_2 or D in the quantitative data analysis) fire up among the closest matches one period before an actual financial downturn has occurred, e.g. reports from Ericsson in the second and third quarters of 2001. The reports contain more conservative expressions with a lot of passive constructions (*we expect, program efficiency)*, and evil terms – nouns and verbs with negative financial connotations (*decrease, slowdown, decline*).

If a company anticipates a worsening of its financial performance in the next quarter, we see more companies with average performance (from Groups C_1 and C_2 in the quantitative data analysis), e.g. a report from Ericsson in the third quarter of 2000, among the closest matches. The tone of the financial report becomes less optimistic and more similar to ones that describe even poorer performance. The style in the report becomes even more conservative, passive, and distanced from the main activities (*we have*, *announced, representing*), using words and short sentence construction with particularly negative financial connotation (*down, sales decline).* The company avoids stating the accomplished results in its quarterly report directly, instead shifting the subjects of emphasis in the report by blaming outside conditions (*sales segment, market share).*

If a company reports average performance that does not change rapidly over time (from Groups C_1, or C_2 in the quantitative data analysis), e.g. Motorola's slow drop from average performance to bad performance, then decision makers see companies with different financial performance among its closest matches. This indicates that a report has no distinctive style, or a sentence construction that might reflect uncertainty about the future of the company.

In conclusion: the proposed framework in Figure 7-1 for mining quarterly reports with a content-based methodology such as the prototype matching method aims at discovering complex patterns in the textual part of the reports and to contribute to making various decisions concerning the financial performance of the analyzed companies by bringing the content of reports to the attention of decision maker in suitable visual format. By looking at the map constructed using SOM clustering, the decision maker determines the relative competitive position of the analyzed company; content-based clustering delivers more insights about what

strategy and direction this company pursues in comparison with other companies from the data collections; finally, collocational networks illustrate the contents of the particular reports that can be visually compared as well.

Collocational networks have illustrated how semantically similar the closest matching reports are to a report-prototype. Because of text multidimensionality, establishing adequate similarities between text documents is hardly achievable. While some collocational networks have outlined the same dimensions of closest-matching reports upon which the prototype-matching method performed its clustering, some other collocational networks have outlined different text dimensions. That led to somewhat discordant results in cross-validation, when occasionally, collocational networks of the closest matches did not resemble each other.

This chapter illustrates the function of the prototype matching method in performing knowledge discovery from financial text collection such as quarterly reports. The utilization of collocational networks helps to accomplish the next step after DM in the KDD process such as representation of the extracted patterns (see Figure 1.1 in Chapter 1) and interpretation of the TM results. Overall, Chapter seven continues to elaborate on the objectives stated in Section 1.3 concerning the combination of the prototype matching method with other methods, and its applicability to the financial domain (research objective *d*). Here I extended the way to exploit the prototype matching method for TM of financial reports, and offered a way to cross validate the results by using linguistic methods.

8. INFORMATION RETRIEVAL BY CONTENT FROM SCIENTIFIC PUBLICATIONS

Prepared as an article compilation, this doctoral dissertation includes a number of separate research projects, all of which have as the common denominator the content-based methodology of prototype matching. This chapter reports the applicability and usability of prototype matching for information retrieval by content from a scientific article collection, which was described briefly in Chapter three. This chapter provides the motivation for this research problem and highlights the major findings that were published in *Papers 3* and *4*.

8.1 Motivation to Information Retrieval by Content of Scientific Publications

During the last years, applied science has become more and more interdisciplinary. In the diverse and multidisciplinary fields such as, for instance, system science, and NL processing, it is often difficult to define clear-cut descriptions of all tracks. The conference chair must be aware of the detailed expertise of the track areas. Track chairs should be proficient in order to route atypical or multidisciplinary papers to tracks with the most appropriate pool of reviewers and prospect audience. Therefore, sorting and allocating papers submitted to a scientific conference in proposed categories and tracks is turning out to be a nontrivial task. First, the conference organizers have a hard time determining and scheduling overlapping sessions successfully. Second, the authors of multidisciplinary research papers face a dilemma concerning the choice of the research track most appropriate for the content of the paper. And third, a conference attendee has a hard time determining which conference sessions are relevant to his research interests. He needs either to browse the entire conference proceedings to identify interesting papers, or to rely on a keyword search. Sometimes even experienced readers, such as track chairmen or members of the organizing committee, encounter difficulties in the determination what track the particular paper should truly belong to. This determination can turn out to be very laborious and time-consuming, and requires a lot of expertise from the conference organizers, who can be considered to be decision makers in the conference setting.

Traditionally, technical or academic papers describe the problem area and the approaches to resolve the problem. Consequently there is more than one main topic that the author introduces and tackles in the paper, for instance, the problem area can be in medicine but the solution can have a mathematical or technical nature. Moreover, the main topic of a paper outlined by an author who submits the paper to a particular track at the conference can belong to several disciplines and, thus, be discussed in more than one proposed conference track. The most popular strategy for routing of conference papers into track is keyword based, which considers keywords as an adequate reflection of the paper's content. Authors are required to first specify topic and subtopic areas for their papers. Authors often have a difficult time selecting precise and informative keywords to adequately describe

their work. As was described in Chapter one, depending on their background, to identify the content of the submitted papers, authors use analogous keywords, which can belong to either the same or different tracks. In addition, the authors and the readers of the scientific articles choose synonymous words for describing the same phenomena, or illustrate different phenomena using polysemous words. While searching a conference text collection for relevant information, readers face problems in constructing smart queries because of word ambiguity, inadequate paper routing into tracks, or because they might not be fully acquainted with the established terminology in a field, or not fully sure about the content of the needed documents. This behavior requires sophisticated IR by content tools that could help users to deal with text collections.

In this chapter, I illustrate the applicability of prototype matching to IR by content from a collection of scientific articles presented at a scientific conference. The system, built on the prototype matching method, has the overall goal of detecting similarities between scientific papers or articles. Scientific articles seem to fit the IR by content task nicely because they are written according to a strict academic writing style, are highly structural, and include well-established scientific jargon shared by people in academia with similar backgrounds. On the one hand, the system aims to assist conference organizers in establishing semantic similarities among the papers automatically. On the other hand, the system aims to assist the attendees in retrieving interesting papers from the conference proceedings based on similarities in their content. A user can take the whole paper, or an abstract from an interesting paper, and use it to construct a smart query. As was mentioned before, the core of the system is "smart" document encoding on different syntactic levels, and document collection clustering. As an experimental data set, 444 scientific papers obtained from The Hawaii International Conference on System Science 2001 (HICSS-34) was chosen. HICSS was selected as a typical example of a multidisciplinary, internationally recognized conference. Scientific textual information along with technical textual information obtained from any conference proceeding is a good sample of explicitly stated material that can potentially suit to the IR by content purpose well. Furthermore, the scientific papers at HICSS-34 were arranged into nine major tracks, which were further divided into 78 minitracks. The organizers made an additional effort to identify six themes that ran across the traditional tracks based on similarities and expansion of the scientific fields in order to help conference participants select the sessions that appealed to them. The outlined six cross-track themes covered 168 papers from 30 conference mini-tracks. The organizing committee assigned a unique identification code to every paper. The code shows what track and minitrack the particular paper belongs to. For instance, the paper "Supporting Reusable Web Design with HDM-Edit" with the conference code INWEB04 was allocated by the conference minitrack chair and authors, into the Web Engineering minitrack from the Internet and Digital Economy track. Besides this track allocation, the paper was classified into an E-commerce Development theme.

In the course of the research on applicability of prototype matching to the task of IR by content from the scientific text collection, a two-level approach was undertaken. First, I considered abstract versions of the papers because normally the

abstracts are submitted earlier to the HICSS minitrack chair. Upon arrival of the abstracts, minitrack chairs made decisions regarding the relevance of the prospective paper to his/her minitrack. Second, I considered full-versions of the papers for analysis. This two-level approach of organizing papers by content repeats the paper routing and allocating processes that the conference organizers and minitrack chairs follow in the determining whether a certain paper belongs to a certain theme or track. The conference organizers and minitrack chairs judge the appropriateness and relevance of a paper by establishing the relevance of its abstract to a track or a theme.

8.2 Abstract-level Analysis

For the pilot explorative study, the abstracts from the entire HICSS-34 conference proceedings database were chosen. Abstracts are designed to condense research for the public eye by offering a preliminary overview of the research in a brief form (dos Santos 1996). At HICSS, track and minitrack chairs collect the abstract versions of the papers to be submitted to their tracks two months prior to full paper submission. Moreover, early submission of abstracts gives the chairs the opportunity to suitably allocate reviewers for the full papers. Based upon the content of the abstracts, they either recommend the authors to submit the full-paper to this particular minitrack or to look for an alternative track/minitrack.

Several separate experiments were conducted to test the ability of the proposed prototype matching method to retrieve papers that are the most similar in meaning from the scientific conference collection. From every abstract I omitted the abstract titles, and author listing as irrelevant and keywords as redundant information. First, I examined the system's ability to retrieve the most similar abstracts from the entire conference collection using any chosen abstract as a prototype query to cluster the collection. The proximity tables were constructed, where the abstract-prototype papers that were semantically closest appeared on the top, and the least similar one closer to the bottom of the table. The abstracts from the top of the proximity table were inspected. Because conference tracks are meant to unite papers from the same research field, the majority of the closest matches to every prototype were assumed to be from the same track ("track" experiment). Second, the consistency of the cross-track themes proposed by the conference organizers was analyzed. Because themes are supposed to unite the papers from different tracks that are semantically similar, I expected that abstracts from the same theme but different tracks would appear as the closest matches to an abstract from a given theme ("theme" experiment). A detailed description of the applied methodology of prototype-matching for IR by content from the HICSS-34 paper collection, the word and sentence histogram creation process, the experiments, and the line of reasoning can be found in (Kloptchenko, Back et al. 2002) and, more briefly in *Paper 4*. The main results and conclusions are presented below.

Table 8-1 contains the results from the first "Track" experiment in the form of hit ratios per track (hit ratio 1 and hit ratio 2), which reflect how many abstracts from the same track were retrieved among the 47 or 25 closest matches on the sentence level. It is believed that the sentence level's clustering conveys a higher

degree of semantics than word usage. Hit ratios were calculated in the same manner as the precision measure described in Chapter four, Section 4.2.1. A hit ratio of 30% for the Emergent technology track means that, on average, every third paper among the 47 closest matches allocated by the system as belongs to this track.

Table 8-1. The results from "Track" experiment

№	Track Title	Number of Papers	Hit ratio 1 (recall window 47)	Hit ratio 2 (recall window 25)
1	Collaboration Systems and Technology	66	25.8%	18.2%
2	Complex Systems	29	27.6%	17.2%
3	Decision Technologies for Management	47	25.5%	19.1%
4	Digital Documents	40	25%	15%
5	Emerging Technology	30	30%	20%
6	Information Technology in Health Care	26	23.1%	19.2%
7	Internet and Digital Economy	68	23.5%	14.7%
8	Organizational Systems and Technology	63	22.2%	15.8%
9	Software Technology	75	21.3%	16%

Other results from the "Theme" experiment are presented in Table 8-2. Here the question of how many papers within a certain theme (their names and sizes are presented in the left columns) have fired as the closest matches to the papers from the same theme on the sentence levels was answered. The hit ratio values show, for example, that the prototype matching method and the conference organizers had clustered 22.2% of papers from E-commerce development cross-track theme with a recall window 47, and 18.5% with a recall window 25, into the same theme

Table 8-2. The results from "Theme" experiment

№	Theme Title	Number of Papers	Hit ratio 3 (recall window 47)	Hit ratio 4 (recall window 25)
1	Knowledge Management	20	20%	20%
2	Data Warehousing/Data Mining	24	20.8%	16.7%
3	Collaborative Learning	22	40.9%	27.3%
4	Workflow	12	25%	16.7%
5	E-commerce Development	54	22.2%	18.5%
6	E-commerce Application	36	25 %	19.4%

Although the values of the hit ratios are rather low, one must understand the nature of comparison that I made between automatic retrieval results and conference track division while computing the values of the hit ratios. The hit ratio values were calculated under the assumption that tracks unite semantically similar papers. Track division is subjective and relative and, thus, it makes a weak reference point for

calculating hit ratio values. However, there are a number of different non-optimal considerations besides the topic of a paper (e.g. conflict of interest, diversity issues, etc.) that influence a track chair's decision in establishing the relevance of the paper to a particular track. Overall, the hit ratios from the "Theme" experiments are slightly higher than these from the "Track" experiments. This could support the plan that themes are constructed from papers with more similar content than the track papers. The deviation in the semantic similarities among theme papers is less than the deviation in the semantics among track papers. Moreover, all the abstracts from the research articles consist of the same components: introduction, method, results, and discussion (dos Santos 1996). The peculiarities of the written style of the scientific abstracts have a significant impact on the clustering ability of the methodology. To test the suitability of the prototype matching method to IR by content from a scientific text collection while eliminating the peculiarities of abstract writing styles, full-paper analysis was undertaken.

8.3 Full-paper Analysis

According to the HICSS tradition, the final decision concerning a paper's acceptance in a minitrack is made by chairs based upon the reviewers' ratings of the paper, the paper's relevance to the track, and the topic-relevance to other accepted papers in the same minitrack. For the second-level analysis, all full papers of about ten pages were chosen from the HICSS-34 collection. The papers were converted from portable-document format (pdf) to plain ASCII text normal format. Distinct regions of the papers (title, authors, abstract, main body, and bibliography) were manually identified and extracted, so only title, abstract, and main body were left remaining in a filtered document collection for further experiments. Another series of experiments conducted on the "new," full-paper version document collection obtained from HICSS-34 tested the system's capability to allocate the scientific conference papers based on their content. The experiments had the same scope as the experiments described in Section 8.2 for abstract level analysis, namely "Track" and "Theme" experiments.

In the first experiment, designed to check the consistency of the conference tracks, I focused on the results retrieved by our systems for every one of the nine tracks. Here, the "hit ratio" values were calculated in the same manner as the hit ratios on the abstract level, with a recall window of 47, as the average size of HICSS tracks, and a recall window of 25. I presented every paper from the collection as a prototype to the system and calculated hit ratios that reflect how often papers from the same track have fired as the closest matches to a presented prototype in a given recall window. Table 8-3 below contains hit ratios per track (hit ratio 5 and hit ratio 6), that reflect how many papers from the same track were retrieved among the 47 or 25 closest matches respectively.

Table 8-3. The results from track division clustering

№	Track Title	Number of Papers	Hit ratio 5 (recall window 47)	Hit ratio 6 (recall window 25)
1	Collaboration Systems and Technology	66	22.7%	15.2%
2	Complex Systems	29	17.2%	13.8%
3	Decision Technologies for Management	47	19.1%	12.8%
4	Digital Documents	40	22.5%	15%
5	Emerging Technology	30	20%	13.3%
6	Information Technology in Health Care	26	23.1%	15.4%
7	Internet and Digital Economy	68	19.1%	13.2%
8	Organizational Systems and Technology	63	22.2%	15.9%
9	Software Technology	75	22.7%	14.7%

Table 8-4 contains the names and sizes of cross-track themes, and values of hit ratio 7, which reflects how many papers from the same theme fired among the 47 closest to a prototype paper from the same theme. The last column (hit ratio 8) shows how many papers have their closest matches from the same theme among the 25 closest matches. For instance, the Collaborative Learning theme has the highest hit ratio 7 at 36.4%. This means that almost every third paper among the closest matches was from the same theme as any chosen prototype paper from the Collaborative Learning theme. The Collaborative Learning theme has the highest hit ratio 8 at 31.8%. It means that almost every third paper among the 25 closest matches was from the Collaboration Systems and Technology or Digital Documents tracks which constitute the chosen theme.

Table 8-4. The results from cross-track theme clustering

№	Theme Title	Number of Papers	Hit ratio 7 (recall window 47)	Hit ratio 8 (recall window 25)
1	Knowledge Management	20	41.6%	33.3%
2	Data Warehousing/Data Mining	24	26.7%	20%
3	Collaborative Learning	22	36.4%	31.8%
4	Workflow	12	36.4%	18.2%
5	E-commerce Development	54	24%	18%
6	E-commerce Application	36	34.8%	26%

The same sizes of recall windows were used to make the results from both experiments comparable. Comparing hit ratios 5 and 6 from Table 8-3 with hit ratios 7 and 8 from Table 8-4, one can conclude that the hit ratios for theme division, in average, were slightly higher than the hit ratios for track division within the same size of recall windows. It demonstrates the stronger semantic similarity among papers from the same cross-track themes, than the semantic similarities

among the papers from the same tracks. The values of hit ratios for full-papers on average were almost the same or slightly higher than the hit ratios for the abstract level analysis.

8.4 Discussion and Evaluation of the Results

The versatile vocabulary, highly ambiguous word usage (i.e. various meanings of the word *system*) and the peculiarities of the conservative academic writing style of the scientific papers have a significant impact on the clustering and retrieval abilities of the prototype matching method. Moreover, mathematical descriptions of the various models introduced in the scientific papers were treated as noise by the prototype matching methodology. The formal descriptions of similar mathematical expressions, equations, and variables, together with graph legends from different papers, are regarded as relevant for some users who are interested in the mathematical side of the problems. However, the descriptions of the mathematical aspects of the papers have changed some sentence constructions and disturbed the retrieval ability of the prototype matching system. All research papers consist of the same components: introduction, method, research background, results, and discussion (Miike, Etsuo et al. 1994). All the papers at HICSS-34 were semantically connected since all of them used approaches from the confluence discipline, similar concepts and vocabulary, e.g. *information system*, *manage*, *approach*, *analyze*, *research*, *objective*, *system*, *information*, etc. Therefore the intervals of the Euclidian distances between prototype-papers and the rest of the documents for full-paper analysis on word and sentence level were very narrow ([0.314...0.773] and [0.23...1.149] respectively). Since authors tend to use similar word order and similar sentence structures to describe their achievements in information system research, e.g. *we present, computer analysis, our paper discusses, construct a model, conduct an experiment, approach is based, process information, in the remainder of the paper, this paper describes, traditional systems, etc*., a particular academic writing style explains the closeness of all papers on the sentence level. Those words and word phrases were not omitted as stop-words by the methodology in the preprocessing and encoding phase, because they bear important meaning.

One can note that the results of automated clustering should be intuitively better match those achieved here ([17.2%…41.6%] for full-paper version recall window=47). However, as Nyberg (2001) has found, even Web sites that get the most positive responses in terms of relevance, like amazon.com and cnn.com, satisfy users' missions only 42 percent of the time at best.

The justification of the ranges is not an obvious task since, for the calculations of matching results, I compared the retrieval results to the track and theme divisions provided by the conference organizing committee. Those divisions can be a product of the message that every paper conveys, the author’s vision of a paper, and a number of non-optimal considerations that conference chairs keep in mind in addition to the topic relevance of a paper submitted to a certain track or theme. As one subcommittee chair has noted, there are a number of other issues and

variables in addition to content relevancy that should be balanced in conference settings. Such variables include conflicts of interest, gender, geography, topic, popularity of a certain research stream or author, etc. (Yarowsky and Florian 1999) noticed that members of conference committees tend to favor the article with more interesting content and findings and route them to their tracks, even if the topics are not relevant. Obviously, the prototype-matching system, which is based on an objective text processing technique, does not consider such issues. The nature of the hit ratios' calculation makes the evaluation of the results very challenging, because the hit ratios were calculated on the strong assumption that a given theme/track division by the HICSS conference committee is the absolutely semantically correct one. However, the HICSS theme/track division is a very weak reference point for comparison because of the issues mentioned above. Moreover, retrieval by content is a human-centered, interactive process, which makes performance in a real-world situation inherently subjective and its evaluation difficult. The ultimate measure of a retrieval system's performance is determined by the usefulness of the retrieved information to the user. Furthermore, relevance is not necessarily a binary concept. It means that the same paper can belong to a number of different minitracks with the same degree of relevance. This type of relevancy can confuse readers, authors, and conference organizers in the determination of a true allocation for a particular paper.

Additionally, there are a number of useful subtasks that our prototype-matching system can handle. It can be used to detect "good candidate" papers to be included in the theme from minitracks that were not included in the theme, and thus, can assist in article routing. The research limitations are thoroughly discussed in *Paper 4*, which was successfully presented and discussed with potential users at HICSS-36.

The selection of the scientific papers for performing IR by content turned out to be disadvantageous for the prototype matching method. First, the encoding of the words does not consider word paraphrasing (word co-occurrence within establish word phrases), which would probably improve retrieving results. Second, the closeness in the vocabulary and semantics of the papers from the emerging closely related themes, such as E-commerce Development and E-commerce application, contribute to low recall. The methodology is incapable of identifying the important patterns in the papers that would distinguish them. Third, the validation of the retrieved papers and their relevance to the theme, requires additional evaluation. In the future, the evaluation method of the applicability of the prototype matching systems for IR by content from scientific collections can be improved in several respects. First, conducting a user-oriented evaluation in addition to using automatic scoring techniques, in order to confirm that increased scores do directly benefit users, seems to be beneficial. However, due to the large quantity of parameters in our system, such an evaluation would need to focus on only a handful of variables. Because a single document frequently belongs to more than one category, automatic clustering of a document to multiple topic clusters can enhance the ability of the system to handle one-to-many relevancy of the papers.

In general, this chapter presents the experiences of the applying the prototype matching method to the task of IR by content. I explored the suitability of

the prototype matching method to IR by content from scientific text collections (research objective *e* from Section 1.3). As a sample scientific text collection I have chosen the collection of scientific abstracts and papers submitted to and accepted by the HICSS-34 conference committee. The exploration of the applicability of the method to IR was performed using the exploratory and constructive research approaches. First, an exploratory study was conducted by applying the prototype matching method to the abstract versions of the papers submitted to HICSS-34. Second, the pilot version of a system based on the prototype matching method was created to run on the full versions of the papers. I elaborated on the prospects and limitations of the methodology in accomplishing the task of IR by content.

9. CONCLUSIONS AND FUTURE RESEARCH

Knowledge workers and decision makers are intuitive thinkers who observe reality, collect and analyze the facts about it, and react upon the produced information. A good knowledge and deep understanding of reality, and observations and predictions about the business environment constitute a basic premise for success. With huge amounts of information available on the Internet and in internal databases, efficient and effective discovery of knowledge has become an imminent issue. Despite the availability and potential benefits of analyzing existing digital information, its usefulness is limited by IT. Data and text mining methods attempt to resolve information overload by finding and delivering valuable "nuggets" to knowledge workers and decision makers automatically. Modern IS aspires to offer its users built-in data and text mining capabilities to discover those value adding "nuggets".

Because of the nature of digitally available written text, effective reduction of textual information overload is often a complex and open-ended process. Effective textual information reduction is supposed to filter out irrelevant information and retrieve chunks of relevant and interesting information to a user. The relevancy of the information is defined by its degree of satisfaction of the information needs of any particular user. It is important to define the information needs accurately, as well as to realize that information needs are always different for one user than for another. Moreover, the efficient articulation of information needs is not always attainable. Constructing smart queries to available databases requires at least some initial understanding of the phenomena described and coded in the data. The automatic discovery and delivery of insights into the phenomena without prior understanding and rigorous analysis can be achieved by combining data and text mining methods.

The dissertation offers ways for extracting insights (knowledge or "nuggets") from low-level data to reduce textual information overload. In the dissertation, I explored the characteristics of the prototype matching method in a number of contexts. I apply the methodology to a collection of scientific papers, and to a collection of financial reports, in order to study the type and value of knowledge to be extracted. Within the scope of this work I explain the nature of the relationship between modern IT and textual information overload, explore the extend of the applicability of the prototype matching method to a collection of financial reports, demonstrate the process of knowledge extraction from numeric and textual data by combining DM methods with the prototype matching method, and determine the suitability of the prototype matching method for retrieving similar scientific papers from a scientific paper collection. The dissertation consists of a detailed summary of the research and seven papers that report the achievements concerning the research objectives. The detailed summary consists of nine chapters describing the context and nature of textual data and TM.

In this chapter, I summarize the findings of the dissertation which are the main contributions. I highlight the interrelationships between research papers, the

objectives of the research and chapters from the summary of the research. I outline the limitations of the studies described in the dissertation, as well as point out some directions for future research.

9.1 Main Contributions of the Dissertation

In the dissertation I have explored the applicability of the prototype matching method for dealing with textual information overload, in addition to reviewing the problem background, related studies, and creating technology summary.

This dissertation contributes to the field of knowledge discovery from textual databases. It researches important and inherent aspects of adequate text processing by communication technology. Textual data inherits several complex and ambiguous attributes of NL, which makes it a comfortable and understandable means for human communication. The huge amount of digitally available text through modern IT creates certain obstacles for reading, analyzing and processing it manually. The discrepancies between human ability for text processing and the amount of available text create the phenomena of textual information overload, which slows down managerial work. IT plays a leading role in the occurrence and reduction of information overload in general. Automatic discovery of useful and novel patterns in text is achievable, because computers do not need to "understand" text in order to look for nuggets in it. The example of text and data mining of quarterly reports confirmed the possibility to discover complex patterns from qualitative and quantitative data that direct toward insights into companies' financial performances.

The objectives of the dissertations were addressed in various research papers and in the chapters of the summary of the dissertation. The intricate interrelationships among the research papers, the research objectives and the chapters of the dissertation are represented in Figure 9-1. The objectives of the dissertation were achieved and reported in the research papers (*Papers 1-7*) and in the chapters of the summary (Chapters 3-8). Chapters one, two and nine introduce supplementary and background ideas to achieve the research objectives. Chapter one provides an introduction to the problem area. Chapter two explains the methods for acquiring knowledge that I followed to achieve the research objectives, i.e. answers the question of how the research is to be conducted, and what kind of paradigm I use to arrive at the conclusion. The objectives of the dissertation were achieved through applying a pluralist research methodology combining interpretive, explorative, and constructive research approaches. Chapter five contributes to attaining all the objectives because it describes the mathematic methodology of the prototype matching used in the research. Chapter nine concludes the research by explaining what was achieved, and how the results expand scientific knowledge.

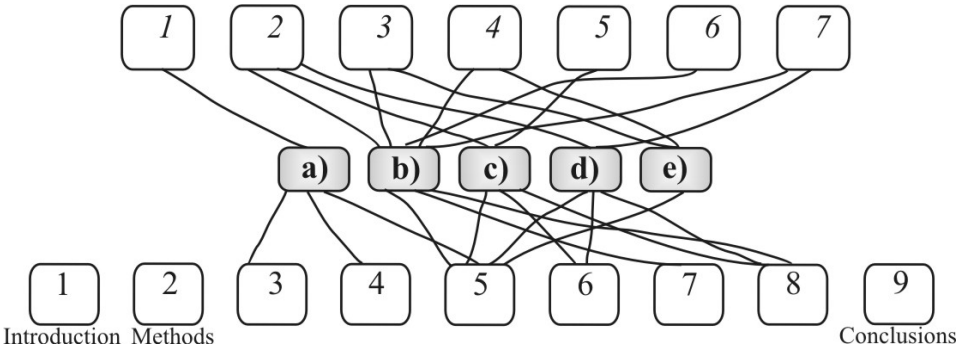


Figure 9-1. Objectives of the dissertation and their realization in the research papers and the chapters of the summary

While addressing the objectives of the dissertation, I explored the strengths of the prototype matching method in performing knowledge discovery from financial and scientific textual collections. Those text collections have inherently different language structure and peculiarities. A financial text collection contains more precise terms and less ambiguous grammatical constructions.

a) I explained the nature of the relationship between textual information overload and IS in Chapters three and four, and in *Paper 1*. Network communication, intelligent agents, database and wireless technologies originate, multiply, and duplicate information causing information overload. At the same time, modern IS, such as TPS, MIS/DSS, OAS, KMS, and multiagent systems, are designed to manage informational flows to deduce important information. I have described the existing OAS with built-in TM capabilities that strive to deduce and deliver important information from crude data to a user. The majority of those systems lack an analytical component. Users are required to know the strengths and weaknesses of the methods built into those systems, as well as some characteristics of the mined data, in order to use those systems effectively. In Chapter four I discussed the mathematical state-of-the art approaches that have been developed by the scientific community to be used in TM systems. I described the methods used for IR, clustering, categorization and results visualization. *Paper 1* offers an example of applying the prototype matching method for authorship attribution, new clustering, financial benchmarking, and allocating conference papers tasks.

b) I explored the extent of the applicability of the prototype matching method for knowledge discovery from a collection of financial reports in Chapters six, seven, and in *Paper 2*, *6*, and *7*. It turned out that the prototype matching method could be used for forecasting the future performance of companies and for competitor benchmarking based on mining quarterly reports from the companies within the same industry, such as the telecommunications sector. The results from applying the prototype

matching method to the collection of quarterly repots were attained by conducting explorative (*Paper 2*) and constructive (*Papers 6* and *7*) studies.

c) I demonstrated how knowledge extraction could be performed from low level textual (qualitative) and numeric (quantitative) data with the help of the prototype matching method in Chapter six, and *Papers 5* and *7*. I presented a conceptual model for a knowledge building system that is based on the integration of data and text mining methods with multiagent technology. I presented the architecture of the system to be used for financial forecasting, news summarization, and financial benchmarking.

d) I suggested ways to utilize the prototype matching method in combination with other DM methods to deliver additional insights to potential users concerning phenomena described in text in Chapters six, seven, and *Papers 2*, and *7*. In *Paper 2*, I explored the applicability of the prototype matching method for the textual parts of the quarterly reports and SOM clustering of financial ratios for financial benchmarking. In *Paper 6*, I suggested linguistic validation of the comparison of the textual parts of quarterly reports, bypassing their manual reading. I used collocational networks to visualize the meaningful concepts in the quarterly reports. In *Paper 7*, I proposed a method for predicting future financial performance based on the combination of the SOM and the prototype matching method suggested in *Paper* 2, with a neural network predictor.

e) I determined the suitability of the prototype matching method to IR by content of the papers from a scientific text collection from HICSS-34 in Chapter eight, and in *Papes 3* and *4*. I started by exploring the suitability of the prototype matching method for the handling scientific text collections by performing IR by content from a collection of abstracts in *Paper 3*. Because it was unclear whether the low precision results are caused by the inability of the prototype matching method to establish semantic similarities in the short articles, or by imperfection of the methodology, I continued by constructing a pilot system based on the prototype matching method for full-paper retrieval in *Paper 4*. The prototype matching system handles the retrieval task from a scientific text collection poorly.

After applying the prototype matching method to the scientific paper collections and the collections of financial reports, two main questions of the dissertation were answered:

1) Is the prototype matching method, which hypothetically identifies semantic similarities among documents in the collections, really discovering relationships among the documents?

The prototype matching method identifies some similarities among financial reports in the sense that the closest matching documents share similar semantic patterns with the prototype. Moreover, those shared patterns could differ from one match to another. The reports from well-performing companies more

likely have as their closest matches reports from other well-performing companies. Whether or not the prototype matching method determines semantic similarities between the conference papers requires further investigation due to the weaknesses in determination of the relevancy of the retrieved papers.

2) What can the user of the potential system built on the prototype matching method learn from the discovered relationships among the documents, based on the experiments on the collections of scientific papers and financial reports?

Using the system built on the prototype matching method for mining a collection of the financial annual or quarterly reports can reveal the benchmarking position of the company against its competitors, determine the accuracy and truthfulness of the facts described in the textual parts of the reports versus financial ratios, and determine the potential future performance of analyzed companies.

The results from prototype-matching clustering of the textual parts of quarterly reports from telecommunications companies have contained the patterns that indicate the operational and partly strategic development of the company. The retrieved TM patterns combined with the quantitative financial data clustering have provided an explanation of the companies' future. These dynamic and not-explicitly stated findings provide the potential users, i.e. decision makers, industrial and market analysts, with inside information on companies' current and prospective financial situations and positions on the market. The combination of the SOM and the prototype matching method for retrieving complex patterns from different data types, contributes to the semi-automatic domain of knowledge integration. The use of linguistic constructions, such as collocational networks, contributes to visualization and explanation of the achieved results from DM and TM. Furthermore, the prediction of future financial performance can be extracted from the qualitative and quantitative parts of the reports with some degree of accuracy.

The investigation of the applicability of the prototype-matching to IR by content from the collection of scientific papers has revealed several problems that might reside in the methodology itself or in universal use of scientific terms. The prototype matching method, as it functions right now, is not suited for the IR by content task. Scientific language is more unclear than the financial language. Frequently used words, such as *model*, *system*, and *information*, were not omitted as stop words because of the importance in their meaning, but the frequency of their appearance in every paper caused even semantically different papers to be treated as similar ones. Mining a text collection in a financial context, such as financial reports, removes word ambiguity automatically. Words such as *bank* in financial reports will most likely be defined only as the financial institution, and not as the side of a river. The problems with suitability of the method to the HICSS-34 collection resulted in low precision. Moreover, the evaluation of the relevancy of a retrieved scientific paper depends on the reader's ability to interpret the content of it in order to state its usefulness. However, the interpretation of the content of information depends on the information receiver's background and level of subject area expertise, and, thus, mathematicians and psychologists will describe the content and meaning of the same article differently.

To sum up, the main contribution of the dissertation is in exploring the characteristics of the prototype matching method (still work-in progress) for working with large masses of text in various contexts and from different perspectives. The dissertation overviews the potential uses of TM and the current status of the research in this area. It positions the nature of TM as a multidisciplinary approach. The prototype matching method supports the proposition by Witten, Bray et al. (1998) stating that we do not need to make computers understand text in order to perform TM successfully. I have learnt that it is more beneficial to apply the method to a collection of financial reports than to a collection of scientific articles. The method is more beneficial in a situation where a user does not know what ought to be discovered. Thus, while mining quarterly reports the reader gets a fast insight and understanding of the financial report by determining the companies' financial positions and forecasting their future financial performance. The method performs, however, poorly for retrieval of scientific publications.

9.2 Limitations and Future Research

The main limitations of the dissertation lie in the sizes and information richness of the explored text collections and the shortcomings of the prototype matching method. The suggestions for improving GILTA were induced from the shortcomings and the results of applying the method to collections of scientific publications and financial reports. The information richness of those collections is an issue in itself, which has not been directly addressed. The explicit link of the dissertation with decision making work suggests that it could have been interesting to include aspects of individual and organizational decision making, especially concerning its impact on managerial productivity. The usefulness of the proposed system for managers seeking to explore large financial text collections, and for conference organizers or attendees concerning large scientific text collections, is worth future investigation. Improvements in the mathematic methods in the methodology are future work. Different encoding techniques can lead to higher precision and recall results from applying the prototype matching method to retrieval of scientific publications. Moreover, the method could be applied to different types of textual collections (legal documents, medical diagnosis) to scrutinize its still undiscovered strengths and weaknesses.

REFERENCES

. Ackoff, R. L., S. K. Gupta and J. S. Minas (1962). *"Scientific Method"*. New York, John Wiley Sons Inc.

Amini, M.-R. and P. Gallinari (2001). "*Self-Supervised Learning for Automatic Text Summarization by Text-span Extraction"*. The 23rd BCS European Annual Colloquium on Information Retrieval, Darmstadt, Denmark.

Anckar, B. (2002). "*Contextual Insights into the Value Creation Process in E-commerce"*. Information Systems. Åbo, Åbo Akademi University.

Anick, P. and S. Vaithyanathan (1997). *"Exploiting Clustering and Phrases for Context-Based Information Retrieval"*. SIGIR 97, Philadelphia, USA, ACM.

Back, B., K. Öström, K. Sere and H. Vanharanta (1998). "*Analyzing Company Performance Using Internet Data"*. The 11th Meeting of the Euro Working Group on DSS, Toulouse, France.

Back, B., J. Toivonen, H. Vanharanta and A. Visa (2001). "*Comparing numerical data and text information from annual reports using self-organizing maps*" International Journal of Accounting Information Systems **2**(4): 249-269.

Baeza-Yates, R. and B. Ribeiro-Neto (1999). "*Modern Information Retrieval*". New York, ACM Press.

Baker, D. and A. McCallum (1998). "*Distributional Clustering of Words for Text Classification*". SIGIR-98, Melbourne, Australia, ACM.

Ball, P. (2003). "*Language evolved in a leap*". Nature News Services.

Bendell, T., L. Boulter and J. Kelly (1998). "*Benchmarking for Competitive Advantage*". London, Pitman Publishing.

Berthold, M. and D. J. Hand (1999). "*Intelligent Data Analysis - An Introduction*". Berlin, Springer Verlag.

Beynon-Davies, P. (2002). "*Information Systems. An Introduction to Informatics in Organizations"*. Nottingham, UK, Palgrave.

Bricker, R. (2000). *"Corporate Communications: NonFinancial Performance Indicators and Operating Measures*". An Introduction to the Literature**:** 1-10.

Broccoli, K. (2001). "Improving Information Retrieval with Human Indexing". Intranet Design: http://www.intranetjournal.com/features/humanindex-1.shtml.

Buckley, C. and J. Walz (1999). *SMART* in TREC 8. TREC - 8.

Burrell, G. and G. Morgan (1979). "*Sociological Paradigms and Organisational Analysis*". London, Heineman.

Chase, V. D. (2003). "*Made to Order: IBM makes sense of unstructured data*". Think Research.

Checkland, P. (1981). "*System Thinking, System Practice*". Chichester, UK, Wiley.

Checkland, P. and S. Holwell (1998). "*Information, Systems, and Information Systems*". UK, John Wiley & Sons.

Cheeseman, P. and J. Stutz (1996). "*Bayesian classification (AuoClass): Theory and results*". Advances in Knowledge Discovery and Data Mining. U. Fayyad, Piatesky-Shapiro,G.., Smyth, P., and Uthurusamy, R. Menlo Park, AAAI Press**:** 153-180.

Chen, H. (1993). "*Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms*". Technical Report

Chiaramella, Y., B. Defude, M. Bruandet and D. Kerkouba (1986). "*IOTA: A Full-text Information Retrieval System"*. ACM Conference on Research and Development in Information Retrieval.

Chien, L. F. (1997). "P*AT-Tree-Based Keyword Extraction for Chinese Information Retrieval*". Proceedings of Special Interest Group on Information Retrieval, SIGIR'97, Philadelphia, USA, ACM Press.

Clarke, R. (2000). "*Appropriate Research Methods for Electronic Commerce*." International Journal of Electronic Commerce.

Computation_Laboratory (1961). "*Information Storage and Retrieval*". Cambridge MA, USA, Harvard University.

Costea, A. and T. Eklund (2003). "*A Two-Level Approach to Making Class Predictions*". 36th Hawaii International Conference on Systems Sciences (HICSS-36), Hawaii, USA, IEEE.

Costea, A., T. Eklund and J. Karlsson (2002). "*A Framework for Predictive Data Mining in the Telecommunications Sector*". WWW/Internet 2002, Lisbon, Portugal, IADIS Press.

Costea, A., A. Kloptchenko and B. Back (2001). "*Analyzing Economical Performance of Central-East-European Countries Using Neural Networks and Cluster Analysis*". The Fifth International Symposium on Economic Informatics.

Cutting, D., Karger, D., Pedersen, J., and Turkey, J. (1992). "*Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections*". 15th Annual International SIGIR'92, Denmark, ACM Press, NY, USA.

Davenport, T. and L. Prusak (1998). "*Working Knowledge - How organizations manage what they know"*. USA, Cambridge, MASS : Harvard Business School.

Davies, J. T. (1973). "*The Scientific Approach*". New York, Academic Press.

Decrop, A. (1999). "*Quantitative Research Methods for the Study of Tourist Behavior*." Consumer Behavior in Travel and Tourism: 335-365.

Deogun, J. and V. Raghavan (1986). "*User-oriented document clustering: a framework for learning in information retrieval*". 1986 ACM conference on Research and development in information retrieval, Pisa, Italy, ACM Press New York, NY, USA.

Dewey, M. (1876). "*A classification and subject index for cataloguing and arranging the books and pamphlets of a library*". Amherst, MA, USA, Case, Lockwood & Brainard Co.

Dörre, J., P. Gerstl and R. Seiffert (1999). "*Text Mining: Finding Nuggets in Mountains of Textual Data*". KDD-99, Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, USA, ACM.

dos Santos, M. (1996). "The textual organization of research paper abstracts in applied linguistics." Text **16**(4): 481-499.

Doukidis, G., F. Land and G. Miller (1989). "*Knowledge-based management support systems*". Chichester, West Sussex, England, Ellis Horwood Limited.

Driver, M., Brousseau, K., and Hunsaker, P. (1993). "*The dynamic decision maker*". San Francisco, Jossey-Bass.

Dubin, D. (1995). "*Document Analysis for Visualization*". SIGIR'95, Seattle, WA, USA, ACM.

Eklund, T. (2002). "*Financial Benchmarking Using Self-Orginizing Maps - A Study of the International Forest Products Industry*". Fepartment of Information Systems. Abo, Abo Akademi University**:** 136.

Eklund, T., B. Back, H. Vanharanta and A. Visa (2002). "*Assessing the Feasibility of Self-Organizing Maps for Data Mining Financial Information*". The Xth European Conference on Information Systems (ECIS 2002), Gdansk, Poland.

El-Hamdouchi, A. and P. Willett (1986). "*Hierarchic Document Clustering Using Ward's Method*". ACM Conference on Research and Development in Information Retrieval, ACM Press.

Fallman, D. and A. Gronlund (2000). "*Rigor and Relevance Remodelled*". IRIS-25, Sweden.

Fan, W., M. Gordon and P. Pathak (2000). "*Personalization of Search Engine Services For Effective Retrieval and Knowledge Management*". 2000 International Conference on Information Systems (ICIS 2000), Brisbane, Australia.

Farhoomand, A. and D. Drury (2002). "*Managerial Information Overload*". Communications of the ACM **45**(10): 127-131.

Fayyad, U. (1996). "*Data Mining and Knowledge Discovery: Making Sense Out of Data*". IEEE Expert: 20-25.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). "*Knowledge Discovery and Data Mining: Towards a Unifying Framework*". The Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, AAAI Press.

Fayyad, U. and R. Uthurusamy (2002). "*Evolving Data Mining into Solutions for Insights*". Communications of the ACM **45**(8): 28-31.

Ferrer i Cancho, R., and Sole, R. (2001). "*The small world of human language*". The Royal Society London **268**: 2261-2265.

FinGloge (2003). "*FinSummarize: Summarizes Key Events on a Set of Companies and/or Set of Years, Global Financial Information Summaries*".

Fitzgerald, B. and D. Howcroft (1998). "*Competing Dichotomies in IS Research and Possible Strategies for Resolution*". International Conference on Information Systems, Helsinki, Finland, ACM.

Flexer, A. (2001). "*On the use of self-organizing maps for clustering and visualization*". Intelligent Data Analysis **5**(5): 373-384.

Foley, J. (1995). "*Managing Information: Infoglut*". Information Week.

Furnas, G. W., T. K. Landauer, L. M. Gomez and S. T. Dumais, . (1987). "*The Vocabulary Problem in Human-System Communication*". Communications of the ACM **30**(11): 964-971.

Gershon, N., S. Eick and S. Card (1998). "*Information Visualization*". ACM Interactions: 9-15.

Girolami, M., A. Vinokourov and K. A. (2000). "*The Organization and Visualization of Document Corpora: A probabilistic Approach*". The 11th International Workshop on Database and Expert Systems Applications (DEXA'00), Greenwich, London, U.K., IEEE.

Goldman, J. A., W. W. Chu, D. S. Parker and R. M. Goldman (1998). "*Term Domain Distribution Analysis, a Data Mining Tool for Text Databases: A Case History in a Thoracic Lung Cancer Text Database*". The first International Conference on Discovery Science.

Goldszmidt, M. and M. Sahami (1998). "*A Probabilistic Approach to Full-text Document Clustering*", SRI International.

Grant, J., T. Fogarty, R. Bricker and G. Previts (2000). "*Corporate Reporting of Nonfinancial Performance Indicators*". Morristown, NJ, Financial Executives Research Foundation.

Hall, H. (1998). "*Networked information: dealing with overload*". Information Scotland, Information Management Publications.

Han, J. and M. Kamber (2001). "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publisher.

Hand, D., H. Mannila and P. Smyth (2001). "*Principles of Data Mining*". Boston, USA, A Bradford Book, The MIT Press, 2001.

Hatzivassiloglou, V., L. Gravano and A. Maganti (2000). An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering. The 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece, ACM Press New York, NY, USA.

Hatzivassiloglou, V., J. Klavans, M. Holcombe, R. Barzilay, M. Kan and K. McKeown (2001). "*SIMFINDER: A Flexible Clustering Tool for Summarization*". NAACL Workshop on Automatic Summarization, Association for Computational Linguistics.

Hawkins, D. F. and B. A. Hawkins (1986). "*The effectiveness of the annual reports as a communication vehicle*". Morristown, NJ, Executive Research Foundation.

Hearst, M. (1999). "*Untangling text Data Mining*". 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Maryland, USA, ACM Press.

Hearst, M. A. (1997). "*Text data mining: Issues, techniques, and the relationship to information access"*. Presentation notes for UW/MS workshop on data mining.

Heylighen, F. (1999). "*Change and Information Overload: negative effects*". Principia Cybernetica Web.

Hidalgo, J.-M.-G. (2002). "*Text Mining and Internet Content Filtering*". ECML/PKDD-2002 Tutorial Notes, Helsinki, Finland, University of Helsinki, Department of Computer Science.

Himberg, J. (2000). "*A SOM based cluster visualization and its application for false coloring*". International Joint Conference on Neural Networks (IJCNN2000).

Hoch, R. (1994). "*Using IR techniques for text classification in document analysis*". 17th Annual international ACM-SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc. New York, NY, USA.

Honkela, T. (1997). "*Self-Organizing Maps in NL Processing*". Espoo, Finland, Helsinki University of Technology.

Hyland, K. (1998). "*Exploring Corporate Rhetoric: Metadiscorse in the CEO's Letter*". The Journal of Business Communication **35**(2): 224-245.

Iivari, J., R. Hirscheim and H. K. Klein (1998). "*A Paradigmatic Analysis Contrasting Information System Development Approaches and Methodologies*". Information Systems Research **9**(2): 164-193.

Isbell, C. L. (1998). "*Restructuring Sparse High Dimensional Data for Effective Retrieval*". Advances in Neural Information Processing Systems.

Jain, A., Murty, M., and Flynn, P. (1999). "*Data Clustering: A Review*". ACM Computing Surveys **31**(3): 265-323.

Järvinen, P. (2001). "*On Research Methods*". Tampere, Opinpaja OY.

Jo, T. (1999). "*Text Categorization Considering Categorical Weights and Substantial Weights of Informative Keywords*". Tokyo, Japan, Samsung SDS**:** 1-17.

Jones, G., A. Robertson, C. Santimetvirul and P. Willett (1995). "*Non-hierarchic Document Clustering Using A Genetic Algorithm*". Information Research **1**(1).

Karanikas, H., Tjortjis, C., and Theodoulidis (2000). "*An Approach to Text Mining using Information Extraction*". Principles and Practice of Knowledge
Discovery in Databases (PKDD-2000), Springer-Verlag Publisher.

Karlsson, J. (2001). "*Financial Benchmarking of Telecommunications Companies. Department of Information Systems*". Turku, Åbo Akademi University.

Karlsson, J., B. Back, H. Vanharanta and A. Visa (2001). "*Analysing Financial Performance with Quarterly Data Using Self-Organising Maps*". Turku, Turku Centre for Computer Science.

Karlsson, J., B. Back, H. Vanharanta and A. Visa (2001). "*Financial Benchmarking of Telecommunications Companies*". Technical Report 430, Turku Centre for Computer Science.

Karlsson, J., T. Eklund, B. Back, H. Vanharanta and A. Visa (2001). "*Transforming Passive Information from the Internet into Refined Information Using Self-Organising Maps*". the 24th Information Systems Research Seminar in Scandinavia (IRIS24), Hardanger, Norway.

Kasanen, E., K. Lukka and A. Siitonen (1993). "T*he Constructive Approach in Management Accounting Research*." Journal of Management Accounting Research **5**: 243-264.

Kaski, S., T. Honkela, K. Lagus and T. Kohonen (1996). "*Creating an Order in Digital Libraries with Self-Organizing Maps*". World Congress on Neural Networks (WCNN96), Mahwah, NJ, USA, INNS Press.

Kaymak, U. and M. Setnes (2000). "*Extended Fuzzy Clustering Algorithms*", Erasmus Research Institute of Management (ERIM), Erasmus University Rotterdam**:** 25.

Keen, P. G. W. (1980). "*MIS Research: Reference Disciplines and a Cumulative Tradition*". The First International Conference on Information Systems (ICIS), Philadelphia, PA, USA.

Kendal, J. (1993). "*Good and evil in chairmen's "boiler plate": an analysis*." Organization Studies **14**: 571-592.

Kim, B., P. Johnson and A. S. Huarng (2002). "*Colored-sketch of Text Information*." Informing Science **5**(4): 163-173.

Klein, H. K. and M. D. Myers (1999). "*A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems*." MIS Quarterly, Special Issue on Intensive Research **23**(1): 67-93.

Kloptchenko, A., B. Back, A. Visa, J. Toivonen and H. Vanharanta (2002). "*A prototype-matching system for scientific abstract collection semantic clustering*", Technical Report 465, Turku Centre for Computer Science

Kloptchenko, A., B. Back, A. Visa, J. Toivonen and H. Vanharanta (2002). "*Toward Content Based Retrieval from Scientific Text Corpora*". 2002 IEEE International Conference on Artificial Intelligence Systems, Divnomorskoe, Russia, IEEE.

Kloptchenko, A., T. Eklund, B. Back, J. Karlsson, H. Vanharanta and A. Visa (2002). "*Combining Data and Text Mining Techniques for Analyzing Financial Reports*". The 8th Americas Conference on Information Systems, Dallas, USA.

Knuth, D.E., J. Morris, and V.R. Pratt. "*Fast pattern matching in strings*". SIAM Journal on Computing, 6(1):323-360, 1977.

Kohavi, R., N. J. Rothleder and E. Simoudis (2002). "*Emerging Trends in Business Analytics*." Communications of the ACM **45**(8): 45-48.

Kohonen, T. (1997). "Self-Organizing Maps". Leipzig, Germany, Springer-Verlag.

Kohonen, T. (1998). "*Self-Organization of Very Large Document Collections: State of the Art*". Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks, Springer, London.

Kohonen, T. (1999). "*WEBSOM*". Helsinki, Helsinki Technological University.

Kohonen, T., S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero and A. Saarela (2000). "Self organization of a massive text document collection." IEEE Transactions on Neural Networks **11**: 574-585.

Kohut, G., and Segars, A. (1992). "*The president's letter to stockholders: An examination of corporate communication strategy*." Journal of Business Communication **29**(1): 7-21.

Kolenda, T. and L. K. Hansen (2002). "*Independent Components in Text*". IEEE Workshop on Neural Networks for Signal Processing XII, IEEE Press.

Kontkanen, P., P. Myllymaki, T. Silander and H. Tirri (1997). "*A Bayesian Approach for Retrieving Relevant Cases*". Artificial Intelligence Applications (The EXPERSYS-97 Conference), Gournay sur Marne, IITT International.

Kontkanen, P., J. Lahtinen, P. Myllymaki, T. Silander and H. Tirri (2000). "*Supervised Model-Based Visualization of High-Dimensional Data*." Intelligent Data Analysis(4): 213-227.

Koskimaki, E., J. Gloos, P. Kontkanen, P. Myllymaki and H. Tirri (1998). "*Comparing Soft Computing Methods in Prediction of Manufacturing Data*". The 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence & Expert Systems (IEA-98-AIE), Castellon, Spain.

Lagus, K. (2000). "*Text Mining with WEBSOM*". Department of Computer Science and Engineering. Espoo, Finland, Helsinki University of Technology**:** 54.

Lagus, K., T. Honkela, S. Kaski and T. Kohonen (1996). "*Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration*". Second International Conference on Knowledge Discovery and Data Mining.

Lam, S. L. Y. and D. L. Lee (1999). "*Feature Reduction for Neural Network Based Text Categorization*". 6th IEEE International Conference on Database Advanced Systems for Advanced Application, DASFAA-99, Hsinchu, Taiwan.

Lam, W., M. Ruiz, and and P. Srinvasan (1999). "*Automatic Text Categorization and ITs Application to Text retrieval*." IEEE Transactions on Knowledge and Data Engineering **11**(6): 865-881.

Landry, M. and C. Banville (1992). "*A disciplined methodological pluralism.*" Accounting, Management and Information Technologies **2**(2): 77-97.

Lansiluoto, A., B. Back, H. Vanharanta and A. Visa (2002). "*Multivariable Business Cycle Analtsis with Self-Organizing Maps - Are the Cycles Similar*". Conference on Accounting information System (ECAIS).

Larsen, B., and Aone, A. (1999). "*Fast and Effective Text Mining Using Linear-time Document Clustering*". KDD-99, San Diego, CA, USA, ACM.

Laurikkala, J. (2001). "*Knowledge Discovery for Female Urinary Incontinence Expert System*". Doctorate Dissertation Department of Computer and Information Sciences. Tampere, Finland, University of Tampere.

Lavrenko, V., M. Schmill, D. Lawrie and P. Ogilvie (2000). "*Mining of Concurrent Text and Time Series*". Text Mining Workshop of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, ACM.

Lawrence, S., K. Bollacker and C. Lee Giles (1999). "*Indexing and Retrieval of Scientific Literature*". 8th International Conference on Information and Knowledge Management, CIKM 99, Kansas City, Missouri, USA, ACM Press.

LeCompte, D. (2000). "*Three numbers that (should) have nothing to do with user interface design*." Internetworking (ITG Publication) **3**(2).

Lee, C., and Yang, H. (1999). "*A Web Text Mining Approach Based on Self-Organizing Map*". WIDM-99, Kansas City, MO, USA, ACM.

Lee, Z., S. Gosain and I. Im (1999). "*Topics of Interest in IS: Evolution of Themes and Differences between Research and Practice*." Information and Management **36**(5): 233-246.

Legge, G. (2001). "*Perception*", Legge, Gordon. **2003**.
Lesser, V. (1995). "*Multiagent Systems: An emerging subdiscipline of AI*." ACM computing surveys **27**(3): 340-342.
Lewis, D. (1992). "*Feature Selection and Feature Extraction for Text Categorization*". Speech and NL Workshop.
Lin, S.-H., C.-S. Shih, M. C. Chen and J.-M. Ho (1998). "*Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach*". SIGIR'98, Melbourne, Australia, ACM Press.
Lin, X. (1995). "*Searching and Browsing on Map Displays*". Proceedings of American Society for Information Science (ASIS-95), Chicago, IL, USA.
Lin, X., D. Soergel and G. Marchionini (1991). "*A Self-organizing Semantic Map for Information Retrieval*". The 14th Annual International ACM/SIGIR conference on Research and development in information retrieval, Chicago, Illinois, USA, ACM Press.
Liu, S. (1998). "*Business Environment Scanner for Senior Managers: Towards Active Executive Support with Intelligent Agents*." Expert Systems with Applications **15**(2): 111-121.
Liu, S. (2000). "*Improving Executive Support in Strategic Scanning with Software Agent Systems*". TUCS. Abo, Abo Akademi University.
Mahling, D. and N. Craven (1995). "*From Office Automation to Intelligent Workflow Systems*." IEEE Expert: 41-47.
Makoto, I. and T. Takenobu (1995). "*Cluster-Based Text Categorization: A Comparison of Category Search Strategies*". Tokyo, Tokyo Institute of Technology.
Manning, C. and H. Shutze (1999). "*Collocations*". Foundations of Statistical NL Processing. Cambridge, MA, The MIT Press**:** 141-177.
Manning, C. and H. Shutze (1999). "*Word Sense Disambiguation*". Foundations of Statistical NL Processing. Cambridge, MA, MIT Press**:** 230-263.
March, S. T. and G. F. Smith (1995). "*Design and natural science research on information technology*." Decision Support Systems **15**(4): 251-266.
Marino, G. (2001). "*Workers Mired in E-mail Wasteland.*" Cnet News.com.
Martín-del-Brío, B. and C. Serrano-Cinca (1993). "*Self-organizing Neural Networks for the Analysis and Representation of Data: Some Financial Cases*." Neural Computing and Applications **1**(2): 193-206.
Mayes, M., B. Drewes and W. Thompson (2002). "*Introduction to Text Mining and SAS Text Miner*". Distilling Textual Data for Competitive Business Advantage. Heldelberg, Germany**:** 20.
McGovern, G. (2002). "*Information technology: Trojan Horse of information overload*." New Thinking Newsletter.
Merkl, D., and Schweighofer (1997). "*En Route to Data Mining in Legal Text Corpora: Clustering Neural Computation, and International Treaties*". 8th International Workshop on database and Expert Systems Applications (DEXA'97), Toulouse, France, IEEE.
MicrosoftCorporation (1999). "*Textual Searches on Database Data*". Textual Searches on Database Data Using Microsoft SQL Server 7.0.

Miike, S., I. Etsuo, K. Ono and K. Sumita (1994). "*A full-text retrieval system with a dynamic abstract generation function*". The seventeenth annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, Springer-Verlag New York, Inc.

Miller, G. (1956). "*The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information*." The Psychological Review **63**: 81-97.

Mingers, J. (2001). "*Combining IS Research Methods: Toward a Pluralist Methodology*." Information Systems Research **12**(3): 240-258.

Mintzberg, H. (1973). The Nature of Managerial Work. New York, Harper and Row.

MIT-Libraries (2001). Corporate Reports.

Mladenic, D. (1999). "*Text-Learning and Related Intelligent Agents: A Survey*." IEEE Intelligent Systems: 44-54.

Moody, D. and A. Buist (1999). "*Improving Links between Information System Research and Practice - Lessons from Medical Profession*". The 10th Australasian Conference on Information Systems, Wellington, New Zealand.

Moxon, B. (1996). "*Defining Data Mining*." DBMS on-line.

Myers, M. D. and G. Walsham (1998). "*Exemplifying interpretive research in Information Systems: an overview*." Journal of Information Technology Theory and Application **13**(4): 233-234.

Nardi, B., Miller, J., and Wright, D. (1998). "*Collaborative, Programmable Intelligent Agents*." Communications of the ACM **41**(3): 96-104.

Navarro, G. (1999). "*Indexing and Searching*". Modern Information Retrieval. Baeza.-Yares and. Ribeiro-Neto, Addison Wesley**:** 191.

Neary, R. (1999). "*Building a Data Warehouse and Data Mining for a Strategic Advantage*." Journal of Information Technology Theory and Application **1**(1): 3-21.

Neto, J. L., A. D. Santos, A. Kaestner and F A. A. (2000). "*Document Clustering and Text Summarization"*. Proceedings 4th Int. Conference on Practical Applications of Knowledge Discovery and Data Mining (PADD-2000)The Practical Application Company, London.

Nie, J.-Y., M. Simard, P. Isabelle and R. Duran (1999). "*Cross-language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web*". SIGIR'99, Berkley, CA, USA, ACM Press.

Nonaka, I., Takeuchi, H. (1995). "*The Knowledge-Creating Company"*, Oxford University Press.

Nunamaker, J. F. J., C. Minder and T. D. M. Purdin (1991). "*Systems Development in Information Systems Research*." Journal of Management Information Systems 7(3): 89-106.

Nyberg, A. (2001). "*Fighting Information Overload*." CFO Magazine.

O'Leary, D. (2000). "*Enterprise Resource Planning Systems: Systems, Life Cycle, Electronic Commerce, and Risk*". Cambridge, UK, Cambridge University Press.

Osborn, J. D., C. I. Stubbart and A. Ramaprasad (2001). "*Strategic Groups and Competitive Enactment: A Study of Dynamic Relationships between Mental Models and Performance.*" Strategic Management Journal **22**: 435-454.

Papadimitriou, J. H., P. Raghavan, H. Tamaki and S. Vempala (1998). "*Latent semantic indexing: A probabilistic analysis*". Proceedings of the 17th ACM Symposium on the Principles of Database Systems, Seattle, USA, ACM Press.

Piatetsky-Shapiro, G. (2000). "*Knowledge Discovery in Databases: 10 years after*." SIGKDD Explorations **1**(2).

Pullum, G., Scholz, B. (2001). "*More than words*." Nature **413**: 367.

ReliaSoft_Corporation (2002). Reliability Glossary.

Riloff, E., and Hollaar, L. (1996). "*Text Databases and Information Retrieval*." ACM Computing Surveys **28**(1): 133-134.

Roussinov, D. and H. Chen (1999). "*Document clustering for electronic meetings: an experimental comparison or two techniques*." Decision Support Systems **27**: 67-79.

Roussinov, D. and J. L. Zhao (2003). "*Message Sense Maker: Engineering a Tool Set for Customer Relationship Management*". Proceedings of the 36th Hawaii International Conference on System Sciences - 2003, Big Island, Hawaii, IEEE.

Ruiz, M. and P. Srinivasan (1998). "*Automatic Text Categorization Using Neural Networks*". Advances in Classification Research, 8th ASIS SIG7CR Classification Research Workshop, NJ, USA.

Sahami, M., S. Yusufali and M. Baldonado (1998). "*SONIA: A Service for Organizing Networked Information Autonomously*". 3rd ACM Conference on Digital Libraries, Pittsburgh, PA, USA.

Salton , G. (1991). "*The smart document retrieval project*". The 14th annual international SIGIR conference on Research and development in information retrieval. Chicago, IL, United States, ACM Press.

Salton, G. and M. McGill (1983). "*Introduction to modern information retrieval*". New York, McGraw-Hill.

Salton, G., A. Wong and C. S. Yang (1975). "*A vector space model for automatic indexing*." Communications of the ACM **18**(11): 613-620.

Sanderson, M. (1994). "*Word Sence Disambiguation and Information Retrieval*". The 17th ACM Conference on Research and Development in IR (SIGIR-94).

Savaresi, S., D. Boley, S. Bittanty and G. Gazzaniga (2002). "*Cluster selection in divisive clustering algorithms*". Second SIAM International Conference on Data Mining, Arlington, VA, USA.

Schutze, H. and C. Silverstein (1997). "*Projection for Efficient Document Clustering*". SIGIR 97, Philadelphia, PA, USA, ACM Press New York, NY, USA.

Semio_Corporation (2001). "*Text Mining and the Knowledge Management Space*".

Siegel, J. (1994). "*Stocks for the long run*". Burr Ridge, IL, New York, NY, IRWIN.

Simon, H. A. (1981). The Sciences of the Artificial. Cambridge, MA, MIT Press.

Sinclair, J. (1991). "*Corpus, Concordance, Collocation*". Oxford, Oxford University Press.

Sparck-Jones, K. (1971). "*Automatic Keyword Classification for Information Retrieval*". Connecticut, Archon Books.

Sparck-Jones, K. (1988). A look back and a look forward. SIGIR 1988, Grenoble, France, ACM Press.

Spengler, S., C. Pinkas, E. Engel, A. Gabric, T. Hansen, C. Hellman, H. McKenzie, L. Powell and A. Thompson (1998). "*The need to KNOW vs. the need to GROW*", Lawrence Berkeley National Laboratory's.

Staw, B. M. (1985). "*Reports on the road to relevance and rigor: Some unexplored issues in publishing organizational research*". Organizational Sciences. L. L. C. P. J. Frost, Homewood Illinois, Richard D. Irwin, Inc.**:** 96-107.

Steinbach, M., G. Karypis and V. Kumar (2000). "*A Comparison of Document Clustering Techniques*". TextMining Workshop (KDD).

Strzalkowski, T., J. Perez-Carballo and M. Marinescu (1996). "*NL Information Retrieval in Digital Libraries*". Digital Libraries (DL96), Bethesda, MD, USA, ACM Press.

Subramanian, R., R. Isley and R. Blackwell (1993). "*Performance and readability: A comparison of annual reports of profitable and unprofitable corporations.*" Journal of Business Communication **30**: 50-61.

Suzuki, J., Yamamoto, Y. (1998). "*Document brokering with agents: Persona approach*". Sixth Workshop on Interactive Systems and Software (JSSST WISS '98), Miyazaki, Japan, ACM.

Szymkowiak, A., J. Larsen and L. K. Hansen (2001). "*Hierarchical Clustering for Datamining*". KES-2001 Fifth International Conference on Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies, Osaka and Nara, Japan.

Tan, A. (1999). "*Text Mining: The state of the art and the challenges*". PAKDD-99, Workshop on Knowledge Discovery from Advanced Databases (KDAD'99), Beijing, China.

Tan, R. G. H., J. van den Berg and W.-H. van den Bergh (2002). "*Credit Rating Classification Using Self-Organizing Maps*". Neural Networks in Business: Techniques and Applications. J. Gupta. Hershey, Idea Group Publishing**:** 140-153.

Tetard, F. (2002). "*Fragmentation of Working Time, and Information Systems*". Turku Centre for Computer Science. Turku, Finland, Abo Akademi University.

Thomas, J. (1997). "*Discourse in the Marketplace: The Making of Meaning in Annual Reports*." Journal of Business Communication **34**: 47-66.

Tirri, H., T. Silander and K. Tirri (1997). "*Using neural networks for descriptive statistical analysis of educational data*". The Annual American Educational Research Association Meeting (AERA'97), SIG Educational Statisticians, Chigago, USA.

Tkatch, D. (1997). "*Text Mining Technology: Turning Information into Knowledge*", IBM Coorp.

Toivonen, J., A. Visa, T. Vesanen, B. Back and H. Vanharanta (2001). "*Validation of Text Clustering Based on Document Contents*". Machine Learning and Data Mining in Pattern Recognition (MLDM 2001), Leipzig, Germany, Springer-Verlag.

Tokunaga T., and M. Iwayama. "*Text categorization based on weighted inverse document frequency*". Technical Report 94 TR0001, Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, 1994.

Tseng, S., C. Yang and C. Hsien (1990). "*An Experimental Model of Chinese Textual Database.*" Journal of the Chinese Institute of Engineers **13**(6): 607-622.

Tull, D. S. and D. I. Hawkins (1987). "*Marketing Research: Measurements and Method*". New York, Macmillan Publishing.

Turney, J. (2001). "*Srings and things*." Nature **410**(873).

Turney, P. (1997). "*Extraction of Keyphrases from Text: Evaluation of Four Algorithms*", National Reseach Council Canada, Institute for Information Technology**:** 1-29.

van Rijsbergen, C. (1979). "*Information Retrieval*" (Second Edition). London:, Butterworths.

Vesanto, J. (1999). "*SOM-based data visualization methods*" Intelligent Data Analysis **3**(3): 111-126.

Vesanto, J., J. Himberg, E. Alhoniemi and J. Parhankangas (1999). "*Self-organizing map in Matlab: the SOM Toolbox*". Matlab DSP Conference 1999, Espoo, Finland.

Visa, A., J. Toivonen, S. Autio, J. Mäkinen, B. Back and H. Vanharanta (2001). "*Data Mining of text as a tool in authorship attribution*". AeroSense 2001, SPIE 15th Annual International Symposium on Aerospace/Defence Sensing, Simulation and Controls, Orlando, Florida, USA.

Visa, A., J. Toivonen, B. Back and H. Vanharanta (2000). "*Knowledge Discovery from Text Documents Based on Paragraph Maps*". The HICSS-33, Hawaii International Conference on System Science, Maui, Hawaii, USA.

Visa, A., J. Toivonen, B. Back and H. Vanharanta (2000). "*A New Methodology for Knowledge Retrieval from Text Documents*". TOOLMET2000 Symposium - Tool Environments and Development Methods for Intelligent Systems.

Visa, A., J. Toivonen, B. Back and H. Vanharanta (2002). "*Contents Matching Defined by Prototypes: Methodology Verification with Books of the Bible.*" Journal of Management Information Systems **18**(4): 87-100.

Visa, A., J. Toivonen, T. Vesanen, J. Mäkinen, B. Back and H. Vanharanta (2002). "*Example Based Text Matching Methodology for Routing Tasks*". The Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland, NIST.

Wainwright, J. and A. Francis (1984). "*Office Automation, Organization and the Nature of Work*", Grower Publishing Company Limited.

Willett, P. (1988). "*Recent Trends in Hierarchic Document Clustering: A Critical Review.*" Information Processing and Management **24**(5): 577-597.

Williams, G. C. (1998). "*Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles*." International Journal of Corpus Linguistics **3**: 151-171.

Winsor, D. (1993). "*Owning corporate texts*." Journal of Business and Technical Communication **7**(2): 179-195.

Wise, J. A. (1999). "*The ecological approach to text visualization*." Journal of American Society for Information Science **50**(13): 1224-1233.

Witten, I., Z. Bray, M. Mahoui and B. Teahan (1998). "*Text mining: A new frontier for lossless compression*". Data Compression Conference '98, IEEE.

Witten, I., C. Nevill-Manning and S. Cunningham (1996). "*Digital Libraries Based on Full-Text Retrieval*". WebNet 96, San Francisco, USA.

Wong, P. C., W. Cowley, H. Foote, E. Jurrus and J. Thomas (2000). "*Visualizing Sequential Patterns for Text Mining*". IEEE Symposium on Information Visualization 2000, Salt Lake City, Utah, USA.

Wuthrich, B., D. Permunetilleke, S. Leung, V. Cho, J. Zhang and W. Lam (1998). "*Daily Prediction of Major Stock Indices from textual WWW data*". American Association for Artificial Intelligence.

Yarowsky, D. and R. Florian (1999). "*Taking the load off the conference chairs: towards a digital paper-routing assistant*". Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora.

Yin, H. (1989). "*Case study research: Design and methods*". Beverly Hills, USA, Sage Publishing.

Zadeh, L. (2000). "*Outline of Computational Theory of Perceptions Based on Computing with Words*". Soft Computing & Intelligent Systems. N.K. Sinha, and L.A. Zadeh. New York, Academic Press**:** 3-22.

Zamir, O., and Etzioni, O. (1998). "*Web Document Clustering: A Feasibility Demonstration*". SIGIR'98, Melbourne, Australia, ACM Press.

Zavrel, J. and J. Veenstra (1995). "*The Language Environment and Syntactic Word-Class Acquisition*". Groningen Assembly on Language Acquisition (GALA 95).

Zikmund, W. G. (2000). "*Business Research Methods*". USA, The Dryden Press/Harcourt College Publisher.

Zipf, G. K. (1972). "*Human behaviour and the principle of least effort. An introduction to human ecology*". New York: Hafner reprint, 1st edn: Cambridge, MA: Addison-Wesley, 1949.

PART II
Original Research Papers

Research Paper 1

Back, B., Kloptchenko, A., Toivonen, J., Vanharanta, H., Visa, A., Prototype-matching methodology applications in Text Mining, In *Proceedings of the International Conference on Information and Knowledge Engineering'02*, ed. by H. Arabnia, Las Vegas, Nevada, CSREA Press, June 24-27, USA, pp. 130-136 , isbn: 1-892512-39-4

PROTOTYPE-MATCHING METHODOLOGY APPLICATIONS IN TEXT MINING

Barbro Back
Åbo Akademi University, IAMSR, Turku, Finland

Antonina Kloptchenko
Turku Center for Computer Science, IAMSR, Åbo Akademi University, Turku, Finland

Jarmo Toivonen
Tampere University of Technology, Department of Information Technology, Tampere, Finland

Hannu Vanharanta
Pori School of Technology and Economics, Pori, Finland

Ari Visa
Tampere University of Technology, Department of Information Technology, Tampere, Finland

Abstract

In modern Western culture, text documents are the most common communication vehicle for the exchange of formal knowledge among people. Text contains a vast range of semantic information that is difficult to decipher automatically. Text mining recognizes that complete understanding of natural language text is a long-standing goal of computer science, and focuses on extracting information from text with high relevance to the user needs. There are a number of tasks that a good text analysis tool should solve. In the following paper we consider the types of problems that can be solved using the prototype-matching methodology for text clustering, introduced by our research group. We describe how the developed methodology can be applicable to solve the diverse range of tasks: organization of scientific conferences, news clustering, analyses of financial information, or authorship attribution.

Keywords: text mining, clustering

1. Introduction

Text is the most frequent and comfortable form of our formal communication. Text is poorly structured and indefinite data that carries different meaning to different users. The advent of online publishing increased the amount of digitalized text and knowledge that resides in it. The Internet, digital libraries, data warehouses, and information organizations generate far more digitally available text than it is possible to process manually. Text as the semi-structured coded information is very difficult to manage and mine by computers. Text Mining (TM) seeks the tools to analyze, and learn the meaning from dynamic poorly structured information. [1]. TM methods and tools strive to accomplish searching, organizing, browsing, and analyzing text collections automatically. TM is about looking for patterns in text and can be defined, according to [2], as the process of analyzing text to extract information that is useful for particular purposes. TM as knowledge management technique deals with learning and discovering information that was previously unknown to its users.

Aiming to present a content of text in manner that is understandable for computers, the authors of the text or the librarian experts introduce mark-ups, indexes or keywords, that outline text main ideas. [3,4,5,6,7] However, the authors and the users of information frequently represent the same semantics with different words or describe different meanings by the same words that have various meanings. These wording peculiarities create a ground for a conflict in the relevance of text mining tools based on predefined text structure. TM

methods best suited for "discovery" purposes should aspire to avoid the subjectivity in digital representation of a document to computers. Building mining on subjectively predefined semantic structure of a document may hide some valuable content that can worsen the creativity.

In the following paper we address the following questions: what kind of informational needs might be satisfied and what kind of real-world problems could be solved using one particular TM methodology and a tool proposed by our research group [8]. First, we give an overview of the related research and methods and briefly describe the objectives of our methodology. In section 3, we introduce the developed methodology, which consists of text preprocessing (filtering), "smart encoding" of an analyzed document on three levels (word, sentence, and paragraph levels), comparison of the frequency distribution of words and prototype matching phases. We outline its novelty toward the existent text mining approaches. In Section 4, we present the text-mining tasks that we have accomplished using the proposed methodology. We have used it for scientific abstract clustering, news clustering, analysis of financial text, and authorship attribution. We present the study with the Bible as a validation of our methodology. Finally, we discuss ongoing researches and the possibility to use the proposed methodology for e-mail handling or search engines and the plans for future research.

2. Related work

Typical technologies for document exploration include topic detection and text categorization, text clustering, summarization, and visualization. [9]. TM methods aim at retrieval relevant to the user need information. Therefore, topic detection and text clustering are the most important TM tasks. The valuable information hidden in the documents, which is not outlined by the conventionally used keywords, cannot be retrieved by user-defined information retrieval methods that usually based on using keywords [4], indexing, or mark-ups. Mantes-y-Gomez et al. has assigned the topic to the text by discovering the associations in news collection with the real-world ones using peaks of topic occurrences [10]. The keywords assigned or extracted by the author, using, for example, PAT tree [4] might characterize the content as well. However, the keyword approach slips into the danger of hiding the meaning of the document from a reader that is not highlighted by keywords. On the other hand, markup utilized for text analysis must be flexible to mark and create tagging to the interesting to the author or to the user things. Markup tells how to display the material, rather than identifying what the material is. User-defined markups help to structure and categorize hypertext documents [6].

There are number of document clustering approaches in TM that are based on statistical clustering techniques [11,12] and neural nets, in form of self-organizing maps [4, 13]. In [4] the authors use SOM on Chinese corpus to build word cluster and sentence cluster maps to cluster the document and label document map according to word co-occurrence. The WebSom approach utilizes SOM for clustering and visualization of text content [13, 14, 9]. The WebSom text categorization does not build text exploration on subjective perceptions of the authors and focuses more on visualization on "understanding" the content of presented documents. Another approach is the analysis of the content of free text introduced by Subasic et al. in [15] is based on fuzzy semantic typing to draw up of the complete fuzzy affect lexicon. Gedeon applied the fuzzy importance measures to retrieve significant "concepts" from the documents in [16]. Similar to our approach, the authors used the entire document as a query vector. The Hyperlink Vector Voting method of indexing and retrieving hypertext documents uses the content of hyperlinks to a document to rank its relevance to the query terms [17].

Our research group introduced the methodology for text clustering that aims to push computers that represent natural language text as a digital array to understand "semantic meaning" hidden in text [8]. The method is free from human judgments in text representation and natural language ties. There is no subjectivity in computer text interpretation, such as humanly detected markups, indexes, keywords or other techniques that glues to the natural language constructions and define the text structure pointing out main "ideas".

3. Method and its procedures

Our research group [8] aims to originate the text mining methodology that could be implemented to the various real-world problems associated with

finding hidden patterns in textual information. The starting point has been to provide a mechanism to enable computers to retrieve those pieces from text that are semantically relevant to each other, the way humans understand them.

The methodology developed for text clustering can be classified as a document's matcher. It has been evolved over the development time and has acquired different trial mathematical techniques, such as SOM and vector quantization algorithms. As a documents' matcher, it has an objective to "match a new document to old documents and to rank the retrieved documents by assigning a score or relevance" [5]. The described methodology is implemented in the prototyping software package GILTA-3.

The methodology for text analysis on the word, sentence and paragraph levels consists from the following steps:

[1] Typically, pre-processing takes place before text documents are represented as vectors in order to act as input to a text clustering system. We perform the text pre-processing by creating an abbreviation table, rounding the numbers, deleting extra carriage returns, composing an abbreviation table and uniting the compound words. The aim of the basic filtering is to modify the text so that every sentence is on its own line. Additionally, there should be only one space between words and an empty line (not used within the original document) between the paragraphs. The punctuation marks are separated by spaces, the numbers are rounded, and dashes and mathematical signs are excluded.

[2] After the basic filtering of the text, we encode the document with the vectors. Each word is analyzed character by character so that a key entry to a code (ASCII) table is calculated. This approach is accurate and sustainable for statistical analysis. It is sensitive to capital letters and conjugations. A word w is transformed into a number according to the following formula (L is the length of the word character string, c_i is the ACSII value of a character within a word w and k is a constant):

$$y = \sum_{i=0}^{L-1} k^i \times c_{L-i}$$

[3] After each word has been converted to a code number we set the minimal and maximal values to the words, and look at the distribution of the words' code numbers. In the training phase the range between the minimal and maximal values of words' code numbers is divided to N_w logarithmically equal bins. First, we calculate the frequency of words belonging to each bin. The bins' counts are normalized according to the quantity of all words in the text. For estimation of the word codes distribution we choose the Weibull distribution. The Best fitted Weibull distribution is to be compared with the code distribution in a sense of the smallest square sum by calculating Cumulative Distribution Function according to following formula (a and b are the parameters be adjusted in Weibull distribution):

$$CDF = 1 - e^{(((-2.6\times\log(y/y_{\max}))^b)\times a)}$$

A number of Weibull distributions are calculated with various possible calculation of a and b's using a selected precision. The size of every bin is $1/N_w$. Every word belongs to a bin that is found using the code number and the best fitting Weibull distribution. The resolution is the best where the words are the most typical to a text (usually 2-5 symbol length words). Rare words are separated in the sense of belonging to bins not so accurately from each other.

[4] Similarly, on the sentence level every sentence is converted into a number. First, every word in a sentence is changed to a bin number in the same way as we did for words. The whole sentence is considered as a sampled signal. We accomplish Discrete Fourier Transform (DFT) to every signal. Since the sentences of the text contain different number of words, the sentence vector's length varies. In the transformation we do not consider all the coefficients, however, we transform bn_i = bin number of the word i into output coefficients from B_0 to B_n to create a cumulative distribution as the one on the word level. The range between the minimal and maximal values on the sentence code numbers is divided to N_s equal size bins. We calculate the frequency count of sentences that belong to the every bin. Then we divide the bins' count by the quantity of all sentences. Finally we find

the best Weibull distribution corresponding to the sentence data.

[5] We convert the paragraphs of the documents into vectors using the code numbers of the sentences. The vectors are Fourier transformed; and the coefficient B_1 represents the paragraph. After that we find the best Weibull distribution corresponding the paragraph data we do the paragraph quantization.

[6] We create the histograms of the documents' word, sentence and paragraph levels according to the corresponding value of quantization (Weibull distribution and common histogram for every level). On the word level word code numbers are quantized using word quantization created with all the words in the database. The histogram consists of N_w bins and is normalized by the total number of words in the document. We create similar histograms for every document in the database for the sentence and the paragraph level.

[7] With the histograms from all the documents in the database, we can analyze documents on the word, sentence or paragraph levels. The information retrieval phase is based on the prototype matching. We calculate the Euclidian distances as a similarity measure between the chosen prototype and the available coded text corpus represented in a form of the sentence or paragraph histograms. Choosing the documents with the closest distances to the prototype completes the retrieval.

4. Applications of the methodology

The prototype-matching methodology can assist in solving various kinds of the real-world problems associated with documents and text processing. The good achievements have been reported with high-content, small and medium sizes text collections. GILTA has been used to cluster a range of text materials, including full-text documents [17,18], abstracts of scientific articles, financial reports [19] and news items with some degree of success. Below, we describe what we feel are the most important applications of the methodology for text mining. We deem that the GILTA tool can be implemented as a module either into document clustering decision support systems, financial analysis tools, or author attribution tools. Additionally, our methodology can be used for the text mining the documents written down in the languages, whose linguistic structures are different from English, e.g. Finnish language.

4.1 Scientific abstracts clustering

Many conference organizers have faced a problem while sorting out the submitted scientific papers to the proposed conference tracks according to the assigned topics. The distinct feature of modern scientific conferences is the cross-topic and interdisciplinary research that makes it hard to decide which track a particular paper belongs to. The conference organizers need a decision support tool for meaningful document clustering. Similar problems can be found in business, when many cross-related reports are submitted to a meeting. We have tried the prototype-matching tool on a text corpus consisting of 444 scientific abstracts obtained from The Hawaii International Conference on System Science 2001 (HICSS-34). It is a time consuming task to proceed this text corpus manually looking for the common topics and links. The taxonomy of the conference grew more complicated than a traditional track division. The scientific papers at HICSS-34 were submitted into 9 major tracks with subdivision into total amount of 78 mini-tracks. And besides the traditional track division of the submitted scientific papers the organizers made an effort to identify six themes that run across the tracks based on the similarities and expansion of the scientific fields. The outlined 6 cross-track themes covered 134 papers in the conference from 26 mini-tracks.

We attempted to justify that the chosen distribution of the papers from the different tracks into a certain theme can be retrieved similarly using GILTA-3 software rather than the keyword approach. The full text of every abstract was encoded into an array of 2080 bins based on common word histogram. We created the sentence histogram of the size of 25 for every abstract. We omitted the paragraph level analysis due to the short length of the presented abstracts. In the experiment we used every abstracts from the same theme to see if the abstracts from the same theme would fire as the closest matches. With a recall window at 25, we discovered that 26% of papers that fire as the closest ones to the papers from the data-mining theme discuss the data/text mining methods applicability and theoretical problems; 12% of papers from e-

commerce development track fired as the closest matches to each other and discuss e-commerce related problems. We compared our clustering results with paper distribution into the tracks, which was performed by the conference organizers. For other cross-track themes, such as knowledge management, collaborative learning, workflow and e-commerce development, the number of papers that fire as the closest one to the papers within a theme was less than 10 %. Our clustering results are somewhat different from the proposed theme division by the organizing committee. Partly they can be explained by the mixed word choice and terminology in the mentioned above fields, more diverse written styles in comparison with the papers from data mining and e-commerce themes.

4.2 News Clustering

We tested the methodology and prototype-matching software on some short news databases [18,19]. We ran the methodology on the Reuters-21578, Distribution 1.0 Database for randomly selected 25 short financial reports from five authors. The reports went through the procedures described in Section 3, i.e. we used preprocessing, quantization and histogram creation processes for word and sentence levels. We examined the two closest prototypes in term of the smallest Euclidian distances for every piece of the news. No ordinary word list was used rather than a list of learned prototypes and scalar quantization. The methodology worked better than random selection for the word level.

4.3 Analyzing qualitative financial data

Traditionally, financial performance of different companies has been analyzed using quantitative data in form of financial ratios. However, some of the valuable financial descriptive data is hidden in texts, e.g. in the annual CEO reports. We undertook the attempt to demonstrate that our methodology can be used for analysis of qualitative data in the form of text documents. We checked the consistency of archived results with the results of the quantitative analysis conducted for the same companies in [20]. We used the database of 234 annual reports of 50 pulp companies from 1985-1989 years. We run the database through the first software version of our methodology. The consistency of the obtained results was not perfect. Some discrepancies between the qualitative and quantitative results related to the companies' performance was found. Partly it can be explained by the tendency of exaggerate actual financial state of the company in the annual report and the use of one-dimensional SOM for define the clusters.

Another study was made for clustering the text parts from quarterly reports from the leaders of the telecommunications sector: Ericsson, Motorola, and Nokia, from the years 2000-2001 [21]. We discovered using more advanced version of our methodology, that annual/quarterly reports tend to have information on future and past performance, so that, tables with financial numbers state the facts about previous performance, and textual descriptive part carries some messages about company future prospective. It explained the dissimilarities in clustering qualitative and quantitative data by the phenomena that exists in qualitative and quantitative parts of every quarter/annual report obtained in [20]. The quantitative part of a report only reflects the past performance of a company. At the same time, the qualitative part of a report holds some message about company's future performances.

4.4 Authorship attribution

We have designed two tests to check how the clustering methodology can obtain the divergences in the text written by different authors. For that purpose we examined three texts from classical authors: William Shakespeare, Edgar Alan Poe, and George Bernard Shaw. [22] The examined pieces contained statistically sufficient amount of text. After the preprocessing, we run vector quantization and created the histograms for the texts on word and sentence levels respectively. We selected bin sizes equal to 2080 on the word level, and 25 on the sentence level. We treated each text piece one by one as a prototype. We matched it against consolidated text from all sources. The results of author attribution were extremely good on the word level. The closest matches occurred among the text pieces written down by the same author. We did not consider any ordinary word list, collocations or typical word phrases. On the sentence level the results were good in general, except for one mismatch from Shaw's Mrs. Warren's Profession and Poe's The Assignation. The developed methodology proved to be quite capable to recognize and distinguish the author styles based on peculiarities of sentence structuring by different authors.

5. Methodology evaluation

We have undertaken one bigger procedure to evaluate our methodology. For the validation purposes we chose the Bible versions in Greek, Latin English, and two versions in Finnish from the years 1933 and 1938 as the test material considered it to be very accurate and bear the same meaning in different languages. We created the word, sentence and paragraph level histograms. The idea was to select a recall window of the closest matches sizes of 10, and to compare all the books in Bible using them as the prototypes against the whole text of different versions of The Bible [9]. We discovered that our methodology found the average of six books out of ten that really belong to the same category of Books in Bible. Our hypothesis was that if we receive similar results with the books in different languages it is evidence that our methodology works. [23]. We considered the co-occurrences of the books not the order within the examined window. There were on the average 4.52 same books within the window in English and Finnish versions based on the word map and 7.94 same books based on the sentence map, and 5.56 same books based on the Paragraph maps. Based on a random sample there should have been only 2, i.e, the results can be considered statistically significant. The tests gave evidence that our methodology works.

6. Current and future work

We are currently investigating many issues concerning applicability, evaluation and improvement of our prototype-matching methodology, examining the constant parameters we had used and justifying the encoding methods. Our research group has attended the Filtering Track of the TREC-2001 Text Retrieval Conference. TREC is co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA). The retrieval task of the Filtering Track consists of building a filtering profile to select the most relevant examples from an incoming stream of documents. We test our methodology on the large document collection, clustering more than 800 000 Reuters text pieces from 84 topics. We explore the opportunity to use the methodology for e-mail sorting and handling, and for search engines as well. In search engines, the valuable information hidden in the documents, which is not outlined by the keywords, cannot be retrieved by user-defined information retrieval methods. As a result, instead of typing the keywords united with Boolean operators in the query line, the whole paragraph of the text interested to the user can be copy-and-pasted there. We believe that the GILTA tool is able to retrieve the documents that are relevant as to their content and we will investigate that in the near future.

7. Acknowledgements

The financial support of TEKES (grant number 40887/97) and the Academy of Finland is gratefully acknowledged.

8. References

[1] DÖRRE, J. GERSTL, P. and R. SEIFFERT (1999). Text Mining: Finding Nuggets in Mountains of Textual Data. In Proceedings of the KDD-99, Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, USA.

[2] WITTEN, I., BRAY Z., MAHOUI, M., and TEAHAN, B. (1998). Text mining: A new frontier for lossless compression. Data Compression Conference '98, IEEE.

[3] LAHTINEN, T. (2000). Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods. PhD thesis, Department of General Linguistics, University of Helsinki, Finland.

[4] CHIENG L. (1997). PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval. In the Proceedings of Special Interest Group on Information Retrieval, SIGIR'97, ACM, Philadelphia, USA.

[5] WEISS, S., WHITE, B., APTE, C., and J. DAMERAU (2000). Lightweight Document Matching for Help-Desk Applications. IEEE Intelligent Systems, March/April

[6] ANDERSON, M. (1999) A Tool for Building Digital Libraries, Journal Review, Vol. 5, Issue 2, February 1999

[7] SALTON, G. (1989). Automatic Text Processing. Addison-Wesley, USA.

[8] VISA, A., BACK, B., and H. VANHARANTA (1999). Toward Text Understanding – Comparison of Text Documents by Sentence Map, In Proceedings of the EUFIT'99, 7th European Congress on

Intelligent Techniques and Soft Computing, CD-ROM, Aachen, Germany.

[9] TOIVONEN, J., VISA, A., VESANEN, T., BACK, B., and H. VANHARANTA (2001). Validation of Text Clustering Based on Document Contents. In Petra Perner and Maria Petrou, editors, Proceedings of MLDM'2001, International Workshop on Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science, Springer-Verlag, Leipzig, Germany.VISA, A. (2001). Technology of Text Mining. In the Proceeding of Machine Learning and Data Mining in Pattern Recognition, (invited paper) Springer Verlag, Leipzig, Germany

[10] MONTES-Y-GOMEZ, M., GELBUKH, A., and LOPEZ-LOPEZ A., Discovering Ephemeral Associations among News Topic. To appear in the Proceedings of the workshop of Adaptive Text Mining, International Joint Conference on Artificial Intelligence IJCAI'2001 , USA.

[11] ZAMIR, O., and ETZIONI, O. (1998). Web Document Clustering: A Feasibility Demonstration. SIGIR'98, Melbourne, Australia, ACM Press

[12] SLONIM, N., and TISHBY, N. (2000). Document Clustering using Word Clusters via Information Bootleneck Method. SIGID 2000, Athens, Greece, ACM Press, NY, USA.

[13] LAGUS, K. (2000). Text Mining with WEBSOM. PhD Thesis, Espoo, Finland.

[14] KOHONEN, T. (1997). Self-Organizing Maps, Springer-Verlag Information Science, p.426;

[15] SUBASIC, P., and A.HUETTNER (2000). Calculus of Fuzzy Semantic Typing for Qualitative Analysis of Text. In the KDD-2000 Workshop on Text Mining, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, USA.

[16] GEDEON, T., SING, S., KOCZY, L., and R. BUSTOS (1996). Fuzzy Relevance Values for Information Retrieval and Hypertext Link Generation, In Proceedings of the EUFIT-96, forth European Congress on Intelligent Techniques and Soft Computing, Aachen, Germany.

[17] LI, Y. (1998). Toward Qualitative Search Engine, IEEE Internet Computing, July-August, USA.

[18] VISA, A., TOIVONEN, J., AUTIO, S., MÄKINEN, J., BACK, B., and VANHARANTA H. (2001b) Data Mining of text as a tool in authorship attribution. In Proceedings of AeroSense 2001, SPIE 15th Annual International Symposium on Aerospace/Defense Sensing, Simulation and Controls. Data Mining and Knowledge Discovery: Theory, Tools, and Technology III, volume 4384, Orlando, Florida, USA.

[19] VISA, A., TOIVONEN, J., BACK, B., and H. VANHARANTA (2000a). Knowledge Discovery from Text Documents Based on Paragraph Maps. In Proceedings of the HICSS-33, Hawaii International Conference on System Science, Maui, Hawaii, USA.

[20] KLOPTCHENKO, A., EKLUND, T., KARLSOON, J., BACK, B., VISA, A., VANHATANTA,H., Combining Data and Text Mining Techniques for Analyzing Financial Reports, to appear in the Proceeding for AMCIS 2002, Texas, USA, 2002

[21] BACK, B., TOIVONEN, J., VANHARANTA, H., and A.VISA (2001). Comparing numerical data and text information from annual reports using self-organizing maps. International Jornal of Accounting Information Systems, 2(2001)

[22] VISA, A., TOIVONEN, J., BACK, B., and H. VANHARANTA (2000b). Toward Text Understanding – Classification of Text Documents by Word Map. In Belur V. Dasathy, editor. In Proceedings of AeroSence 2000, SPIE 14th Annual International Symposium on Aerospace/Defense Sensing, Simulating and Controls. Data Mining and Knowledge Discovery: Theory, Tools, and Technology II, vol. 4057, Orlando, USA.

[23] VISA, A., TOIVONEN, J., VANHARANTA, H. and BACK B. (2001a). Prototype-matching – Finding Meaning in thee Books of the Bible, In Proceedings of the HICSS-34, Hawaii International Conference on System Science, Maui, Hawaii, USA.

Research Paper 2

Kloptchenko, A., Eklund, T., Karlsson, J., Back B., Vanharanta, H., Visa, A., Combining Data and Text Mining Techniques for Analyzing Financial Reports, In *Proceedings of 2002 Americas Conference on Information Systems (AMCIS 2002),* Dallas, USA, 8-11 August, 2002, pp. 20-28. Accepted for publication in Journal of Information Systems in Accounting, Finance, and Management (IJISAFM)

COMBINING DATA AND TEXT MINING TECHNIQUES FOR ANALYZING FINANCIAL REPORTS[1]

Antonina Kloptchenko
Turku Centre for Computer Science and IAMSR / Åbo Akademi University, Turku, Finland
Antonina.Kloptchenko@abo.fi

Tomas Eklund[2]
Turku Centre for Computer Science and IAMSR / Åbo Akademi University, Turku, Finland
Tomas.Eklund@abo.fi

Jonas Karlsson
IAMSR / Åbo Akademi University, Turku, Finland
Jonas.Karlsson@abo.fi

Barbro Back
Åbo Akademi University, Department of Information Systems, Turku, Finland
Barbro.Back@abo.fi

Hannu Vanharanta
Pori School of Technology and Economics, Pori, Finland
Hannu.Vanharanta@pori.tut.fi

Ari Visa
Tampere University of Technology, Department of Information Technology, Tampere, Finland
Ari.Visa@tut.fi

Acknowledgements

The financial support from TEKES (grant number 40943/99) and the Academy of Finland is gratefully acknowledged. We are grateful to Jarmo Toivonen for his contributions in the early stages of this research.

---

[1] A previous version of this paper has been presented at the 8th Americas Conference on Information Systems (AMCIS2002), Dallas, Texas, August 9-11, 2002

[2] Corresponding author: Tomas Eklund, IAMSR / Åbo Akademi University, Lemminkäisenkatu 14 B, 20520 Turku, Finland. Phone: +358 2 215 3352, Fax: +358 2 215 4809

COMBINING DATA AND TEXT MINING TECHNIQUES FOR ANALYZING FINANCIAL REPORTS

ABSTRACT

There is a vast amount of financial information on companies' financial performance available to investors in electronic form today. While automatic analysis of financial figures is common, it has been difficult to extract meaning from the textual part of financial reports automatically. The textual part of an annual report contains richer information than the financial ratios. In this paper, we combine data and text mining methods for analyzing quantitative and qualitative data from financial reports, in order to see if the textual part of the report contains some indication about future financial performance. The quantitative analysis has been performed using self-organizing maps, and the qualitative analysis using prototype-matching text clustering. The analysis is performed on the quarterly reports of three leading companies in the telecommunications sector.

***Keywords**: Self-organizing map, text mining, annual reports, prototype-matching clustering*

1 INTRODUCTION

A huge amount of electronic information concerning companies' financial performance is available in databases and on the Internet today. This information is potentially very valuable to companies' decision makers, their partners, competitors, and shareholders. The important task for them is to extract relevant information for decision-making purposes from the available data storages on time and, preferably, by the click of a mouse button. Data and text mining methods for discovering hidden patterns in various types of data aim to offer this opportunity.

We use data and text mining methods to retrieve non-obvious indications about the future financial performance of companies from the quantitative and qualitative parts of their annual/quarterly financial reports. Annual/quarterly reports are one of the most important external documents that reflect companies' strategy and financial performance. Annual/quarterly reports are therefore an important medium for the company's communication with its investing public.

This research continues and builds on the work of Visa et al. (1999, 2000) and Back et al. (1998, 2001), which was devoted to studying various types of financial data from financial reports. In this paper, we carry out an investigation of the textual parts of the financial reports in combination with the study of companies' financial ratios. It is believed that text in a particular context bears more diverse information than dry numbers do. Therefore, we perform clustering of quantitative data in the form of financial ratios using the Self-Organizing Map (SOM), and clustering of qualitative data in the form of the textual part of quarterly reports using a prototype-matching approach. We discovered that annual/quarterly reports tend to state information about company's past performance, but also contain indications of its future performance, i.e. the tables

with financial numbers indicate how well a company has performed, while the linguistic structure and written style of the text may indicate how well a company will do.

We begin our explanation by providing a short overview of studies relating to analyzing financial reports. Then, we provide a description of the methodology for quantitative financial data clustering and the choice of appropriate financial ratios. Next, we propose a text analysis method. We then relate quantitative and qualitative analysis by reviewing an example of using both for analysis of telecommunications companies' performance. As a sample data set, we have chosen quarterly reports from the leaders in the telecommunications sector: Ericsson, Motorola, and Nokia, from the years 2000-2001. We review the results and their evaluation, and compare our findings with those of Back et al. (2001). Finally, we highlight a number of issues for further investigation.

2 BACKGROUND

Neural networks, in the form of self-organizing maps, provide a good tool for clustering and visualization of large amounts of numeric information. An early example of the application of neural networks for financial analysis is the study by Martín-del-Brío & Serrano-Cinca (1993). Martín-del-Brío & Serrano-Cinca used self-organizing neural networks to study the financial state of Spanish companies, and to attempt to predict bankruptcies among Spanish banks during the 1977-85 banking crisis. The authors found the SOM to be a "very interesting tool for financial decision making".

Back et al. (1998) compared 120 companies in the international pulp and paper industry. The study was based on standardized financial statements for the years 1985-89. The objective of the study was to investigate the potential of using self-organizing maps in the process of analyzing large amounts of quantitative financial data. The results of the study indicate that self-organizing maps could be feasible tools for processing vast amounts of financial data.

Several studies have been made on the relationship between the readability of the annual reports and the financial performance of a company (Subramanian et al, 1993). As research has shown, the annual reports of the companies that performed well were easier to read than those of companies that did not perform well. Moreover, writers of annual reports see the message they put in the report as a representation of their personality (Winsor, 1993). A close look at the language structure in the letters to stockholders made in (Thomas, 1997) showed that the structure of the financial reports might reveal some things that the company may not wish to announce directly to its outside audience. Another conclusion of this study was the confirmation of the Pollyanna Hypothesis that earlier had an intuitive character. It states that regardless of the financial state of the company, the language in the annual letters will be predominantly positive. The annual reports are not only the best possible description of a company, but are also a description of a company's managerial priorities. Thus, communication strategies hidden in annual reports differ in terms of the subjects emphasized when the company's performance worsens (Kohut and Segars, 1992).

One of the first attempts to semi-automatically analyze a company's performance, by examining quantitative and qualitative information from annual reports, was made in the study (Back et al., 2001). The authors compared numeric and textual information for 76 companies in the international pulp and paper industry for the period of 1985-1989 using self-organizing maps. They indicated that the differences in qualitative and quantitative data clustering results are due to a slight tendency to

exaggerate the performance in the text. A conclusion of the research was a proposition: "an experiment with putting in a lag and analyzing whether the text corresponds better to the next year's numerical data would also be interesting". We decided to continue the research in combining data and text mining techniques for financial analysis, using an improved document clustering method, and a different hypothesis and data set.

3 METHODOLOGY

Our methodology section builds on two steps in order to analyze two types of data: quantitative and qualitative (Back et al., 2001). We use the SOM clustering ability (Kohonen, 1997) for financial benchmarking of financial quantitative data. We use the prototype-matching text clustering methodology proposed by Visa et al. 2001 for qualitative data analysis.

3.1 Financial Data Clustering

Self-Organizing Maps

The SOM technique creates a two-dimensional map from n-dimensional input data. This map resembles a landscape in which it is possible to identify borders that define different clusters (Kohonen 1997). These clusters consist of input variables with similar characteristics.

The methodology used when applying the self-organizing map is as follows (Back et al., 1998). Firstly, the data material is chosen. It is often advisable to pre-process the input data so that the learning task of the network becomes easier (Kohonen, 1997). Next, the network topology, learning rate, and neighborhood radius are chosen. Thirdly, the network is constructed by showing the input data to the network iteratively using the same input vector many times, the so-called training length. The process ends when the average quantization error is small enough. Finally, the best map is chosen for further analysis, and the clusters are identified using the U-matrix, and interpreted (assigned labels) using the feature planes. From the feature planes we can read per input variable per neuron the value of the variable associated with each neuron.

The network topology refers to the form of the lattice. There are two commonly used lattices, rectangular and hexagonal. In a rectangular lattice a node has four neighbors, while in a hexagonal lattice, it has six. This makes the hexagonal lattice preferable for visualization purposes (Kohonen, 1997). The learning rate refers to how much the winning input data vector affects the surrounding network. The neighborhood radius refers to how much of the surrounding network is affected. The average quantization error indicates the average distance between the best matching units and the input data vectors. Generally speaking, a lower quantization error indicates a better-trained map.

To visualize the final self-organizing map we use the unified distance matrix method (U-matrix). The U-matrix method can be used to discover otherwise invisible relationships in a high-dimensional data space. It also makes it possible to classify data sets into clusters of similar values. Feature planes (Figure 1), representing the values in a single vector column, are used to identify the characteristics of these clusters.

Choice of Data and Information

In Karlsson et al. (2001a, 2001b), data in the form of companies' annual reports was collected, primarily through the Internet. For the study, a SOM was created for the years 1995-99, and the different clusters on the map were isolated and analyzed. In this study, the same map has been used to study the performance of three leading international telecommunications manufacturers during four quarters of the year 2000, and during the first three quarters of 2001.

Thus, the dataset was collected from seven quarterly reports for three telecommunications manufacturers: Nokia, Ericsson, and Motorola. The dataset consists of both quantitative and qualitative data. The quantitative data consist of a number of calculated financial ratios, and the qualitative data of the textual discussion from each report.

Choice of Financial Ratios

In order to make the quantitative data comparable, financial ratios had to be calculated. The selection of relevant financial ratios was based on an empirical study by Lehtinen (1996), in which international accounting differences were analyzed in greater detail, especially concerning the reliability and validity of the ratios. We selected and calculated seven financial ratios, which fulfilled the criteria of good validity and reliability, for each of the companies. The financial ratios can be divided into four different classes: profitability ratios, liquidity ratios, solvency ratios and efficiency ratios. It is common to choose ratios that measure different aspects of financial behavior. Our emphasis in this study was on profitability, and therefore, we selected three profitability ratios; Operating Margin, Return on Total Assets (ROTA) and Return on Equity (ROE). One liquidity ratio, Current Ratio, was used. We measured the solvency of the companies using the ratios Equity to Capital and Interest Coverage. Finally, we chose Receivables Turnover to measure the efficiency of the companies.

Table 1. The financial ratios and their formulas

Profitability Ratios	***1. Operating Margin***	$\frac{Operating\ \ Profit}{Net\ \ Sales} * 100.$
	2. ROTA	$\frac{Total\ \ Income + Interest\ \ Expense}{(Total\ \ Assets)\ Average} * 100.$
	3. ROE	$\frac{Net\ \ Income}{(Share\ \ Capital + Retained\ \ Earnings)\ Average} * 100.$
Liquidity Ratios	***4. Current Ratio***	$\frac{Current\ \ Assets}{Current\ \ Liabilitie\ s}.$
Solidity Ratios	***5. Equity to Capital***	$\frac{Share\ \ Capital + Retained\ \ Earnings}{(Total\ \ Assets)\ Average} * 100.$
	6. Interest Coverage	$\frac{Interest\ \ Expenses + Income\ \ Taxes + Net\ \ Income}{Interest\ \ Expenses}.$
Efficiency Ratios	***7. Receivables Turnover***	$\frac{Net\ \ Sales}{(Accounts\ \ Receivable)\ Average}.$

Pre-processing of the Quantitative Data

In order to ease the learning process of the SOM, the data have to be pre-processed. In addition to cleaning the data, this also implies standardization. The choice of standardization method is a very tricky issue, since it is very important for the outcome

of the map. In this experiment, we standardized the data by scaling the variables according to the variance.

In addition, a limit was placed on the extreme values, as is recommended by among others (Johnson and Wichern, 1997). This was because the network was initially placing too much emphasis on extreme values. The result was a flat map with one region of extreme values. In this case, the limits were set at –50 respectively 50.

3.2 Prototype-matching Text Clustering

We have used the methodology for text prototype matching to create the textual clusters. The clusters contain the reports that are the closest in meaning to a chosen prototype report. The methodology is based on textual collection preprocessing, i.e. word and sentence level processing. We transform every word into a number, taking into account word length in ASCII symbols, and the ASCII value of every character in a word. We encode every text document by creating a common word histogram for the entire text collection, choosing a suitable cumulative distribution. We chose the Weibull distribution[3], since it is a versatile distribution that can take on the characteristics of other types of distributions, based on the value of the shape parameter. In the training phase we divide the best fitted Weibull distribution into a number of logarithmically equal bins, the number of which is equal to the number of all words in the text collection. Each word after quantization is presented as a bin number and the values of the best-fitted Weibull distribution. Thus, we have performed text quantization on the word level, by creating a common word histogram for the entire text collection. The most common words in the text gain a dense resolution in the histogram bins.

Similar procedures for converting every word into a bin number on the sentence level are performed, in order to present the whole sentence as a vector. Hereafter, we consider the Fourier transformed encoded sentences as input vectors. After converting every sentence into a number, we create a cumulative distribution in the same way as on the word level. We divide the distribution into logarithmically equal bins, the number of which is equal to the number of all sentences in the text collection. Next, we count the frequency of sentences belonging to each bin, and find the best-fitted Weibull distribution based on the cumulative distribution of the coded sentences and their scalar quantization to equally distributed bins.

In the next phase, we construct individual sentence and word histograms for each document in the collection according to the documents' word and sentence code numbers and the corresponding value of quantization (Toivonen et al., 2001). Having sentence and word level histograms allows us to compare documents to each other simply by calculating the Euclidian distances between their histograms. The smallest Euclidian distance between word histograms indicates a common vocabulary of the reports. The smallest Euclidian distance between sentence histograms indicates similarities in written style and/or content of the reports (Visa et al., 2001).

---

[3] http://www.weibull.com/LifeDataWeb/the_weibull_distribution.htm

4 RESULTS

4.1 Quantitative Data Analysis

The map was created using SOM_PAK, a SOM training software package developed at the Helsinki University of Technology (Kohonen et al., 1996). The U-matrix map is visualized using the software Nenet v1.1a. The trained map from Karlsson et al. (2001a) was also used in the experiment, and the relevant data were mapped on to this existing map. The training parameters for the map are illustrated in Table 2.

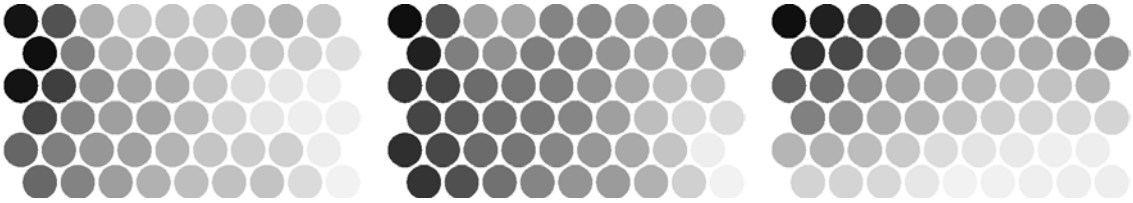


Figure 1. The Feature Planes for the Ratios Operating Margin, Return on Total Assets, and Equity to Capital.

Table 2. The Used Training Parameters.

Network size:		9 × 6	
Training length of first part:	5,000	Training length of second part:	50,000
Neighborhood radius of first part:	9	Radius of second part:	0.02
Learning rate of first part:	0.05	Training rate of second part:	1

Defining the Clusters

By carefully analyzing the output map, six major clusters of companies were identified. To identify the clusters we used both the U-matrix map and the individual feature planes. Figure 1 shows examples of the feature plane maps, in this case for the ratios Operating Margin, ROTA and Equity to Capital. High values are indicated by lighter shades, and lower values by darker shades. By analyzing the shades of the borders between the hexagons on the U-matrix map, it is possible to find similarities as well as differences. Furthermore, the values of the neurons have been evaluated in order to determine that the clusters are correct. The identified clusters are presented in Figure 2, in the form of a U-matrix map.

Group A1 and Group A2 represent the best-in-class companies. For the companies situated in subgroup A1, profitability is very good, with very high values in the financial ratios Operating Margin, ROTA, and ROE. Solvency is decent, i.e. the values of the Equity to Capital ratio and the Interest Coverage ratio vary from good to average. Group A2 is the second subgroup of the best in class group. The companies situated in this group are characterized by slightly lower profitability than Group A1, but instead, liquidity and solvency are much better. These companies generally have the best values in Current Ratio on the map.

Group B is where the companies with slightly poorer performance than those in Group A1 and A2 are situated. These companies are distinguished by good profitability, and particularly ROE values are excellent. These companies also have somewhat poorer liquidity and solvency than the companies in Group A.

Group C1 is the better of the two subgroups in Group C. In this group, companies possess decent profitability, good liquidity, and also good values in Equity to

Capital. Group C2 is the slightly poorer of the two middle groups. These companies have decent profitability, but poor liquidity. Interest Coverage and Receivables Turnover are also poor, but Equity to Capital, on the other hand, is very good.

Group D is the poorest group, containing the companies with the worst financial performance. Distinguishing features are poor profitability and solvency. Liquidity is average, and Receivables Turnover varies from very good to poor.

Analysis

The map that we have analyzed in this report (Figure 2) consists of the three most important manufacturers of telecommunications equipment in the world: Nokia, Ericsson and Motorola.

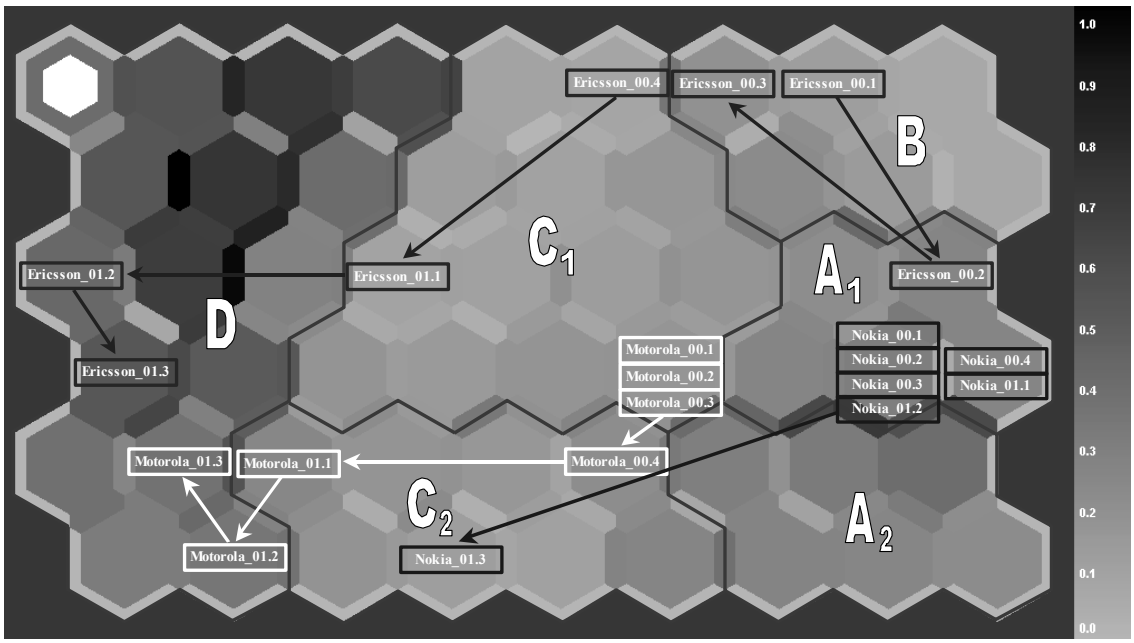


Figure 2. The identified clusters and the quarterly movements of Ericsson, Motorola, and Nokia.

Nokia moves very little during the first six studied quarters, remaining in Group A1. The values in the financial ratios are at almost the same level as they have been for the past six years, although a small decrease can be noted for the last quarters. Nokia performs excellently during the entire experiment. In the third quarter 2001, however, Nokia drops from the best group, entering Group C2. This is a result of Nokia making a one-time charge of EUR 714 million to cover Nokia Networks receivables related to a defaulted financing to Telsim, a cellular operator in Turkey, and to the insolvency of Dolphin in the UK.

Ericsson's performance differs considerably from Nokia's. During the first and third quarters of 2000, Ericsson is situated in Group B, the same group as for the past six years. The second quarter was particularly good for Ericsson, and the company moved into Group A1. During this quarter Ericsson shows significantly increased values in the financial ratios. In the fourth quarter Ericsson begins to experience difficulties, backtracking into Group C1. This is mainly due to decreased profitability and solvency. In 2001, Ericsson experiences severe difficulties on the

telecommunication market; almost all financial ratios show decreased values during the first quarter, and Ericsson drops within Group C1, ending up close to Group D. In the second quarter 2001, results are very poor, and terrible profitability and solvency places Ericsson in the poorest group, Group D. The only class that improves during this period is liquidity. In the third quarter 2001, the negative result is slightly smaller, and Ericsson shows a minor improvement in the financial ratios. However, the company remains in Group D.

Motorola performs consistently during the first three quarters of 2000, remaining situated in the same neuron within Group C1, very close to Group A1. Similarly to Ericsson, Motorola shows signs of worsening performance during the fourth quarter 2000, dropping into Group C2. The largest change can be found in the ROTA, ROE, and Interest Coverage ratios. In the first quarter 2001, Motorola drops even further within Group C2, ending up very close to Group D. In the second quarter, Motorola drops into Group D. During the third quarter profitability is even poorer, resulting in a decrease in three of the financial ratios. This results in a small movement within Group D. Motorola experienced a negative net income during the three first quarters in 2001, while Ericsson only experienced a negative net income during the second and third quarter. However, both companies display very similar performance during the last three quarters of the study.

4.2 Qualitative Data Analysis

All the reports were processed according to the procedures described in the Methodology section. In other words, we encoded every word from the reports, and constructed a common word histogram. We then encoded each sentence from the reports and constructed a common sentence histogram, and a unique sentence histogram for every report. We present the results from the qualitative data clustering for Nokia, Ericsson and Motorola in Table 3. Each column in Table 3 contains the prototype-report in the header and its four closest matches. The bold letters by the report codes denote the cluster from the quantitative clustering that a particular report belongs to. We highlighted the reports from Motorola 2001, quarter three for reasons stated later.

In order to obtain these results, we match every quarterly report against the entire data collection. We compare all of the quarterly reports in our data collection by calculating the Euclidian distance between their sentence histograms. For example, for the Ericsson report from 2000, quarter one, the closest report by content on the sentence level is from Nokia, 2000, quarter one. The second closest is the report form Nokia, 2000, quarter three. This means that the Nokia reports from 2000, quarters one and three and the Ericsson report from 2000, quarter one have similarities in sentence construction and word choice, which constitutes the language structure and written style. Word choice has a smaller impact on the clustering results on the sentence level than sentence construction does. Quarter labels and proper names, e.g. Nokia, Motorola or Ericsson, do not determine the clusters on the sentence level. On word level clustering, on the other hand, the clusters consisted primarily of reports by the same company.

Table 3. The closest matches for every report in the collection (Sentence level)

Ericsson2000Q1 B	**Ericsson2000Q2 A_1**	**Ericsson2000Q3 B**	**Ericsson2000Q4 C_1**	**Ericsson2001Q1 C_1**	**Ericsson2001Q2 D**	**Ericsson2001Q3 D**
Nokia2000Q1 **A_1**	Ericsson2000Q3 **B**	Ericsson2000Q4 **C_1**	Ericsson2000Q3 **B**	Ericsson2001Q2 **D**	Nokia2001Q3 **C_2**	Ericsson2001Q1 **C_1**
Nokia2000Q3 **A_1**	Nokia2000Q2 **A_1**	Motorola2001Q3 **D**	Motorola2001Q2 **C_2**	Ericsson2001Q3 **D**	Ericsson2001Q1 **C_1**	Ericsson2001Q2 **D**
Motorola2001Q3 **D**	Ericsson2000Q1 **B**	Ericsson2000Q2 **A_1**	Motorola2001Q3 **D**	Nokia2001Q3 **C_1**	Ericsson2001Q3 **D**	Nokia2001Q3 **C_2**
Motorola2001Q2 **C_2**	Ericsson2000Q4 **C_1**	Ericsson2000Q1 **B**	Nokia2000Q1 **A_1**	Motorola2001Q3 **D**	Nokia2001Q1 **A_1**	Nokia2001Q2 **A_1**

Motorola2000Q2 C_1	**Motorola2000Q3 C_1**	**Motorola2000Q4 C_2**	**Motorola2001Q1 C_2**	**Motorola2001Q2 D**	**Motorola2001Q3 D**
Motorola2001Q3 **D**	Ericsson2001Q2 **D**	Motorola2001Q3 **D**	Motorola2000Q2 **C_2**	Ericsson2000Q4 **C_1**	Motorola2000Q2 **C_1**
Motorola2001Q2 **D**	Nokia2000Q2 **A_1**	Nokia2000Q4 **A_1**	Motorola2001Q2 **C_1**	Motorola2001Q3 **D**	Nokia2000Q1 **A_1**
Nokia2000Q2 **A_1**	Nokia2000Q1 **A_1**	Nokia2000Q1 **A_1**	Nokia2001Q2 **A_1**	Motorola2000Q2 **C_1**	Nokia2001Q3 **C_2**
Nokia2000Q4 **A_1**	Nokia2001Q3 **C_2**	Ericsson2001Q2 **D**	Nokia2001Q3 **C_2**	Ericsson2000Q1 **B**	Ericsson2000Q1 **B**

Nokia2000Q1 A_1	**Nokia2000Q2 A_1**	**Nokia2000Q3 A_1**	**Nokia2000Q4 A_1**	**Nokia2001Q1 A_1**	**Nokia2001Q2 A_1**	**Nokia2001Q3 C_2**
Ericsson2000Q1 **B**	Nokia2001Q2 **A_1**	Nokia2001Q3 **C_2**	Nokia2001Q1 **A_1**	Ericsson2000Q1 **B**	Nokia2000Q2 **A_1**	Ericsson2001Q2 **B**
Motorola2001Q3 **D**	Motorola2001Q3 **D**	Ericsson2000Q1 **B**	Ericsson2000Q1 **B**	Nokia2000Q4 **A_1**	Nokia2001Q3 **C_2**	Nokia2000Q3 **A_1**
Nokia2000Q2 **A_1**	Nokia2000Q1 **A_1**	Motorola2001Q2 **C_1**	Motorola2001Q3 **D**	Ericsson2001Q2 **D**	Motorola2000Q2 **C_1**	Motorola2001Q3 **D**
Nokia2000Q3 **A_1**	Motorola2000Q2 **C_1**	Nokia2000Q1 **A_1**	Motorola2000Q2 **C_1**	Nokia2000Q1 **A_1**	Motorola2001Q1 **C_2**	Ericsson2001Q1 **B**

4.3 Combining Quantitative and Qualitative Analysis

Nokia performed very well during the analyzed period of time. Only during the last analyzed quarter did the financial performance decrease. The closest matches for the first four quarters were from Groups A-B, and later on from Group C (excluding the report by Motorola in 2001, quarter three, which appears to contain some linguistic peculiarities). The only closest match from Group D, apart from Motorola quarter three, appeared two quarters before the financial downturn actually happened, in quarter one of 2001. The second quarter 2000 sees an increase in the amount of average matches, with three reports from Group C among the closest matches. Quarter two 2000 shows a very small decrease in performance, but the real drop occurs in the third quarter.

Ericsson indicated the whole range of financial performance, moving from a good group to the best and, finally, to the worst financial group. Ericsson's performance improved slightly in the second quarter of 2000, as was indicated by the closest matches of the previous period. However, the tone changes somewhat in the second quarter and the company moves back into Group B in the third quarter. The first indications of worsening performance occurred in the third analyzed quarter, when reports indicating average financial performance started to fire as the closest matches. Even though Ericsson performed at the average level, and was in Group C1 in the first quarter of 2001, three out of the four closest matches were from Group D. For 2001 the poor reports were dominant among the closest matches.

Consistently, Motorola had performed poorly over the whole analyzed period, moving slowly from group C1 to D. The reports from the companies with the poorest performance appear as the closest matches to the Motorola reports in the second and third quarters of 2000 in anticipation of worsening performance in the fourth quarter of 2000, when Motorola ended up in Group C2. Most of the reports that fired as the closest matches to the Motorola reports in the first two quarters of 2001 were from companies from groups C and D. This indicated the financial downturn in the second and third quarters of 2001.

After carefully reading the quarterly financial reports, and analyzing the qualitative and quantitative information in them, we uncovered a pattern in quarterly movements that was confirmed with existing domain knowledge from (Karlsson et al., 2001a, 2001b).

Analysis

Information gathered from all occurring matches in combination with quantitative data clustering makes it possible to conclude that our analysis schema has captured a tendency: the text reports tend to foresee changes in the financial state of the company, before those changes influence the financial ratios. One must remember that the closest matching report may come from a company currently displaying good financial performance, but which may already be displaying negative consequences during the following quarter. An example of such a report is Nokia's report for 2001 quarter two. Although this report is an A2 report, the overall tone of the report is negative, as is displayed by its closest matches.

Roughly speaking, we notice a tendency among the reports to indicate the level of performance in the following quarter. We realize that the size of our data collection in the qualitative analysis is the biggest limitation to this conclusion. Ideally, we need to have the reports from the past of the companies in the data collection in order to identify indications of their future performance.

If a company reports good, steady performance over a certain period of time (from Groups A1, A2, or B in the quantitative data analysis) then we see the reports from companies with similar performance among the closest matches to the analyzed quarterly report, e.g. a report from Nokia 2000, quarter one as its closest match. When a company performs well, and expects to continue doing so, the tone of the report is positive with extensive use of optimistic vocabulary (increase, share growth, higher, our profitability, new, strong demand), active verbs (doing, not being), and clause constructions (operating margin, demand growth, increased share).

If a company reports an abrupt worsening of its performance, we see more companies with poor performance (from Groups C2 or D in the quantitative data analysis) among the closest matches one period before an actual financial downturn has occurred. The reports contain more conservative expressions (we expect, program efficiency), and nouns and verbs with negative financial connotations (decrease, slowdown, decline).

If a company anticipates a worsening of its financial performance in the next quarter, we see more companies with average performance (from Groups C1 and C2 in the quantitative data analysis), e.g. a report from Ericsson 2000, quarter three, among the closest matches. The tone of the financial report becomes less optimistic and more similar to ones that describe poorer performance. The style in the report becomes even more conservative (we have, announced, representing), using words and short sentence construction with particularly negative financial connotation (down, sales decline). The company avoids directly stating the accomplished results in its quarterly reports, instead shifting the subjects of emphasis in the report (sales segment, market share).

If a company reports average performance that does not change rapidly over time (from Groups C1, or C2 in the quantitative data analysis), e.g. Motorola's slow drop from average performance to poor performance, then we see companies with different financial performances among its closest matches. This might indicate that a report has no distinctive style, or a sentence construction that might reflect uncertainty about the future of the company. It is notable that Motorola's report from 2001, quarter three, behaves in an interesting way. It has fired among the four closest matches to 11 quarterly reports in our experiment. The tone of the Motorola's report is very neutral and requires further linguistic analysis.

5 COMPARISON WITH THE PREVIOUS STUDY

There are a number of differences between the current study and the previous research by Back et al. (2001) that are worth noting. Firstly, Back et al. (2001) have based their quantitative analysis on data from standardized annual reports, whereas we used non-standardized quarterly reports in this study. Secondly, we used a different methodology for analyzing the qualitative data. Back et al. (2001) have clustered the encoded text from annual reports by building SOMs on each of the levels (word, sentence, and paragraph) to create document histograms. We used a different text clustering method to create histograms of word and sentence levels for the quarterly reports. Instead of SOM based clustering we applied a vector quantization algorithm based on choosing the best-fitted Weibull distribution for the word and sentence vectors.

Both results, from our and previous studies, indicate the differences in the clustering of quantitative and qualitative data from the reports. Back et al. (2001) have explained this discrepancy by a slight tendency to exaggerate the performance in the textual part of reports in comparison to the real financial status of company in quantitative terms. Additionally, the authors of the previous study have suggested introducing a lag to see if the results would correspond better to the financial performance of the following year. In the current study we observed how fluctuations in real financial performances documented in qualitative parts influence the quantitative part of reports within some existing time lag. Moreover, time lag is of an individual length to every company, i.e. for Ericsson it lasts one quarter, for Motorola it can last two quarters. We did not consider using our prototype-matching clustering method as an information retrieval tool for searching and browsing, as was suggested in Back et al. (2001). Instead, we used it for text mining to discover non-obvious information.

6 CONCLUSIONS

It was a desire to come up with a better scheme to provide a way of finding hidden indications about a company's future financial movements that first motivated the work described in this paper. We specifically looked to benchmarking techniques based on SOM clustering that can classify and visualize the performance of the companies. Then, we analyzed the textual parts of quarterly reports for the same period of time, in order to reveal the heuristic relationship between the written style and facts stated by the numbers.

As the findings of Back et al. (2001) indicated, our study also showed that clusters from qualitative and quantitative analysis did not coincide. We explain the dissimilarities in clustering qualitative and quantitative data by the discovered phenomena that exists in qualitative and quantitative parts of every quarterly/annual report. The quantitative part of a report only reflects the past performance of a company by stating past facts. At the same time, the qualitative part of a report holds some message about future company performance and managerial expectations. We think that the sophisticated semi-automatic analysis of the style of the financial report helps to reveal insiders' moods and anticipations about the future performance of their company. The tone of a written report tends to change some time before the actual financial changes that influence quantitative part of the report occur.

The results that we have obtained after analyzing the qualitative and quantitative information from quarterly reports have proven that some future changes in financial performance can be anticipated by analyzing text from reports. Before a dramatic

change occurs in a company's financial performance, we see a change in the written style of a financial report. The tone tends to be closer to the company's future performance. If the company's position will be poorer in quantitative terms during the next quarter, the report of the current quarter tends to become more pessimistic, even though the actual financial performance remains the same.

The strongest limitation in our study is the small size of the data collection in text clustering. The limited vocabulary (terms related to finance and the telecommunications sector), extensive use of proprietary names (such as Motorola, Nokia, and Ericsson), and indications of time period (quarter, year, annual), might have slightly influenced the clustering ability in our qualitative analysis, although much less than on the word level. We plan to expand the study to a larger text collection.

Industrial analysts have methods to uncover indications and hints about the future financial performance of the company by reading their financial reports and making "professional guesses". We tried to retrieve those hints semi-automatically and wish to continue this study by comparing our predictions from mining tools with predictions made by industry analysts. The availability of computerized solutions for detecting companies' future financial intentions can be used in two ways: it can enlighten the work load of analytics saving them money and effort, but it can also conceivably help companies' officials to manipulate the public by faking written style.

REFERENCES

Back B, Sere K, Vanharanta H. 1998. Managing complexity in large data bases using self-organizing maps. *Accounting Management and Information Technologies* **8**(4): 191-210.

Back B, Toivonen J, Vanharanta H, Visa A. 2001. Comparing numerical data and text information from annual reports using self-organizing maps. *International Journal of Accounting Information Systems* **2**(4): 249-269.

Johnson RA, Wichern DW. 1997. *Business Statistics: Decision Making with Data*. John Wiley & Sons, Inc: New York, N. Y.

Karlsson J, Back B, Vanharanta H, Visa A. 2001a. *Financial Benchmarking of Telecommunications Companies.* TUCS Technical Report No. 395. Turku Centre for Computer Science: Turku.

Karlsson J, Back B, Vanharanta H, Visa A. 2001b *Analysing Financial Performance with Quarterly Data Using Self-Organising Maps.* TUCS Technical Report No. 430. Turku Centre for Computer Science: Turku.

Kohonen T, Hynninen J, Kangas J, Laaksonen J. 1996. *SOM_PAK: The Self-Organizing Map Program Package*: Helsinki University of Technology: Espoo.

Kohonen T. 1997. *Self-Organizing Maps*. Springer-Verlag: Leipzig.

Kohut G, Segars A. 1992. The president's letter to stockholders: An examination of corporate communication strategy. *Journal of Business Communication* **29**(1): 7-21.

Lehtinen J. 1996. *Financial Ratios in an International Comparison*. Acta Wasaensia: Vaasa.

Martín-del-Brío B, Serrano-Cinca C. 1993. Self-organizing neural networks for the analysis and representation of data: Some financial cases. *Neural Computing and Applications* **1**(2): 193-206.

Subramanian R., Isley R., Blackwell R. 1993. Performance and readability: A comparison of annual reports of profitable and unprofitable corporations. *Journal of Business Communication* **30**: 50-61.

Thomas J. 1997. Discourse in the marketplace: The making of meaning in annual reports. *Journal of Business Communication* **34**: 47-66.

Toivonen J, Visa A, Vesanen T, Back B, Vanharanta H. 2001. Validation of text clustering based on document contents. In *Machine Learning and Data Mining in Pattern Recognition (MLDM 2001)*, Perner P (ed). Springer-Verlag: Leipzig.

Visa A, Back B, Vanharanta H. 1999. Toward Text Understanding - Comparison of Text Documents by Sentence Map. In proceedings of *The 7th European Congress on Intelligent Techniques and Soft Computing (EUFIT'99)*, 13-16 September, Aachen, Germany.

Visa A, Toivonen J, Back B, Vanharanta H. 2000. A New Methodology for Knowledge Retrieval from Text Documents. In proceedings of the *TOOLMET2000 Symposium - Tool Environments and Development Methods for Intelligent Systems*, 13-14 April. Oulu, Finland.

Visa A, Toivonen J, Vanharanta H, Back B. 2001. Prototype-matching - Finding Meaning in the Books of the Bible. In *Proceedings of the Hawaii International Conference on System Science (HICSS-34)*, Maui, Hawaii.

Winsor D. 1993. Owning corporate texts. *Journal of Business and Technical Communication* **7**(2): 179-195.

Research Paper 3

Kloptchenko A., Back B., Visa, A., Toivonen, J., Vanharanta, H., Toward Content Based Retrieval from Scientific Text Corpora, In *Proceedings of 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS)*, Divnomorskoe, Russia, 5-10 September, 2002, pp. 444-449, isbn: 0-7695-1733-1/02

Toward Content based retrieval from Scientific Text Corpora

Antonina Kloptchenko
Barbro Back
Turku Center for Computer Science, IAMSR,
Åbo Akademi University
Turku, Finland
E-mail: {Antonina.Kloptchenko,
Barbro.Back}@abo.fi

Ari Visa
Jarmo Toivonen
Tampere University of Technology,
Department of Information Technology,
Tampere, Finland
Phone: +358 3365 438
E-mail: {Ari.Visa, Jarmo.Toivonen }@tut.fi

Hannu Vanharanta
Pori School of Technology and
Economics, Pori, Finland
Phone: +358 2 627 2759
E-mail: Hannu.Vanharanta@pori.tut.fi

Abstract

The growth of digitally available text information has created a need for effective information retrieval and text mining tools. We have used a content-based retrieval method that is built on a prototype-matching technique for clustering scientific text corpora, which in our case are the abstracts from The Hawaii International Conference on System Science 2001. Our aim is to retrieve the documents from a conference paper collection according to similarities in their contents and semantic structures. The method consists of "smart" document encoding on word and sentence levels, creating common word and sentence histograms using a vector quantization algorithm, and matching those histograms for every for document retrieval. In the paper, we position our methods among the existing document clustering methods, explain the motivation behind the clustering of scientific conference papers, and give an example of using our prototype tool for content-based retrieval on the scientific abstract collection. The method offers a promising alternative for retrieval by content.

Keywords: information retrieval, prototype matching, text

1. INTRODUCTION

The Internet, digital libraries, data warehouses, and information organizations generate and carry far more available text information than it is possible for anyone to process manually(Aslam 1999). During the last years the taxonomy of scientific conferences has grown very complicated, due to the blurred borders of modern research fields. The task of how to sort out the papers submitted to a scientific conference in the proposed categories and tracks is not trivial any more. Text is unstructured and indefinite data that carries different meaning to different users. The authors and the readers of the scientific articles frequently represent the same semantics using different words (synonmy) or describe different meanings using words that have various meanings (polysemy). Authors use similar keywords for identifying the content of the presented papers, which can belong to either the same or different tracks. Sometimes, even experienced readers, such as track chairmen, encounter certain difficulties with the determination of what track a particular paper belongs to. In this paper, we offer a prototype matching clustering system for text retrieval by content. We illustrate it using a scientific conference abstract collection from The Hawaii International Conference on System Science 2001. The system is based on "smart" document encoding and collection clustering. It aims to help the conference organizers and attendees to retrieve the papers from the conference proceeding based on their semantic content similarities. We suggest that the user take an abstract from an interesting paper, and use this paper prototype as a query. (dos Santos 1996)

The material presented in the remainder of this paper is organized as follows. In Section 2, we review the related work in using clustering for information retrieval and text mining purposes. In Section 3, we describe the document clustering methodology based on document encoding, creating word and sentence histograms, and prototype matching steps. In Section 4, we provide our motivation to perform a task of the prototype matching clustering on a scientific conference corpus and describe our experimental data set. In Section 5, we give a brief exposition of our experiments. Section 6 presents a discussion about the results. Finally, in Section 7, we provide some conclusions and suggestions for future work.

2. BACKGROUND

Document clustering and its applications in the information retrieval (IR) domain have been extensively explored. Clustering in TM strives to create a subset from a collection of documents, so that a cluster represents a group of documents having features that are similar, compared to the features of other groups (Hand D. 2001). Clustering does not require any predefined categories for grouping the documents (Jain 1999). The central assumption proposed by Van Rijsbergen in 1979, and known as Cluster Hypothesis, has made document clustering a powerful method for IR (van Rijsbergen 1979). It states that a document relevant to a request is more likely to be similar to one another than to non-relevant documents. Hierarchical, K-means and Binary Relational Clustering are the most known text clustering methods. (Karanikas 2000). Hierarchic document clustering using Ward's method based upon a series of nearest neighbor searches was addressed in (El-Hamdouchi 1986). Cutting (1992), Schutze (1997) suggested clustering algorithms for real-time computations and IR.

In (Lee 1999), a SOM-based clustering method based on word co-occurrences was presented for retrieval on a Chinese corpus from the web. Clustering for organizing the retrieval results on the Web using snippets, not a full text, was studied in (Zamir 1998). Text categorization according to natural topic structure using dense subgraph structure was accomplished in (Aslam 1999).Anick (1997) studied a document clustering approach for retrieval by content. The main points of this approach were to exploit clustering and paraphrases of term occurrence. Merkl (1997) used another clustering approach for retrieving by content and organizing legal text corpora. It was based on SOM as a clustering mechanism, and aimed at the detection of similarities between documents. In a majority of those algorithms, the user participates actively in the whole clustering process, controlling the fulfillment of his/her information needs.

There are a number of primary challenges in textual data clustering for retrieval by content, i.e. the effective representation of text, the determination of similarity, and the high dimensionality of document collections. The effective solutions for those challenges are discussed in (Schutze 1997), (Salton. G. 1983), (Hand D. 2001), and (Anick 1997).

We designed our prototype-matching clustering approach for a purpose of retrieval by content. It differs from the methods mentioned above because it does not focus on words or their co-occurrences (Lee 1999), or on feature extraction

(Larsen 1999), and does not create a high dimensional vector space to represent the whole collection (Cutting 1992). It takes into consideration that sentence structure; word order and paragraph structure carry just as much important semantic information to a reader as word appearances.

3. METHODOLOGY

The prototype-matching clustering methodology has been evolved over the development time and has acquired different clustering techniques (SOM and vector quantization algorithm), and currently consists of the following steps:

1. Pre-processing and basic filtering take place before text documents are presented to the text clustering system. Compiling the abbreviation file performs synonym or compound word filtering. Punctuation marks are separated by spaces. Numbers are rounded, and extra carriage returns, mathematical signs, and dashes are excluded. We do not perform stemming to keep our method language independent.
2. After basic filtering of the text, we encode the document on the word level. A word w is transformed into a number according to the following formula:

$$y = \sum_{i=0}^{L-1} k^i \times c_{L-i} \qquad (1)$$

where L is the length of the word character string, c_i is the ACSII value of a character within a word w and k is a constant. Every word and single punctuation mark are encoded to individual feature word vectors. This approach is accurate and sustainable for statistical analysis, although it is sensitive to capital letters and conjugations.
3. After each word has been converted to a code number we set the minimal and maximal values for the words, and look at the distribution of the words' code numbers for the entire document collection. In the training phase, the range between the minimal and maximal values of words' code numbers is divided into N_w logarithmically equal bins. We calculate the frequency of words belonging to each bin. For estimation of the word codes' distribution, we chose the Weibull distribution - one of the most widely used lifetime versatile distributions in reliability engineering (www.weibull.com, 1998). A number of parameters for Weibull distributions are calculated with various possible values for a and b using a selected precision. The best fitting Weibull distribution is to be compared with the code distribution in a sense of the smallest square sum by calculating the Cumulative Distribution Function according to:

$$CDF = 1 - e^{(((-2.6\times\log(y/y_{\max}))^b)\times a)} \qquad (2)$$

where a and b are the parameters to be adjusted in Weibull distribution. The size of every bin is $1/N_w$. Hereby, we have created a common word histogram for the entire document collection. Every word belongs to a bin that can be found using the code number and the parameters of the best fitting Weibull distribution. The quantization is the best where the words are the most typical to a document collection (usually 2-5 symbol length words).
4. On the sentence level every sentence is converted into a number after word coding. The whole sentence is considered as a sampled signal. We apply Discrete Fourier Transformation (DFT) to every coded sentence in a collection. Since the sentences in the text contain different numbers of words, the sentence vector's lengths vary. In the transformation we do not consider all of the coefficients, however, we transform bin number of the word i into output coefficients from B_0 to B_n to create a cumulative distribution like the one on the word level. The range between the minimal and maximal values of the sentence code numbers is divided into N_s equally sized bins. We calculate the frequency of sentences belonging to each bin. Then we divide the bins' counts with the total quantity of sentences. Finally, we find the parameters for the best Weibull distribution corresponding to the sentence data.
5. We examine every document in a collection by creating the histograms of the documents' word, and sentence code numbers (levels), according to the corresponding value of quantization. On the word level the filtered text from the document is encoded word by word. Each word code number is quantified using word quantization created with all the words in the database. The histogram consists of N_w bins and is normalized by the total number of words in the document. We created similar histograms for every document in the database for the sentence level.
6. Using the word and sentence histograms of all the documents in the database, we can analyze the single documents' text on the word and sentence levels, and compare them using any distance measures (e.g. Euclidian proved to be the best choice). The closest in terms of the smallest Euclidian distance form a cluster. Choosing the documents with the closest distances to the prototype completes the retrieval.

4. DESCRIPTION OF TASK

One of the distinct features of many modern conferences is cross-topic and interdisciplinary research. This feature creates certain obstacles within decision-making concerning what track a particular paper belongs to. Authors, conference organizers and attendees face difficulties in the conference setting while choosing an appropriate track. The conference organizers have repeatedly faced that there are similarities in the submitted papers that run across the traditional tracks.

We offer our user the opportunity to input into the system the abstract from a conference paper he/she has an interest in, and, thereby, to retrieve the papers that are semantically close to it. The user can insert a whole abstracts instead of spending time on constructing a smart query in prototype software we had created.

As an experimental data set, we have chosen 444 scientific abstracts obtained from The Hawaii International Conference on System Science 2001 (HICSS-34). Abstracts are designed to project research for the public eyes by offering a preliminary overview of the research in brief form (dos Santos 1996). The average length of HICSS abstracts is 300 words. The scientific papers at HICSS-34 were arranged into 9 major tracks, which were further divided into 78 mini-tracks. The organizers made an effort to identify six themes that run across the tracks based on the similarities and expansion of the scientific fields besides the traditional track division. Table 1 contains the taxonomy of the HICSS-34 conference.

5. EXPERIMENTS

We examined the system's ability to retrieve the most similar abstracts from the entire conference abstract collection. We have used any chosen abstract as a prototype query, trying to retrieve the abstracts of papers that are the most semantically similar to a prototype from a collection. We have expected the retrieval results to be from the same tracks, since tracks are the subsets of thematically similar research papers.

In the experiment, we have studied every abstract from the conference collection and their closest matches. We have performed clustering by calculating the Euclidian distances between the sentence histograms of an abstract-prototype and other abstracts, concentrating our attention on the abstract appearance in our clusters and in conference track division. We report our results for the recall window 47, which is equal to the average number of papers in the tracks. We did not consider order within a recall window, only paper co-occurrence.

№	Track Title /№ papers /№ Minitracks
1	Collaboration Systems and Technology /66 /9
2	Complex Systems /29 /5
3	Decision Technologies for Management /47 /7
4	Digital Documents /40 /6
5	Emerging Technology /30 /4
6	Information Technology in Health Care /26 /5
7	Internet and Digital Economy /68 /12
8	Organizational Systems and Technology /63 /14
9	Software Technology /75 /13
№	**Theme Title /№ papers in it**
1	Knowledge Management/20
2	Data Warehousing-Data Mining/24
3	Collaborative Learning/22
4	Workflow/12
5	E-commerce Development/54
6	E-commerce Application/36

Table 1. HICSS-34 Taxonomy

6. RESULTS AND DISCUSSIONS

We explain the results obtained from our system and a line of our reasoning on the example of the paper "Supporting Reusable Web Design with HDM-Edit" (INWEB 04) from "Web Engineering" minitrack, in "Internet and the Digital Economy" track. The paper analyzes the requirements and a design of a web-publishing tool. It sketches and describes HDM-editor, discusses the experiences of it use, and finally compares the requirements of the current version of the tool. The conference organizers had classified INWEB 04 into "Web Engineering" minitrack from "Internet and Digital Economy" track and additionally, into the Cross-Track Theme 5 "E-commerce Development". The theme unites the abstracts from 2 tracks: "Software Technology Track" and "Internet and Digital Economy", divided into a total number of 9 minitracks. Table 3 contains the distances between our prototype and the abstracts that are similar to it. The left column contains the codes of the papers that are the first 18 matches out of 443 possible ones in a recall window 47. The right column contains the distances. We used the italic font to outline the papers that belong to the same track as INWEB04.

After we read carefully every abstract from the top of a distance proximity table we have noticed, that the first nearest abstracts to INWEB04 discuss the problems related to collaboration support tools for web-based cooperation ("Experiences with Collaborative Applications that Support Distributed Modeling" (CLUSR23) from Collaboration Systems and Technology Track), coordination of shared software space ("Lost and Found Software Space" (ST3SE06) from the Software Engineering Tools Track). Those papers coincide with some of the ideas from INWEB04, such as a need for a support tool, its development, design and reuse. The closest matches are from the different fields of management information systems, namely software engineering (ST3SE06), groupware (CLUSR23) and business modeling ("Operations Centers for Logistics: General Concepts and the Deutsche Post Case" (DTIST04)), but they address the same problems of collaboration and tool reuse, either in software design or organizational structures.

INWEB04	0
CLUSR23	0.671421
ST3SE06	0.706321
DTIST04	0.773758
DDOML11	0.787317
OSOST06	0.789265
DDPTC06	0.796077
ST2EA03	0.83283
ST4TI08	0.83283
INBTB05	0.843204
OSTOI02	0.849694
INEEC06	0.857373
OSDWH01	0.857373
CLUSR05	0.870564
DDOML08	0.891668
INIEB03	0.898571
ST1MA01	0.898571
INWRK05	0.91045
INWRK02	0.910451

Table 3. A fragment of the proximity table to INWEB04
Recall window = 47

Table 4 contains a fragment of a proximity table for 5 papers: Impact of Renewable "Distributed Generation on Power Systems" (CSSAR01), "Multi-Area Probabilistic Reliability Assessment" (CSSAR02), "Min-max Transfer Capability: A New Concept" (CSSAR04), "Network Control as a Distributed, Dynamic Game" (CSSAR05), "Power System State Estimation: Modeling Error Effects and Impact on System Operation" (CSSAR06). They belong to "Security, Reliability and Control" minitrack of "Complex Systems" track.

After the detailed inspection of the distance proximity table for those papers, we discovered that some of the papers, being from the different tracks, have tendency to fire as the closest matches to the papers from this minitrack. For instance, the paper "Empirical Norms as a Lever for On-line Support of General Practice" (HCDMG08) being from "Information Technology in Health Care" track discusses problems of complex system model building, its sustainability and usage that are semantically similar to problems addressed in previous papers. Reasoning as follows, if paper A is close in meaning to paper C, and paper B is close to the same paper C, then paper A and B are semantically close, we induced the sustainability of our retrieval results. We highlighted those cross-referring papers by italic font in Table 4. Using gray background we outlined the papers "Collective Memory Support in Negotiation: A Theoretical Framework" (CLNSS05) and "Multi-level Web Surfing" (ETWFW05) that make the semantic similarity between CSSAR05 and CSSAR06 stronger. By Cosmic Sans font we highlighted the papers from the same "Complex Systems" track. We reasoned similarly for analyzing the retrieval by

CSSAR01	CSSAR02	CSSAR04	CSSAR05	CSSAR06
DDUAC06	OSKBE03	*DTUML06*	*ST3DS03*	DDTEC02
HCIST03	OSCIS01	HCTMD04	*CLUSR04*	*HCDMG08*
ST3SE03	CLUSR09	HCTMD05	*INMIW05*	OSSCI01
ST2EA04	*DTABS01*	ST2CP03	*OSOST09*	*CLALN02*
CLUSR16	DTIST02	DDOML06	*OSPMT06*	CSMAE02
DTMKI05	ST3SE02	DTDMK01	*ST2EA04*	INCRM04
CSMAE02	CLUSR19	INBTB04	CLALN05	*CLNSS05*
INIEB04	HCDMG01	DTABS03	*HCDMG08*	*ETWFW05*
CSIMG04	ST1MA02	*OSPMT06*	ST2CP04	*ST3DS03*
DDPTC08	ST3SA01	INCRM04	*DTUML06*	*DTIST01*
OSINF05	OSTTA07	DTABS04	CSIMG04	*CLNGL01*
ST4TI05	CSHDS02	CLDGS02	CSSOC03	*ST2WS01*
CLUSR02	INCRM03	CLENG01	CSSAR06	HCHIS01
INCRM05	ST1QS02	INCDE06	*CLNSS05*	*INEEC03*
ST2CP01	ST3SE01	INMAR04	*ETWFW05*	*INIEB04*
ST4NI03	HCDAM03	INMIW07	CLALN02	*DTUML06*
CLENG02	CLUSR23	CSIMG01	CSSAR04	CSHDS03
CLUSR08	OSETH03	*DDUAC04*	OSINF04	ST3SA06
ST2WS01	CSMAE07	*OSINF05*	CSIMG01	*ST2EA04*
ST3SA02	ETWFW03	CSSAR05	INEEC03	CSSAR08
HCDMG08	*OSOST09*	*ETSIT06*	DDUAC04	CSSAR05
DTABS01	CLUSR13	*INMIW05*	OSPMT04	*DTABS04*

Table 4. A Fragment from a Proximity Table for 5 papers from "Complex Systems" Track

content results for every track.

The hit ratios, that show how often the papers from the same track have fired on the top of a distance proximity table to a prototype from the same track, are presented on Table 5 for a recall window 47. Before warning, that the values of hit ratios are rather low one should understand the nature of comparison that we made between automatic retrieval results and conference track division while calculating hit ratio values. The hit ratio values are calculated in the assumptions that tracks unite semantically close paper. Track division is subjective and makes a weak reference point for calculating hit ration values very relative. As was noticed in (Yarowsky 1999), there are number of different issues expect topic of a paper, e.g. conflict of interest, to be considered while routing an article to a particular track in a conference settings.

№	Track Title	№ Papers	Hit ratio
1	Collaboration Systems and Technology	66	25.8%
2	Complex Systems	29	27.6%
3	Decision Technologies for Management	47	25.1%
4	Digital Documents	40	19%
5	Emerging Technology	30	30%
6	Information Technology in Health Care	26	23.1%
7	Internet and Digital Economy	68	23.5%
8	Organizational Systems and Technology	63	22.2%
9	Software Technology	75	21.3%

Table 5. The results from track clustering (Recall window = 47)

We have noticed that word usage and peculiarities of the academic written style of the scientific abstracts have a significant effect on the clustering ability of our methodology. Therefore the ranges of distance measures on word and sentence level were so narrow ([0.484344...1.246202] and [0.38517...1.414215] respectively). The majority of abstracts contain words such as *paper, analysis, discusses, present, the, result, system, model, process, information*, which makes abstract vocabulary very specific and versatile. The meaning of the text plays an important role in the clustering results as well. The evidence to this conclusion is strong on the sentence level analysis. The closeness of all abstracts on the sentence level can be explained by a particular academic writing style with specific sentence structure, e.g. *we present, our paper discusses, this paper describes*.

As for the limitations of our study, we can consider the critique toward the scalability of the methodology, limited experimental data collection and result evaluation. However, the methodology evaluation was offered in (Visa 2002) by examining the similarities in different translation of the books of Bible. The scalability of the method was already examined on TREC data (Visa 2001).

7.CONCLUSIONSAND FUTURE WORK

In this paper we have retrieve the semantically close abstracts from a scientific text corpus from the Hawaii International Conference on System Science-34 using to the prototype-matching clustering method. We aimed at establishing the semantic similarities among the conference papers by clustering the abstracts from them. Our prototype-matching clustering method consists of text filtering, "smart" document encoding on word and sentence levels, creating word and sentence level histograms, and prototype matching steps. We form clusters according to the Euclidian distances between the text of a prototype and the rest of a document collection.

Even though our clustering results turned out to be somewhat different from the track division offered by the conference organizers, our method was able to capture some semantic similarities between the scientific abstracts. The specific limited vocabulary and conservative academic style of the abstracts had a strong impact on our clustering results.

We suggest the use of our system's prototype-matching clustering ability, when the decision makers need to process a big number of text documents during the limited period of time. Reading some of the chosen papers in each cluster can provide the decision maker with the main ideas of all the documents from this cluster. As future work, we will consider to try out the method on the full-text articles from the HICSS-34 document collection.

8. ACKNOWLEDGEMENT

We gratefully acknowledge the financial support of TEKES (grant number 40887/97) and the Academy of Finland.

9. REFERENCE

Anick, P., Vaithyanathan, S. (1997). Exploiting Clustering and Phrases for Context-Based Information Retrieval. SIGIR 97, Philadelphia, USA, ACM.

Aslam, J., Pelekhov, K., and Rus, D. (1999). A Practical Clustering Algorithms for Static and Dynamic Information Organization. ACM-SIAM Symposium on Discrete Algorithms, ACM Press.

Cutting, D., Karger, D., Pedersen, J., and Turkey, J. (1992). Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. 15th Annual International SIGIR'92, Denmark, ACM Press, NY, USA.

dos Santos, M. (1996). "The textual organization of research paper abstracts in applied linguistics." Text **16**(4): 481-499.

El-Hamdouchi, A., and Willett, P. (1986). Hierarchic Document Clustering Using Ward's Method. ACM Conference on Research and Dvelopment in Information Retrieval, ACM Press.

Hand D., M. H., and Smyth P. (2001). Principles of Data Mining. Boston, USA, A Bradford Book, The MIT Press, 2001.

Jain, A., Murty, M., and Flynn, P. (1999). "Data Clustering: A Review." ACM Computing Surveys **31**(3): 265-323.

Karanikas, H., Tjortjis, C., and Theodoulidis (2000). An Approach to text Mining using Information Extraction. Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), Springer-Verlag Publisher.

Larsen, B., and Aone, A. (1999). Fast and Effective Text Mining Using Linear-time Document Clustering. KDD-99, San Dieago, CA, USA, ACM.

Lee, C., and Yang, H. (1999). A Web Text Mining Approach Based on Self- Organizing Map. WIDM-99, Kansas City, MO, USA, ACM.

Merkl, D., and Schweighofer (1997). En Route to Data Mining in Legal Text Corpora: Clustering Neural Computation, and International Treaties. 8th International Workshop on database and Expert Systems Applications (DEXA'97), Toulouse, France, IEEE.

Salton. G., a. M., M. (1983). Introduction to modern information retrieval. New York, McGraw-Hill.

Schutze, H., and Silverstein, C. (1997). Projection for Efficient Document Clustering. SIGIR 97, Philadelphia, PA, USA, ACM Press New York, NY, USA.

van Rijsbergen, C. (1979). Information Retrieval (Second Edition). London:, Butterworths.

Visa, A., Toivonen, J., Autio, S., Mäkinen, J., Back, B., and Vanharanta H. (2001). Data Mining of text as a tool in authorship attribution. AeroSense 2001, SPIE 15th Annual International Symposium on Aerospace/Defense Sensing, Simulation and Controls, Orlando, Florida, USA.

Visa, A., Toivonen, J., Back, B., and Vanharanta,H. (2002). "Contents Matching Defined by Prototypes: Methodology Verification with Books of the Bible." Journal of Management Information Systems **18**(4): 87-100.

Yarowsky, D. a. R. F. (1999). Taking the load off the conference chairs: towards a digital paper-routing assistant. Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora.

Zamir, O., and Etzioni, O. (1998). Web Document Clustering: A Feasibility Demonstration. SIGIR'98, Melbourne, Australia, ACM Press.

Research Paper 4

Kloptchenko, A., Back, B., Vanharanta, H., Toivonen, J., Visa, A., Prototype-matching System for Allocating Conference Papers, In *Proceedings of The Hawaii International Conference on System Science 2003 (HICSS-36)*, Hawaii, Big Island, USA, 6-9 January, 2003

Prototype-Matching System for Allocating Conference Papers

Kloptchenko Antonina *, Back Barbro*, Vanharanta Hannu, Toivonen Jarmo**, Visa Ari**

*Turku Center for Computer Science, Åbo Akademi University, Turku, Finland
Pori School of Technology and Economics, Pori, Finland
**Tampere University of Technology, Tampere, Finland
Phone*: (358)2215-3319, fax*: (358)2215-4809
Antonina.Kloptchenko@abo.fi, Barbro.Back@abo.fi, Hannu.Vanharanta@pori.tut.fi,
Jarmo.Toivonen@tut.fi, Ari.Visa@tut.fi

Abstract

Conferences on applied research require more complicated taxonomy than traditional organization of conferences by tracks. A topic of a paper, submitted to a conference on the applied research and the keywords, outlined by authors can be discussed in more than one proposed conference track. Sorting out the papers submitted to a scientific conference in the proposed categories and tracks is becoming a nontrivial task. Conference organizing committees try to schedule submitted papers very carefully to increase the success rate of the conference. For example, the organizers of The Hawaii International Conference on System Science 2001(HICSS-34) allocated the theme similarities in papers that were submitted into different tracks and identified 6 cross-track themes to schedule them appropriately.

In this paper, we offer a prototype matching system for text retrieval by content and try it out on the HICSS 34 conference proceeding. On the one hand, the system assists the conference organizers to automatically establish semantic similarities among papers and allocate them into common themes. On the other hand, the system assists the attendees to retrieve the papers from the conference proceedings based on their content similarities. A user can take an abstract or a paragraph from an interesting paper, and use it as a prototype query. The information system is based on document preprocessing, "smart" document encoding and prototype-matching clustering of a text collection.

1. Introduction

Traditionally many scientific conferences are organized into tracks. Conferences on applied research have a complicated taxonomy, because of the overlapping borders of applied research fields. The topic of a paper, submitted to a conference on applied research, can belong to several disciplines and be discussed in more than one proposed conference track. On the one hand, the conference organizers have a hard time determining and scheduling overlapping sessions successfully. On the other hand, a conference attendee has a hard time determining which conference sessions are relevant to his/her research interests. He/she needs either to browse the entire conference proceedings to identify interesting papers or to rely on a keyword search, considering keywords as a reflection of the paper content. Authors often use analogous keywords, which can belong to either the same or different tracks to identify the content of the submitted papers. Moreover, the authors and the readers of the scientific articles can represent the same semantics using different words (synonymy) or describe different meanings using words that have various meanings (polysemy) [8]. Sometimes, even experienced readers, such as track chairmen, encounter certain difficulties in the determination of what track the paper should truly belong to.

The amount of text in large conference proceedings requires a new generation of techniques and tools to support scientists in finding critical nuggets of useful knowledge. Quantity, quality and ambiguous structure of available text create many obstacles in working with it. Browsing, searching and organizing text collections turn out to be time consuming and costly procedures. Text mining (TM) methods in form of information retrieval (IR) by content tools strive to assist user information needs. TM is the process of analyzing text to extract information that is useful for particular purposes [24]. While searching text collections for relevant information, users face problems in constructing smart queries because they might not be fully acquainted with the established terminology in a field, or not fully sure about the content of the needed documents. This behavior requires sophisticated IR-by-content tools that could help users to deal with text collections.

The Hawaii International Conference on System Science (HICSS)[1] is a general-purpose conference that has served the computer society for over three decades. Contrary to many other conferences that have a focus on a specific subject or topic, HICSS addresses a wide range of issues from computer science, computer engineering, and information systems. The objective of HICSS is to "provide a unique environment in which researchers, academicians and practitioners in the information, computer and system sciences can exchange ideas, techniques and applications" [21]. The organizing committee of HICSS tries to build a workshop-like setting at the conference and schedule all the sessions carefully to create a high degree of interaction and discussion among the conference participants. Over the past year it has become clear to organizers that there are similarities or common themes in the papers that were submitted into different tracks. In 2001 the conference organizing committee of HICSS-34 had identified six cross-track themes that united some minitrack from different tracks. The organizers scheduled the papers from those themes very carefully to help conference attendees to participate in all relevant sessions. This case shows that sorting out the papers submitted to a scientific conference in the proposed conference tracks develops into a rather complicated task.

In this paper, we propose a prototype matching system for text retrieval by content. The prototype is a document or a part of it, which is of interest to a particular user. This prototype is matched with an existing document collection. We illustrate the system using a scientific conference collection from The Hawaii International Conference on System Science 2001. On the one hand, the system aims to assist the conference organizers to establish semantic similarities among the papers automatically. On the other hand, the system aims to assist the attendees to retrieve interesting papers from the conference proceeding based on their content similarities. A user can take the whole paper or an abstract from an interesting paper, and use it to construct a smart query. The core of the system is "smart" document encoding on different syntactic levels, and document collection clustering.

The material presented in the remainder of this paper is organized as follows. In Section 2 we review the related work in using text-clustering techniques for organizing text collections to enable information retrieval (IR) by content. In Section 3, we explain the methodology of a prototype-matching system that consists of document encoding, prototype matching and retrieval parts. The prototype-matching part is based on creating histograms for word and sentence levels of every document in a collection. In Section 4 we describe our motivation for creating the prototype matching system for identifying the scientific papers relevant to the user in a conference collection. In Section 5, we give a brief description of HICSS 34 scientific paper collection that we have worked with. Section 6 contains an description of our experiments, and discussion about the results. Finally, in Section 7, we provide some conclusions and suggestions for future work.

2. Background and related studies

The prototype-matching system that we use in our experiments is an IR by content technique based on document collection clustering. As a good IR system, our system directs a user to the semantically relevant document to satisfy his/her information needs. The characterization of relevance is complex, and thus, for the reasons of efficiency, IR systems use simplistic representation of document content and user information need. Good IR systems, by any mechanism available, should discover this dichotomy [5]. Text collection clustering and term-based approaches for the IR domain have been extensively explored to cope with this complicated task. Below we give a brief overview of the studies made in those approaches.

2.1. Text Clustering for IR by content

Organizing text collections for enabling the retrieval by content [2, 14, 15, 16] and searching [4] can be accomplished by using text collection clustering. Cutting et al. used a clustering technique that supports an iterative searching interface by dynamically scattering a document collection into smaller semantic clusters. A user navigated the document search space by selecting relevant documents among the clusters to regroup the results [4]. Anick and Vaithyanathan exploited document clustering and paraphrasing of term occurrence for document retrieval by content [2]. Merkl and Schweighofer used a different approach for detection of the similarities between documents in organized legal text corpora to enable document retrieval by content [16]. They combined a vector space model, cluster analysis and Self-Organizing Maps (SOM) to organize the legal text corpora as the hypertext and knowledge base of descriptors, probabilistic context-sensitive rules and meta-rules of legal concepts. Lee and Yang presented a SOM-based clustering approach based on word co-occurrences for IR on a Chinese corpus from the web [14]. SOM is a general unsupervised tool popular for

[1] http://www.hicss.org/history.pdf

clustering and visualization of very large document collections [11]. SOM organizes high-dimensional input data so that similar inputs are mapped close to each other. The WebSom system is based on SOM clustering and allows browsing and retrieval of the resulting matching list, allowing the user to navigate a multi-level search of text collection [12]. Lin et al. explored the potentials of the SOM semantic map as a retrieval interface for an online bibliographic system [15]. In a majority of the above-mentioned algorithms for IR by content, the user participates actively in the clustering and navigating processes, controlling the fulfillment of his/her information needs.

The effective representation of text, the determination of similarity, and the high dimensionality of document collections are primary challenges in text collection clustering for retrieval by content. Effective solutions to these challenges are discussed in [2, 8, 18], and [19]. Robustness and expendability are important for practical use of IR by content methods. The majority of content-based retrieval systems are based on computational linguistic approaches and linguistic knowledge about the text collection. Hatzivassiloglou et al. noticed that linguistically motivated features in conjunction with full word vectors increase the overall clustering performance [9]. Miike et al. developed a Japanese full-text retrieval system that analyzes text and enables the user to generate an abstract interactively [17]. The system was based on linguistic knowledge and clues, such as idiomatic expressions and was domain independent but required a dictionary of 60,000 entries for morphological analyzes of sentences.

2.2. Term-based approaches for IR

Other commonly used approaches for IR by content are based on user-defined *term-based* methods, such as keywords [3, 6, 10, 20], indexing [13, 23], or mark-ups [1]. Conversely, some valuable information hidden in the documents, which is not outlined by manually or automatically chosen keywords, indexes and markups, cannot be retrieved.

Keyword based clustering approaches have been studied by Sparck-Jones in [20]. Keywords from the Dewey decimal classification in content of books in the United States characterize text well but lack in accuracy [6]. Chien used Patricia tree for extracting the assigned by the author keywords and characterizing the content [3]. Jo assigned categorical substantial weights for informative, functional and alien keywords for text categorization [10].

C. van Rijsbergen studied the classical indexing IR approaches [22]. Lawrence created a full-text index of scientific literature on the web aiming at dissemination, retrieval and accessibility of the scientific literature [13]. The authors used the standard practice of indexing by building hash-table of words (inverted index) that contained a compressed version of the word and a pointer to a block of a record file corresponding to the positions in a matching document.

Markup tells how to display the material, rather than identifying what the material is. User-defined markups help to structure and categorize hypertext documents [1]. Keywords, headings and indexes can be used to mark and to create tags to the interesting document, i.e. flexible markups.

We designed our prototype-matching language independent clustering system to enable IR by content. It differs from the methods mentioned above because it does not focus on word co-occurrences [14], and does not create a high dimensional vector space to represent the whole document collection [4]. We tried to keep as much information from the original text of every document without modifications as possible. Our method takes into consideration that sentence structure and word order carry just as much important semantic information to a reader as word appearances. Our approach differs from the article routing method because it can be used without any human annotation and specially constructed profiles of expertise [25].

3. Prototype-matching system and its methodology

Our prototype-matching system is a simple content-based IR system. The system aims to retrieve the documents that contain *the same meaning* from the entire document collection. The prototype-matching system analyzes a document collection structure and thus is domain adjusting. The system consists of three parts: document collection preprocessing and encoding, document processing and matching, and document retrieval.

3.1. Document collection preprocessing and encoding

a. Pre-processing takes place before text documents are presented to the text clustering system. We do a basic filtering so that every sentence occupies its own line. Compiling the abbreviation file performs synonym and compound word filtering. We round numbers, separate punctuation marks with spaces, and exclude extra carriage returns, mathematical signs, and dashes. We kept the stop words.

b. After basic filtering of the documents in a text collection, we perform bag-of-word encoding of every document in it. Although this encoding approach is accurate and sustainable for statistical analysis, it is sensitive to capital letters and conjugations. Every word w in a document is transformed into a unique number according to the following formula:

$$y = \sum_{i=0}^{L} k^i \times c_{L-i} \quad (1),$$

where L is the length of the word character string, c_i is the ACSII value of a character within a word w, and k is a constant. We empirically choose k equal 256 since we are using 8-bit ACSII character set. The encoding algorithm produces a unique number for each word disregarding word stems, capitalization and synonyms, so that only the same word can get an equal number. The codes of every word and every single punctuation mark from every document formed feature word vectors representing individual documents stored in the file that corresponds to this document.

3.2. Document processing and matching

We have used the text clustering methodology and vector quantization algorithm for document processing and matching on word and sentence levels [23] that currently consist of the following steps:

a. We look at distribution (a set of word code numbers from 3.1b) for the entire document collection and compute the minimal and maximal parameters (a and b) for the word codes. In the training phase, we divide the range between the minimal and maximal values of words' code numbers into N_w logarithmically equal bins. We normalize the bins' counts according to the quantity of all words in the text. For estimation of the word codes' distribution, we chose the Weibull distribution. The Weibull distribution - one of the most widely used lifetime distributions in reliability engineering[2] - is a versatile distribution that can take on the characteristics of other types of distributions based on the value of the shape parameter. A number of Weibull distributions are calculated with various possible values for a and b using a selected precision. The best fitting Weibull distribution is compared with the code distribution by calculating the Cumulative Distribution Function according to:

$$CDF = 1 - e^{(((-2.6\times\log(y/y_{\max}))^{b})\times a)} \quad (2),$$

where a and b are the parameters of adjusted Weibull distribution. The size of every bin is $1/N_w$.

[2] http://www.weibull.com (1998)

Hereby, we have created a common word histogram for the entire document collection. The quantization is the best where the words are the most typical to a text (usually 2-5 symbol words - the most widespread length of English words). The distribution and thus quantization of longer words is sparser.

b. Similarly to the word level, we convert every sentence into a number on the sentence level. First, every word in a sentence is changed to a bin number (bn_i) in the same way as we did for words. The whole sentence is considered as a sampled signal. Since the sentences in the text contain different numbers of words, the sentence vector's lengths vary. To get past this fact we apply Discrete Fourier Transformation (DFT) to convert every sentence vector in a collection into input signal. Then we select coefficient B_i (i=1..n) to represent the transformation and the sentence signal. For these coefficients we create quantization like the one on the word level. The range between the minimal and maximal parameters of the sentence code distribution is divided into N_s equally sized bins. We calculate the frequency of sentences belonging to each bin. Then we divide the bins' counts with the total number of sentences in a collection. Finally, we find the best Weibull distribution corresponding to both cumulative distributions. A graphical representation of a sentence quantization process is given in [23].

c. Furthermore, we examine every document in a collection by creating the histograms of the documents' word and sentence code numbers (levels), according to the corresponding values of quantization. We encode the filtered document from a collection word by word on the word level. Each word code number is quantified using word quantization created with all the words in the database. The histogram consists of N_w bins and is normalized by the total number of words in the document. We create similar histograms for every document in the database for the sentence level.

3.3. Document retrieval

Using the histogram of all the documents in the collection, we analyze the single documents' text on the word and sentence levels, in order to compare them using any distance measures. The closest documents in terms of the smallest Euclidian distance between them form a cluster. To complete the retrieval part we choose the documents with the smallest distances to the prototype. The system creates a distance the proximity table of all distances among the documents in a collection. We retrieve the documents from the top of proximity table to every prototype document presented to the system within the set recall window.

4. Description of Task and Prototype Software

One of the distinct features of the Hawaii International Conference on System Science is cross-topic research. The interdisciplinary nature of research creates certain obstacles within decision-making concerning what track a particular paper belongs to.

The conference organizers have noticed the similarities that run across the traditional tracks in the submitted papers and tried to schedule those papers carefully to avoid conflicts. Trying to save the efforts of processing submitted papers manually, the conference organizers often rely on an authors' presentation of the keywords and headings as a reflection of the main topic of a paper, or on authors' choice of the submission minitrack, or on the track leaders who decide whether a particular paper is relevant to a track stream. All of those approaches risk the occurrence of incorrect paper classification and decrease the conference attendees' satisfaction. We offer our user the opportunity to input into the system the paper (prototype) from a conference collection he/she has an interest in, and thereby, to retrieve the papers that are semantically close to it. The user uses a prototype (a whole paper, or its abstract) instead of spending time on constructing a smart query.

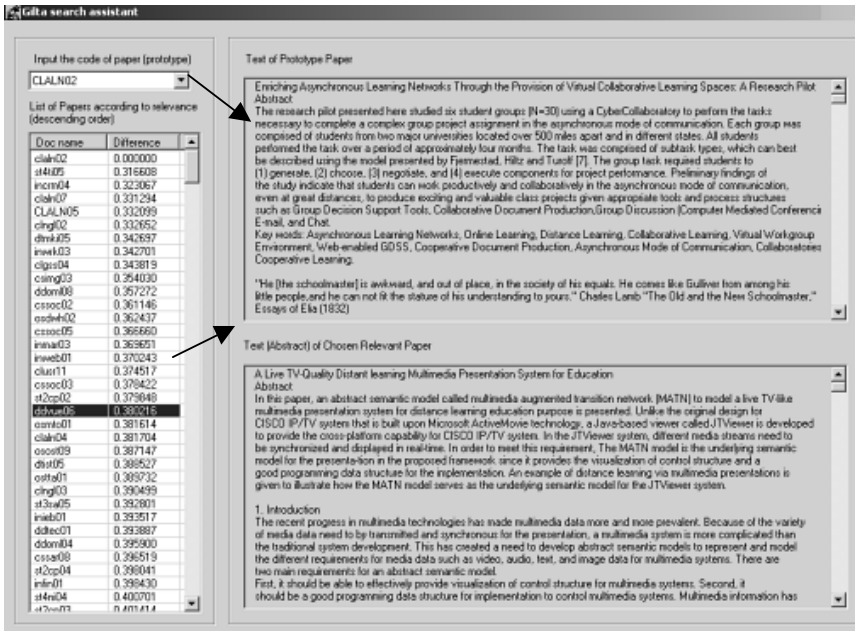


Figure 1. User Interface

The interface of our running prototype software is depicted in Figure 1. On the left panel we present the conference codes of a prototype paper (e.g. "Enriching Asynchronous Learning Networks Through the Provision of Virtual Collaborative Learning Spaces: A Research Pilot" (CLALN 02)) and the top of its distance proximity table with the codes of closest submitted papers (e.g. "webXice: an Infrastructure for Information Commerce on the WWW" (ST4TI05), "Conceptualizing Trust: A Typology and E-Commerce Customer Relationships Model" (INCRM04), etc.). We present the unfiltered texts of the query (a prototype paper) and a chosen retrieved paper ("A Live TV-Quality Distant learning Multimedia Presentation System for Education" (DDVUE06) that is close to a prototype on the upper and lower right panels respectively.

Our task is similar to an article routing task from [25] with n classes and a set of m articles. The task attributes each of m articles to one of the n classes, i.e. each paper has to go to one track/theme. The mathematical tasks of article routing and paper allocation for retrieval by content are similar. However, our interpretation of the task is different in the sense that automatic routines do not seek to establish similarities between m classes (tracks/theme) and n articles as we do. The differences between m classes in article routing task are known or can be learned from a previously constructed profile of expertise. We rely solely on our system's retrieval ability to determine paper content similarities and establish tracks and themes.

5. Description of Data Collection

As an experimental data set for our prototype-matching system, we have chosen all 444 scientific papers obtained from HICSS-34, consisted of about 4440 pages of scientific information, and occupied approximately 14 MB of hard drive space. Looking for the common topics and cross-links in this collection manually would be a time-consuming task. The scientific papers at HICSS-34 were arranged into nine major tracks, which were further divided into seventy-eight minitracks. The organizers made an effort to identify six themes that run across the traditional tracks based on the similarities and expansion of the scientific fields. Table 1 presents the taxonomy of the HICSS-34 conference. The outlined six *cross-track themes* are listed at the bottom of Table 1. They covered 168 papers in the conference from

30 mini-tracks. An organizing committee assigned a unique identification code to every paper. The code shows what track and minitrack the particular paper belongs to. We processed papers in portable-document format, which were converted to plain ASCII text. Distinct regions of the papers (title, authors, abstract, main body and bibliography) were manually identified and extracted, so only title, abstract and main body were left remaining in a filtered document collection for further experiments.

№	**Track Title /Number of papers /Number of**
1	Collaboration Systems and Technology /66 /9
2	Complex Systems /29 /5
3	Decision Technologies for Management /47 /7
4	Digital Documents /40 /6
5	Emerging Technology /30 /4
6	Information Technology in Health Care /26 /5
7	Internet and Digital Economy /68 /12
8	Organizational Systems and Technology /63 /14
9	Software Technology /75 /13
№	**Theme Title /Number of papers in it**
1	Knowledge Management/20
2	Data Warehousing-Data Mining/24
3	Collaborative Learning/22
4	Workflow/12
5	E-commerce Development/54
6	E-commerce Application/36

Table 1. HICSS-34 Tracks and Themes Taxonomy

6. Experiments

We have conducted several separate experiments to test the prototype-matching system capability to allocate the scientific conference papers based on their content. In our experiments, we have used the methodology described in Section 3 and the HICSS text collection described above. We have examined the ability of the prototype-matching system to retrieve the most similar papers to a presented prototype from the entire conference collection. We have expected to have papers either from the same track or from the same cross-track theme among the closest matches to a certain prototype. We have tried different sizes of recall window. We have not considered the order within a window, only paper co-occurrence, using precision and recall as effectiveness measures [22].

6.1 Experimental Settings

In the first experiment we have examined the consistency of the tracks proposed by conference organizers using every paper from every track as a prototype and a query. We have evaluated how often papers from the same track fire as the closest matches to the prototype-paper from the same track. We have analyzed tracks one by one, making a comparison between our prototype-matching clustering and the track division, proposed by the conference organizers.

The second experiment had a slightly different scope. We have examined the consistency of the cross-track themes, proposed by conference organizing committee as the semantic subsets of the papers from different track, with different key words and lexis. The division was meant to unite the papers with different terminologies and headings from different tracks into one interdisciplinary research theme. We have used every paper in every theme as a prototype and a query, trying to retrieve the most semantically similar papers. We have analyzed the cross-track themes one by one, comparing our prototype-matching clustering with the conference theme division. Because we built our system so that it could detect not only word co-occurrence but also main semantic similarities, we anticipate the second experiment give clustering results close to the conference theme division. We have expected to have the closest matches from different tracks but from the same cross-track theme.

6.2 Results

After presenting every paper as a prototype to the system we have obtained the proximity tables. The proximity table is a matrix of distances between a prototype and the rest of the papers in a collection. We present the fragment of a proximity table and the line of reasoning about some results below. Table 2 contains an example of a proximity table (recall window 23) for "Enriching Asynchronous Learning Networks Through the Provision of Virtual Collaborative Learning Spaces: A Research Pilot" (CLALN02), "Studies of ALN: An Empirical Assessment" (CLALN05), "CTER OnLine: Evaluation of an Online Master of Education Focusing on Curriculum, Technology and Education Reform" (CLALN06), and "A comparative Content Analysis of Face-to-Face vs. ALN-Mediated Teamwork" (CLALN07). All those papers belong to the Collaboration Systems and Technology Track, the Asynchronous Learning Networks Minitrack and, additionally, are assigned to Collaborative Learning Theme. The taxonomy of this theme is presented in Table 3. In Table 2 we highlighted the codes of papers that belong to the same cross-track theme as our sample

papers using gray background. We used the italic font to outline the codes of papers from the minitracks of the tracks that form the Collaborative Learning cross track theme. Paper CLGSS04 is underlined for the reasons stated in the discussion section.

One can notice that the paper CLALN02 has CLALN05 and CLALN07 among its closest matches, CLALN05 has CLALN02 among its closest matches, but at the same time, CLALN07 has none of those papers among its closest matches. The distances between CLALN07 and its closest matches in comparison with distances between CLALN02 or CLALN05 and their closest matches respectively, show that CLALN07 has papers other than CLALN02 and CLALN05 that are semantically closer. The distance range for CLALN 02 is [0.317..0.387], but at the same time, the distance range for CLALN07 is [0.249..0.329] (recall window=23). By the implication logic, the paper CLGSS04's closeness to CLALN02, CLALN05, and CLALN07 proves that all those papers are semantically similar.

Codes of the papers that are closest matches to a prototype	Codes of prototype-papers							
	CLALN02	0	**CLALN05**	0	**CLALN06**	0	**CLALN07**	0
	ST4TI05	0.316	**ST3SE03**	0.265	**ST4NI04**	0.304	***CLGSS04***	0.248
	INCRM04	0.323	**ST4TI05**	0.322	***CLDGS07***	0.334	***DDOML08***	0.254
	CLALN07	0.331	***DDPTC09***	0.328	**OSTTA05**	0.337	***DTMKI03***	0.270
	CLALN05	0.332	**CLALN02**	0.33	***DDOML04***	0.338	**ST2CP07**	0.271
	CLNGL02	0.332	***DDOML11***	0.334	**OSRMA02**	0.339	**CSSOC03**	0.279
	DTMKI05	0.342	**INWEB01**	0.334	**ST2IM01**	0.339	**OSTOI02**	0.282
	INWRK03	0.343	**OSMTO01**	0.341	***DTMKI03***	0.340	**INCRM01**	0.292
	CLGSS04	0.344	***CLUSR23***	0.346	**OSSCI03**	0.347	**ST2CP04**	0.295
	CSIMG03	0.354	***CLUSR14***	0.347	**ST2CP04**	0.348	***DTMKI05***	0.296
	DDOML08	0.357	**ST2CP03**	0.358	**ETEGV05**	0.355	**ST3SE01**	0.297
	CSSOC02	0.361	**CSSOC03**	0.359	**INMAR05**	0.356	***DTUML04***	0.306
	OSDWH02	0.362	**DDVUE05**	0.364	***DDOML08***	0.359	***DDUAC03***	0.307
	CSSOC05	0.366	**INWRK03**	0.365	**OSINF04**	0.360	***CLUSR17***	0.309
	INMAR03	0.369	**CLNGL02**	0.366	***DTUML03***	0.361	**CSSOC02**	0.311
	INWEB01	0.370	***DTMKI05***	0.368	**INWEB06**	0.362	**INWRK03**	0.312
	CLUSR11	0.374	***DTMKI03***	0.369	**CLALN07**	0.364	**OSPMT05**	0.313
	CSSOC03	0.378	***CLCDV06***	0.370	**INCRM04**	0.367	**ST2CP01**	0.315
	ST2CP02	0.379	**CSSAR08**	0.373	***DTUML04***	0.368	**INCDE05**	0.317
	DDVUE06	0.380	**ST3SA05**	0.373	***DTMKI05***	0.370	**OSCIS01**	0.318
	OSMTO01	0.381	***CLGSS04***	0.376	**OSETH01**	0.371	**OSPMT01**	0.323
	CLALN04	0.382	***DDOML08***	0.377	**INCRM01**	0.375	**ST1MA05**	0.326
	OSOST09	0.387	**CLNGL03**	0.378	**INWEB05**	0.381	**INWEB01**	0.328

Table 2. A Fragment from a Proximity Table for papers CLALN02, CLALN05, CLALN06, CLALN 07 from "Collaborative Learning Theme" (Recall window =23)

Name of the Track	Name of the Minitrack within a Theme (code of papers in it)
Collaboration Systems and Technology Track	Asynchronous Learning Networks (CLALN01-07)
	Next Generation of Learning Platforms (CLNGL01-03)
	Technology Supported Learning (CLTSL01-03)
Digital Documents Track	Digital technology and Educational Culture (DDTEC01-03)
	Digital Documents in the Office and Education (DDVUE01-06)

Table 3 Taxonomy of Collaborative Learning Theme

6.2.1. The First ("Track") experiment. In the first experiment, we have detected all closest matches to every paper from the conference collection. Aiming to check the consistency of conference tracks, we have focused on the results retrieved by our systems for every of nine tracks. We have presented every paper from the collection as a prototype to the system and calculated the "hit ratio", which reflects how often a paper from the same track has fired as the closest match to the presented prototype in a given recall window. We assumed that the tracks should be somewhat balanced by the number of papers in them, and used a recall window of 47, as the average size of HICSS tracks and at 25 for reasons stated later. Table 4 contains hit ratios per track (hit ratio1 and hit ratio2), that reflect how many papers from the same track were retrieved among the 47 or 25 closest matches respectively.

Track Title	Number of Papers	Max Hit ratio1	Max Hit ratio 2
Collaboration Systems and Technology	66	22.7%	15.2%
Complex Systems	29	17.2%	13.8%
Decision Technologies for Management	47	19.1%	12.8%
Digital Documents	40	22.5%	15%
Emerging Technology	30	20%	13.3%
Information Technology in Health Care	26	23.1%	15.4%
Internet and Digital Economy	68	19.1%	13.2%
Organizational Systems and Technology	63	22.2%	15.9%
Software Technology	75	22.7%	14.7%

Table 4. The results from track division clustering (Recall window = 47 and 25)

6.2.2 The Second ("Theme") experiment. In the second experiment we have focused on the closest matches to 168 papers that were arranged by the conference organizers into six cross-track themes. We expected to have many closest matches from the same cross-track theme in the recall window, because themes are meant to

unite semantically close papers. Assuming that the conference committee wishes to have roughly balanced themes, we set a recall window at 25. Table 5 contains the name and sizes of cross-track themes, and hit ratios that reflect how many papers from the same theme fired among the 25 closest ones to a prototype paper from the same theme. The last column (hit ratio4) shows how many papers have their closest matches from the same track, some minitracks of which have formed the certain theme. For instance, the "Collaborative learning" theme has the highest hit ratio at 31.8%. This means that almost every third paper among the closest matches was from the same theme as a prototype paper. The Collaborative Learning theme has the highest hit ratio4 at 23.2%, stating that almost every fourth paper among the 25 closest matches was from Collaboration Systems and Technology or Digital Documents tracks, maybe from minitracks different than mentioned in Table 3.

Theme Title	Number of Papers	Max Hit ratio3	Max Hit ratio4
Knowledge Management	20	25%	8.46%
Data Warehousing/ Data Mining	24	20.8%	9.7%
Collaborative Learning	22	31.8%	23.2%
Workflow	12	18.2%	10.5%
E-commerce Development	54	18.5%	11.4%
E-commerce Application	36	17.1%	8.6%

Table 5. The results from cross-topic theme clustering (Recall window = 25)

6.3. Discussions

We set the same size of recall windows to make the results from both experiments comparable. Looking at hit ratio2 from Table 4 and hit ratio3 from Table 5 we conclude that the hit ratios for theme division, on average, were slightly higher than hit ratios for track division with the same size recall window. This demonstrates the stronger semantic similarity among the papers from the same cross-track themes than the semantic similarities among the papers from the same tracks. However, the hit ratio for "E-commerce Development" and "Workflow" themes are rather low in comparison with hit ratios of the other themes. After analyzing those themes we discovered that some papers in them contain the initial data that was treated as noise by our system. Although the paper "Workflow Analysis Using Attributed Metagraphs" (ST2IM05) has fired in the bottom of a proximity table to all 11 papers from the same theme, it clearly belongs to the Workflow theme, because it discusses presentation and formal analysis of workflows as metagraphs with specified temporal constraints for time-critical tasks, i.e. for generalization of traditional network scheduling methods used in project management. The paper is full of technical details and formulas that our system, apparently, failed to understand. The "E-commerce Development" theme has several outliers that do belong to the theme semantically, but use very diverse lexicon (e.g. on-line markets instead of electronic or e-market), such as "Second-Degree Price Discrimination for Information Goods Under Nonlinear Utility Functions" (INEEC06) and "Transforming Financial Markets to Retail Investors: A Comparison of the U.S. and the German On-line Brokerage Market" (INFIN02). The paper INEEC06 contains a mathematical description of a model for price discrimination followed by heavy mathematical reasoning from one formula to another that our system fails to recognize. The paper INFIN02 has fired at the bottom of the proximity table to 16 papers from the same cross-track theme because it uses many unique proper adjectives, such as German, European and American that influences the sentence construction. The change in sentence construction has an impact on the retrieval ability of our system. In order to enhance our IR system's abilities to handle synonymous attributes and disambiguation we plan to construct an extensive synonym table for filtering.

We have noticed that word usage and some peculiarities of the academic written style of the scientific papers in a collection have a significant impact on the clustering ability of our method. All research papers consist of the same components: introduction, method, research background, results and discussion [7]. Therefore the intervals of distance measures on word and sentence level are so narrow ([0.314...0.773] and [0.23...1.149] respectively). A particular academic writing style, since authors tend to use similar word order and similar sentence structures to describe their achievements in information system research, e.g. *we present, computer analysis our paper discusses, construct a model, approach is based, process information, in the remainder of a paper, this paper describes, traditional systems, etc*. explains the closeness of all papers on the sentence level. Finally, we have discovered that our prototype-matching clustering of the scientific text corpus is somewhat different from the theme division proposed by the organizing committee. It can be explained either by poor allocation of papers to minitracks, tracks or themes by the HICSS organizers, or by poor performance of the proposed method.

One can notice that the good matching results of automated clustering with human's manual selection should be in the range [40%…60%] versus our experimental results with matching range [13%…32%]. The justification of the ranges is not an obvious task since for the calculations of matching results we compared our retrieval results to the track and theme divisions provided to us by the conference organizing committee. Those divisions can be a product of the message that every paper conveys, the author's vision of a paper and a number of non-optimal considerations that conference chairs keep in mind in addition to the topic relevance of a paper submitted to a certain track or theme. As one subcommittee chair has noticed, there are a number of other issues in addition to content relevancy that should be balanced in conference settings, such variables as conflict of interests, gender, geography, topic, etc. Yarowsky and Florian noticed that members of conference committees tend to favor the article with the most interesting content and findings and route them to their tracks, even if the topics are not so relevant [25]. Obviously, our prototype-matching system that is based on objective text processing technique does not consider those issues. The nature of hit ratios' calculation makes the evaluation of out results very challenging. We calculated the hit ratios on the strong assumption that a given theme/track division by the HICSS committee is the semantically absolutely correct one. However, the HICSS theme/track division is a very weak reference point for comparison because of the issues mentioned above. For instance, CLALN02 and DDVUE06 papers that belong to the same cross-track theme "Collaborative Learning", apparently convey different messages (see Figure 1).

Additionally, there are a number of useful subtasks that our prototype-matching system can handle. It can be used to detect "good-candidate" papers to be included in the theme from minitracks that were not included in it. For example, the paper "The Mindpool Hybrid: Theorizing a New Angle on EBS and Suggestion Systems" (CLGSS04) from the Group Support Systems minitrack has fired at the top of the proximity table to almost every third paper from Collaborative Learning theme (shown in Table 2). Careful reading of this paper has shown that it could be included in the theme because it discusses the improvements possible that can be achieved by using computer support in electronic brainstorming, which is a collaborative technique in nature.

Our system can have another use, similar to article routing discussed in [25]. The prototype-matching system can mark submitted papers and send them to the appropriate conference subcommittee or minitrack. Then we need to add into our document collection the descriptions of minitracks that are usually provided by minitrack chairs. There are fewer possibilities for conflicts of interest involving the balance between the issues discussed above. The users of the system could use this method in a semi-automatic setting where the program makes recommendations for paper rerouting but it is up to the conference or track chairs to make the final decision.

As for the limitations of our study, we can consider the critique toward the scalability of the methodology, small experimental data collection and result evaluation. However, the methodology evaluation was offered in [23] by examining the similarities in different translation of the books of Bible. The question is how well does the new proposed system compare to the standard (manual) methods of doing this, both in terms of costs (manual work by the chairs, subcommittee chairs, authors, etc) and benefits has to be explored.

7. Conclusions and Future Work

In this paper we have clustered a scientific text collection from the Hawaii International Conference on System Science-34 using the prototype-matching IR system. The conference organizers of HICSS-34 had offered non-traditional cross-track theme classification of the submitted papers to help the conference attendees to visit all sessions relevant to their research needs. We aimed to allocate semantically close papers from HICSS according the non-traditional conference taxonomy. Our prototype-matching IR system consists of document encoding, prototype matching and retrieval parts. The core of prototype-matching system of text filtering, "smart" document encoding on word and sentence levels, creating word and sentence level histograms, and prototype matching phases. We form clusters according to the Euclidian distances between the prototype article and the rest of a document collection.

In the paper we have presented two experiments from the clustering sessions on a scientific collection. We tested the system's ability to retrieve the closest papers according to content from the whole document collection. In our first experiment we examined track consistency with respect to the retrieved results from our system. In the second experiment, we examined the semantic closeness of the papers within every cross-track theme by studying their retrieved closest matches. Even though our clustering results turned out to be somewhat different from the cross-track division offered by the conference organizers, our method was able to capture some semantic similarities between the scientific papers. The

specific limited vocabulary and conservative academic style of the scientific papers had a strong impact on our clustering results.

We suggest the use of our system's prototype-matching clustering ability for processing a big number of text documents during the limited period of time. Reading some of the chosen papers in each cluster can provide the decision maker with the main ideas of all the documents from this cluster. As future work, we consider evaluation of the system's effectiveness with the help of expert-readers, methods for improvements, and exploration of additional uses of the system. In the long-term prospective, the system can be used as an on-line retrieval system that can help conference participants to choose the most interesting conference venues.

8. Acknowledgement

We gratefully acknowledge the financial support of TEKES (grant number 47 533) and the Academy of Finland.

9. Reference

[1]. Anderson, N., "A Tool for Building Digital Libraries", *Digital Library Journal Review* 5, (2), 1999

[2]. Anick, P., Vaithyanathan, S., "Exploiting Clustering and Phrases for Context-Based Information_Retrieval" SIGIR 97, Philadelphia, USA, ACM,.1997

[3]. Chien, L. F., "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval", Proceedings of Special Interest Group on Information Retrieval, SIGIR'97, Philadelphia, USA, ACM Press, 1997.

[4]. Cutting, D., Karger, D., Pedersen, J., and Turkey, J., "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections", 15th Annual International SIGIR'92, Denmark, ACM Press, NY, USA, 1992

[5]. Deogun, J., Raghavan, V., User-oriented document clustering: a framework for learning in information retrieval, ACM conference on Research and development in information retrieval, Pisa, Italy, ACM Press New York, NY, USA, 1986

[6]. Dewey, M., *A classification and subject index for cataloguing and arranging the books and pamphlets of a library,* Amherst, MA, USA, Case, Lockwood & Brainard Co., 1876

[7]. dos Santos, M., "The textual organization of research paper abstracts in applied linguistics", *Text* 16(4), 1996, pp 481-499

[8]. Hand D., M. H., and Smyth P., *Principles of Data Mining. Boston*, USA, A Bradford Book, The MIT Press, 2001.

[9]. Hatzivassiloglou, V., Gravano, L., Maganti, A., "An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering", The 23rd ACM/SIGIR conference on Research and development in IR, Athens, Greece, ACM Press New York, USA, 2000

[10]. Jo, T., "Text Categorization considering Categorical Weights and Substantial Weights of Informative Keywords", Tokyo, Japan, Samsung SDS, 1999, pp1-17

[11]. Kohonen, T., "Self-Organization of Very Large Document Collections: State of the Art", Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks, Springer, London, 1998

[12]. Kohonen, T., *WEBSOM*. Helsinki, Helsinki Technological University, Finland, 1999

[13]. Lawrence, S., Bollacker, K., Lee Giles, C., "Indexing and Retrieval of Scientific Literature", The 8th International Conference on Information and Knowledge Management (CIKM 99), Kansas City, MO, USA, ACM Press, 1999

[14]. Lee, C., and Yang, H., "A Web Text Mining Approach Based on Self- Organizing Map", WIDM-99, Kansas City, MO, USA, ACM Press, 1999

[15]. Lin, X., Soergel, D., and Marchionini, G., "A Self-organizing Semantic Map for Information Retrieval", The 14th ACM/SIGIR conference on Research and development in IR, Chicago, IL, USA, ACM Press, 1991.

[16]. Merkl, D., and Schweighofer, "En Route to Data Mining in Legal Text Corpora: Clustering Neural Computation, and International Treaties", The 8th International Workshop on Database and Expert Systems Applications (DEXA'97), Toulouse, France, IEEE, 1997

[17]. Miike, S., Etsuo, I., Ono, K., Sumita, K., "A full-text retrieval system with a dynamic abstract generation function", The 17th ACM/SIGIR conference on Research and development in IR, Dublin, Ireland, Springer-Verlag New York, Inc., 1994

[18]. Salton. G., a. M. McGill (1983). *Introduction to modern information retrieval.* New York, McGraw-Hill.

[19]. Schutze, H., and Silverstein, C., "Projection for Efficient Document Clustering", ACM/SIGIR-97, Philadelphia, PA, USA, ACM Press New York, USA, 1997

[20]. Sparck-Jones, K., *Automatic Keyword Classification for Information retrieval,* Connecticut, Archon Books, 1971

[21]. Sprague, R. H., Jr., "Preface to The Hawaii International Conference on System Science 2001", HICSS-34, Maui, Hawaii, 2001

[22]. van Rijsbergen, C., *Information Retrieval* (Second Edition), London: Butterworths, 1979

[23]. Visa, A., Toivonen, J., Back, B., and Vanharanta, H., "Contents Matching Defined by Prototypes: Methodology Verification with Books of the Bible." *Journal of Management Information Systems* 18(4), 2002, pp 87-100

[24]. Witten, I., Bray Z., Mahoui, M., and Teahan, B., "Text mining: A new frontier for lossless compression",. Data Compression Conference '98, IEEE, 1998

[25]. Yarowsky, D. and R. Florian, "Taking the load off the conference chairs: towards a digital paper-routing assistant",

Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora, ACM Press, 1999

Research Paper 5

Kloptchenko, A., T. Eklund, A. Costea, B. Back (2003), A Conceptual Model for a Multiagent Knowledge Building System, in *Proceedings of the 5th International Conference on Enterprise Information Systems (ICEIS 2003)*, Angers, France, April 23-26, 2003, Vol. 2, pp. 223-228

A CONCEPTUAL MODEL FOR A MULTIAGENT KNOWLEDGE BUILDING SYSTEM

Antonina Kloptchenko, Tomas Eklund, Adrian Costea, Barbro Back
Turku Centre for Computer Science and IAMSR / Åbo Akademi University, Lemminkäisenkatu 14 B, 20520 Turku, Finland
Email: akloptch@abo.fi, toeklund@abo.fi, acostea@abo.fi, bback@abo.fi

Keywords: Software agents and multiagent systems, financial analysis, data mining, text mining

Abstract: Financial decision makers are challenged by the access to massive amounts of both numeric and textual financial information made achievable by the Internet. They are in need of a tool that makes possible rapid and accurate analysis of both quantitative and qualitative information, in order to extract knowledge for decision making. In this paper we propose a conceptual model of a knowledge-building system for decision support based on a society of software agents, and data and text mining methods.

1 INTRODUCTION

A huge amount of electronic information concerning different companies' financial performance and market situation is available in various databases and on the Internet today. This information can potentially be very valuable to companies' decision makers, their partners, competitors, investors, analysts, and stakeholders. These individuals want to extract relevant information for decision-making purposes from the widely available data storages on time and, preferably, by the click of a mouse button. The enormous supply of data available often exceeds our capacity to analyze it, leading to information overload. Users need to transform new data into valuable knowledge very quickly in order to react to rapidly changing conditions and make crucial decisions in time.

Although there are a number of methods and technologies available for creating, storing, and monitoring new data, there are not very many comprehensive and popular techniques for transforming all data into valuable information and knowledge. The fields of *knowledge discovery in databases (KDD), data mining (DM), and text mining (TM)* have provided a number of new approaches for analysis of large databases of financial data. KDD is the entire process of discovering interesting knowledge, such as patterns, associations, changes and anomalies, and significant structures from large amounts of stored data, while DM refers to the actual use of data mining tools for identifying patterns in the data (Fayyad et al. 1996).

Most data mining techniques for financial applications deal with quantitative data. The analysis of qualitative information (company strategy, economic market outlook, i.e. the textual parts of financial statements, as well as information from outside sources) is very important and can be done using text mining approaches. TM refers to the nontrivial extraction of implicit, previously unknown, and potentially useful information from large textual datasets (Dorre et al., 1999). Unlike numeric data, textual statements contain not only the factual event but also the explanation for why it happens (Wuthrich et al. 1998).

The individuals are fortunate if the valuable data that they need are already stored in one available database on the web. More often the data are located on a number of different sites. An emerging problem is how to find and collect these data and process them so that they provide additional valuable knowledge. The majority of data mining techniques are meant for extracting meaningful patterns from numeric, well-structured databases. At the same time, ambiguously structured text databases grow large in size and significance, and require effective text mining techniques. A multi-agent software system consisting of a collection of individual software agents, each of which provides a certain task (Lesser 1995) and/or uses different data mining techniques, can be a possible solution for accomplishing this task.

In this paper we create a conceptual model of a knowledge building system based on a society of software agents, and data and text mining methods. Each agent exhibits intelligence by using different

data and text mining methods. We believe that software agents, which are able to execute tasks on behalf of a business process, computer application, or an individual, are well suited to dealing with collecting, processing, and compiling vast volumes of dynamic data from distributed sources. The system could monitor new financial updates from a variety of sources, and calculate financial ratios for different companies. These data could be used for various tasks, for example, financial benchmarking and assessing creditworthiness of different companies.

Our model suggests the integration of several computing techniques, namely self-organizing maps for clustering quantitative information, decision trees and/or multinomial logistic regression for classifying new cases into previously obtained clusters, prototype-matching for semantic clustering qualitative information, and various techniques for text summarization. We have previously tested some of the techniques in certain modules of the conceptual model.

The paper is organized as follows: In Section 2 we describe the problem area and the approaches used in financial data analysis for solving the discussed problems and provide an overview of literature and related work in multiagent system design. We describe the conceptual model of our multiagent decision support system in Section 3. We explain the methodological issues of the different computational techniques we propose in Section 4. We discuss the possible limitations and difficulties associated with building and using the proposed system in Section 5. Section 6 contains our conclusions and the directions of future work.

2 DESCRIPTION OF PROBLEM AREA AND RELATED WORK

Financial analysis is very important in today's global economy. Access to more information should be beneficial to any investor or financial stakeholder. Financial benchmarking is an important and valuable tool for assessing the actual financial performance of a company. Financial benchmarking is the process of comparing a number of competitors according to, most commonly, a number of financial ratios, chosen based on the motive for the benchmarking (for example, to compare profitability, efficiency, etc). This type of benchmarking is often external, and does not require the participation of the benchmarked companies. Indeed, financial benchmarking is often performed by consulting companies, or business or industry-specific journals (such as *Pulp and Paper International*). Financial benchmarking can also be used by individual investors seeking to evaluate the actual financial performance or state of an investment object in comparison to competing investment opportunities.

An assessment of the creditworthiness of debt-issuing companies is based on the financial statements of the issuer and on expectations of future economic development using a combination of qualitative and quantitative analysis (Tan et al. 2002). Credit rating agencies (e.g. Moody's Investor Services, Standard & Poor Corp., FLIP) are commercial firms that receive payment for publishing an evaluation of the creditworthiness of their clients. Creditworthiness information is especially useful when borrowing takes place through the issue of securities, rather than by bank loans, since buyers of securities do not know the issuers as well as banks usually know their customers.

The idea of a society of software agents was introduced in Wang et al. (2002) for monitoring and detection of financial risk. In a society of software agents each agent carries out different functions autonomously. We use a multiagent approach for building our knowledge creating system.

There have been a number of attempts to use multiagent systems to support business processes and deal with business environment. Liu (1998) suggested a software agent approach in environmental scanning activities for senior managers. An agent system developed by PriceWaterhouseCoopers, called EdgarScan, scans the financial reports in the Securities and Exchange Commission's database (EDGAR). The agent works by scanning the document for tags that indicate certain financial data. The system also includes a basic graphical benchmarking system, which only allows the user to compare companies by one ratio or value at a time. The agent can be found at http://www.pwcglobal.com/gx/eng/ins-sol/online-sol/edgarscan. Nelson et al. (2000) have proposed an auditing system (FRAANK) based on an agent that retrieves financial information from the EDGAR database.

One popular data mining technique for quantitative data analysis is the *self-organizing map (SOM)* (Kohonen 1997). The SOM has been used for a variety of tasks relating to financial analysis, for example, credit analysis (Martín del-Brío and Serrano-Cinca 1993; Back et al. 1995; Serrano-Cinca 1996; Kiviluoto 1998; Tan et al. 2002), financial benchmarking (Back et al. 1998; Karlsson et al. 2001; Eklund et al. 2002), and macro level economic environment analysis (Kaski and Kohonen 1996). Tan et al. (2002) studied the rating process using Self-organizing maps for clustering and

visualizing the financial ratios. Lavrenko et al. (2000), Back et al. (2001), and Kloptchenko et al. (2002) have combined quantitative and qualitative financial data using quantitative and qualitative clustering techniques for knowledge discovery.

3 THE CONCEPTUAL MODEL

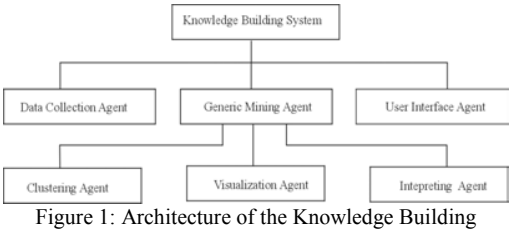


Figure 1: Architecture of the Knowledge Building System.

The proposed conceptual model of the knowledge-building system is depicted in Figure 1. It consists of six agents, i.e. *the Data Collection Agent, the Generic Mining Agent, the User Interface Agent, the Clustering Agent, Visualization Agent* and *the Interpreting Agent*. Each agent carries out its own functions and uses information provided by other agents connected to it. These agents handle three main activities (that are provided by three autonomous agents): data collection and storage (Data Collection Agent), searching for hidden patterns (Generic Mining Agent), and user-interface design (User Interface Agent).

The *Data Collection Agent* is intended to collect, assemble, and sort the quantitative and qualitative data from various Internet resources, such as Bloomberg, Reuters, Wall Street Journal, MSNBC, and individual companies' web sites. These data consist of, for example, market updates, quotes, financial reports, market reports, etc.

The *User Interface Agent* is intended to be responsible for providing the communication channel between the system and the human user that chooses the goal for the system. It should offer the choice of a number of tasks defined by the user in their setup of the system. For example, two possible applications are financial benchmarking and credit rating. These tasks are defined by the data (numeric and textual) included, as well as by the importance placed on each piece of data (for example, the importance of a particular financial ratio). In short, the agent should present the system options, receive the user input commands, and show the final results after it has interacted with the other agents.

The *Generic Mining Agent* is intended to include at least three activities in data processing (see Figure 1.): clustering of the data, visualization of the intermediary results of the previous process, and interpretation of the final results. The clustering techniques are instance dependent, in the sense that we can apply different clustering algorithms when performing data and text mining. We have three agents for the three distinct steps in data processing: the *Clustering Agent, the Visualization Agent*, and the *Interpretation Agent*.

Depending on what mining techniques and data are used, there are two main instances of the Generic Mining Agent: *Data Mining Agent* (Figure 2.) and *Text Mining Agent* (Figure 3.). We see the Generic Mining Agent as a generic class (in programming language understanding), which does not exist physically, but rather is an abstract class that is implemented via its instances. A distinction between

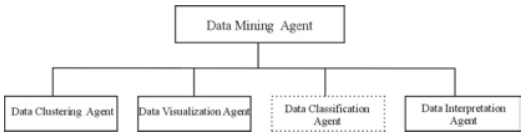


Figure 2: Data Mining Instance of the Generic Mining Agent.

the two instances of the Generic Mining Agent is based on the types of data they mine: Data Mining Agent (for processing numeric data) and Text Mining Agent (for processing text data).

In addition to the activities that are common for both the Data and Text Mining Agents, there are other activities that can be implemented, for example, constructing classification models in the case of the Data Mining Agent and information summarization for the Text Mining Agent. Two new agents can perform these two different activities: the *Data Classification Agent* (see Figure 2, dot-line rectangle) and the *Summarization Agent* (see Figure 3, dot-line rectangle).

The *Knowledge Building System* aims at creating new knowledge by consolidating the obtained new information from the Data Mining and Text Mining Agents. The Knowledge Building System will behave reactively to the goal of the system.

The *Data Mining Agent* would be responsible for numeric data processing and pattern discovery. The Data Mining Agent should provide the Knowledge Building System with the cluster that a company (or other data, depending upon the intended goal) belongs to, as well as the characteristics of the clusters (high profitability, low solvency, etc.), i.e. the results of the entire clustering. The *Data Clustering Agent* should calculate the chosen financial ratios for the chosen companies, standardize the data, and cluster them using self-organizing maps. Finally, the *Data Visualization Agent* visualizes the results.

After we visualize the map clusters provided by the Data Clustering Agent we could use the *Data Classification Agent* that creates a decision tree and/or a multinomial logistic regression model for classifying new financial data (Costea and Eklund 2003). The Data Classification Agent might also use other classifiers. Among these, the agent should use the model that achieves the highest accuracy in training and the best prediction performance.

Then, using all the information from the previous agents, combined with knowledge from other agents in the system, the *Data Interpretation Agent* would attempt to explain the findings. For example, in quantitative clustering, it is important to find explanations for a particular event, such as decreased profitability. This type of information can

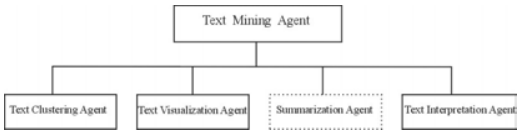


Figure 3: Text Mining Instance of the Generic Mining Agent.

be found in the textual part of the annual report.

The *Text Mining Agent* is intended to be responsible for processing textual information, and choosing the essential indications in it. It could use the *Summarization Agent* that deals with domain information, creating news summaries for any chosen company, or general market information for any chosen time period, and reports it to the user. Then, the *Text Clustering Agent* would perform financial statement clustering by using the prototype-matching methodology (Visa et al. 2002; Back et al. 2001) and reports which financial reports are close in meaning to each other. The *Text Visualization Agent* would present a visual U-matrix map with cluster representation and labels of the companies, which are clustered according to the similarity of their financial statements. The *Text Interpretation Agent* would have the same functionality as the Data Interpretation Agent, the difference being the type of data that is processed.

The Knowledge Building System combines the information from the Data and Text Mining Agents, i.e. it reports to the user how well the chosen company is performing in light of the chosen task, what level of performance the company displays in comparison with other companies in the analysis (clusters), and explains why (text summaries and clusters). The outputs of the two "instance" agents (Data and Text Mining Agents) can be validated one against the other, and the Knowledge Building System can do this automatically, alarming the user if the results are not convergent.

4 METHODS USED BY THE AGENTS

Our agents use several specific data mining techniques for clustering, visualization, and classification of quantitative and qualitative data.

We have used the SOM for clustering the quantitative data. The SOM is an unsupervised neural network for exploratory data analysis. The SOM takes multidimensional numeric data and clusters them on a two-dimensional topological map. Kiang and Kumar (2001) made a comparison between self-organizing maps and factor analysis and K-means clustering. The authors compared the tool's performances on simulated data, with known underlying factor and cluster structures. The results of the study indicate that self-organizing maps can be a robust alternative to traditional clustering methods.

Once trained SOM models are created, the problem of dealing with new data arises. Instead of time consuming retraining, a different method was proposed in Costea and Eklund (2003). The authors suggest a two-level methodology including initial clustering using SOM, and decision tree or multinomial logistic regression classification models trained on the original SOM model. This way the user is able to deal with new data without retraining maps. We have compared the two classification techniques in terms of their accuracy rates and class predictions and reached the conclusion that choosing among possible classifiers is problem dependent. We can extend the number of variables used for training the SOM maps, since the algorithm does not have restrictions from this point of view. Conversely, this methodology can be used as an alternative way of assessing the creditworthiness of companies as opposed to that provided by, say, Standard & Poor's (Tan et al. 2002).

We have tested the use of the prototype-matching approach for text clustering. This method is based on textual collection processing on word and sentence level processing (Visa et al. 2002; Toivonen et al. 2001). The prototype is a document, or a specific part of it, which is of interest to a particular user. A prototype is matched with an existing text collection to obtain a cluster of semantically similar documents. The methodology is based on text preprocessing, and word and sentence level text encoding and histogram creation.

The text summarization algorithm should extract the most relevant sentences from one or multiple documents with regard to a query. Therefore, we propose the use of a text clustering algorithm (e.g. prototype-matching or bisect k-means) for organizing one or more relevant documents into a

tight cluster, and a feature extraction algorithm (e.g. occurrence of cue words, frequent words and proper nouns, position of the sentence with them in the text, sentence length, etc.) and classification algorithm (e.g. Naïve-Bayes classifier, C4.5) for extracting relevant sentences in the relevant documents. The combination of the mentioned techniques requires thorough study for successful summarization. We realize that straightforward word matching is not enough for effective detection of similarity between text pieces.

5 LIMITATIONS AND DIFFICULTIES

There are, of course, a number of problems associated with building a system of this complexity based on data that are freely presented on the Internet. We can divide system limitations in, at least, two categories: limitations that are specific for each individual agent and limitations regarding the integration of different agents. The data collection agent's ability to automatically retrieve financial data from Internet resources is severely hampered by a lack of standard for online financial reporting. A possible future solution to this problem is XBRL (eXtensible Business Reporting Language). XBRL is an XML (eXtensible Markup Language) standard created specifically to address the problem of online business reporting. Currently, there is no way for collection agents to automatically retrieve financial data from diverse web sites without specifically coding the agent for a specific page. (Debreceny and Gray 2001)

Another type of limitation of the system is due to the limitations of the deployed DM and TM techniques (Data and Text Mining Agents). For example, with all its advantages over standard clustering techniques, the SOM has one major drawback: verification of the achieved clustering results. This issue is addressed in Wang (2001), in which the author proposes a number of techniques for verifying clustering results. Similar techniques will have to be used in the system we are proposing.

Text mining techniques have a number of disadvantages due to the highly dimensional structure of text. Two textual pieces can often be nearest neighbors in terms of using similar vocabulary, without actually belonging to the same semantic class. Prototype-matching clustering is an exploratory technique that possesses some difficulties with determination of the clusters, and with their comparison with quantitative clustering. Although, theoretically, text implies richer information about an event than a numerical snapshot of the fact does, this is difficult to verify. Even having excellent text mining techniques on hand that could mine the indications of future financial performances of the company, those indications can be easily concealed by smart word choice and sentence construction.

Also, as was illustrated by the Enron and WorldCom scandals, the financial information presented in annual reports is not always reliable. Of course, if this incorrect information is inserted into our system, the results will also be incorrect. Moreover, there might be unintentional mistakes in the data. Therefore, some kind of error detection and handling capabilities should be built into the system. This is also required by the actual definition of KDD, which includes data cleaning and error detection (Fayyad et al. 1996).

The integration limitations are closely related to the individual agents limitations, e.g.: because of the lack of standard of financial information available on the Internet, the Data Collection Agent might not be able to provide the data that we need to address a specific problem, which makes its integration with the Knowledge Building System extremely difficult.

6 CONCLUSIONS AND FUTURE WORK

In the current research paper we introduced a conceptual model of a system based on different data/text mining methods for knowledge building from freely available data distributed on the web. The system aims to automatically perform different tasks such as data collection, financial benchmarking, assessing creditworthiness of companies, and finding hidden patterns in unordered and unstructured text data. The system uses two types of data (numeric and textual) and data processing techniques (data and text mining techniques) to support and explain the phenomena.

In this paper we discussed the operational facilities of the proposed system that will be accomplished by text and data mining methods. The system knowledge base, system external interface and limitations should be researched further.

As further research problems we could investigate new methods for collecting the input information for the Data and Text Mining Agents (that is improve the Data Collection Agent), extend the conceptual model to include subagents that perform tasks for their "parent" agents: Data Cleaning Agent, Data Aggregator Agent (aggregates information find on different web sources and presents this information further to Data Collection Agent).

REFERENCES

Back, B., G. Oosterom, K. Sere, m. van Wezel, 1995. Intelligent Information Systems within Business:Bankruptcy Predictions Using Neural Networks. In *The 3rd European Conference on Information Systems (ECIS'95)*, Athens, Greece.

Back, B., K. Sere, H. Vanharanta, 1998. Managing complexity in large data bases using self-orginizing maps. In *Accounting Management and Information Technologies* 8(4): 191-210.

Back, B., J. Toivonen, H. Vanharanta, A. Visa, 2001. Comparing numerical data and text information from annual reports using self-orginizing maps. In *International Journal of Accounting Information Systems* 2: 249-269.

Costea, A. and T. Eklund, 2003. A Two-Level Approach to Making Class Predictions. In *The 36th Hawaii International Conference on Systems Sciences (HICSS-36)*, Hawaii, USA, IEEE.

Debreceny, R. and G. L. Gray, 2001. The production and use of semantically rich accounting reports on the Internet: XML and XBRL. In *International Journal of Accounting Information Systems* 2(1): 47-74.

Dorre, J., Gerstl, P., and R. Seiffert, 1999, Text Mining: Finding Nuggets in Mountains of Textual Data, In *Proceedings of th 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, USA

Eklund, T., B. Back, H. Vanharanta, A. Visa, 2002. Assessing the Feasibility of Self-Organizing Maps for Data Mining Financial Information. In *The Xth European Conference on Information Systems (ECIS 2002)*, Gdansk, Poland.

Fayyad, U., G. Piatetsky-Shapiro, P. Smythe, 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *The Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, AAAI Press.

Karlsson, J., B. Back, H. Vanharanta, A. Visa, 2001. *Financial Benchmarking of Telecommunications Companies*. TUCS Technical Report No. 395, Turku Centre for Computer Science. Turku.

Kaski, S. and T. Kohonen, 1996. Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World. In *The Third International Conference on Neural Networks in the Capital Markets*, World Scientific.

Kiang, M. and A. Kumar, 2001. An Evaluation of Self-Organizing Map Networks as a Robust Alternative to Factor Analysis in Data Mining Applications. In *Information Systems Research* 12(2): 34-41.

Kiviluoto, K., 1998. Predicting bankruptcies with the self-organizing map. In *Neurocomputing* 21(1-3): 191-201.

Kloptchenko A., T. Eklund., B. Back, J. Karlsson, H. Vanharanta, A. Visa, 2002. Combining Data and Text Mining Techniques for Analyzing Financial Reports. In *The 8th Americas Conference on Information Systems (AMCIS2002)*, Dallas, USA.

Kohonen, T., 1997. *Self-Organizing Maps*, Springer-Verlag. Leipzig, 2nd edition.

Lavrenko, V., M. Schmill, D. Lawrie, P. Ogilvie, 2000. Mining of Concurrent Text and Time Series. In *Text Mining Workshop of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA. ACM.

Lesser, V., 1995. Multiagent Systems: An emerging subdiscipline of AI. In *ACM Computing Surveys* 27(3): 340-342.

Liu, S., 1998. Business Environment Scanner for Senior Managers: Towards Active Executive Support with Intelligent Agents. In *Expert Systems with Applications* 15: 111-121.

Martín-del-Brío, B. and. C. Serrano-Cinca, 1993. Self-organizing Neural Networks for the Analysis and Representation of Data: Some Financial Cases. In *Neural Computing and Applications* 1: 193-206.

Nelson, K. M., A. Kogan, R. P. Srivastava, M. A. Vasarhelyi, H. Lu, 2000. Virtual auditing agents: the EDGAR Agent challenge. In *Decision Support Systems* 28(3): 241-253.

Serrano-Cinca, C., 1996. Self organizing neural networks for financial diagnosis. In *Decision Support Systems* 17(3): 227-238.

Tan, R., J. den Berg, W. den Bergh, 2002. Credit Rating Classification Using Self-Organizing Maps. In *Neural Networks in Business: Techniques and Applications*, ed. by K. Smith and J. Gupta, Idea Group Publishing. Hershey.

Toivonen, J., A. Visa, H. Vanharanta, B. Back, 2001. Validation of Text Clustering Based on Document Contents. In *Machine Learning and Data Mining in Pattern Recognition (MLDM 2001)*, Leipzig, Germany. Springer-Verlag.

Visa, A., J. Toivonen, B. Back, H. Vanharanta, 2002. Contents Matching Defined by Prototypes: Methodology Verification with Books of the Bible. In *Journal of Management Information Systems* 18(4): 87-100.

Wang, S., 2001. Cluster Analysis Using a Validated Self-Organizing Method: Cases of Problem Identification. In *International Journal of Intelligent Systems in Accounting, Finance and Management* 10(2): 127-138.

Wang, H., J. Mylopoulos, S. Liao, 2002. Intelligent Agents and Financial Risk Monitoring Systems. In *Communications of the ACM* 45(3): 83-88.

Wuthrich, B., D. Permunetilleke, S. Leung, V. Cho, J. Zhang, W. Lam, 1998. Daily Prediction of Major Stock Indices from textual WWW data. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98),* New York, NY, USA, AAAI Press.

Research Paper 6

Kloptchenko A., Magnusson C., Back B., Visa A., Vanharanta H**,** **"**Mining Textual Contents Of Quarterly Reports", presented at the *XXVI Annual Congress of European Accounting Association*, 2-4 April, 2003, Seville, Spain – TUCS Technical report 515, isbn: 952-12-1138-5. Accepted for publication in the International Journal of Digital Accounting Research (IJDAR)

Mining Textual Contents of Quarterly Reports

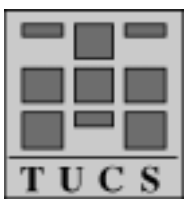
Antonina Kloptchenko
Turku Centre for Computer Science, Institute for Advanced Management System Research, Åbo Akademi University, Lemminkäisenkatu 14, FIN-20520 Turku, Finland

Camilla Magnusson
Department of General Linguistics, University of Helsinki, Helsinki, Finland

Barbro Back
Turku Centre for Computer Science, Institute for Advanced Management System Research, Åbo Akademi University, Lemminkäisenkatu 14, FIN-20520 Turku, Finland

Ari Visa
Tampere University of Technology, Department of Information Technology, Tampere, Finland

Hannu Vanharanta
Pori School of Technology and Economics, Pori, Finland



Turku Centre for Computer Science
TUCS Technical Report No 515
May 2002
ISBN 952-12-1138-5
ISSN 1239-1891

Abstract

A huge amount of electronic information concerning companies' financial performance is available in organizational databases and on the Internet today. Numeric financial information is important for many stakeholders and has been extensively analyzed for many decades with advanced computational methods. Textual financial reports and news contain not only the factual information about events, but also explain why they have happened. Exploiting finance and business related textual information in addition to numeric financial information could potentially increase the quality of decision-making. Researchers are searching for effective and computationally fairly simple tools that would be able to handle sophisticated text-related tasks without thorough linguistic preprogramming.

The message, stylistic focus, language and readability of financial reports are good indicators of the perspectives and developments of any company. These indicators can guide companies' decision makers to more efficient actions in the dynamic business environment. Although, financial experts and experienced readers can detect them and make more precise financial decisions, the manual analysis of textual reports require a lot of time, and time is a costly asset in a financial community. Text Mining methods aim to offer opportunities for automatic analyzing and discovering previously unknown patterns in text. Therefore, the less expensive computer-based solutions for mining financial texts for hidden indications of companies' perspectives are needed.

In this paper, we have studied the language and contents of quarterly reports using linguistic and text mining methods. We have compared the results obtained from linguistic analysis of quarterly reports by means of collocational networks and the results obtained from automatic text mining analysis of quarterly report by means of prototype matching clustering. Our objective was to study how well the computer-aided text-mining tool can perceive the content of quarterly reports in comparison with linguistically motivated collocational networks that outline the most frequent and significant words in the texts. The purpose was to see how meaningful the prototype matching clustering is from the perspective of collocational networks linguistic analysis. We performed the study on the quarterly reports from three leading companies in the telecommunications sector, Motorola, Ericsson and Nokia, for the years 2000-2001.

Our results are somewhat controversial. Some of the reports from the companies have as their closest matches the reports with similar collocational networks and some do not have.

Keywords: text mining, annual reports, prototype-matching clustering, collocational networks, collocations

TUCS Laboratory
Data Mining and Knowledge Management Laboratory

1. Introduction

A huge amount of electronic information concerning companies' financial performance is available in organizational databases and on the Internet today. Numeric financial information is important for many stakeholders and has been extensively analyzed for many decades with advanced computational methods. Textual financial reports and news contain not only the factual information about events, but also explain why they have happened. Exploiting finance and business related textual information in addition to numeric financial information should increase the quality of decision-making. Constantly updated text collections have grown so large that there is not enough time to read and analyze them manually. Additionally, the ambiguous structure of texts makes their analysis rather complicated. Researchers are searching for elegant and computationally fairly simple tools that would be able to handle sophisticated text-related tasks without thorough linguistic preprogramming.

The message, stylistic focus, language and readability of financial reports are good indications about the perspectives and developments of any company. These indications can guide companies' decision makers to more efficient acts on the market. Although, financial experts and experienced readers can detect those indications and make more precise financial decisions, the manual analysis of textual reports requires a lot of time, and time is a costly asset in a financial community. Text Mining methods aim to offer an automatic way for analyzing and discovering previously unknown patterns in text Hearst (1999). Therefore, less expensive computer-based solutions for mining financial texts for hidden indications of companies' perspectives are needed.

The most typical company report is without doubt the annual report, which has received a certain amount of attention from linguists and financial specialists. Annual reports, while being important documents to stockholders and financial communities are controversial. They generate disagreement regarding audience, objectives and credibility Thomas (1997). As a genre, annual reports resemble quarterly reports closely. The same writers produce quarterly and annual reports for the same readers within the same community. The reports have a similar structure, conventions, basic functions and communicative purposes but the time spans are different. The study of the linguistic contents of quarterly reports has nevertheless been overlooked in favour of the study of the language of annual reports. In the short-term perspective quarterly reports are important means for companies in appraising past performance and projecting future opportunities to the readers, who primarily consist of investors and analysts. The beginning of every report, known as the manager's/president letter/message to stockholders, contains management's strategy, summary of the financial performance for the year and an attempt to put in perspective the success or failure of the various initiatives of the company Thomas (1997).

In this paper, we study the language and contents of quarterly reports using linguistic and text mining methods. We compare the results obtained from linguistic analysis of quarterly reports by means of collocational networks and the results obtained from automatic text mining by means of prototype matching clustering. Our objective is study how well a computer-aided text-mining tool can perceive the content of quarterly reports in comparison with linguistically motivated collocational networks, which outline the most frequent and significant words in the texts. The purpose is to investigate how meaningful the prototype matching clustering is from the perspective of

collocational networks linguistic analysis. We perform the study on the quarterly reports from three leading companies in the telecommunications sector, Motorola, Ericsson and Nokia, for the years 2000-2001.

The rest of the paper is organized as follows. We start our explanation by giving a short overview of studies relating to analyzing the language of annual reports, as the closest alternative to quarterly reports. Then, we provide a description of the prototype-matching method that we have used for automatic text mining analysis. Next, we introduce collocational networks as a method we used for linguistic analysis. We relate our results from text mining and linguistic analysis by reviewing an example using both methods for analysis of telecommunications companies' quarterly reports. We analyze the results achieved by the two methods and compare them to each other. We conclude with some suggestions for further research.

2. Related Studies

Since the language of quarterly reports has not been studied, our literature review is based on a broad body of literature on the language of annual reports, conducted both within linguistics and business communication studies. A common feature for the studies mentioned here is that they have only concentrated on one part of the reports, the manager's/president letter/message to the shareholders. Our study is different in this respect, as we focus on the whole body of the reports.

Thomas (1997) concentrated on transitivity, thematic structure, context, cohesion and condensation in the language used in the reports. Thomas studied the annual reports of a machine tool manufacturer during a period, which began with prosperity and ended with severe losses. During the time frame of the analysis, the structure of the language used in the reports had changed. According to Thomas' study, an increase in the use of passive constructions can be seen as the profits decrease. There is also an increase in verbs that present the actor (i.e. the company) as "being" rather than as "doing". This indicates that management is trying to present itself as a victim of unfortunate circumstances. This creates an impression of objectivity for the reader, as if the management was presenting plain facts on recent events. On the other hand, when the company was making more profit, it presented itself as aggressive and forward moving through the use of active voice and verbs with both an actor and a goal. A close look at the language structure in the letters to stockholders made by Thomas (1997) showed that the structure of the financial reports might reveal some things that the company may not wish to announce directly to its outside audience. Another conclusion of this study was the confirmation of the Pollyanna Hypothesis[1].

Kendal (1993) introduced the concept of drama when she noticed a similar opposition between the actions of the company and circumstances created by nonhuman agents. Kendall has classified the words and phrases describing actors and objects in the drama into two groups, *God terms* and *Devil terms*. Some examples of god terms are growth, increased sales and competitive position. These words represent concepts that

[1] By studying negative and positive words in annual reports, Hildebrandt and Snyder (1981) induced Pollyanna Hypothesis (Hildebrandt H.H. and Snyder R., (1981), The Pollyanna hypothesis in business in business writing:. Initial results, suggestions for research. The Journal of Business Communication, **18** (1), 5-15). It states that regardless of the financial state of the company, the language in the annual letters will be predominantly positive.

are unquestionably good in the eyes of the company. Devil terms, on the other hand, are terms like losses, decline in sales and regulations.

Other studies have been made with a focus on the relationship between the readability of the annual reports and the financial performance of a company (Subramanian et al. 1993). The annual reports of the companies that performed well were easier to read than those that originated from companies that did not perform well. Studies have also shown that writers of annual reports see the message they put in the report as their personal representation (Winsor 1993). The annual reports are not only the best possible description of a company, but are also a description of a company's managerial priorities. Thus, the communication strategies hidden in annual reports differ in terms of the subjects emphasized when the company's performance worsens (Kohut and Segars 1992). After performing computer-aided content analysis of more than four hundred president's letter to shareholders and examining empirical linkages between themes in annual reports and companies' performances, Osborn et al. (2001) conclude that the text in annual reports reflects the strategic thinking of the management of a company.

Attempts to semi-automatically analyze a company's performance by examining quantitative and qualitative parts from annual reports have been done by (Back et al. (2001) and Kloptchenko et al. (2002). Back et al. (2001) indicated that there are differences in qualitative and quantitative data clustering results due to a slight tendency to exaggerate the performance in the text. Kloptchenko et al. (2002) attempted to explain this tendency using quantitative analysis by means of self-organizing maps for financial ratio clustering, and qualitative analysis by means of the prototype-matching for quarterly report text clustering. In both studies the researchers noticed that the combination of two mining techniques for two different types of data describing the same phenomena could bring additional knowledge to a decision maker. While annual/quarterly reports explicitly state information about a company's past performance, they also contain some indications of future performance, i.e. the tables with financial numbers indicate how well a company has performed, while the linguistic structure and written style of the text may tell what a company intends to do. The study has shown that the sophisticated semi-automatic analysis of the style and content of the financial reports help to reveal insiders' moods and anticipations about the future performance of their company.

3 Methodology

Our methodology section builds on two different methods with the intention of performing two types of text analysis. We use the prototype-matching method proposed by Visa et al. (2000) for computer-aided text mining, and collocational networks proposed by Magnusson (2002) for linguistic analysis.

3.1 Text Mining with Prototype-matching

The prototype is a document or a part of it, which is of a particular interest to a particular user. This prototype is matched with an existing text collection in order to

obtain a cluster of semantically similar documents. The methodology is based on textual collection preprocessing, i.e. word and sentence level processing. We transform every word into a number, taking into account word length in ASCII symbols, and the ASCII value of every character in a word. We create a common word histogram for the entire text collection and choose a suitable Weibull cumulative distribution. Each word after quantization is presented as a bin number and the values of the best-fitted Weibull distribution. We have performed text quantization on the word level, by creating a common word histogram for the entire text collection. The most common words in the text gain a dense resolution in the histogram bins.

We perform similar procedures for converting every word into a bin number on the sentence level, in order to present the whole sentence as a vector. Hereafter, we consider the Fourier transformed encoded sentences as input vectors and choose a cumulative distribution the same way as on the word level. We divide the distribution into logarithmically equal bins, the number of which is equal to the number of all sentences in the text collection. The best-fitted Weibull distribution is found based on the cumulative distribution of the coded sentences and their scalar quantization to equally distributed bins.

In the next phase, we construct individual sentence and word histograms for each document in the collection according to the documents' word and sentence code numbers and the corresponding value of quantization Toivonen et al. (2001). Having sentence and word level histograms allows us to compare documents to each other simply by calculating the Euclidian distances between their histograms. The smallest Euclidian distance between word histograms indicates a common vocabulary of the reports. The smallest Euclidian distance between sentence histograms indicates similarities in written style and/or content of the reports Visa et al. (2001).

3.2 Linguistic Analysis with Collocational Networks

In order to visualise the central concepts and their connections within a quarterly report we used a method originally devised by Williams (1998). In his study, Williams uses the network as a way of exploring the language of science in order to create specialised dictionaries. The main concept in this method is a collocational network. For the purposes of this study, *collocation* was interpreted by Sinclair (1991)simply as "the occurrence of two or more words within a short space of each other in a text". Collocational networks are visual constructions of collocations forming the unique frame of reference for any "word" within a given sub language (Furnas 1987). Collocational networks outline the central concepts in a text, and their textual connections to each other.

It should be noted that the contents of each report were analysed separately. This means that pairs of words which are referred to as collocations in this study are patterns which occur within a single text, and therefore cannot be considered to be typical for English or even business English.

An important factor in this method is the concept of *significant collocation*. Significant collocation takes place when two or more words occur together more frequently than would be expected by coincidence. Following Williams (1998), significant collocation is measured using the *Mutual Information* or *MI score*. The MI score, an information theoretic concept introduced in linguistics by Church and Hanks

(1990), compares the frequency of co-occurrence of node and collocate with the frequency of their occurrence independently of each other. A more thorough description and evaluation of the MI score can be found in Stubbs (1995). It has the same value regardless of which word of a pair is the collocate and which is the node. The MI score is also sensitive to changes in the absolute number of collocates, when the relative proportion of joint occurrences compared to independent occurrences remains the same. In these cases it works "counter-intuitively": decreasing as the absolute number of collocates increases. This means that two words which always occur together get a higher MI score if they occur only once than if they occur more frequently. Because of these drawbacks an alternative to the MI score might be considered for further development of this.

Collocational networks give us an opportunity to examine which concepts are emphasised by the company in a particular report and how these concepts are reflected through the words that constitute the nodes of the network. We can examine which concepts are most frequently linked to each other, by revealing which words regularly appear within a close proximity to each other. This method does not always bring out combinations of words that are perceived by speakers of the language to belong together as phrases or compound words, such as *balance* and *sheet*, unless they occur very frequently in the text. Because the aim of this study is not to find collocations that are typical for business language in general, but central collocations for analysed reports, this limitation is not considered to be a problem.

4 Results

4.1 Text Mining of Quarterly Reports

We encoded every word from the reports, and constructed a common word histogram as the first step of text clustering. Then, we encoded each sentence from the reports and constructed a common sentence histogram, and a unique sentence histogram for every report. In order to obtain the clusters of the closest reports, every quarterly report must be treated as a prototype and matched against the entire report collection. By calculating the Euclidian distance between reports' sentence histograms we can compare all of the quarterly reports in our data collection. For example, for the Ericsson report from 2000, quarter 1, the closest report by content on the sentence level is from Nokia, 2000, quarter 1. The second closest is the report from Nokia, 2000, quarter 3. This means that the Nokia reports from 2000, quarters 1 and 3 and the Ericsson report from 2000, quarter 1 have similarities in sentence construction and word choice, which constitutes the language structure and written style.

The results from text mining of quarterly reports for Nokia, Ericsson and Motorola are presented in Table 1. Each column in Table 1 contains the report-prototype in the header and the four closest matches to it. Quarter names and proper names, e.g. Nokia, Motorola or Ericsson, did not determine the clusters. Thus, the closest matches to Ericsson prototype-reports are not always other reports from Ericsson. The same is true for Nokia and Motorola reports because word choice has a smaller impact on the formed clusters than the sentence construction. Therefore, we

consider the text mining results from sentence level, attempting to justify on what kind of linguistic basis the computer-aided tool chooses the closest matches.

Table 1. The closest Matches to every report in the collection (Sentence level)

Ericsson2000Q1	Ericsson2000Q2	Ericsson2000Q3	Ericsson2000Q4	Ericsson2001Q1	Ericsson2001Q2	Ericsson2001Q3
Nokia2000Q1	Ericsson2000Q3	Ericsson2000Q4	Ericsson2000Q3	Ericsson2001Q2	Nokia2001Q3	Ericsson2001Q1
Nokia2000Q3	Nokia2000Q2	Motorola2001Q3	Motorola2001Q2	Ericsson2001Q3	Ericsson2001Q1	Ericsson2001Q2
Motorola2001Q3	Ericsson2000Q1	Ericsson2000Q2	Motorola2001Q3	Nokia2001Q3	Ericsson2001Q3	Nokia2001Q3
Motorola2001Q2	Ericsson2000Q4	Ericsson2000Q1	Nokia2000Q1	Motorola2001Q3	Nokia2001Q1	Nokia2001Q2

Motorola2000Q2	Motorola2000Q3	Motorola2000Q4	Motorola2001Q1	Motorola2001Q2	Motorola2001Q3
Motorola2001Q3	Ericsson2001Q2	Motorola2001Q3	Motorola2000Q2	Ericsson2000Q4	Motorola2000Q2
Motorola2001Q2	Nokia2000Q2	Nokia2000Q4	Motorola2001Q2	Motorola2001Q3	Nokia2000Q1
Nokia2000Q2	Nokia2000Q1	Nokia2000Q1	Nokia2001Q2	Motorola2000Q2	Nokia2001Q3
Nokia2000Q4	Nokia2001Q3	Ericsson2001Q2	Nokia2001Q3	Ericsson2000Q1	Ericsson2000Q1

Nokia2000Q1	Nokia2000Q2	Nokia2000Q3	Nokia2000Q4	Nokia2001Q1	Nokia2001Q2	Nokia2001Q3
Ericsson2000Q1	Nokia2001Q2	Nokia2001Q3	Nokia2001Q1	Ericsson2000Q1	Nokia2000Q2	Ericsson2001Q2
Motorola2001Q3	Motorola2001Q3	Ericsson2000Q1	Ericsson2000Q1	Nokia2000Q4	Nokia2001Q3	Nokia2000Q3
Nokia2000Q2	Nokia2000Q1	Motorola2001Q2	Motorola2001Q3	Ericsson2001Q2	Motorola2000Q2	Motorola2001Q3
Nokia2000Q3	Motorola2000Q2	Nokia2000Q1	Motorola2000Q2	Nokia2000Q1	Motorola2001Q1	Ericsson2001Q1

Ericsson. In the majority of the cases, the closest matches to Ericsson quarterly reports are the Ericsson and Nokia reports from other periods of time. However, for Ericsson report from quarter 1 (Q1), 2000 the four closest matches are reports from Nokia and Motorola only. The Ericsson report from quarter 3, 2000 has three closest reports from Ericsson from different quarters of year 2000. The most similar report is the one from the next quarter. Furthermore, the report from Ericsson quarter 3, 2000 is the most similar match to the Ericsson report from quarter 4, 2000. Noticeable, Nokia reports disappear from the closest matches for Ericsson, 2000 quarter 3.

Motorola. The majority of the closest matches to Motorola quarterly reports are the Motorola and Nokia reports from different time periods than Motorola prototype-reports. A minority of the closest matches to Motorola quarterly reports are Ericsson reports. The report from Motorola quarter 3, 2001 has fired as the closest one to eleven different reports from Ericsson, Motorola and Nokia, four of those eleven have appeared as the closest ones to Motorola report itself because of the symmetry of Euclidian distance used in the method. The Motorola reports are shorter than Ericsson and Nokia ones.

Nokia. The majority of the closest matches to Nokia quarterly reports are Nokia reports from other time periods. The next closest are the reports from Motorola, and a minority of the closest matches are the reports from Ericsson. The length of Nokia quarterly reports varies significantly, from 2989 words (Nokia Q1, 2000) to 5463 words (Nokia Q4, 2000). Ericsson reports disappear from being among the closest matches to reports from Nokia 2000, quarter 2 and Nokia 2001 quarter 2, but reappears among the closest matches for Nokia report for 2001, quarter 3.

4.2 Collocational Networks for Quarterly Reports

Before the actual analysis could take place we carried out some preliminary measures, i.e. we removed all tables that could easily be separated from the text, and left out some minor tables. This was compensated during the drawing of the networks by leaving out words such as *adjusted*, *operational* and *non-operational*, which occurred in these tables. During the drawing of the networks a number of other words with little relevance for the report as a whole were left out as well. Words left out were low-content words such as prepositions, articles, conjunctions, words referring to the time span of the report (*quarter*, *first*, *second* etc.) as well as words referring to figures or currency.

The initial stage of the analysis consisted of calculating the Mutual Information (MI) score for all words occurring within a span of four words, as recommended by Sinclair (1991) for studying collocations in English. With text sizes of approximately 4000 words, an MI score of 2.00 was found to produce a network of a size suitable for this study. Lowering the score would have brought in words which occur together only occasionally, whereas a higher limit would have produced a network with only the most frequent combinations, leaving out many of the interesting changes which occur among the mid-frequency words.

We approach the collocational networks produced from the quarterly reports from each company in sequence. There are two points of interest in particular where stability or changes can be seen: the structures of the networks and the words they contain. A closer look will be taken at both of these points in the networks created out of each company's quarterly reports.

Ericsson

A brief overview of the collocational networks based on Ericsson's quarterly reports shows that they never exhibit similarity in their architecture. During the period studied here, both the structure and the content of the networks vary considerably. This is also obvious when looking at the text in the reports: during this period the reports undergo several structural changes along with the worsening of Ericsson's financial performance. New headings are introduced and old ones are abandoned or reorganised.

A particularly remarkable change in the networks happens between the third and fourth report for 2000. Structurally, these networks are completely different. There is also a significant difference between the number of lexical items and the lexical items themselves used in the networks. The collocational networks for Ericsson reports from quarter 3 and 4, 2000 are presented in Figure 1 and 2 respectively.

The network for quarter 3, 2000 starts with the most frequent word, *Ericsson*, which is linked to five collocates. One of these collocates, *increased*, is linked to *sales*, which has four other collocates of its own. One of these collocates, *systems*, is linked to *mobile*, which has five more collocates. These linkages mean that the main network for quarter 3, 2000 consists of three parts, connected by collocational pairs. In addition to this, there are several separate collocational pairs and small networks outside the main network.

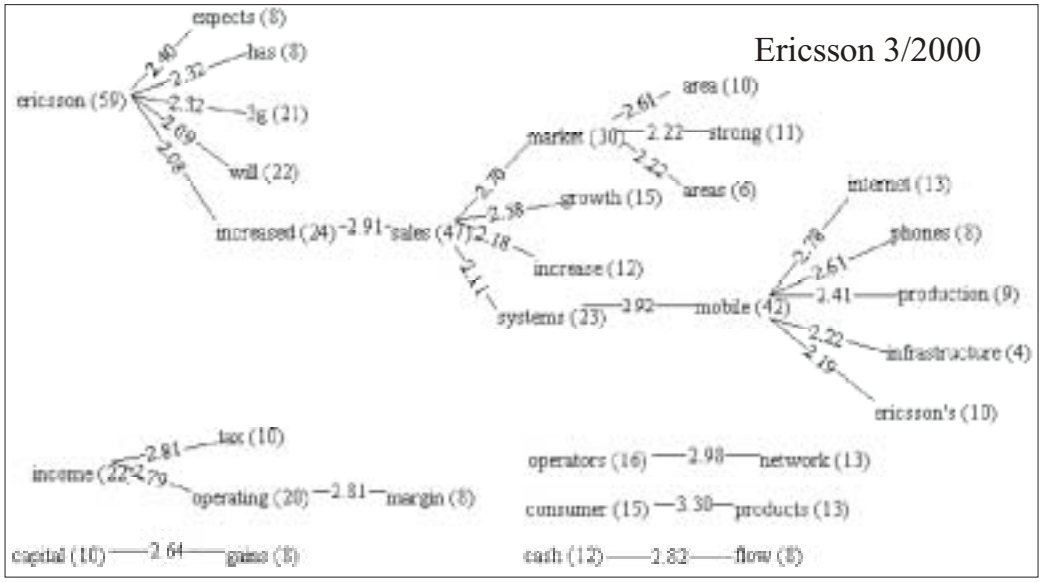


Figure 1. Collocational Network for Ericsson report from quarter 3, 2000

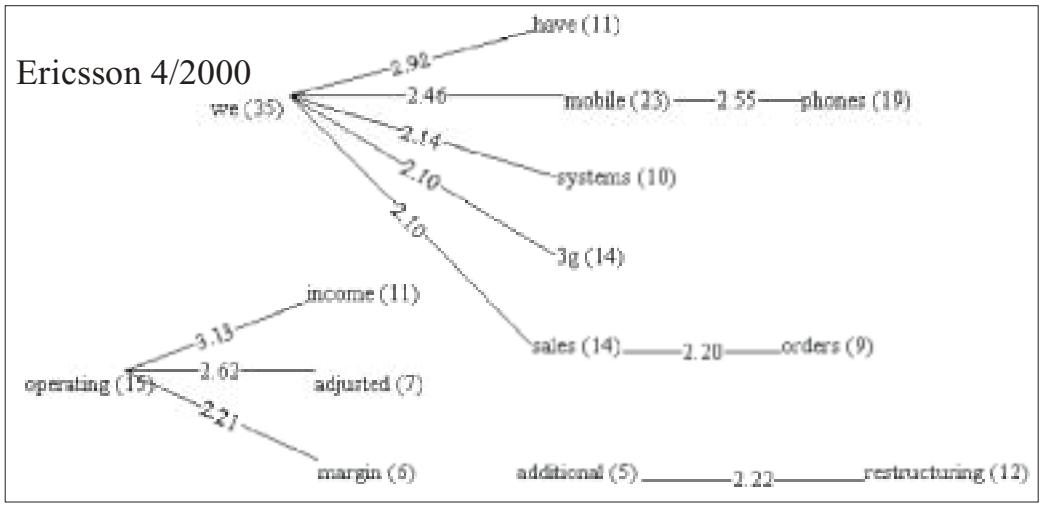


Figure 2. Collocational Network for Ericsson report from quarter 4, 2000

The structure of *network quarter 4, 2000* is very different from the structure of reports from quarter 3, 2000. It consists of a main network attached to the most frequent word, *we*, and a smaller, separate network around *operating*. *We* is a new word in this network, and the most frequent word in network from the report of quarter 3, 2000, while the main word *Ericsson*, has disappeared. Starting from this network, the company now refers to itself using a pronoun instead of the name *Ericsson*. In addition to these two major networks, there is one separate collocational pair, consisting of two new words, *additional* and *restructuring*.

The number of words in the network is much smaller than in the previous network (33 vs. 14), and the structure is much less complex. The most obvious reason for this is

the fact that report quarter 3, 2000 consists of approximately 3600 words, whereas report quarter 4, 2000 consists of approximately 2100 words.

In the following network, quarter 1, 2001, the change continues. This network contains even fewer words than the previous one. Now there is only one word, *expect*, connected to *we,* as opposed to five collocates in the previous network. A new addition is the collocation *efficiency program*, a term bearing obvious negative connotations.
In the last three networks of 2001, more words start to appear and the structures become more complicated. Structurally these networks resemble the networks representing early 2000. Looking at the words they contain, however, they are very different. These networks contain collocations such as *restructuring charges*, *increased borrowing* and *efficiency program*, all of these pointing to a negative development within the company.

Nokia

The collocational networks for Nokia reports from quarter 1 and 2, 2001 are presented in Figure 3 and 4 respectively. The networks are almost identical, containing the name *Nokia* as a central node with links to words referring to the company's business segments such as *Networks* or *Mobile Phones* or general nouns used in business texts such as *sales*, *market* and *growth* .

However, quite a remarkable change can be seen to take place between the first and the second report for 2001. Structurally, networks quarter 1, 2001 and quarter 2, 2001 look quite similar. They both have just one central word, *Nokia*, around which most of the other words are attached. In addition, there are two collocational pairs outside the main structure. Two words, which appear in quarter 2, 2001, marking the change in the networks, are *decline* and *decreased*. Neither of these words appear in the previous network. In the text of the report for quarter 1, 2001 *decline* does not appear and *decreased* only appears twice, thus making the sudden increase to five and sixteen occurrences respectively is quite an eye-catching. At the same time words bearing positive connotations, such as *growth* and *increased* disappear. The connection between these events is made explicit by the fact that *sales*, a word which is linked to *increased* in the first network, is linked to *decline* in the second.

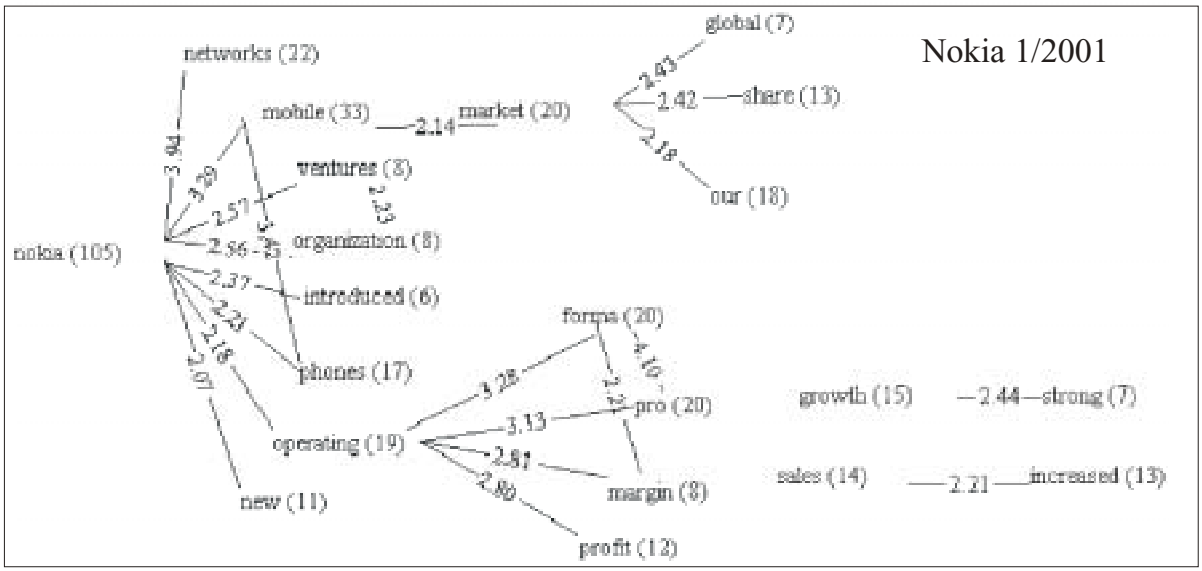


Figure 3. Collocational Network for Nokia report from quarter 1, 2001

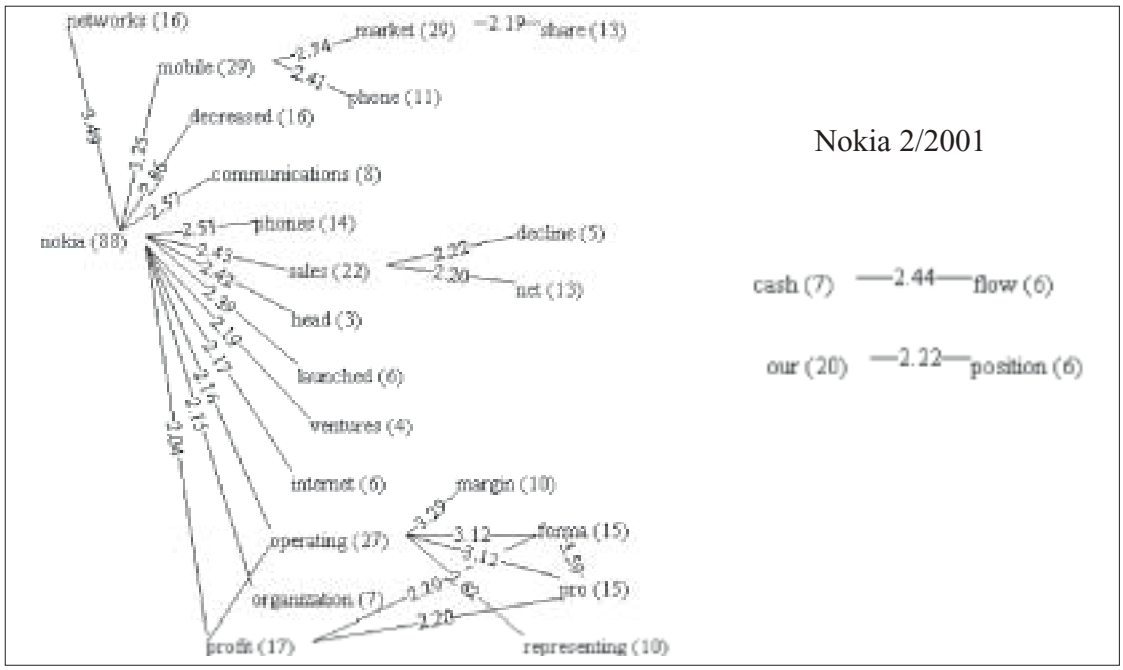


Figure 4. Collocational Network for Nokia report from quarter 2, 2001

Mototrola

The collocational networks for Motorola reports from quarter 4, 2000 and quarter 1, 2001 are presented in Figure 5 and 6. The networks created out of Motorola's reports show less uniformity than the Nokia networks. Still, the networks for the year 2000 resemble each other quite closely. They consist of one main network with the word *sales* as a central node. Linked to it are words like *increased*, *higher*, *orders* and *systems*. Interestingly, the word *lower* appears in these networks.

A budding change can be seen in the first network for 2001. The main network still concentrates around *sales*, but there is also another smaller network around the word *Motorola*, which is the most frequent word in the text and is linked to the collocates *announced* and *new*. It seems as if the company is trying to emphasise the announcements of new innovations. Interestingly, at the same time positive words like *higher* and *increased* have disappeared from the network.

Network 2/2001 looks quite similar. *Motorola* is still the most frequent word, but it is now only linked to *announced*. The word *decline* has also appeared as a collocate to *sales*.

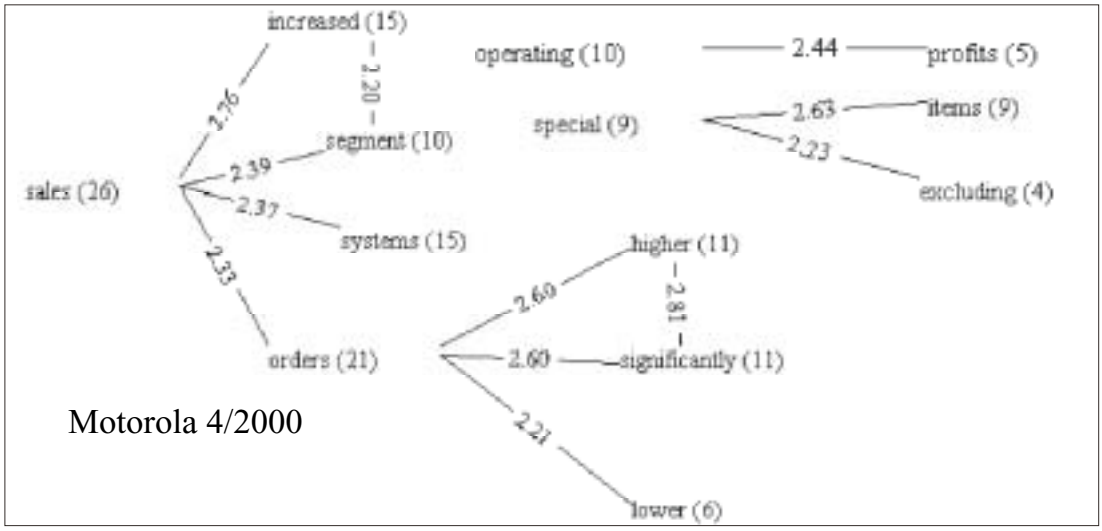


Figure 5. Collocational Network for Motorola report from quarter 4, 2000

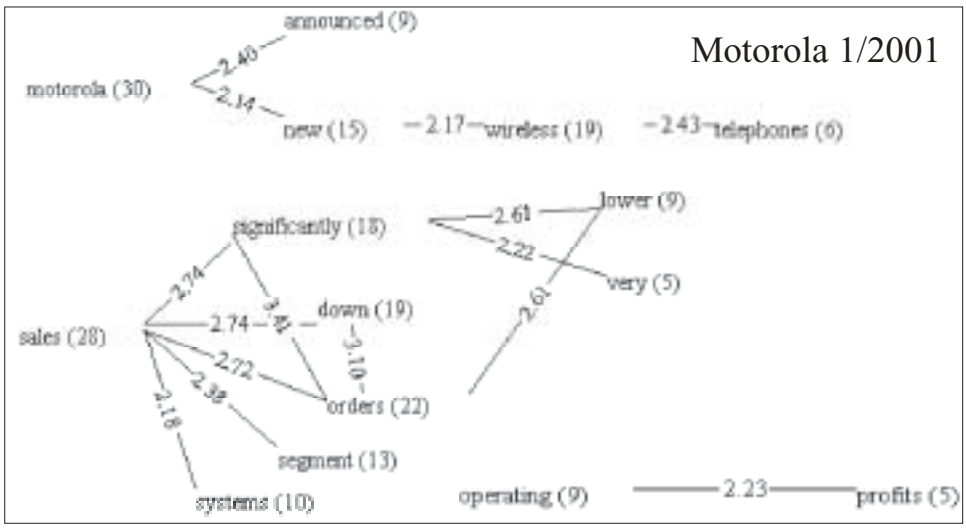


Figure 6. Collocational Network for Motorola report from quarter 1, 2001

In the third network for 2001, illustrated in Figure 7, the changes continue. This network looks very different from the previous ones, as it doesn't contain any structure resembling a network, only pairs of collocations. *Sales*, which was a central node in the previous networks, is now only linked to *segment*. *Motorola* is still linked to *announced*. What makes the contents of this network particularly different from previous networks, is the complete lack of words describing the financial developments of the company, such as *increased*, *decreased*, *higher* and *lower*.

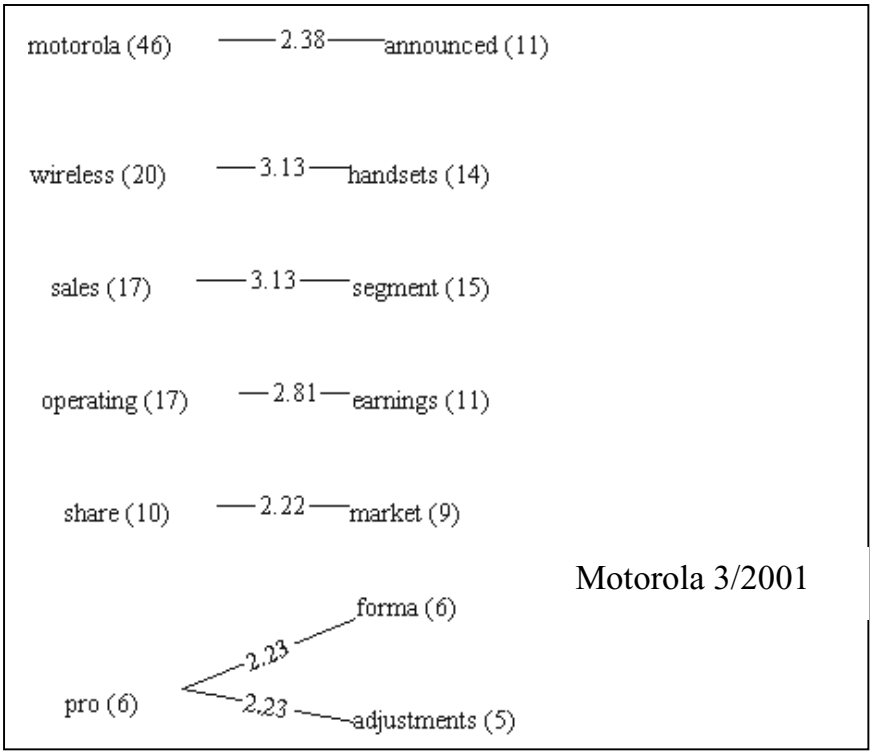


Figure 7.Collocational Network for Motorola report from quarter 3, 2001

4.3 Combining Text Mining Results and Linguistic Analysis

Ericsson

Contradicting to the results of collocational network analysis, Ericsson reports from 2000 quarter 3 and 4 are determined to be the closest matches in text mining analysis.

The collocational networks of the closest matches to Ericsson report from quarter 3, year 2000 are presented in Figure 8. They are the reports from Ericsson 4/2000, Motorola 3/2001, and Ericsson 1,2/2000.

The common parts of the collocational networks for those reports are circled. There is not much similarity between the collocational networks of Ericsson 3/2000 and Ericsson 4/2000 reports as was noticed in the previous section. The common collocates for those reports are *mobile-phone*, *operation*-margin, and *operating-income*. There is no resemblance of the collocational networks between the analyzed report of Ericsson 3/2000 and its other closest match from Motorola 3/2001, because Motorola report has too few links between collocates, that makes the entire network structure weak and links between collocates insignificant for the analysis.

Although the layouts of the collocational networks of the analyzed report and Ericsson reports from quarters 1 and 2/2000 are different, the resemblance is much higher with many common collocates: *sales-grow, sales-increase, sales-increased, sales-market-area, mobile-internet, mobile-infrastructure, income-taxes, income-operating, margin, capital-gains, cash-flow.* The architectures of the collocational networks of Ericsson reports from quarters 1 and 2/2000 are very similar.

The prototype matching method is not able to detect the differences in Ericsson reports from quarter 3 and 4, year 2000, however Nokia disappears from the closest

matches to those reports. The prototype-matching method shows that we have a change in the formation of the closest matches for the report from Ericsson 2000, quarter 2 since other Ericsson reports start to fire among the closest matches. The collocational networks derive this change only by the quarter 4, 2000, by noticing the dissimilarities in collocational network layouts for Ericsson reports from quarter 3 and 4, year 2000.

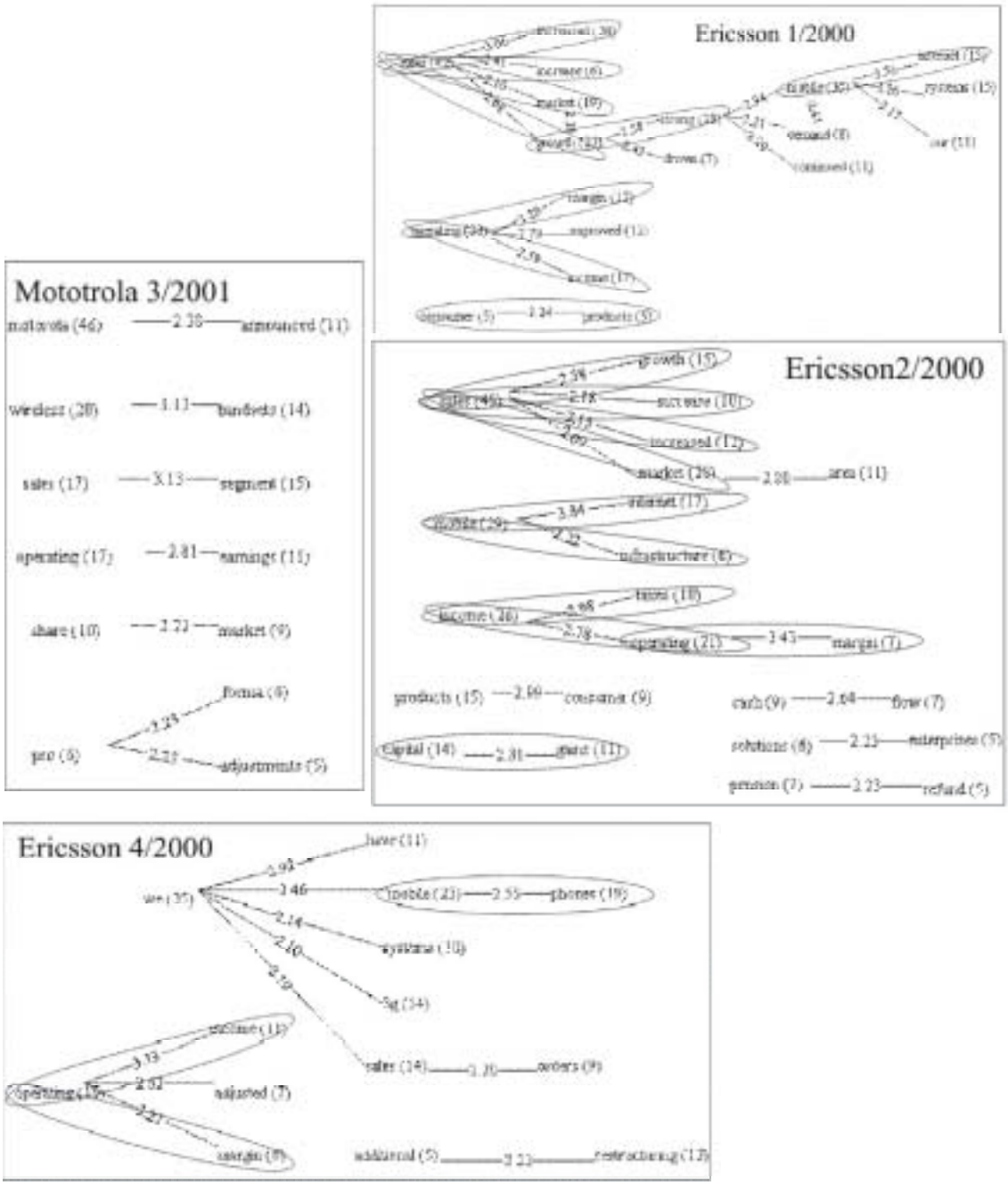


Figure 8. The collocational networks of four closest matches to Ericsson 3/2000 report

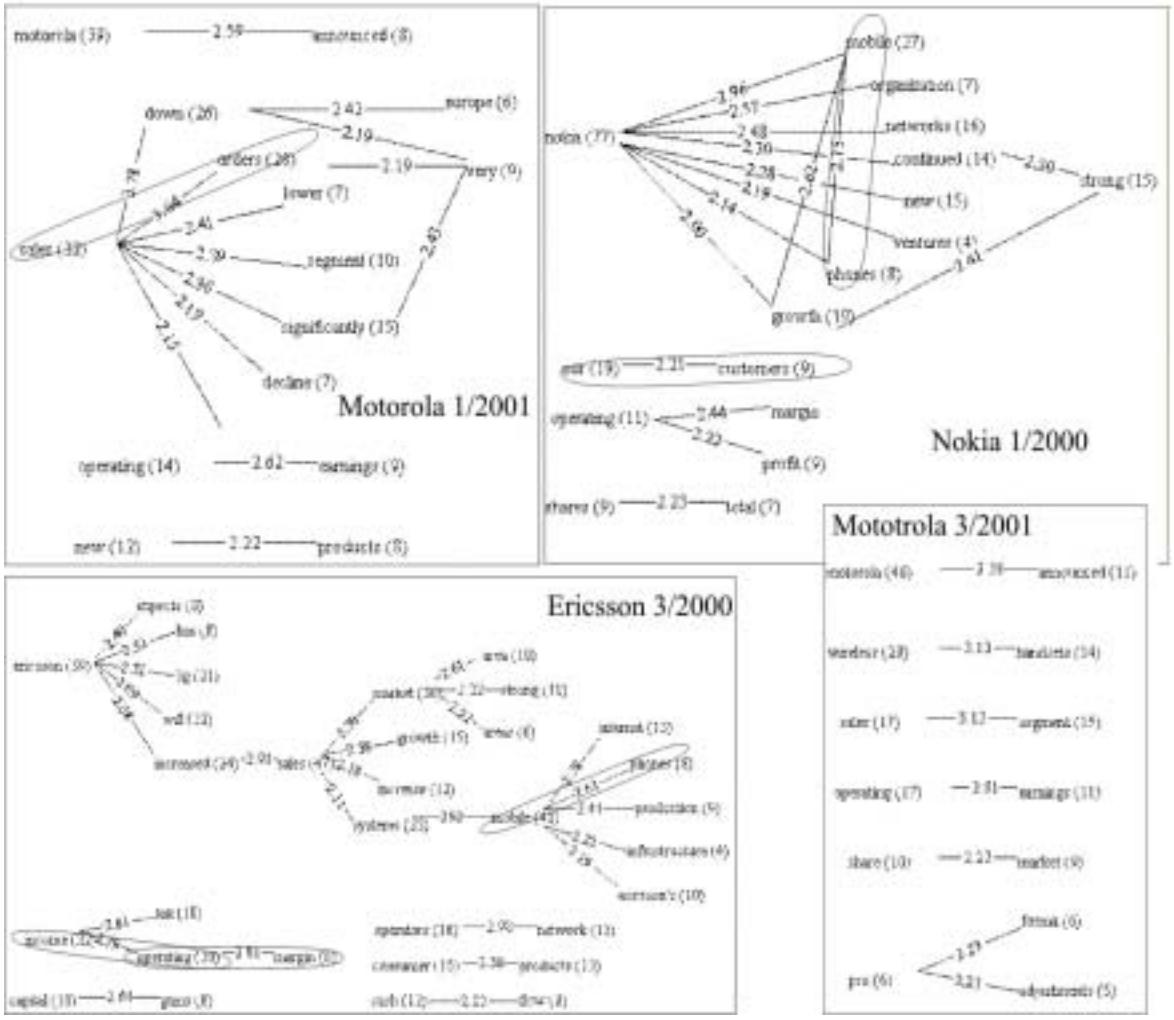


Figure 9. The collocational networks of four closest matches to Ericsson 4/2000 report

There are many more common collocates: *mobile-internet, consumer-products, operating-margin, sales-increase, sales-growth,* and *sales-market.* The common collocates between Ericsson report 3/2000 and collocational networks of Ericsson 4/2000, Motorola 2/2001, Motorola 3/2001, and Nokia 1/2000 reports are circled in Figure 9. Only one (for Motorola 1/2001) or two (Nokia 1/2000 and Ericsson 3/20000) collocates were spotted by the prototype-matching method in the multidimensional structure of quarterly reports. This might mean that either the collocation networks or the prototype matching method omitted some important information coded in the reports, or that different parts of information that was influencing the methods were equally important but simply did not coincide.

The architecture of the Motorola report from 3/2001 has very sparse structure, and thus, was picked out by the prototype-matching method as the closest match to eleven prototype-reports. It means that having only word pairs that were outlined in the collocational networks are not the dominating dimensions upon which the prototype-matching method had performed its clustering.

Nokia

The collocational networks of the closest matches to the collocational network for the Nokia report from quarter 1, 2001 are presented in Figure 10. The common collocates of Ericsson 1/2000, Nokia 4/2000, Ericsson 2/2001, Nokia 1/2000 and the analyzed Nokia report of Nokia1/2001 represented in Figure3, are circled. The resemblance between Ericsson 1/2000 Nokia 1/2001 lies in the following collocates: *sales-increased, growth strong, operating-margin.* There is not much in common between the collocational networks of the analyzed Nokia report and the Ericsson 2/2001 report (only the *operating-margin* collocate). The architectures of the collocational networks of Nokia 1,4/2000 and Nokia 1/2001 are very similar (word Nokia has a central position). Therefore, there are even more common collocates for Nokia 4/2000 and Nokia 1/2001: *mobile-market*, *Nokia-mobile, Nokia-new, Nokia-networks, Nokia-ventures, Nokia-phones, Nokia-organization, Nokia-introduces, Nokia-operating-profit, market-share.* The resemblance between the analyzed report and the next close match from Nokia 1/2000 is strong: *Nokia-mobile, Nokia- networks, Nokia-new, Nokia-ventures, strong-growth, operating-margin, operating-profit.*

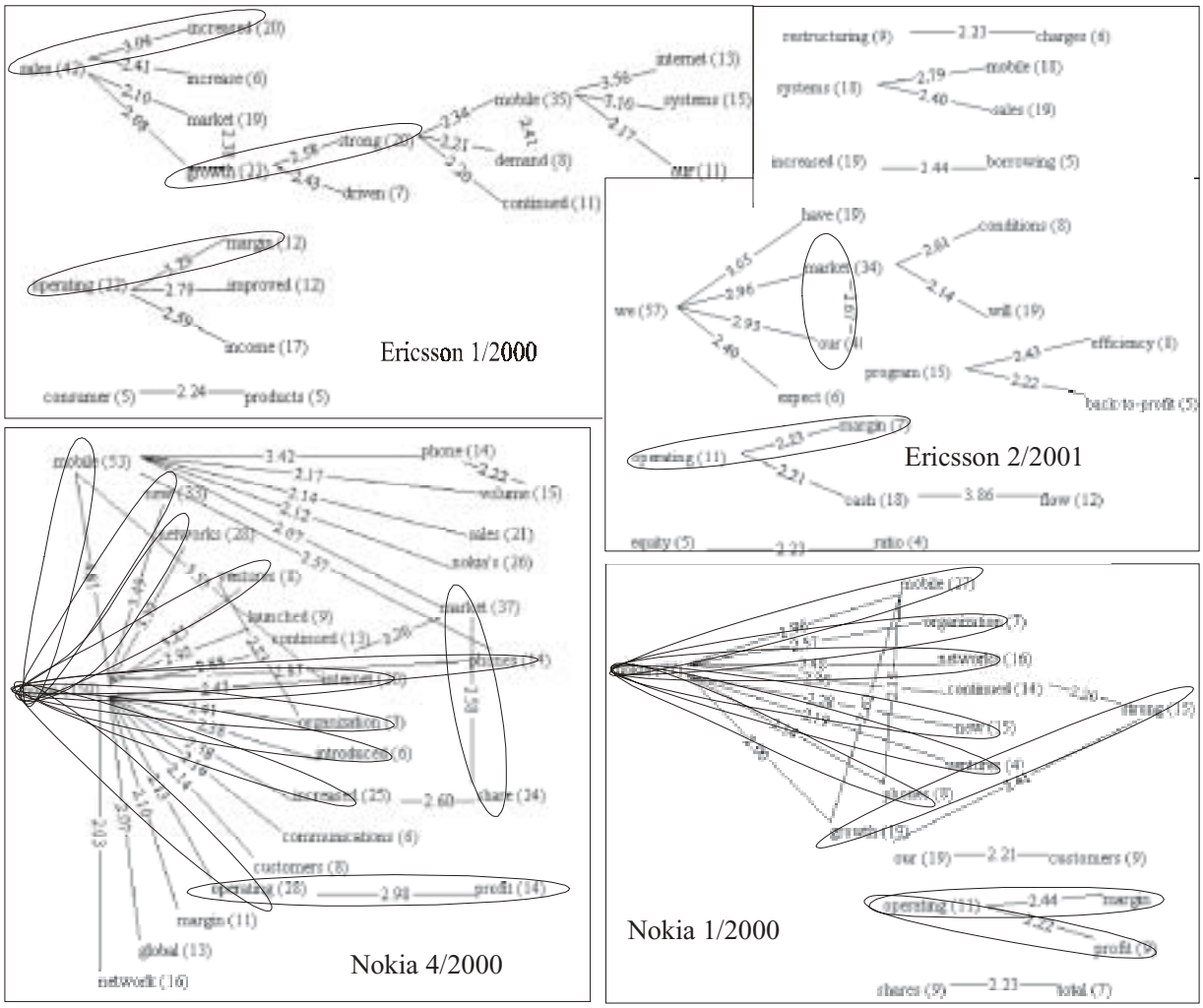


Figure 10. The collocational networks of four closest matches to Nokia 1/2001 report

The collocational networks of the closest matches to the Nokia report from quarter 2, 2001 are illustrated in Figure 11. They are the collocational networks of the Nokia 2/2000, Nokia 3/2001, Motorola 2/2000, and Motorola 1/2001 reports. The common collocates of those reports and the analyzed ones are circled. It is notable that the reports fro Nokia quarter 2 and 3/2001 have a block of similar construction of collocates *operating-margin- forma-pro-representing, profit- forma-pro, operating-representing.* Structurally, the collocational networks of those two closest matches are similar.

Although the collocational networks of all Nokia reports are quite similar, the closest to them, identified by text mining analysis, are not necessarily the reports from Nokia. It appears that the collocational networks of the analyzed Nokia reports bear more resemblance to the collocational networks of Ericsson reports than to Motorola ones, despite the fact that both Motorola and Ericsson reports are among the closest matches.

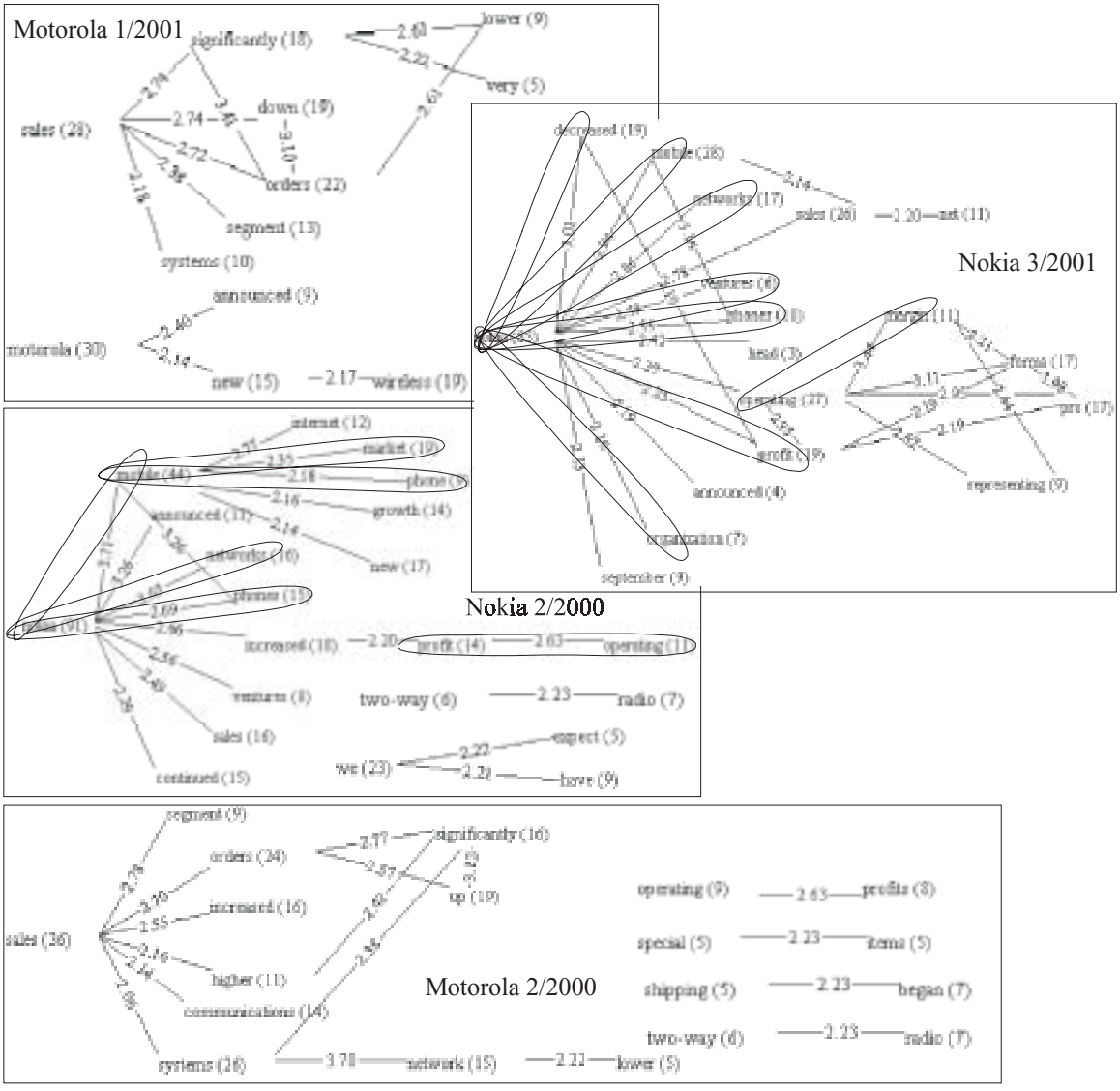


Figure 11. The collocational networks of four closest matches to Nokia 2/2001 report

Motorola.

The collocational networks of the analyzed Motorola reports and their closest matches do not have many common collocates. The common parts between the Motorola 4/2000 and its closest matches: Motorola 3/2001, Nokia 1 and 4/2000 and Ericsson 2/2001 are circled in Figure 12. The resemblance between their corresponding collocational networks is very low. The layouts of the collocational networks of Motorola report 1/2001 and the collocational networks of two of its closest matches from Motorola 2/2000 and Motorola 2/2001 are rather similar (word *sales* has a central position). The common collocates are *sales-segment, sales-systems, operating-profits* and *sales-orders-significantly* (for the Motorola 2/2000 report only)*, Motorola-announced* and *significantly-very* (for the Motorola 2/2001 report only). However, there is almost no resemblance between the networks of other closest matches.

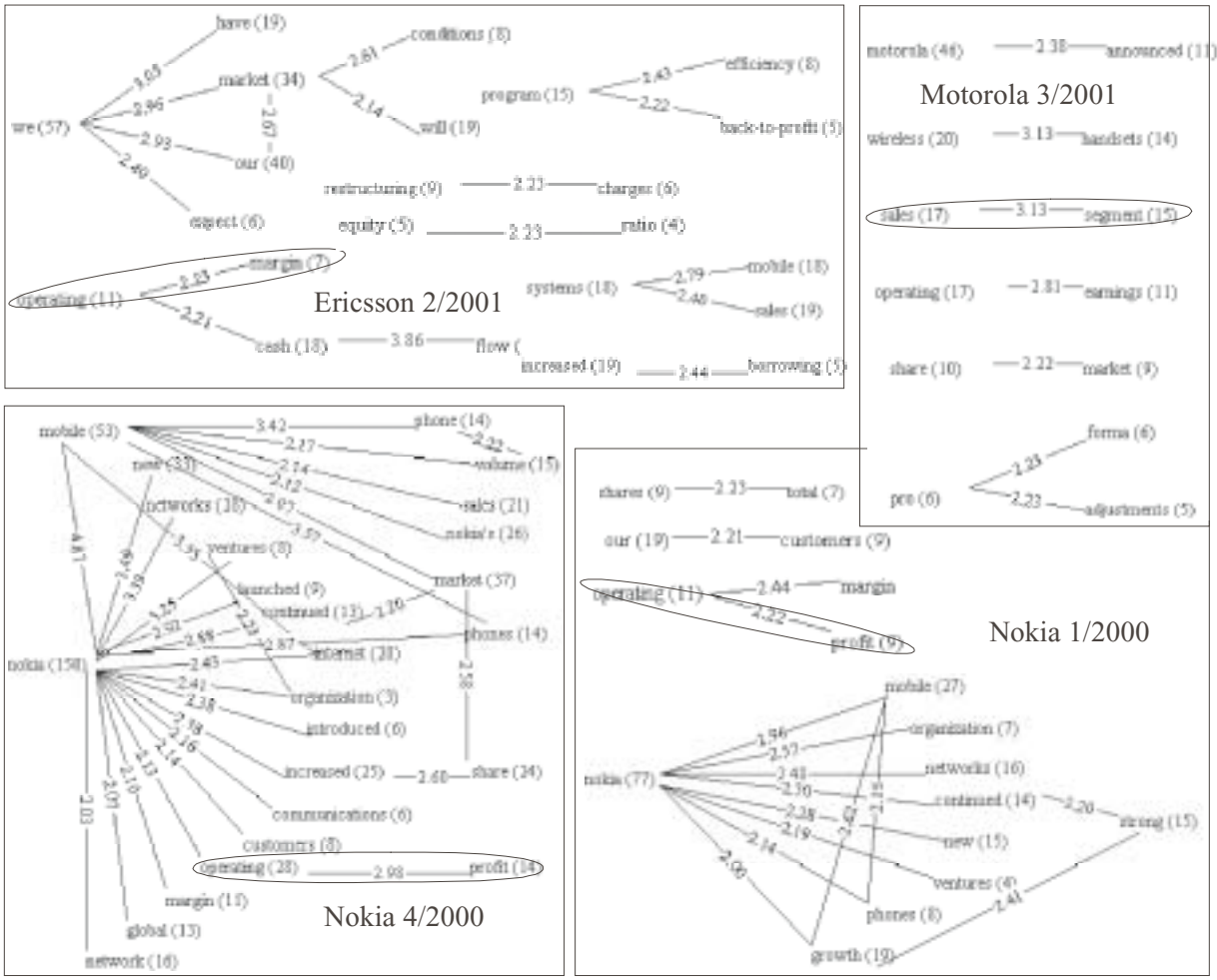


Figure 12. The collocational networks of four closest matches to Motorola 4/2000 report

The common collocates of those reports and the analyzed one are circled and presented in Figure 13. It is notable that Motorola's report from 2001, quarter 3, behaves in an interesting way by firing among four closest matches to eleven quarterly reports in our experiment. This happens because of very many undistinguished collocations with no distinctive network at all. The tone of the Motorola's report is very

neutral and requires further linguistic analysis. The report from the third quarter of year 2001 has no long collocational dependencies and its collocational network looks very different from the rest of the networks. The lack of strong connections between the important terms in this report has resulted in the fact that Motorola 3/2001 has fired as the closest match to eleven analyzed reports (see Table 1).

Collocational networks have illustrated how semantically similar the closest matches reports are to a report-prototype. Because of text multidimensionality establishing adequate similarities between text documents is hardly achievable. Therefore, only two dimensions (*operating* and *margin*) were detected as a similarity reference for Motorola 4/2000, Nokia 4/2000 and Ericsson 2/2001. At the same time, another dimension was spotted by the prototype-matching method for relating Motorola 3/2001 to the analysed Motorola report from 4/2000 – *sales* and *segment*. Although the overall resemblance in words and architecture among the collocational networks of closest matches for Motorola 4/2000 is weak, there is still several semantic similarities between them, upon which the prototype-matching method formed the clusters.

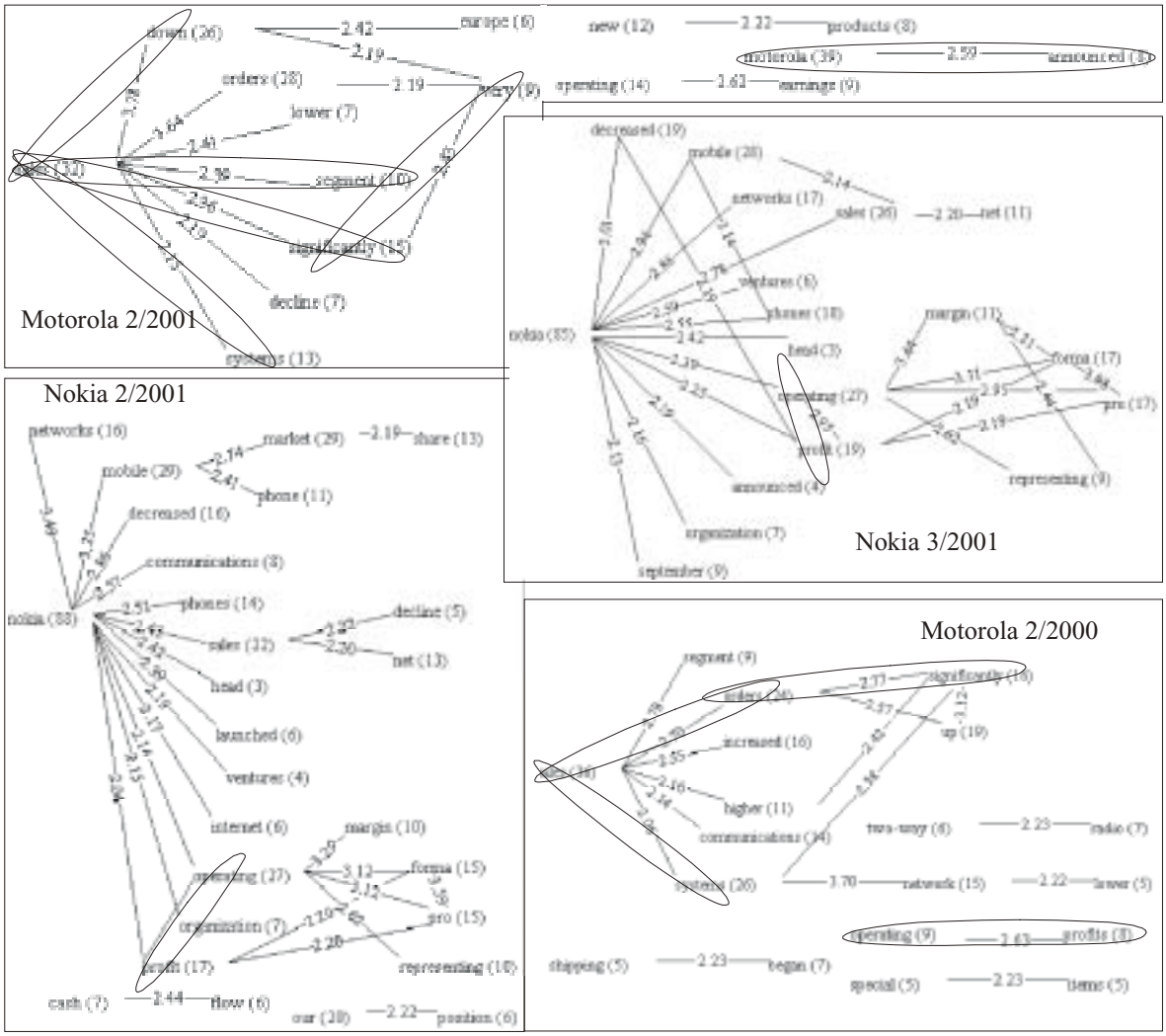


Figure 13. The collocational networks of four closest matches to Motorola 1/2001 report

5. Conclusions and Discussion

Information gathered from all occurring matches from the text mining part of study and the similarity of their collocational networks make it possible to conclude that the computer-aided text mining schema has captured a tendency of changing content in the reports relatively well. Our results from two parts of the analysis have shown to be somewhat controversial. Some of the reports from the companies have as their closest matches reports with similar collocational networks and some do not. Nevertheless, collocational networks and prototype matching text mining aim at presenting and visualizing textual information in a format that can be intuitively recognizable by decision makers. In other words, instead of reading all the reports and trying to compare the companies achievements and determining companies strategies, the decision maker can quickly browse the structure of collocational network or look at what types of reports are similar to the analyzed one.

We realize that the size of our text collection is the biggest limitation for drawing general conclusion. There are some limitations in constructing the collocational networks that affect the ability of the network to outline the central concepts within a report. The comparison results are controversial, because text is a multidimensional data that different readers understand differently. While collocational networks outlined one dimension in text, based on the parameters we had chosen, computer-based analysis took into account several text dimensions.

It will be beneficial to analyze why the closest matches to any chosen report have somewhat different collocational networks. Maybe the closest matches are capturing something more than the structure of the most commonly used terms in the reports. We plan to combine the results from our comparison with the analysis of the actual financial performance of the analyzed companies, i.e. using financial ratios and domain knowledge. The occurrence of the closest matches to any chosen prototype-report possibly contains information on companies future financial performance. In other words, if one knows that Nokia is a financially well performing company, than having closest matches from Nokia to any chosen prototype-report can indicate the semantic similarities that outline good performance. Contrarily, the disappearance of Nokia reports from the list of closest matches, such as for Ericsson report 2000, quarter 3 can imply a decrease or structural change in its financial performance, which was actually detected by the change in the collocation networks of the analyzed Ericsson report and consequent one.

The prototype-matching method compares textual reports by detecting only several similar dimensions from them. While some collocational networks have outlined the same dimensions of closest-matches reports upon which the prototype-matching method performed its clustering, some other collocational networks have outlined different text dimensions. That led to somewhat discordant results in cross-validation, when occasionally collocational networks of the closest matches did not resemble each other.

For future work, a study on the usability of the prototype-matching method and collocation networks that extract the essential key terms from long financial reports by managers can be performed.

6. Acknowledgements

The research was presented at at the XXVI Annual Congress of the European Accounting Association, 2-4 April, 2003, Seville, Spain. We gratefully acknowledge the financial support of TEKES (grant number 47 533). We are grateful to Antti Arppe for his valuable comments and suggestions.

References

Back, B., J. Toivonen, H. Vanharanta and A. Visa (2001). "Comparing numerical data and text information from annual reports using self-organizing maps." International Journal of Accounting Information Systems **2**(4): 249-269.

Church, K. and P. Hanks (1990). "Word association norms, mutual information, and lexicography." Computational Linguistics **16**: 22-29.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T., . (1987). "The Vocabulary Problem in Human-System Communication." Communications of the ACM **30**(11): 964-971.

Hearst, M. (1999). Untangling Text Data Mining. 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Maryland, USA, ACM Press.

Kendal, J. (1993). "Good and evil in chairmen's "boiler plate": an analysis." Organization Studies **14**: 571-592.

Kloptchenko, A., T. Eklund, B. Back, J. Karlsson, H. Vanharanta and A. Visa (2002). Combining Data and Text Mining Techniques for Analyzing Financial Reports. The 8th Americas Conference on Information Systems, Dallas, USA.

Kohut, G. and A. Segars (1992). "The president's letter to stockholders: An examination of corporate communication strategy." Journal of Business Communcation **29**(1): 7-21.

Osborn, J. D., C. I. Stubbart and A. Ramaprasad (2001). "Strategic Groups and Competitive Enactment: A Study of Dynamic Relationships between Mental Models and Performance." Strategic Management Journal **22**: 435-454.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford, Oxford University Press.

Stubbs, M. (1995). "Collocations and semantic profiles. On the cause of the trouble with quantitative studies." Functions of Language **2**: 23-55.

Subramanian, R., R. Isley and R. Blackwell (1993). "Performance and readability: A comparison of annual reports of profitable and unprofitable corporations." Journal of Business Communication **30**: 50-61.

Thomas, J. (1997). "Discourse in the Marketplace: The Making Meaning of Annual Reports." Journal of Business Communication **34**: 47-66.

Toivonen, J., A. Visa, T. Vesanen, B. Back and H. Vanharanta (2001). Validation of Text Clustering Based on Document Contents. Machine Learning and Data Mining in Pattern Recognition (MLDM 2001), Leipzig, Germany, Springer-Verlag.

Visa, A., J. Toivonen, S. Autio, J. Mäkinen, B. Back and H. Vanharanta (2001). Data Mining of text as a tool in authorship attribution. AeroSense 2001, SPIE 15th Annual International Symposium on Aerospace/Defense Sensing, Simulation and Controls, Orlando, Florida, USA.

Visa, A., J. Toivonen, B. Back and H. Vanharanta (2000). A New Methodology for Knowledge Retrieval from Text Documents. TOOLMET2000 Symposium - Tool Environments and Development Methods for Intelligent Systems.

Williams, G. C. (1998). "Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles." International Journal of Corpus Linguistics **3**: 151-171.

Winsor, D. (1993). "Owning corporate texts." Journal of Business and Technical Communication **7**(2): 179-195.

Research Paper 7

Kloptchenko, A. (2003), Determining Companies' Future Financial Performance from Their Past Quarterly Reports, accepted at the First Annual Pre-ICIS Workshop on Decision Support Systems, December, 14, Seattle, USA

Determining Companies' Future Financial Performance from Their Past Quarterly Reports

Antonina Kloptchenko
TUCS/Åbo Akademi University
akloptch@abo.fi

1. Introduction

Prediction or forecast in business and finance for decision support purposes is a far–reaching goal. Most of the prediction studies in IS were made using quantitative data in form of company financial ratios, S&P index, current stock prices, monthly growth, changes of macroeconomic factor, changes in market volatility, and other ratios. However, some important information that influence market can be coded not in numbers, but in text, for instance, the forecasts made by expert top-analysts, CEO's interviews, market analysis, quarterly and annual reports. Processing the huge amount of financial-related textual and numeric information looking for the hints of companies' future financial performance is not easily attainable task for human. We decided to use the combination of artificial neural networks (ANNs), data and text mining methods for determination of the financial performance of a company based on results from clustering qualitative and quantitative data from its quarterly reports. Quarterly reports were chosen as one of the most significant and strictly regulated companies' external documents that reflect on companies' strategy and the financial performance to stakeholders that are easily available online for decision makers.

There are a number of computer-based methods for analyzing quantitative accounting and finance related data ranging from spreadsheets to neural networks. However, there are not so many computer-based methods for analyzing qualitative data, and there are even fewer that combine quantitative and qualitative data to forecast the companies' future performance. Data and text mining methods offer the opportunity to discover hidden patterns that can be useful for particular purposes from huge amount of data (Fayyad 1996). The combination of data and text mining methods allows discovering more complex patterns in business-related data and brings additional understanding of real-world business situations to the decision makers.

Here we confirm the earlier discovery of (Kloptchenko, Eklund et al. 2002) that annual and quarterly reports contain information on both future and past performance and that text bears more diverse information than dry numbers do, by not only stating the facts but also explaining why they have happened. We create a methodology that mines the quarterly reports of companies-competitors or partners for forecasting their future financial performance. Firstly, we cluster the quantitative data from quarterly reports in form of financial ratios using the Self-Organizing Map (SOM). Secondly, we perform content-based clustering of qualitative data from textual part of quarterly reports using a prototype-matching methodology. Thirdly, we train the ANNs and use them as classifiers of results from quantitative and qualitative clustering. Aiming at making all the major steps in the forecasting methodology automatic we use feedforward multilayered neural networks (MFNN) with backpropagation learning as the final step. The reminded manual steps in the methodology are associated with the nature of ANNs, and require a human expertise in choosing appropriate architectures of SOM and MFNN, and the interpretation of quantitative data clustering results using domain knowledge.

We use a sample dataset consisting of both quantitative and qualitative data obtained from the Internet to illustrate the proposed forecasting methodology. The SOM created by

(Karlsson, Back et al. 2001) based on 99 telecommunication companies' annual reports for the years 1995-99 was used as a base to study the quarterly performance of Nokia, Ericsson, and Motorola during 2000-02. The quantitative data consist of a number of calculated financial ratios, and the qualitative data of the textual description from each report.

The rest of the paper continues as follows. Section 2 provides the background for the current study in form of overview the related studies. Section 3 describes the methods used in the methodology, namely the SOM for quantitative data clustering, the prototype matching algorithms for qualitative data clustering and the MFNN with backpropagation learning algorithm for forecasting. Section 4 offers an example of using quantitative and qualitative data analysis on a sample data set and the use of the clustering results as an input data for the neural networks. We review the results of forecast and compare them with the actual companies' performance. Section 5 discusses a number of limitations of the current exploratory study. Finally, in Section 6, we highlight some issues for further investigation.

2. Background

Since the language of quarterly reports has not been studied to any larger extent, our literature review is based on a broad body of literature written on the language of annual reports, conducted within linguistics, business communication and financial decision making studies. The annual and quarterly reports have a similar structure, conventions and communicative purposes. In short-term perspective quarterly reports are the important means for companies in appraising past performance and projecting companies' future opportunities to the readers, who primarily consist of investors and analysts. While studying the relationship between readability of annual reports and financial performance of the companies, (Subramanian, Isley et al. 1993) had shown that the annual reports of the well-performed companies were easier to read than those of poorly performed companies. (Thomas 1997) showed that the structure of language in the financial reports might unintentionally reveal things that the company may not have wished to reveal to its audience. The annual reports are not only the best possible description of a company, but are also a description of a company's managerial priorities. According to (Kohut and Segars 1992) communication strategies hidden in annual reports differ in terms of the subjects emphasized when the company's performance worsens.

The first attempts to semi-automatically analyze and compare the information from quantitative and qualitative parts of annual reports were made by (Back, Toivonen et al. 2001), (Back, Vanharanta et al. 1999). Their results indicated that there are differences in clustering results of qualitative and quantitative data due to a slight tendency to exaggerate the real financial performance in the text. It was proposed that text might correspond better to the next year's numerical data. (Kloptchenko, Eklund et al. 2002) continued the research in this field, using an improved document clustering method, a different hypothesis and data set. They discovered that quarterly reports tend to contain information on both future and past performance so that the tables with financial numbers indicate what a company has done and linguistic structure and written style in textual part indicate what a company will do. They suggest using these indications for forecasting purposes.

3. Methodology

Our methodology section builds on tree steps: two steps of mining quantitative and qualitative data (Back et al., 2001) and the final step of classifying retrieved findings to determine future

financial performance. The schematic overview of our methodology is given in Figure 1, where the first two steps can be performed consequently.

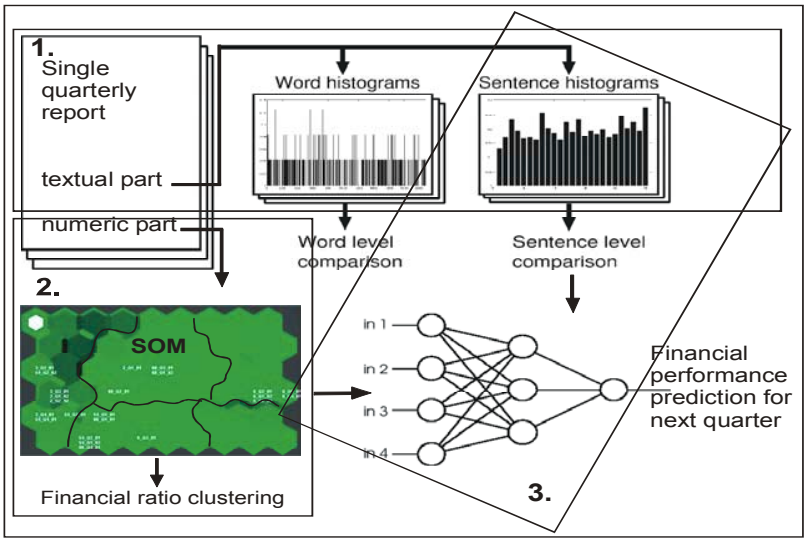


Fig. 1 The three-step methodology for forecasting financial performance from quarterly reports.

We use the SOM clustering ability (Kohonen, 1997) for financial benchmarking of quantitative data. SOMs are useful tools for exploratory data analysis that create a two-dimensional map from highly-dimensional input data. This map resembles a landscape in which it is possible to identify borders to differentiate different clusters that consist of input variables with similar characteristics (Honkela, Kaski et al. 1997). In order to make the quantitative data comparable, seven selected financial ratios were calculated (Lehtinen, 1996). Seven financial ratios,: *Operating Margin*, *Return on Total Assets* (ROTA) and *Return on Equity* (ROE), one liquidity ratio, *Current Ratio*, two solvency ratios *Equity to Capital* and *Interest Coverage,* and only one efficiency ratio *Receivables Turnover* fulfilled the criteria of good validity and reliability. The formulas for the ratios and data standardization can be found in (Kloptchenko, Eklund et al. 2002).

The prototype-matching text clustering methodology proposed by Visa et al. 2001 was used for qualitative data analysis. The prototype is a document or a part of it, which is of specific interest to a particular user. The chosen prototype is matched with an existing document collection to find the most similar documents in a collection. Constructing the histograms of the documents' word and sentence code numbers according to the corresponding value of quantization (Toivonen et al., 2001) allowed us to compare documents to each other simply by calculating the Euclidian distances between their histograms. The smallest Euclidian distance between word histograms indicates a common vocabulary of the reports. The smallest Euclidian distance between sentence histograms indicates similarities in written style and/or content of the reports (Visa et al., 2001).

We built supervised MFNN and trained it by backpropagation algorithm (Rumelhart, Widrow et al. 1994) using clustering results from two previous steps as an input data. The feedforward ANNs are used for nonlinear transformations of a multidimensional input variable into another multidimensional output variable (Safer and Wilamowski 1999). A measure of performance indicates how well a neural network has learned the relationships in the data. In prediction problems this measure is an error between the predicted outputs and the actual

desired outputs. In the study we followed the recommendations given by (Walczak and Cerpa 1999) to choose size of test and train data set, appropriate learning algorithm, network architecture, number of hidden nodes and type of activation function to avoid the overfitting of MFNN.

4. Results

4.1 Quantitative data analysis

By carefully analyzing the output map, six major clusters of companies were identified by using both the U-matrix map and the individual feature planes (Kloptchenko, Eklund et al. 2002). By analyzing the shades of the borders between the hexagons, we found similarities as well as differences among them. The identified on a base of evaluated neurons clusters are presented in Figure 2 in the form of a U-matrix map. Group A_1 and Group A_2 represent the class of the best-performed companies. For the companies situated in subgroup A_1, profitability is very good, with very high values in Operating Margin, ROTA, and ROE ratios. Group B represents the companies with slightly lower performance. These companies are distinguished by good profitability. ROE values are excellent. These companies have lower liquidity and solvency ratios than the companies in Groups A. Companies from groups C_1 and C_2 have moderate performance. In C_1 group, companies possess decent values in profitability, liquidity, and Equity to Capital ratios. Companies from C_s subgroup have decent values of profitability, but low liquidity, Interest Coverage and Receivables Turnover ratios. The values of Equity to Capital ratio, on the other hand, are good. Group D contains the class of the companies with bad performance. Their distinguishing features are low values of profitability and solvency ratios. At the same time, values of liquidity are average, and Receivables Turnover varies from very good to bad.

Figure 2. The identified clusters and the quarterly movements of Ericsson(3), Motorola(54), and Nokia(6).

The map reads that for instance, **Ericsson**, is facing severe financial difficulties during 2001-2002. Previously they have been situated close to or inside the above average groups, but during the last two years they are showing much poorer result, and are situated in the worst group.

Ericsson2000RQ1 **B**
Nokia2000RQ1 **A_1**
Motorola2001RQ3 **D**
Motorola2000RQ2 **C_1**
Ericsson2000RQ3 **B**
Ericsson2000RQ2 **A_1**
Ericsson2000RQ3 **B**
Nokia2001RQ4 **A_1**
Nokia2000RQ3 **A_1**
Ericsson2000RQ1 **B**
Ericsson2000RQ3 **B**
Ericsson2000RQ2 **A_1**
Ericsson2000RQ4 **C_1**
Ericsson2000RQ1 **A_1**
Nokia2001RQ4 **A_1**
Ericsson2000RQ4 **C_1**
Ericsson2000RQ3 **B**
Motorola2001RQ2 **D**
Ericsson2000RQ1 **B**
Motorola2001RQ3 **D**
Ericsson2001RQ1 **C_1**
Ericsson2001RQ3 **D**
Ericsson2001RQ2 **D**
Nokia2001RQ4 **A_1**
Nokia2001RQ3 **C_2**
Ericsson2001RQ2 **D**
Nokia2001RQ3 **C_2**
Motorola2001RQ4 **D**
Nokia2001RQ4 **A_1**
Ericsson2001RQ1 **C_1**

Table 3. The sample of the closest matches to Ericsson reports

4.2 Qualitative data analysis

The example of the obtaining results from qualitative data clustering for Ericsson is presented in Table 3. The column contains a report-prototype in the gray-shaded header and the four closest matches to it in the consequent rows. The bold letters by the report codes denote the cluster from the quantitative clustering to which a particular report belongs.

It reads, for example, that the Ericsson report from 2000, quarter 1 belonging to group B companies, at the sentence level comparison has the closest report by content from Nokia, 2000, quarter 1 belonging to best performing companies from group A_1. The second closest is the report from Motorola 2001, quarter 3 from the group D and so one. This means that the reports from Nokia 2000, quarter 1, Motorola 2001, quarter 3 and the Ericsson report from 2000, quarter 1 have similarities in sentence construction and word choice, which constitutes the language structure and written style. Word choice has a small impact on determining the closest matches that form clusters than the sentence construction. As an evidence of that, quarter names and proper names, e.g. Nokia, Motorola or Ericsson, did not determine the clusters.

As a general observation, the reports from the companies with good and steady financial performance have the reports from the well-performed companies among their closest matches, i.e. appearance of Nokia report from 2001, quarter 4 among the closest matches. If a company's performance worsens in the future, than the reports from average or badly performed companies fire among the closest matches, i.e. Ericsson report from 2000, quarter 4. The report from Motorola year 2000 quarter 3 has linguistically peculiarities and can be disregarded from the quantitative analysis.

4.3 Forecasting from a combination of quantitative and qualitative data mining

We constructed an individual MFNN for every analyzed company because every company has its own trend, cycle of development and speed of applying changes to affect its performance. Additionally, every company has a unique style to describe its evolution during reported period in every quarterly report. Training one network for all the reports from three companies did not give any robust results due to these individual trends that the companies follow. Because of the limitation of our data set and the nature of our research, where we want to find out whether the company worsens or improves its performance, we merge clusters A_1 and A_2 into cluster A (best performers), and clusters C_1 and C_2 into cluster C (moderate performers) resulting into 4-level performance scale. In order to simplify the learning for MFNN, we decode the inputs on the scale from 1 to 4, so that cluster A corresponds to 1, B to 2, C to 3, and D to 4. The number the hidden layer that seemed to work in our study the best was one with one neuron in it.

We use the financial position of the company obtained from SOM clustering (from 4.1), and positions of the closest matches obtained from qualitative clustering (from 4.2) as input variables for MFNNs. The four closest matches are taken from Table 3 along with corresponding letters of their financial performance from quantitative clusters, as well as previous financial position. For every company data we train three-layered MFNN with one node in hidden layer. Table 4 presents the sample of Ericsson data set that was fed into training the MFNN for predicting Ericsson performance in 2002, quarter 1.

Similar data sets for Motorola and Nokia were used to create separate MFNNs. The activation functions used in the hidden and final layers were sigmoidal. We have tried to reduce the dimensionality of the MFNN by decreasing the number of input variables, but according to the domain knowledge, four first matches capture the tendency of future financial performance the optimal way. Therefore, despite the overfitting problem, we trained 5-1-1 MFNN for the analyzed companies. Although we did not use any validation sets in our exploratory study, the fact that financial performance for all three companies could be predicted by MFNN with similar architecture intuitively proves the initial idea.

decoded data	Financial position	1st match	2nd match	3rd match	4th match	Target Output	Output	Flag
Ericsson2000RQ1	2	1	4	3	2	A	A	TRAIN
Ericsson2000RQ2	1	2	1	1	2	B	B	TRAIN
Ericsson2000RQ3	2	2	1	1	2	C	C	TRAIN
Ericsson2000RQ4	3	2	4	2	4	C	C	TRAIN
Ericsson2001RQ1	3	4	4	1	3	D	D	TRAIN
Ericsson2001RQ2	4	3	4	1	3	D	D	TRAIN
Ericsson2001RQ3	4	3	4	4	1	D	D	TRAIN
Ericsson2001RQ4	4	4	3	1	4	D	D	TRAIN
Ericsson2002RQ1	4	4	4	3	1	D	D	TEST
						A,B,C,D		SYMBOL

Table 4. Sample of training and testing data and architecture of the MFNN for Ericsson

5. Limitations

As the strongest limitation of our exploratory study we consider the small size of data collection in text clustering that created danger of MFNN overfitting and restricted the validation. We realize that our sample dataset is too limited to draw a general accurate conclusion on predicting power of MFNN over the future financial performance, because ANNs by nature are the best tools to work with "big" data. However, in the dynamic business environment there is a place to the forecasting situation then the data set grows as time passes by. Moreover, sometime the accuracy of prediction results can be even reduced by large historic data, if the business cycle and outside market condition change multiple times during analysis. The limited vocabulary (terms related to finance and the telecommunications sector), extensive use of proprietary names (such as Motorola, Nokia, and Ericsson), and indications of time period (quarter, year, annual), slightly influenced the clustering ability in our qualitative analysis. We plan to expand the study to a larger collection of quarterly/annual reports and try out to train ANN with different learning algorithm, such as radial basis ANN.

6. Conclusions and Future Work

The proposed methodology was designed for performing competitor or industry analysis, as well as determining future financial performance of the companies within the same line of business. We clustered the financial ratios of the companies using SOM and visualized this classification. Then, we analyzed the textual parts of quarterly reports for the same period of time, in order to reveal the heuristic relationship between the written style and facts stated by the numbers (ratios).

This research extended the work of (Kloptchenko, Eklund et al. 2002) by predicting the future financial performance of the companies based on the hidden stylistic indications in textual parts of the reports by means of MFNN. Before a dramatic change occurs in company financial performance, we see a change in the written style of a financial report. The tone tends to be closer to the next company performance. If the company's position is worse during next quarter, the report of the current quarter gets more pessimistic, even though the actual financial performance remains the same.

Our future work is directed toward trying out the methodology on the more extensive data set, and for companies from different lines of business. The desired ultimate output on the research is to come up with user-friendly prototype-system with automatic data gathering, choosing parameters for SOM and MFNN, and automatic cluster detection in SOM. The quantitative data mining using SOM, qualitative data mining using prototype-matching methodology and classifying the clustering results using MFNN provide a sound framework for future extension and experimentation.

7. Acknowledgements

The financial support from TEKES (grant number 40943/99) and the Academy of Finland is gratefully acknowledged. I am grateful to Tomas Eklund, Barbro Back, and Jonas Karlsson for their contributions at the earliest stages of the research.

References

Back, B., J. Toivonen, et al. (2001). "Comparing numerical data and text information from annual reports using self-organizing maps." International Journal of Accounting Information Systems **2**(4): 249-269.

Back, B., H. Vanharanta, et al. (1999). Knowledge Discovery in Analyzing Texts in Annual Reports. IFORS SPC-9, Intelligent Systems and Active DSS, Turku, Finland.

Fayyad, U. (1996). "Data Mining and Knowledge Discovery: Making Sense Out of Data." IEEE Expert: 20-25.

Honkela, T., S. Kaski, et al. (1997). WEBSOM - Self-Organizing Maps of Document Collections. WSOM'97: Workshop on Self-Organizing Maps, Espoo, Helsinki University of Technology.

Karlsson, J., B. Back, et al. (2001). Financial Benchmarking of Telecommunications Companies. Turku, Turku Centre for Computer Science.

Kloptchenko, A., T. Eklund, et al. (2002). Combining Data and Text Mining Techniques for Analyzing Financial Reports. The 8th Americas Conference on Information Systems, Dallas, USA.

Kohut, G. and A. Segars (1992). "The president's letter to stockholders: An examination of corporate communication strategy." Journal of Business Communication **29**(1): 7-21.

Rumelhart, D., B. Widrow, et al. (1994). "The Basic Ideas in Neural Networks." Communications of the ACM **37**(3): 87-92.

Safer, A. and B. M. Wilamowski (1999). Using neural networks to predict abnormal returns of quarterly earnings. International Joint Conference on Neural Networks - IJCNN'99, Washington, DC.

Subramanian, R., R. Isley, et al. (1993). "Performance and readability: A comparison of annual reports of profitable and unprofitable corporations." Journal of Business Communication **30**: 50-61.

Thomas, J. (1997). "Discourse in the Marketplace: The Making Meaning of Annual Reports." Journal of Business Communication **34**: 47-66.

Walczak, S. and N. Cerpa (1999). "Heuristics principles for the design of Artificial Neural Networks." Information and Software Technology **41**(2): 109-119.

Turku Centre for Computer Science

TUCS Dissertations

15. **Tomi Pasanen**, In-Place Algorithms for Sorting Problems
16. **Mika Johnsson**, Operational and Tactical Level Optimization in Printed Circuit Board Assembly
17. **Mats Aspnäs**, Multiprocessor Architecture and Programming: The Hathi-2 System
18. **Anna Mikhajlova**, Ensuring Correctness of Object and Component Systems
19. **Vesa Torvinen**, Construction and Evaluation of the Labour Game Method
20. **Jorma Boberg**, Cluster Analysis. A Mathematical Approach with Applications to Protein Structures
21. **Leonid Mikhajlov**, Software Reuse Mechanisms and Techniques: Safety Versus Flexibility
22. **Timo Kaukoranta**, Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization
23. **Gábor Magyar**, On Solution Approaches for Some Industrially Motivated Combinatorial Optimization Problems
24. **Linas Laibinis**, Mechanised Formal Reasoning About Modular Programs
25. **Shuhua Liu**, Improving Executive Support in Strategic Scanning with Software Agent Systems
26. **Jaakko Järvi**, New Techniques in Generic Programming - C++ is more Intentional than Intended
27. **Jan-Christian Lehtinen**, Reproducing Kernel Splines in the Analysis of Medical Data
28. **Martin Büchi**, Safe Language Mechanisms for Modularization and Concurrency
29. **Elena A. Troubitsyna**, Stepwise Development of Dependable Systems
30. **Janne Näppi**, Computer-Assisted Diagnosis of Breast Calcifications
31. **Jianming Liang**, Dynamic Chest Image Analysis: New Model-based Methods for Dynamic Pulmonary Imaging and Other Applications
32. **Tiberiu Seceleanu**, Systematic Design of Synchronous Digital Circuits
33. **Tero Aittokallio**, Characterization and Modelling of the Cardiorespiratory System in Sleep-disordered Breathing
34. **Ivan Porres**, Modeling and Analyzing Software Behavior in UML
35. **Mauno Rönkkö**, Stepwise Development of Hybrid Systems
36. **Jouni Smed**, Production Planning in Printed Circuit Board Assembly
37. **Vesa Halava**, The Post Correspondence Problem for Marked Morphisms
38. **Ion Petre**, Commutation Problems on Sets of Words and Formal Power Series
39. **Vladimir Kvassov**, Information Technology and the Productivity of Managerial Work
40. **Franck Tétard**, Managers, Fragmentation of Working Time, and Information Systems
41. **Ján Maòuch**, Defect Theorems and Infinite Words
42. **Kalle Ranto**, $\mathbf{Z}_4$-Goethals Codes, Decoding and Designs
43. **Arto Lepistö**, On Relations Between Local and Global Periodicity
44. **Mika Hirvensalo**, Studies on Boolean Functions Related to Quantum Computing
45. **Pentti Virtanen**, Measuring and Improving Component-Based Software Development
46. **Adekunle Okunoye**, Knowledge Management and Global Diversity: A Framework to Support Organisations in Developing Countries
47. **Antonina Kloptchenko**, Text Mining Based on the Prototype Matching Method

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematics

Åbo Akademi University
- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration
- Institute of Information Systems Science