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.

Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS'02)
0-7695-1733-1/02 $17.00 © 2002 IEEE

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.

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

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.

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.

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

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 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	19.1%
4	Digital Documents	40	25%
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

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

[2]. 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.

[3]. 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.

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

[5]. 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.

[6]. Hand D., M. H., and Smyth P. (2001). Principles of Data Mining. Boston, USA, A Bradford Book, The MIT Press, 2001.

[7]. Jain, A., Murty, M., and Flynn, P. (1999). "Data Clustering: A Review." ACM Computing Surveys **31**(3): 265-323.
[8]. 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.
[9]. Larsen, B., and Aone, A. (1999). Fast and Effective Text Mining Using Linear-time Document Clustering. KDD-99, San Dieago, CA, USA, ACM.
[10]. Lee, C., and Yang, H. (1999). A Web Text Mining Approach Based on Self- Organizing Map. WIDM-99, Kansas City, MO, USA, ACM.
[11]. 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.
[12]. Salton. G., a. M., M. (1983). Introduction to modern information retrieval. New York, McGraw-Hill.
[13]. 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.
[14]. 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.
[15]. 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.
[16]. 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.
[17]. Zamir, O., and Etzioni, O. (1998). Web Document Clustering: A Feasibility Demonstration. SIGIR'98, Melbourne, Australia, ACM Press.