Πανεπιστήμιο Πατρών

Τμήμα Μαθηματικών και

Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Διατμηματικό Μεταπτυχιακό Πρόγραμμα Ειδίκευσης

«Μαθηματικά των Υπολογιστών και των Αποφάσεων»

Διπλωματική Εργασία

του Θωμά Α Παπαστεργίου

«Μέτρα ομοιότητας στην τεχνική ομαδοποίησης (Clustering): Εφαρμογή σε ανάλυση κειμένων (Text Mining)»

Επιβλέπων καθηγητής:

Βασίλης Βουτσινάς

Πάτρα 2005

Η διπλωματική εργασία πάνω στο θεματικό πεδίο της Εξόρυξης Κειμένου, (Text Mining) εκπονήθηκε από τον μεταπτυχιακό φοιτητή Θωμά Παπαστεργίου στα πλαίσια του Διατμηματικού Μεταπτυχιακού Προγράμματος Ειδίκευσης «Μαθηματικά των Υπολογιστών και των Αποφάσεων» των τμημάτων Μαθηματικών και Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής του Πανεπιστήμιου Πατρών. Επιβλέπων καθηγητής ήταν ο κ. Βασίλης Βουτσινάς και η τριμελής επιτροπή συμπληρώθηκε από τον κ. Ιωάννη Χατζηλυγερούδη και τον κ.Γαρουφαλάκη.

Πρόλογος

Η παρούσα διπλωματική εργασία που αφορά στον θεματικό χώρο της «Εξόρυξης Κειμένου» εκπονήθηκε στα πλαίσια του Διατμηματικού Μεταπτυχιακού Προγράμματος Ειδίκευσης «Μαθηματικά των Υπολογιστών και των Αποφάσεων».

Σε αυτήν την διπλωματική εργασία ο αναγνώστης εισάγεται, μετά από μια σύντομη εισαγωγή, στις βασικές έννοιες της ομαδοποίησης. Εκεί εξετάζονται σχετικά αναλυτικά τα διάφορα μέτρα ομοιότητας και ανομοιότητας με έμφαση σε αυτά που αφορούν τα κατηγορικά δεδομένα. Στο επόμενο κεφάλαιο εξετάζονται διάφοροι δημοφιλείς αλγόριθμοι ομαδοποίησης με έμφαση πάλι σε αυτούς που αναφέρονται σε κατηγορικά δεδομένα ή σε αυτούς που μπορούν να μετατραπούν έτσι ώστε να ομαδοποιούν κατηγορικά δεδομένα. Στην συνέχεια γίνεται μια επισκόπηση στις βασικές τεχνικές που χρησιμοποιούνται στην εξόρυξη κειμένου. Ιδιαίτερη έμφαση δίνεται στον τομέα της αναπαράστασης κειμένου καθώς και στον τομέα της ομαδοποίησης κειμένου.

Η συνεισφορά μας στη θεματική περιοχή της εξόρυξης κειμένου έγκειται στην ανάπτυξη ενός καινούργιου μέτρου ανομοιότητας μεταξύ κατηγορικών δεδομένων το οποίο βασίζεται σε οντολογίες οριζόμενες από τον χρήστη που αναπαριστούν την γνώση του στο πεδίου εφαρμογής. Το μέτρο αυτό ενσωματώνεται στον πολύ δημοφιλή αλγόριθμο ομαδοποίησης αριθμητικών δεδομένων k – means έτσι ώστε αυτός να μπορεί να ομαδοποιεί και κατηγορικά δεδομένα. Επίσης εξετάζεται η ενσωμάτωση του προτεινόμενου μέτρου ανομοιότητας σε μια σειρά άλλων αλγορίθμων ομαδοποίησης αριθμητικών δεδομένων όπως παραδείγματος χάριν ο k – windows, ο z – windows και άλλοι. Στην συνέχεια ελέγχεται το προτεινόμενο μέτρο ανομοιότητας συγκρίνοντας τα αποτελέσματα των ομαδοποιήσεων του k – means με ενσωματωμένο το προτεινόμενο μέτρο ανομοιότητας, με γνωστά μέτρα ανομοιότητας. Για τους ελέγχους χρησιμοποιούνται το μέτρο ανομοιότητας του k – modes καθώς και το μέτρο ανομοιότητας της εννοιολογικής ομαδοποίησης που προτάθηκε από τον Kodratoff ενσωματωμένο στον αλγόριθμο k – means. Οι έλεγχοι έγιναν σε μια βάση δεδομένων συνθεμένη από πραγματικά δεδομένα.

Επιπλέον η συνεισφορά έγκειται σε μια ακόμα εφαρμογή του προτεινόμενου μέτρου ανομοιότητας. Αυτή η εφαρμογή αφορά στη λύση του προβλήματος πατρότητας κειμένων το γνωστό authoring – attribution problem. Η εφαρμογή αφορά ποιήματα του έλληνα ποιητή Ντίνου Χριστιανόπουλου καθώς και των Ιαπώνων ποιητών Μπασό και Μπουσόν. Χρησιμοποιώντας το προτεινόμενο μέτρο ανομοιότητας ενσωματωμένο στον k – means κατορθώσαμε να διαχωρίσουμε αρχικά τα ελληνικά ποιήματα του Χριστιανόπουλου από τα ποιήματα των Ιαπώνων και στην συνέχεια εκλεπτύνοντας όλο και πιο πολύ την διαμέριση των ποιημάτων να παίρνουμε ομάδες ποιημάτων που μπορούν να περιγραφούν εύκολα με την κοινή λογική. Θα πρέπει να αναφέρουμε ότι τα ποιήματα διαλέχτηκαν στην τύχη χωρίς να υπάρχει μια προκαθορισμένη δομή των ομάδων στην βάση δεδομένων που χρησιμοποιήθηκε. Παρόλα αυτά όμως μπορέσαμε να ανακαλύψουμε πολύ καλά δομημένες και πολύ εύκολα επεξηγήσιμες ομάδες στα δεδομένα μας, κάτι που αποτελεί μια σαφή ένδειξη για τις δυνατότητες του προτεινόμενου μέτρου ανομοιότητας.

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να ευχαριστήσω ιδιαιτέρως τον επιβλέποντα καθηγητή κ. Βουτσινά Βασίλη για την ουσιαστική βοήθεια μου προσέφερε με την άψογη συνεργασία, τις πολύτιμες συμβολές και την καθοδήγησή του σε όλη την διάρκεια της εκπόνησης αυτής της διπλωματικής εργασίας, καθώς η βασική ιδέα της ερευνάς του ανήκει.

Επίσης ευχαριστώ του καθηγητές κ. Ιωάννη Χατζυλιγερούδη και κ. Ιωάννη Γαρουφαλάκη για τις παρατηρήσεις και τις υποδείξεις τους στη βελτίωση της εργασίας.

Τέλος θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Γεράσιμο Αντζουλάτο για την ανεκτίμητη βοήθειά του κάθε φορά που την χρειάστηκα καθώς και την Κική Παπαστεργίου, την Δανάη Θεοχάρη και την Νίκη Τριανταφυλλίδου για την πραγματικά ανεκτίμητη βοήθειά τους της τελευταίας στιγμής.

1 Εισαγωγή

1.1 Εισαγωγή

Η πρακτική της κατάταξης των αντικείμενων σε κλάσεις σύμφωνα με κάποιες αντιλαμβανόμενες ομοιότητες είναι ένα αρκετά μεγάλο πεδίο της επιστήμης. Η οργάνωση των δεδομένων σε αισθητές κατηγορίες είναι ένας από τους πιο θεμελιώδεις τρόπους με τους οποίους οι άνθρωποι αντιλαμβάνονται τον κόσμο μαθαίνουν και σκέφτονται. Τα παιδιά μαθαίνουν από πολύ μικρή ηλικία να ταξινομούν τα αντικείμενα του περιβάλλοντος τους και να συνδέουν τις προκύπτουσες κατηγορίες με ουσιαστικά της γλώσσας τους. Παραδείγματος χάριν όταν βλέπουμε μία γάτα σε ένα πεζοδρόμιο αμέσως αναγνωρίζουμε την οντότητα «γάτα» ως ένα αντικείμενο της κατηγορίας «γάτα» και έτσι μπορούμε να συμπεράνουμε ότι το αντικείμενο που εντοπίσαμε θα νιαουρίζει χωρίς να το ακούσουμε να το κάνει. Προφανώς το συμπέρασμά μας βασίζεται στο ενοποιητικό χαρακτηριστικό των αντικειμένων της κατηγορίας «γάτα».

Ο όρος *Ομαδοποίηση* (Clustering) ή *Ανάλυση Ομαδοποίησης* (Cluster Analysis) όπως αλλιώς ονομάζεται, είναι ένα γενικό όνομα για μια ευρεία ποικιλία διαδικασιών που χρησιμοποιούνται για την δημιουργία μιας ταξινόμησης [1].

Το κύριο μέλημα της ομαδοποίησης είναι η ανάκτηση «λογικών» ομάδων (Clusters) οι οποίες προϋπάρχουν στα δεδομένα και οι οποίες επιτρέπουν την ανακάλυψη ομοιοτήτων και διαφορών ανάμεσα στα δεδομένα έτσι ώστε να παράγονται χρήσιμα συμπεράσματα για αυτά. Η Ομαδοποίηση χρησιμοποιείται για τον χειρισμό του μεγάλου πλήθους των δεδομένων που λαμβάνουν καθημερινά οι άνθρωποι. Η εξντλητική επεξεργασία της πληροφορίας αυτής είναι αδύνατη. Συνεπώς είναι απαραίτητη η κατηγοριοποίηση των οντοτήτων, δηλαδή των αντικειμένων, των προσώπων, των γεγονότων, σε ομάδες. Κάθε ομάδα χαρακτηρίζεται μέσω του κοινού χαρακτηριστικού των οντοτήτων που περιέχει.

Τα αντικείμενα που θα καταταχτούν σε ομάδες περιγράφονται είτε από ένα σύνολο μετρήσεων είτε από τις σχέσεις που έχει αυτό το αντικείμενο με τα άλλα υπό εξέταση αντικείμενα. Η απουσία ετικετών κατηγορίας είναι αυτό που ξεχωρίζει την Ανάλυση ομαδοποίησης από την διακριτή ανάλυση, από την αναγνώριση προτύπων και από την ανάλυση αποφάσεων. Το αντικείμενο της Ανάλυσης Ομαδοποίησης αφορά στο να βρει κανείς μια πρόσφορη και έγκυρη οργάνωση των δεδομένων.

Θα δώσουμε ένα παράδειγμα για να διασαφηνίσουμε εντελώς την διαφορά μεταξύ της θεωρίας αποφάσεων (decision making) και της διαδικασίας ομαδοποίησης (clustering). Ας υποθέσουμε ότι θέλουμε να ομαδοποιήσουμε σε δυο ομάδες τα μεταπτυχιακά προγράμματα σπουδών που υπάρχουν στην Ευρωπαϊκή ένωση βασιζόμενοι στα εξής γνωρίσματα (attributes): μέγεθος του τμήματος, τεχνικός εξοπλισμός, εξωτερική ερευνητική υποστήριξη και δημοσιεύσεις του τμήματος. Στην περίπτωση της λήψης αποφάσεων (decision making) κάποιος ειδικός (expert) θα πρέπει εκ των προτέρων να ορίσει αυτές τις δυο κατηγορίες. Αυτές οι δυο κατηγορίες θα οριστούν παίρνοντας προκαταταγμένα παραδείγματα από ένα σύνολο παραδειγμάτων, το οποίο στην ορολογία της θεωρίας αποφάσεων λέγεται σύνολο

εκπαίδευσης (training set). Οι οντότητες τώρα αυτού του συνόλου εκπαίδευσης θα χρησιμοποιηθούν για να κατασκευαστούν σύνορα απόφασης (decision boundaries) ή με άλλα λόγια κατώφλια απόφασης στα γνωρίσματα των οντοτήτων, τα οποία θα ξεχωρίζουν τα δυο είδη των μεταπτυχιακών προγραμμάτων. Αφού θα έχουν κατασκευαστεί τα όρια απόφασης, τα άλλα μεταπτυχιακά προγράμματα που δεν έχουν χαρακτηριστεί από τον ειδικό ακόμα, θα καταχτούν σύμφωνα με τον αλγόριθμο στις δυο ομάδες. Στην περίπτωση της ομαδοποίησης τώρα δεν έχουμε την ανάγκη από έναν ειδικό για να μας ορίσει εκ των προτέρων τις δυο ομάδες. Η επιδίωξή μας είναι να προσδιορίσουμε εάν μια διαμέριση του αρχικού συνόλου σε δυο ομάδες που να βασίζεται στα χαρακτηριστικά των οντοτήτων είναι λογική και αν ναι, να μπορέσουμε να καθορίσουμε τα μέλη αυτών των δυο ομάδων. Αυτό μπορεί να γίνει εάν ορίσουμε ομοιότητες μεταξύ των προγραμμάτων μεταπτυχιακών σπουδών που να βασίζονται στα χαρακτηριστικά τους και μετά να κατασκευάσουμε ομάδες τέτοιες ώστε οι ομοιότητες των οντοτήτων εντός αυτών των ομάδων να είναι μεγαλύτερες από τις ομοιότητες των οντοτήτων που ανήκουν σε διαφορετικές ομάδες.

Η ανάλυση ομαδοποίησης είναι ένας κλάδος της διερευνητικής ανάλυσης δεδομένων (exploratory data analysis), με τον οποίο προσπαθούμε να κρισάρουμε τα δεδομένα, τα οποία συλλέγονται με κάθε τρόπο, για να βγάλουμε κάποιο νόημα από αυτά. Η πληροφορία που κερδίζουμε από τα δεδομένα εάν εφαρμόσουμε μια ομαδοποίηση θα πρέπει να ωθεί την δημιουργικότητα του χρήστη, να υποδεικνύει νέα πειράματα, να προσφέρει μια νέα πιο βαθιά ματιά στα αρχικά «άναρχα» δεδομένα ή τέλος να μας δίνει έστω μια πιο κατανοητή μορφή των δεδομένων που έχουμε.

Δεν είναι τυχαίο ότι η ομαδοποίηση χρησιμοποιείται σε μια πλειάδα ετερόκλιτων μεταξύ τους επιστημών. Έχει χρησιμοποιηθεί στις επιστήμες ζωής (βιολογία, ζωολογία), στις ιατρικές επιστήμες (ψυχολογία, παθολογία), στις κοινωνικές επιστήμες (κοινωνιολογία, αρχαιολογία), στις επιστήμες της γης (γεωγραφία, γεωλογία) αλλά και στις επιστήμες των μηχανών [4]. Ο όρος «ομαδοποίηση» παίρνει διαφορετικά ονόματα στις διαφορετικές επιστήμες. Έτσι παραδείγματος χάριν στην αναγνώριση προτύπων ονομάζεται εκπαίδευση χωρίς επίβλεψη (unsupervised learning) ή ως εκπαίδευση χωρίς δάσκαλο (learning without a teacher), στην βιολογία και στην οικολογία αναφέρεται ως αριθμητική ταξινόμηση (numerical taxonomy), στις κοινωνικές επιστήμες ως τυπολογία (typology), ενώ στην θεωρία γράφων ως τμηματοποίηση (partition). Τέλος στο πεδίο της εξόρυξης δεδομένων ονομάζεται και μη κατευθυνόμενη ανακάλυψη γνώσης (undirected knowledge discovery) [43].

Αναφέραμε παραπάνω ότι ο άνθρωπος είναι από την φύση του πάρα πολύ ικανός στην δημιουργία κατηγοριοποιήσεων, στην κατάταξη οντοτήτων σε διαφορετικές ομάδες. Σε τι μας χρησιμεύει λοιπόν η ανάλυση της ομαδοποίησης;

Αρχικά, θα πρέπει να αναφέρουμε, ότι ένα πρόγραμμα ομαδοποίησης μπορεί να χρησιμοποιήσει ένα συγκεκριμένο αντικειμενικό κριτήριο κατάταξης πολύ πιο συνειδητά απ’ ότι ένας άνθρωπος. Οι άνθρωποι είναι εξαιρετικοί στην ανακάλυψη ομάδων σε δυο ή και σε τρεις διαστάσεις, αλλά διαφορετικά άτομα δεν θα αναγνωρίσουν πάντα τις ίδιες ομάδες στα ίδια σύνολα δεδομένων. Το μέτρο ομοιότητας που θα χρησιμοποιηθεί από τα διαφορετικά άτομα εξαρτάται από τον εκπαιδευτικό και πολιτιστικό περίγυρο του ατόμου. Έτσι είναι πολύ συχνό, δυο διαφορετικά άτομα να αναγνωρίσουν διαφορετικές ομάδες στο ίδιο σύνολο δεδομένων, ιδιαίτερα εάν οι ομάδες δεν είναι καλά διαχωρισμένες μεταξύ τους. Ως

ένα ακραίο παράδειγμα θα αναφέρουμε την κατάταξη ενός αγράμματου χωρικού όταν του δόθηκε ένα σύνολο από τρία δέντρα και μία τσάπα. Ο άνθρωπος αυτός κατέταξε την τσάπα μαζί με το ένα από τα δέντρα σε ένα σύνολο και τα άλλα δύο δέντρα σε ένα άλλο σύνολο, δικαιολογώντας αυτήν του την επιλογή λέγοντας ότι νοιώθει πιο κοντά σ' αυτό το δέντρο και την τσάπα αφού αυτά γνωρίζει, ενώ δεν νοιώθει οικεία με τα άλλα δυο δέντρα!

Κατά δεύτερο λόγο, αλλά όχι και λιγότερο σημαντικό, θα πρέπει να αναφέρουμε ότι ένας υπολογιστής μπορεί ανακαλύψει ομάδες πολύ πιο γρήγορα σ' ένα σύνολο δεδομένων απ' ότι ένας άνθρωπος, ειδικά εάν οι οντότητες που υπάρχουν στο σύνολο των δεδομένων μας, χαρακτηρίζονται από έναν μεγάλο αριθμό χαρακτηριστικών. Η ταχύτητα, η αξιοπιστία και η συνέπεια με την οποία ένας αλγόριθμος ομαδοποίησης θα ανακαλύψει ομάδες σε ένα τεράστιο σύνολο από δεδομένα είναι ένα ισχυρό κίνητρο για να τον χρησιμοποιήσουμε. Ένας αλγόριθμος ομαδοποίησης θα απελευθερώσει τον μηχανικό δεδομένων από την κουραστική δουλειά της ομαδοποίησης έτσι ώστε να μπορέσει αυτός να ξοδέψει περισσότερο χρόνο στην ανάλυση των ομάδων που θα παραχθούν.

Η ομαδοποίηση είναι επίσης πάρα πολύ χρήσιμη στην υλοποίηση της τεχνικής του διαίρει και βασίλευε (divide and conquer) ώστε να ελαττωθεί η πολυπλοκότητα διάφορων αλγορίθμων της θεωρίας αποφάσεων στην αναγνώριση προτύπων. Για παράδειγμα, η τεχνική του πλησιέστερου γείτονα (nearest neighbor) είναι μια πολύ διαδεδομένη τεχνική στην αναγνώριση προτύπων [39]. Μερικές φορές όμως είναι πολύ χρονοβόρο το να βρεις τον κοντινότερο γείτονα ενός προτύπου, ειδικά εάν τον σύνολο όλων των προτύπων είναι τεράστιο. Ο Fukunga και ο Narendra [50] χρησιμοποίησαν τον γνωστό αλγόριθμο ομαδοποίησης ISODATA, για να αποσυνθέσουν τα πρότυπα και έπειτα χρησιμοποιώντας την μέθοδο Branch an Bound κατασκεύασαν έναν γρήγορο αλγόριθμο υπολογισμού του πλησιέστερου γείτονα. Παρόμοια ο Fukunga και ο Short [51] χρησιμοποίησαν την ομαδοποίηση στο problem localization, όπου ένας απλός κανόνας απόφασης μπορεί να κατασκευαστεί στις τοπικές περιοχές, σε ομάδες δηλαδή του χώρου των προτύπων. Και οι εφαρμογές της ομαδοποίησης ολοένα και αυξάνονται.

Η ανάλυση ομαδοποίησης είναι παιδί της ανάπτυξης της τεχνολογίας των υπολογιστών. Η μεγάλη ανάπτυξη των αλγορίθμων ομαδοποίησης έγινε τις τελευταίες τέσσερις δεκαετίες, κυρίως μετά την δημοσίευση του βιβλίου των Robert Sokal και Peter Sneath το 1963 [1]. Μετά από αυτό, άρχισε η μεγάλη αύξηση του πλήθους των δημοσιευμένων εφαρμογών της ομαδοποίησης σε πολλά, διαφορετικά μεταξύ τους επιστημονικά πεδία. Δύο ήταν οι κύριοι λόγοι που αναπτύχθηκε τόσο πολύ η βιβλιογραφία γύρω από αυτό το επιστημονικό πεδίο [1, 34, 121]:

1. Η ανάπτυξη υπολογιστών μεγάλης υπολογιστικής ισχύς. Οι μέθοδοι ομαδοποίησης είναι εξαιρετικά δύσκαμπτες, υπολογιστικά χρονοβόρες και δύσκολες ειδικά όταν εφαρμόζονται σε μεγάλες βάσεις δεδομένων. Η ανάπτυξη λοιπόν ισχυρών υπολογιστικών μηχανών μπόρεσε να αμβλύνει κάπως αυτά προβλήματα και να δώσει μια άλλη ώθηση σ' αυτό το επιστημονικό πεδίο.

2. Η θεμελιώδης σημασία της ταξινόμησης σε όλους σχεδόν τους κλάδους της επιστήμης. Οι ταξινόμηση είναι βασική για την κατασκευή της περιοχής

έρευνας κάθε επιστήμης αφού περιέχει τις κυριότερες έννοιες που χρησιμοποιούνται στον εκάστοτε κλάδο της επιστήμης. Παραδείγματος χάριν, η ταξινόμηση των ασθενειών παρέχει την δομική βάση της επιστήμης της ιατρικής.

1.2 Τι είναι ομάδα: Ορισμοί.

Ας προσπαθήσουμε τώρα να δώσουμε έναν ορισμό για το τι είναι *Ομάδα* (Cluster). Θα μπορούσαμε να πούμε ότι Ομάδα είναι μια συλλογή από όμοια αντικείμενα που έχουν μαζευτεί ή συγκεντρωθεί μαζί. Ο Everitt [42] δίνει μεταξύ άλλων τους εξής ορισμούς της Ομάδας:

- Ομάδα είναι το σύνολο των οντοτήτων που είναι *όμοιες*, και οντότητες από διαφορετικά σύνολα δεν είναι όμοιες.

- Ομάδα είναι μια συσσωμάτωση σημείων του πεδίου αναφοράς τέτοια ώστε η απόσταση κάθε δυο σημείων της Ομάδας να είναι μικρότερη από την απόσταση μεταξύ κάθε σημείου της Ομάδας και οποιουδήποτε άλλου σημείου.

- Οι ομάδες μπορούν να περιγραφούν ως συνδεδεμένες περιοχές ενός πολυδιάστατου χώρου, που περιέχουν σχετικά υψηλή πυκνότητα σημείων και που είναι χωρισμένες μεταξύ τους με περιοχές, η πυκνότητα των οποίων είναι σχετικά μικρότερη.

Οι δυο τελευταίοι ορισμοί υποθέτουν ότι τα αντικείμενα της Ομάδας παριστάνονται ως σημεία ενός μετρικού χώρου, πράγμα που δεν είναι απαραίτητο να ισχύει σε όλες της εφαρμογές όπως θα δούμε και παρακάτω.

Μπορούμε πολύ εύκολα να αναγνωρίσουμε μια Ομάδα όταν την δούμε στο επίπεδο αλλά δεν είναι ξεκάθαρο το πως το κάνουμε. Ενώ είναι σχετικά εύκολο να δώσουμε έναν λειτουργικό ορισμό για την Ομάδα εντούτοις δεν είναι και τόσο εύκολο να δώσουμε έναν ορισμό που θα μας αποκαλύπτει τον τρόπο παραγωγής μιας Ομάδας. Αυτό οφείλεται στο γεγονός ότι τα αντικείμενα μπορούν να κατηγοριοποιηθούν με πολλούς έγκυρους τρόπους ανάλογα με αυτό που έχουμε κάθε φορά κατά νου. Για να πάμε το πρόβλημα ένα βήμα παρακάτω, θα πρέπει να αναφέρουμε ότι σε μερικές περιπτώσεις οι Ομάδες μπορούν να αλλάζουν με τον χρόνο, όπως παραδείγματος χάριν στην περίπτωση ομάδων αστέρων[33]. Πολλές φορές οι ομάδες εξαρτώνται και από την λεπτομέρεια με την οποία κοιτάμε τα δεδομένα μας. Η Εικόνα 1.1 παρουσιάζει την παραπάνω ιδέα με σημεία στο δισδιάστατο επίπεδο. Αν παρατηρήσουμε την εικόνα με όχι μεγάλη λεπτομέρεια θα διακρίνουμε 4 ομάδες. Αν τώρα παρατηρήσουμε τα δεδομένα μας με μεγαλύτερη ακρίβεια θα τα χωρίζαμε σε 12 ομάδες. Βλέποντας λοιπόν τα δεδομένα υπό διαφορετικές κλίμακες μας βοηθάει να κατανοήσουμε καλύτερα την δομή τους. Έτσι ένα από τα κρίσιμα προβλήματα στην εύρεση των Ομάδων είναι να καθορίσουμε τι είναι εγγύτητα και να βρούμε έναν τρόπο να την μετρήσουμε.

Εικόνα 1.1: Παραδείγματα ομάδων στο επίπεδο

Θα επιμείνουμε λίγο ακόμα στην εξερεύνηση του τι είναι ομάδα, αφού είναι ένα από τα πιο σημαντικά αλλά και από λιγότερο αποσαφηνισμένα αντικείμενα που πραγματευόμαστε. Το γεγονός αυτό οφείλεται κατά πολύ στην έλλειψη ενός και μοναδικού ορισμού του τι είναι ομάδα. Παραπάνω είδαμε δυο ορισμούς. Ας παραθέσουμε τώρα και μερικούς ακόμα.

Ομάδα

- Είναι ένα σύνολο από συνεχόμενα στοιχεία ενός στατιστικού πληθυσμού, για παράδειγμα το σύνολο από ανθρώπους που κατοικούν σε ένα μόνο σπίτι, μια συνεχή ροή παρατηρήσεων σε μια ταξινομημένη σειρά και ένα σύνολο από γειτονικά κομμάτια σε ένα πεδίο.(Kendall και Buckland Dictionary of Statistical Terms).
- Είναι ένα σύνολο περισσότερο όμοιων μεταξύ τους οντοτήτων σε σχέση με οντότητες από άλλες ομάδες.
- Είναι ένα υποσύνολο από οντότητες οι οποίες μπορούν να χρησιμοποιηθούν κατά μια έννοια ως ισοδύναμες.(Wallace και Boulton 1968)

Το κοινό χαρακτηριστικό όλων αυτών των ορισμών είναι ότι χρησιμοποιούν τις έννοιες της ομοιότητας, της απόστασης και της ταύτισης χωρίς να ορίζονται εκ των προτέρων αυτές οι έννοιες. Για τον λόγο αυτό ο Bonner (1964) πρότεινε ως βασικό κριτήριο για τον καθορισμό τέτοιων εννοιών, όπως ομάδα και ομοιότητα, να είναι στη κρίση του χρήστη[1, 34, 42].

Παρά το γεγονός όμως πως δεν υπάρχει ένας κοινά αποδεχτός ορισμός της ομάδας, εντούτοις θα πρέπει οι ομάδες να έχουν κάποιες κοινές ιδιότητες. Αυτές

περιγράφονται από τους Sneath και Sokal (1973). Οι πιο σημαντικές από αυτές τις ιδιότητες αναφέρονται στο [1] και παρατίθενται και εδώ:

Πυκνότητα (Density): είναι εκείνη η ιδιότητα που καθορίζει την ομάδα ως ένα σχετικά παχύ σμήνος από σημεία σε έναν χώρο, συγκρινόμενη με άλλες περιοχές του χώρου, που μπορεί να έχουν λιγότερα αν όχι καθόλου σημεία. Δεν υπάρχει απόλυτο μέτρο πυκνότητας.

- **Διασπορά (Variance):** είναι ο βαθμός διασκόρπισης των σημείων σε σχέση με το κέντρο της ομάδας. Η ιδιότητα της ομάδας θα μπορούσε να θεωρηθεί ως η σχετική εγγύτητα των σημείων στον χώρο. Επομένως οι ομάδες χαρακτηρίζονται ως «σφικτές» ("tight") όταν όλα τα σημεία της ομάδας είναι συγκεντρωμένα κοντά στο κέντρο ή ως «χαλαρές» ("loose") όταν τα σημεία της ομάδας είναι διασκορπισμένα σε σχέση με το κέντρο.
- **Διάσταση (Dimension):** είναι μια ιδιότητα αρκετά συγγενής με την διασπορά. Αν μια ομάδα αναγνωριστεί τότε μπορεί να μετρηθεί η ακτίνα της. Αυτή η ιδιότητα χρησιμοποιείται μόνο για ομάδες που σχηματίζουν υπερσφαίρες σε έναν πολυδιάστατο χώρο που ορίζεται από τις μεταβλητές.
- **Σχήμα (Shape):** είναι ο σχηματισμός που έχουν τα σημεία στον χώρο. Υπάρχουν πολλών ειδών σχήματα όπως υπερσφαιρικά, ελλειψοειδή, επιμήκη… Αν οι ομάδες σχηματίζονται με τέτοιον τρόπο ώστε η έννοια της διαμέτρου ή της ακτίνας να μην έχει νόημα τότε μπορεί να υπολογιστεί η έννοια της συνεκτικότητας (connectivity) των σημείων της ομάδας, η οποία είναι ένα σχετικό μέτρο της απόστασης μεταξύ αυτών.
- **Διαχωρισμός (Separation):** είναι ο βαθμός υπερκάλυψης των ομάδων. Οι ομάδες μπορεί να υπερκαλύπτονται (overlap) ή να κείνται χωριστά στον χώρο. Παραδείγματος χάριν οι ομάδες μπορεί να είναι η μία κοντά στην άλλη χωρίς να έχουν πολύ ξεκάθαρα όρια ή να βρίσκονται σχετικά μακριά μεταξύ τους με ξεκαθαρισμένα όρια.

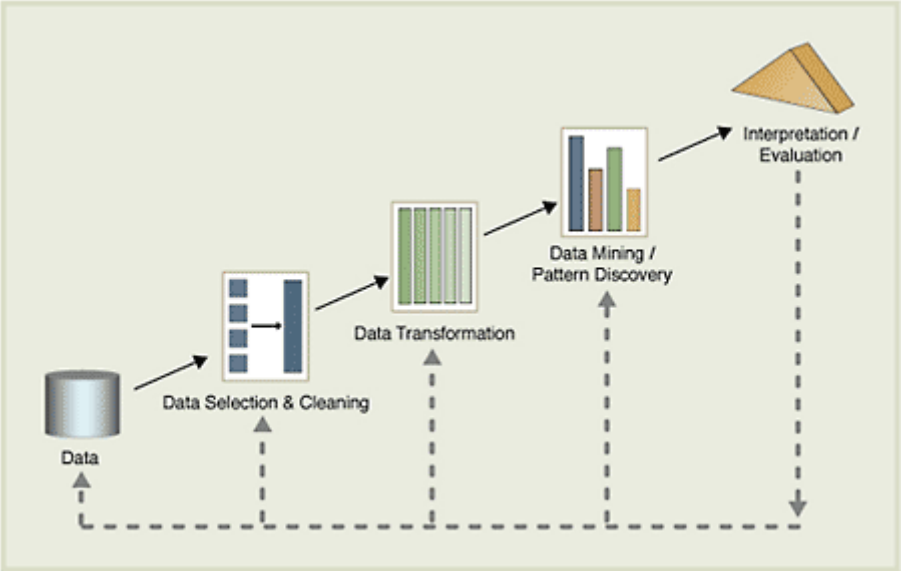
Λαμβάνοντας υπ' όψιν του παραπάνω ορισμούς, ο Everrit έδωσε τον ακόλουθο ορισμό για το τι είναι ομάδα.

> **Ομάδα (cluster):** είναι μια συνεχής περιοχή του χώρου που περιέχει μια σχετικά υψηλή πυκνότητα από σημεία και χωρίζεται από άλλες περιοχές με σχετικά μεγάλη πυκνότητα σημείων, με περιοχές που περιέχουν χαμηλή πυκνότητα σημείων.

Έτσι, οι Ομάδες που περιγράφονται με αυτόν τον τρόπο ονομάζονται **φυσικές ομάδες (natural clusters)**. Ένα από τα πλεονεκτήματα αυτού του τρόπου ορισμού των ομάδων είναι ότι δεν περιορίζεται το σχήμα των ομάδων όπως συμβαίνει με άλλους ορισμούς [42]. Με άλλα λόγια οι φυσικές ομάδες δεν επιβάλουν κανέναν εκ των προτέρων περιορισμό στην δομή της ομάδας [34].

1.3 Διαδικασία Ομαδοποίησης

Θα ασχοληθούμε τώρα με τα βήματα που πρέπει να ακολουθηθούν με σκοπό την δημιουργία μιας ομαδοποίησης. Τα βήματα αυτά φαίνονται στην Εικόνα 1.2 και συνοψίζονται στα [1, 56, 136].



Εικόνα 1.2: Τα βασικά βήματα της διαδικασίας εξόρυξης δεδομένων.

1. **Επιλογή χαρακτηριστικών.** Από όλο το σύνολο των δεδομένων επιλέγονται εκείνες οι οντότητες οι οποίες θα ομαδοποιηθούν, δηλαδή το δείγμα. Θα πρέπει καταρχάς να επιλεγούν κατάλληλα, ποια χαρακτηριστικά θα συμμετέχουν στην διαδικασία της ομαδοποίησης. Θα πρέπει να διαλεχτούν εκείνα που να κωδικοποιούν κατάλληλα την πληροφορία που αναφέρεται στο πρόβλημα που θέλουμε να λύσουμε. Αυτά τα δεδομένα συνήθως παριστάνονται ως d - διάστατα διανύσματα χαρακτηριστικών. Αν είναι αναγκαίο τα δεδομένα που θα επιλεγούν μπορεί να υποστούν μια διαδικασία προεπεξεργασίας όπως φαίνεται στην Εικόνα 1.2.

2. **Επιλογή αλγορίθμου ομαδοποίησης.** Σ' αυτό το βήμα επιλέγεται το αλγοριθμικό σχήμα που θα χρησιμοποιηθεί για την ανακάλυψη της δομής που περιέχεται στα δεδομένα. Ουσιαστικά αυτό σημαίνει την επιλογή ενός μέτρου εγγύτητας και ενός αλγορίθμου ομαδοποίησης.

    a. **Επιλογή του μέτρου εγγύτητας.** Το μέτρο εγγύτητας είναι ένα μέτρο που ποσοτικοποιεί την ομοιότητα ή την ανομοιότητα μεταξύ δύο, κάθε φορά, οντοτήτων των δεδομένων.

    b. **Επιλογή του κριτηρίου ομαδοποίησης.** Η επιλογή του κριτηρίου ομαδοποίησης εξαρτάται από τον τύπο των δεδομένων που έχουμε (συνεχή, δυαδικά, διακριτά), καθώς και από την δομή των ομάδων που υπάρχουν μέσα στα δεδομένα μας. Παραδείγματος χάριν, συμπαγείς ομάδες μπορεί να είναι ευαίσθητες σύμφωνα με ένα κριτήριο εγγύτητας, ενώ επιμήκεις ομάδες να είναι ευαίσθητες σύμφωνα με

κάποιο άλλο κριτήριο. Το κριτήριο ομαδοποίησης εκφράζεται μέσω μια συνάρτησης κόστους ή μέσω ενός διαφορετικού κανόνα.

Οι αλγόριθμοι ομαδοποίησης μπορούν να χωριστούν σε δύο μεγάλες κατηγορίες:

- **Παραμετρικές (parametric) μέθοδοι ομαδοποίησης:** Στις παραμετρικές μεθόδους ομαδοποίησης καθορίζεται εκ των προτέρων κάποιο κριτήριο ομαδοποίησης και τα δοσμένα δείγματα ομαδοποιούνται σε έναν προκαθορισμένο αριθμό ομάδων έτσι ώστε να βελτιστοποιείται το κριτήριο που έχει επιλεγεί. Σ' αυτές τις μεθόδους θα πρέπει οπωσδήποτε να γνωρίζουμε εκ των προτέρων το πλήθος των ομάδων που θέλουμε. Το σχήμα των ομάδων καθορίζεται από το επιλεγμένο κριτήριο. Μια άλλη προσέγγιση των παραμετρικών μεθόδων ομαδοποίησης είναι η υπόθεση ενός μαθηματικού τύπου που εκφράζει την κατανομή των δεδομένων μας. Ένα από τα πιο τυπικά παραδείγματα είναι το άθροισμα κανονικών κατανομών. Σε μια τέτοια περίπτωση το πρόβλημα έγκειται στην εύρεση εκείνων των παραμετρικών τιμών που εκφράζουν καλύτερα τα δεδομένα. [52]

- **Μη παραμετρικές (non parametric) μέθοδοι ομαδοποίησης:** Οι μη παραμετρικές μέθοδοι ομαδοποίησης δεν προϋποθέτουν εκ των προτέρων τη γνώση του πλήθους των ομάδων ή της υπονοούμενης δομής των ομάδων. Στην κατηγορία αυτή ανήκουν μέθοδοι οι οποίες δεν χρησιμοποιούν κάποιο κριτήριο ομαδοποίησης ούτε και προϋποθέτουν κάποιον μαθηματικό τύπο για την κατανομή των δεδομένων.

3. **Αξιοπιστία.** Οι ομάδες που θα προκύψουν θα πρέπει να αξιολογηθούν για την ορθότητά (correctness) τους, καθώς οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται κατά κύριο λόγο όταν έχουμε λίγη ή καθόλου εκ των προτέρων γνώση. Έτσι αναζητούνται συνήθως εκείνες οι τεχνικές που θα μπορέσουν να διακρίνουν τις «αληθινές» από τις «ψεύτικες» δομές. Όλες αυτές οι τεχνικές ονομάζονται **τεχνικές αξιοπιστίας (validation techniques)** [1, 38, 56, 136]. Και υπάρχουν τρεις κυρίως μέθοδοι για τον έλεγχο της αξιοπιστίας των ομάδων.

   a. **Εξωτερικά κριτήρια (External Criteria).** Οι έλεγχοι σημαντικότητας (significant tests) που εκτελούνται βασίζονται σε εξωτερικούς παράγοντες.

   b. **Εσωτερικά κριτήρια (Internal Criteria).** Οι έλεγχοι σημαντικότητας που εκτελούνται βασίζονται στης μεταβλητές των χαρακτηριστικών των οντοτήτων που χρησιμοποιήθηκαν.

   c. **Σχετικά κριτήρια (Relative Criteria).** Η ομαδοποίηση συγκρίνεται με αποτελέσματα άλλων σχημάτων ομαδοποίησης που προκύπτουν με την εκτέλεση του ίδιου αλγόριθμου ομαδοποίησης χρησιμοποιώντας όμως διαφορετικές παραμέτρους.

4. **Ερμηνεία των αποτελεσμάτων.** Μετά την διαδικασία ανακάλυψης των δομών των ομάδων που υπάρχουν μέσα στα δεδομένα, θα πρέπει να ερμηνευτούν οι εξαγόμενες ομαδοποιήσεις για να λάβουμε τα τελικά συμπεράσματα από την διαδικασία. Η ερμηνεία των αποτελεσμάτων γίνεται συνήθως από τους επιστήμονες του εκάστοτε τομέα οι οποίοι χρησιμοποίησαν την διαδικασία ομαδοποίησης.

1.4 Εφαρμογές της Ομαδοποίησης

Αναφερθήκαμε γενικά σε ποιες περιοχές της επιστήμης χρησιμοποιείται ως εργαλείο η ομαδοποίηση. Εδώ θα εξετάσουμε κάπως αναλυτικότερα τις διάφορες εφαρμογές της ομαδοποίησης δεδομένων.

Η ομαδοποίηση χρησιμοποιείται, εκτός από της προαναφερθείσες εφαρμογές στην στατιστική ανάλυση δεδομένων, στην επεξεργασία εικόνας (image possessing), στην ανάκτηση πληροφοριών (data retrieval), στην επιχειρησιακή έρευνα (marketing research) – για την ομαδοποίηση παραδείγματος χάριν πελατών σε ομάδες, στην χημεία (chemistry) – για την ομαδοποίηση στοιχείων σε ομάδες σύμφωνα με τις ιδιότητες τους, στην εξόρυξη δεδομένων (data mining), στην αναγνώριση προτύπων (pattern recognition) και στον καθαρισμό δεδομένων (data cleansing). [1, 4, 12, 34, 42, 52, 59, 64, 63, 136].

Κάποιες επιπλέον εφαρμογές της ομαδοποίησης είναι και οι παρακάτω:.

1. **Μείωση Δεδομένων (Data redaction):** Το πρόβλημα του μεγάλου όγκου των παρατηρήσεων εμφανίζεται σε πολλά ερευνητικά πεδία. Αυτό είναι ένα αρκετά δυσεπίλυτο πρόβλημα εάν τα δεδομένα δεν οργανωθούν και ταξινομηθούν σε πιο εύχρηστες ομάδες, οι οποίες θα μπορούσαν να χρησιμοποιηθούν ως μονάδες. Οι τεχνικές της ομαδοποίησης μπορούν να χρησιμοποιηθούν για την μείωση των δεδομένων. Όλη η προσπάθεια λοιπόν είναι στο να μειώσουμε την πληροφορία που μας παρέχουν τα πλήθους N δεδομένα, στην πληροφορία που μας δίνουν τα $g < N$ δεδομένα χωρίς σημαντικές απώλειες πληροφορίας. Με άλλα λόγια, αναζητούμε εκείνη την απλοποίηση των δεδομένων που θα μας αποφέρει την μικρότερη απώλεια πληροφορίας. Παραδείγματος χάριν, στην μεταφορά της πληροφορίας καθορίζεται ένας αντιπρόσωπος για κάθε μια από της ομάδες στις οποίες έχουμε χωρίσει τα δεδομένα. Έτσι αντί να μεταφέρουμε όλο το πλήθος των δεδομένων, αρκεί να μεταφέρουμε μόνο έναν κωδικό που αντιστοιχεί στον αντιπρόσωπο κάθε ομάδας. Με αυτόν τον τρόπο μπορούμε να επιτύχουμε συμπίεση των δεδομένων.
2. **Έλεγχος υποθέσεων (hypothesis testing):** Η ομαδοποίηση χρησιμοποιείται ακόμα και για τον έλεγχο της αξιοπιστίας υποθέσεων. Ας θεωρήσουμε παραδείγματος χάριν την ακόλουθη υπόθεση: «Οι μεγάλες εταιρίες επενδύουν στο εξωτερικό». Μπορούμε να εφαρμόσουμε την ομαδοποίηση σε ένα μεγάλο αντιπροσωπευτικό δείγμα εταιριών. Κάθε εταιρία μπορεί να παριστάνεται από το μέγεθος της, τις δραστηριότητες της στο εξωτερικό και την ικανότητά της να διεκπεραιώνει επιτυχώς τα ληφθέντα έργα εφαρμοσμένης έρευνας. Αν η εφαρμογή της ομαδοποίησης μας επιστρέψει μια ομάδα που περιέχει τις μεγάλες εταιρίες που επιχειρούν στο εξωτερικό αδιαφορώντας για την

ικανότητα της επιτυχούς διεκπεραίωσης του έργου, τότε η παραπάνω υπόθεση είναι αληθής αφού υποστηρίζεται από την ομαδοποίηση.

3. **Παραγωγή υποθέσεων (hypothesis generation):** Μια άλλη εφαρμογή της ομαδοποίησης μπορεί να είναι η παραγωγή (εξαγωγή) υποθέσεων που αφορούν την φύση των δεδομένων. Έτσι η ομαδοποίηση μπορεί να χρησιμοποιηθεί ως όχημα για να προταθεί μια υπόθεση. Παραδείγματος χάριν θα μπορούσαμε να προτείνουμε την υπόθεση «Οι μεγάλες εταιρίες επενδύουν στο εξωτερικό», αν η εφαρμογή της ομαδοποίησης μας δώσει μια τέτοια ομάδα. Εδώ θα πρέπει να σημειώσουμε ότι για τον έλεγχο αυτών των υποθέσεων θα πρέπει οπωσδήποτε να χρησιμοποιηθούν καινούργια δεδομένα και όχι εκείνα τα οποία σχετίζονται με τη παραγωγή της συγκεκριμένης υπόθεσης.
4. **Πρόβλεψη βασισμένη σε ομάδες.** Στην περίπτωση αυτή η ομαδοποίηση εφαρμόζεται σε ένα σύνολο προτύπων και κάθε ομάδα χαρακτηρίζεται από τα χαρακτηριστικά των προτύπων τα οποία ανήκουν σε κάθε ομάδα. Αν ένα άγνωστο πρότυπο εμφανιστεί, τότε αποφασίζεται σε ποια ομάδα είναι πιο πιθανό να ανήκει εξετάζοντας την ομοιότητα του με κάποιον προκαθορισμένο αντιπρόσωπο της ομάδας. Έτσι το πρότυπο αυτό αναγνωρίζεται και χαρακτηρίζεται σύμφωνα με τα χαρακτηριστικά της ομάδας. Ας υποθέσουμε πως σε ένα σύνολο από ασθενείς, μολυσμένους από την ίδια αρρώστια, εφαρμόζουμε την ομαδοποίηση. Το αποτέλεσμα θα είναι ένα σύνολο ομάδων ασθενών που θα χαρακτηρίζονται από τις αντιδράσεις των ασθενών σε συγκεκριμένες φαρμακευτικές αγωγές. Αν τώρα ένας νέος ασθενής εμφανιστεί τότε θα πρέπει να αναγνωριστεί η πιο πιθανή ομάδα στην οποία θα μπορούσε να ανήκει και με αυτόν τον τρόπο αποφασίζεται ποια φαρμακευτική αγωγή θα του χορηγηθεί.

1.5 Ομαδοποίηση στην εξόρυξη δεδομένων

Το συνεχές αυξανόμενο πλήθος των δεδομένων και η συνεπακόλουθη αύξηση του μεγέθους των βάσεων δεδομένων, όπου αυτά είναι αποθηκευμένα, κάνουν επιτατική την ανάγκη για ανάπτυξη νέων εργαλείων και νέων τεχνικών για την ανάλυση και την εξαγωγή (εξόρυξη) γνώσης από αυτά. Η εποχή της πληροφορίας στην οποία ζούμε και το τεράστιο μέγεθος των αποθηκευμένων πληροφοριών που συνεπάγεται, κάνουν τις παραδοσιακές στατιστικές τεχνικές ανάλυσης των δεδομένων ανεπαρκείς και χρονοβόρες και για αυτό νέες τεχνικές για την ανακάλυψη γνώσης θα πρέπει να βοηθούν τους ανθρώπους στην ανάλυση τεράστιων ποσοτήτων πληροφορίας. Αυτές οι τεχνικές και τα εργαλεία είναι το αντικείμενο ενός νέου πεδίου που καλείται ανακάλυψη γνώσης σε βάσεις δεδομένων (knowledge discovery in databases - KDD) [43].

Ένας ορισμός για το τι είναι ανακάλυψη γνώσης σε βάσεις δεδομένων που προτάθηκε από τους Fayyad, Piatesky – Shapino, Smyth και Uthurusamy το 1996 είναι ο εξής:

> **Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases)** είναι μια μη – τετριμμένη διαδικασία αναγνώρισης έγκυρων νέων, ενδεχομένως χρήσιμων και τελικά κατανοήσιμων προτύπων στα δεδομένα [43].

Υπάρχει μια σημαντική διαφορά ανάμεσα στους όρους KDD και Data mining. Ο όρος KDD αναφέρεται στην ολική διαδικασία ανακάλυψης γνώσης από τα δεδομένα, σε αντίθεση με τον όρο Data Mining ο οποίος αναφέρεται κυρίως στην εφαρμογή των αλγορίθμων για την εξαγωγή της γνώσης από τα δεδομένα χωρίς να αναφέρεται και στα άλλα στάδια της KDD διαδικασίας. Ένας ορισμός που δίνεται στο [43] για το τι είναι εξόρυξη δεδομένων δίνεται παρακάτω.

> **Εξόρυξη Δεδομένων (Data Mining)** είναι ένα βήμα της KDD διαδικασίας που αναφέρεται στην ανακάλυψη και εφαρμογή αλγορίθμων για την ανάλυση δεδομένων, οι οποίοι κάτω από υπολογιστικά αποδεκτούς περιορισμούς, παράγουν μια συγκεκριμένη απαρίθμηση (enumeration) των προτύπων στα δεδομένα [43].

Δύο είναι οι βασικοί στόχοι της εξόρυξης δεδομένων: Η πρόβλεψη (prediction) και η περιγραφή (description). Η πρόβλεψη εμπλέκει κάποιες μεταβλητές ή κάποια πεδία της βάσης δεδομένων έτσι ώστε να προβλεφθούν άγνωστες ή μελλοντικές τιμές ή και άλλες μεταβλητές ενδιαφέροντος. Η περιγραφή από την άλλη πλευρά εστιάζει κυρίως στην ανακάλυψη προτύπων στα δεδομένα τα οποία εύκολα μπορούν να ερμηνευτούν και να τα περιγράψουν [43].

Οι κυριότερες λειτουργίες στην εξόρυξη δεδομένων είναι η ταξινόμηση (classification) και η ομαδοποίηση (clustering). Σκοπός της ταξινόμησης είναι η παραγωγή κανόνων από μεγάλες σχεσιακές βάσεις δεδομένων που να μπορούν να ταξινομήσουν καινούργια άγνωστα δεδομένα σε προκαθορισμένες κλάσεις οι οποίες να περιγράφονται από ένα σύνολο χαρακτηριστικών. Η εξαγωγή των κανόνων γίνεται με την χρήση μεθόδων μάθησης με επίβλεψη (supervised learning methods) [43]. Από τους πιο γνωστούς αλγόριθμους κατάταξης είναι οι ID3, C45, CN2 και CART. Το βασικό μειονέκτημα αυτών των μεθόδων είναι η υπολογιστική δυσκαμψία και η συνεπακόλουθη δαπάνη σε υπολογιστικό χρόνο. Από την άλλη πλευρά όμως, η εξόρυξη δεδομένων απαιτεί την επεξεργασία μεγάλων ποσοτήτων δεδομένων με ικανοποιητική ακρίβεια. Για αυτό και η πρόκληση είναι η ανάπτυξη μεθόδων που θα επεξεργάζονται τεράστια ποσά δεδομένων που υπερβαίνουν κατά πολύ την μνήμη ενός επεξεργαστή, με αποτελεσματικό και χρονικά ικανοποιητικό τρόπο. Σ' αυτήν την κατεύθυνση έχουν αναπτυχθεί σύγχρονες μέθοδοι όπως η στρατηγική meta – classifying ή meta – learning [19, 24, 23].

Η δεύτερη βασική λειτουργία στην εξόρυξη δεδομένων είναι η ομαδοποίηση των εγγραφών μια βάσης δεδομένων σε υποομάδες (clustering). Η ομαδοποίηση είναι μια περιγραφική λειτουργία που σκοπό έχει την ανίχνευση ενός πεπερασμένου πλήθους ομάδων ή κατηγοριών (clusters) που περιέχονται στα δεδομένα [43]. Όπως έχει τονιστεί και παραπάνω πολλές διαδικασίες εξόρυξης προϋποθέτουν ένα προκαταταγμένο σύνολο δεδομένων εκπαίδευσης και επιχειρούν την ανάπτυξη ενός αξιόπιστου μοντέλου ικανού να προβλέπει την κατάταξη ενός καινούργιου μη καταταγμένου αντικειμένου. Στην ομαδοποίηση δεν υπάρχουν ούτε εκ των προτέρων ταξινομημένα δεδομένα αλλά ούτε και διαχωρισμοί μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Εδώ αναζητούνται όμοιες ομάδες εγγραφών (clusters) με την ελπίδα αυτές να έχουν όμοιες ιδιότητες και να περιγράφουν τα δεδομένα μας [14].

Θα πρέπει να αναφέρουμε εδώ ότι οι αλγόριθμοι ομαδοποίησης διαχειρίζονται μεγάλο πλήθος δεδομένων και απαιτούν έναν αρκετά μεγάλο αριθμό υπολογισμών. Συνεπώς οι η πολυπλοκότητά τους εξαρτάται από το πλήθος των δεδομένων που επεξεργάζεται ο κάθε αλγόριθμος. Από την άλλη, το τεράστιο μέγεθος των δεδομένων που αποθηκεύονται στις βάσεις δεδομένων ωθεί σήμερα το ερευνητικό ενδιαφέρον κυρίως σε αλγορίθμους ομαδοποίησης, που μπορούν αν χειριστούν δεδομένα πολύ μεγαλύτερα από την κύρια μνήμη ενός επεξεργαστή. Για την αντιμετώπιση αυτού του προβλήματος έχει προταθεί μια επαναληπτική διαδικασία, που βασίζεται στην τμηματοποίηση του συνόλου των δεδομένων σε υποσύνολα. Στην πρώτη φάση, κάθε υποσύνολο δίνεται ως είσοδός σε κάθε έναν αλγόριθμο ομαδοποίησης. Κατά την δεύτερη φάση, τα μερικά αποτελέσματα σχηματίζουν ένα σύνολο δεδομένων το οποίο τμηματοποιείται σε ομάδες τις καλούμενες και μετα – ομάδες (meta – clusters). Κάτω από ορισμένες συνθήκες οι ομάδες αυτές αποτελούν τις επιθυμητές ομάδες [18].

2 Βασικές έννοιες ομαδοποίησης

2.1 Εισαγωγή

Όπως αναφέραμε και παραπάνω η βασική αρχή της ομαδοποίησης είναι η ανακάλυψη των φυσικών ομάδων που προϋπάρχουν στα δεδομένα, έτσι ώστε τα αντικείμενα κάθε ομάδας να είναι όμοια μεταξύ τους ενώ τα αντικείμενα διαφορετικών ομάδων να είναι ανόμοια.

Θα αναφέρουμε εδώ τους βασικούς όρους που χρησιμοποιούνται στο πεδίο της ομαδοποίησης. Έτσι λοιπόν με τους όρους **περίπτωση** (**case**), **οντότητα** (**entity**), πρότυπο, (**pattern)** αναφέρεται ό,τι πρόκειται να ομαδοποιηθεί, ενώ οι όροι **μεταβλητές** (**variables**), **χαρακτηριστικό** (**attribute** ή **feature** ή **character**) αναφέρονται στις μεταβλητές της κάθε οντότητας. Με τον όρο **πίνακας δεδομένων** (**raw data**) αναφερόμαστε στον N x ρ πίνακα των περιπτώσεων των μεταβλητών τους, προτού λάβει χώρα ο υπολογισμός κάποιου μέτρου ομοιότητας ή ανομοιότητας.

Ο πίνακας αυτός **X** θα έχει την εξής μορφή:

$$X = \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1\rho} \\ \dots & \dots & \dots \\ x_{\mathrm{N}1} & \dots & x_{\mathrm{N}\rho} \end{bmatrix} \quad (2.1)$$

Το x_{ij} είναι η τιμή της j – οστής μεταβλητής της i – οστής περίπτωσης. Ο όρος **πίνακας εγγύτητας** (**proximity matrix**) αναφέρεται στον N x N πίνακα εγγύτητας (ομοιότητας ή ανομοιότητας) των περιπτώσεων που προκύπτει μετά την εφαρμογή κάποιου μέτρου εγγύτητας στον πίνακα δεδομένων.

Θα μπορούσαμε να διατυπώσουμε πιο τυπικά το πρόβλημα της ομαδοποίησης ως εξής:

> Έστω ένα δοσμένο σύνολο δεδομένων X το οποίο αποτελείται από N αντικείμενα, το καθένα αποτελούμενο από ρ μεταβλητές. Το πρόβλημα της ομαδοποίησης έγκειται στην παραγωγή ενός σχήματος κατηγοριοποίησης C, το οποίο θα ομαδοποιεί τα N αντικείμενα σε g μη τεμνόμενες ομάδες (clusters) C_1, C_2, …, C_g, σύμφωνα με κάποιο κριτήριο, έτσι ώστε να ικανοποιούνται οι ακόλουθες τρεις ιδιότητες:
>
> - $C_i \neq 0, i = 1,..., g$ (2.2)
> - $\bigcup_{i=1}^{g} C_i = X$ (2.3)
> - $C_i \bigcap C_j = 0, i \neq j, i, j = 1,..., g$ (2.4)

2.2 Επιλογή χαρακτηριστικών

Η επιλογή των χαρακτηριστικών (μεταβλητών) που θα χρησιμοποιηθούν για να περιγράψουν το πρόβλημά μας, είναι ένα από τα πιο κρίσιμα αλλά και από τα λιγότερο κατανοητά βήματα της διαδικασίας ομαδοποίησης. Το κυριότερο χαρακτηριστικό των δεδομένων είναι η διάστασή τους (dimensionality). Με τον όρο διάσταση εννοούμε το πλήθος των ανεξάρτητων παραμέτρων (μεταβλητών) που απαιτούνται για την περιγραφή των δεδομένων [36].

Τα δεδομένα της ομαδοποίησης είναι ένα σύνολο από N οντότητες οι οποίες περιέχουν ρ χαρακτηριστικά. Η αρχική επιλογή του συνόλου των χαρακτηριστικών που χρησιμοποιείται για την περιγραφή κάθε οντότητας αποτελεί το πλαίσιο αναφοράς μέσα στο οποίο δημιουργούνται οι ομάδες [42].

Το βασικό πρόβλημα είναι να βρεθεί εκείνο το σύνολο των μεταβλητών που αναπαριστά καλύτερα την έννοια της εγγύτητας. Η επιλογή αυτή αντανακλά την κρίση των ερευνητών σχετικά με τους σκοπούς της ομαδοποίησης. Επομένως το ερώτημα που προκύπτει είναι αν η επιλογή των μεταβλητών είναι ορθή, αν δηλαδή οι μεταβλητές είναι σχετικές με τον τύπο της ομαδοποίησης που αναζητείται [42].

2.3 Τύπος και Κλίμακα Χαρακτηριστικών

Τα δεδομένα χαρακτηρίζονται από τον τύπο (type) και την κλίμακα (scale) τους. Ένα χαρακτηριστικό λοιπόν μπορεί να λάβει τιμές από ένα συνεχές διάστημα των πραγματικών αριθμών ή από ένα πεπερασμένο σύνολο αριθμών. Στην πρώτη περίπτωση το χαρακτηριστικό καλείται συνεχές (continuous) ενώ στην δεύτερη περίπτωση το χαρακτηριστικό καλείται διακριτό (discrete) ή πολλαπλών τιμών (multi – valued). Στην περίπτωση που το πεπερασμένο διακριτό σύνολο έχει μόνο δύο στοιχεία, τότε η μεταβλητή καλείται δυαδική (binary) ή διχοτόμος (dichotomous) [36, 136].

Τα χαρακτηριστικά χρησιμοποιούνται ανάλογα με την κλίμακά τους [36, 62, 64, 136]. Θα μπορούσαμε να ξεχωρίσουμε τέσσερις κατηγορίες κλίμακας των χαρακτηριστικών:

1. Nominal. Οι τιμές των χαρακτηριστικών της κλίμακας αυτής χαρακτηρίζουν κωδικοποιημένες καταστάσεις. Παραδείγματος χάριν ένα χαρακτηριστικό που αντιστοιχεί στο φύλο ενός ανθρώπου. Οι πιθανές τιμές του χαρακτηριστικού αυτού είναι δύο: αρσενικό και θηλυκό. Θα μπορούσαμε δηλαδή να κωδικοποιήσουμε την τιμή αρσενικό με τον αριθμό 1 ενώ θα μπορούσαμε να κωδικοποιήσουμε την τιμή θηλυκό στο 0. Είναι φανερό ότι οποιαδήποτε ποσοτική σύγκριση μεταξύ των τιμών ενός nominal χαρακτηριστικού δεν έχει κανένα νόημα. Επίσης είναι εντελώς προφανές η διάταξη των τιμών τέτοιου είδους χαρακτηριστικών κανένα νόημα δεν έχει.

2. Ordinal. Στην κατηγορία αυτή ανήκουν χαρακτηριστικά των οποίων η διάταξη των τιμών τους έχει νόημα. Παραδείγματος χάριν μια μεταβλητή που χαρακτηρίζει την επίδοση ενός μαθητή σε κάποιο μάθημα. Οι πιθανές τιμές του χαρακτηριστικού αυτού είναι 4, 3, 2, 1 και αντιστοιχούν στις βαθμολογίες

«άριστα», «πολύ καλά», «καλά» και «όχι καλά». Προφανώς αυτές οι τιμές μπορούν να διαταχθούν σε μια σειρά, όπως παραπάνω, η οποία να έχει νόημα. Ωστόσο η διαφορά μεταξύ δυο τέτοιων τιμών δεν έχει καμιά ποσοτική σημασία, αφού μια τέτοια σύγκριση δεν έχει νόημα.

3. Interval – scaled. Αν για ένα χαρακτηριστικό, η διαφορά μεταξύ δυο τιμών έχει νόημα ενώ ο λόγος τους δεν έχει, τότε το χαρακτηριστικό αυτό ανήκει στην κατηγορία αυτή. Ένα τυπικό παράδειγμα είναι η μέτρηση της θερμοκρασίας. Αν παραδείγματος χάριν η θερμοκρασία στο Παρίσι είναι 5 βαθμοί Κελσίου ενώ η θερμοκρασία στην Αθήνα είναι 10 βαθμοί Κελσίου τότε έχει νόημα να πούμε ότι η θερμοκρασία στην Αθήνα είναι 5 βαθμούς πιο μεγάλη απ' ότι στο Παρίσι. Από την άλλη μεριά όμως δεν έχει νόημα να πούμε ότι στο Παρίσι κάνει δυο φορές πιο πολύ κρύο απ' ότι στην Αθήνα.

4. Ratio – scaled. Αν ο λόγος μεταξύ δυο τιμών ενός χαρακτηριστικού έχει νόημα τότε το χαρακτηριστικό ανήκει στην κατηγορία αυτή. Ένα παράδειγμα είναι το βάρος ενός ανθρώπου, αφού έχει νόημα να πούμε πως ένα άτομο που ζυγίζει 100 κιλά είναι δυο φορές βαρύτερο από ένα άτομο που ζυγίζει 50 κιλά.

Θα πρέπει να σημειώσουμε εδώ πως κάθε κλίμακα περιέχει όλες τις ιδιότητες της προηγούμενής της. Ένα ratio – scaled χαρακτηριστικό περιλαμβάνει όλες τις ιδιότητες ενός interval – scaled, και αυτό με την σειρά του περιλαμβάνει όλες τις ιδιότητες ενός ordinal χαρακτηριστικού και ούτω καθεξής. Θα μπορούσαμε να πούμε ότι οι παραπάνω χαρακτηρισμοί βρίσκονται σε μια ιεραρχική σχέση με κάθε χαρακτηριστικό να κληροδοτεί τις ιδιότητες του στα επόμενα χαρακτηριστικά.

Ορισμένες φορές στην βιβλιογραφία τα χαρακτηριστικά διακριτού τύπου, (nominal ή ordinal) αναφέρονται και ως κατηγορικά χαρακτηριστικά (categorical) [145, 146]. Αυτήν την ορολογία θα χρησιμοποιούμε και εμείς από δω και περά. Μπορούμε επίσης να αναφερθούμε στα συνεχή, στα διακριτά και στα interval – scaled χαρακτηριστικά ως ποσοτικά (quantitative), ενώ μπορούμε να αναφερόμαστε στα nominal και ordinal χαρακτηριστικά και ως ποσοτικά (qualitative) [64].

2.4 Μέτρα εγγύτητας

2.4.1 Εισαγωγή

Η επιλογή και ο υπολογισμός του μέτρου εγγύτητας (proximity measure) είναι πολύ βασικός για την διαδικασία ομαδοποίησης. Η επιλογή του μέτρου εγγύτητας είναι ένα ιδιαίτερα ενδιαφέρον πρόβλημα της ομαδοποίησης [34]. Είναι εξαιρετικά σημαντικό ποια μέθοδο θα διαλέξουμε για να συγκρίνουμε τα δεδομένα μας. Εδώ θα πούμε μερικά πράγματα για το τι είναι και πως ορίζονται τα μέτρα εγγύτητας και θα περιγράψουμε μερικά υπάρχοντα μέτρα εγγύτητας. Η έμφαση μας θα δοθεί σε μέτρα εγγύτητας κατηγορικών δεδομένων.

Στην πλειοψηφία τους οι τεχνικές ομαδοποίησης αρχίζουν με τον υπολογισμό ενός τετραγωνικού πίνακα ομοιότητας (similarity matrix) μεταξύ των οντοτήτων. Αυτός ο συμμετρικός πίνακας μας δίνει την ομοιότητα ή την ανομοιότητα μεταξύ όλων των οντοτήτων που λαμβάνουν μέρος στην ομαδοποίηση. Πολλές μέθοδοι ομαδοποίησης μπορούν να θεωρηθούν ως προσπάθειες να συνοψιστεί η πληροφορία που

αναφέρεται στις σχέσεις μεταξύ των οντοτήτων και η οποία περιέχεται στον πίνακα ομοιότητας, έτσι ώστε οι σχέσεις αυτές να μπορούν να γίνουν πιο εύκολα κατανοητές και να μπορούν να ερμηνευτούν πιο εύκολα. Είναι φανερό ότι η έξοδος ενός αλγορίθμου ομαδοποίησης, οι ομάδες δηλαδή στις οποίες έχουν χωριστεί τα δεδομένα, θα είναι τόσης σημασίας όσης είναι και οι ομοιότητες και οι αποστάσεις εισόδου που δίνονται από την πίνακα ομοιότητας [42].

Ένα συγγενικό πρόβλημα με αυτό της επιλογής του μέτρου ομοιότητας είναι και η αξιολόγηση (weighting) των μεταβλητών. Η αξιολόγηση είναι η διαδικασία, κατά την οποία δίνονται διαφορετικά βάρη στις μεταβλητές κατά τον προσδιορισμό του μέτρου ομοιότητας, έτσι ώστε να δοθεί μεγαλύτερη ή μικρότερη σημασία σε κάποιες μεταβλητές που θεωρούνται μεγαλύτερης ή μικρότερης σημασίας αντίστοιχα στην ομαδοποίηση. Το ερώτημα που εγείρεται όμως είναι κατά πόσο μπορεί να θεωρηθεί αξιόπιστη η αξιολόγηση των μεταβλητών σε μια διαδικασία ομαδοποίησης. Δυο είναι κυρίως οι λόγοι που μας υπαγορεύουν μια τέτοια αμφιβολία. Πρώτον η των μεταβλητών θέτουν μια υποκειμενική κρίση για το τι είναι και τι δεν είναι σημαντικό και δεύτερον τα βάρη απεικονίζουν μόνο τις υπάρχουσες ομαδοποιήσεις των δεδομένων, διαφορετικά βάρη προφανώς θα δώσουν διαφορετικές ομαδοποιήσεις. Οι τεχνικές της αξιολόγησης από την άλλη πλευρά, χρησιμοποιούνται στα δεδομένα με την ελπίδα να αναδυθούν χρήσιμες ομάδες οι οποίες πριν δεν είχαν εμφανιστεί [42].

2.4.2 Μέτρα ομοιότητας Μέτρα ανομοιότητας: Γενικοί ορισμοί

Τα μέτρα εγγύτητας χωρίζονται σε μέτρα ομοιότητας (similarity measures) και μέτρα ανομοιότητας (dissimilarity measures). Παρακάτω θα δώσουμε τους αυστηρούς ορισμούς των μέτρων ομοιότητας και ανομοιότητας [136].

Δοθέντος λοιπόν ενός συνόλου δεδομένων X

Ένα μέτρο ομοιότητας s στο X είναι μια συνάρτηση

$$s : X \times X \to \Re$$

όπου R είναι το σύνολο των πραγματικών αριθμών, έτσι ώστε

$$\exists s_0 \in \Re : -\infty < s(x, y) \le s_0 < +\infty, \forall x, y \in X \quad (2.5)$$

$$s(x,x) = s_0, \forall x \in X \quad (2.6)$$

και

$$s(x, y) = s(y, x), \forall x, y \in X \quad (2.7)$$

Αν επιπλέον ισχύουν

$$s(x, y) = s_0 \text{ αν και μόνο αν } x = y \quad (2.8)$$

και

$$s(x,y)s(y,z) \le \left[s(x,y)+s(y,z)\right]s(y,z), \forall x,y,z \in X \quad (2.9)$$

τότε το μέτρο ομοιότητας s είναι μια μετρική (metric similarity measure).

Ένα μέτρο ανομοιότητας d στο X είναι μια συνάρτηση

$$d : X \times X \to \Re$$

έτσι ώστε

$$\exists d_0 \in \Re : -\infty < d_0 \le d(x,y) < +\infty, \forall x,y \in X \quad (2.10)$$

$$d(x,x) = d_0, \forall x \in X \quad (2.11)$$

και

$$d(x,y) = d(y,x), \forall x,y \in X \quad (2.12)$$

Αν επιπλέον ισχύουν

$$d(x,y) = d_0 \text{ αν και μόνο αν x = y (2.13)}$$

και

$$d(x,y) \le d(x,y) + d(y,z), \forall x,y,z \in X \, (2.14)$$

τότε το μέτρο ανομοιότητας d είναι μια μετρική (metric dissimilarity measure).

Από τους παραπάνω ορισμούς μπορούμε να συμπεράνουμε μερικά ενδιαφέροντα γεγονότα:

- Τα μέτρα ομοιότητας μπορούν να λάβουν θετικές αλλά και αρνητικές τιμές όπως φαίνεται από τις σχέσεις (2.5) και (2.10).

- Από την εξίσωση (2.8) μπορούμε να συμπεράνουμε ότι η μέγιστη τιμή της ομοιότητας μεταξύ δυο διανυσμάτων του X επιτυγχάνεται όταν αυτά ταυτίζονται.

- Από την εξίσωση (2.13) προκύπτει ότι η ελάχιστη τιμή ανομοιότητας μεταξύ δυο διανυσμάτων του X επιτυγχάνεται όταν αυτά ταυτίζονται.

- Γενικά θα μπορούσαμε να πούμε ότι τα μέτρα ομοιότητας είναι αντίθετα από τα μέτρα ανομοιότητας. Εύκολα μπορούμε να αποδείξουμε ότι αν το μέτρο ανομοιότητας d είναι μια μετρική, με $d(x,y) > 0, \forall x,y \in X$ τότε το μέτρο ομοιότητας $s = \frac{a}{d}$ με α > 0 είναι και αυτό μετρική. Επίσης εύκολα μπορούμε

να αποδείξουμε ότι το μέτρο ομοιότητας $d_{max} - d$ είναι μια μετρική, όπου d_{max} συμβολίζει την μέγιστη τιμή του d ανάμεσα σε όλα τα ζεύγη σημείων του X.

Θα μπορούσαμε να ομαδοποιήσουμε τα μέτρα εγγύτητας σε τέσσερις μεγάλες κατηγορίες:

1. **Μέτρα απόστασης (Distance Measures).**
2. **Συντελεστές Σχέσης (Association Coefficients).**
3. **Συντελεστές Συσχέτισης (Correlation Coefficients).**
4. **Πιθανοτικοί Συντελεστές Ομοιότητας** (**Probabilistic Similarity Measures**).

Θα αναλύσουμε τώρα λίγο περισσότερο τα διαφορετικά είδη των μέτρων εγγύτητας.

2.4.3 Μέτρα απόστασης

Τα μέτρα απόστασης (distance measures) είναι πολύ διαδεδομένα και μπορούν να διακριθούν σε μέτρα ομοιότητας και μέτρα ανομοιότητας. Μια άλλη κατηγοριοποίηση σύμφωνα με το είδος των τιμών που συγκρίνουν είναι σε μέτρα πραγματικών τιμών και μέτρα διακριτών τιμών.

Έστω X ο l – διάστατος χώρος των δεδομένων μας. Δυο οντότητες (περιπτώσεις) είναι ταυτόσημες όταν κάθε μια περιγράφεται μέσω μεταβλητών με την ίδια σημαντικότητα. Σ' αυτήν την περίπτωση το μέτρο ανομοιότητας είναι 0. Όταν το μέτρο ανομοιότητας έχει μικρή τιμή τότε τα οι οντότητές μας έχουν υψηλό βαθμό συσχέτισης και αντίστροφα όταν το μέτρο ανομοιότητας έχει μεγάλη τιμή η συσχέτιση των οντοτήτων είναι μικρότερη. Ακριβώς το αντίθετο συμβαίνει στα μέτρα ομοιότητας. Για τα μέτρα απόστασης ισχύουν δύο πράγματα. Πρώτον έχουν άνω φράγμα και δεύτερον εξαρτώνται από την κλίμακα. [1, 4, 12, 42, 64, 121, 136].

2.4.3.1 Μέτρα Ανομοιότητας πραγματικών τιμών.

Θα αναφερθούμε τώρα στα πιο γνωστά και στα πιο ευρέως χρησιμοποιούμενα μέτρα ανομοιότητας πραγματικών τιμών.

- Διατιμημένη (weighted) l_p μετρική

Η διατιμημένη l_p μετρική δίνεται από τον τύπο:

$$d_p(x,y) = \left(\sum_{i=1}^{l} w_i |x_i - y_i|^p \right)^{1/p} \quad (2.15)$$

όπου x_i και y_i είναι η i – οστή συντεταγμένη των x και y με i =1,…,l και $w_i>0$ είναι ο i – οστός συντελεστής βάρους. Αν $w_i = 1$ τότε η παραπάνω η παραπάνω μετρική ονομάζεται μη διατιμημένη μετρική l_p ή μετρική Minkofski.

- Όταν το p λαμβάνει την τιμή 2 τότε η παραπάνω μετρική καλείται Ευκλείδεια απόσταση ή l_2 μετρική.

$$d_2(x,y) = \left(\sum_{i=1}^{l} (x_i - y_i)^2 \right)^{1/2} \quad (2.16)$$

Για την αποφυγή της τετραγωνικής ρίζας, η τιμή της απόστασης τετραγωνοποιείται και συμβολίζεται ως d_2^2. Η έκφραση αυτή αναφέρεται και ως Τετραγωνική Ευκλείδεια Απόσταση.

Η διατιμημένη l_2 μετρική μπορεί να γενικευτεί με τον ακόλουθο τρόπο:

$$d(x,y) = \sqrt{(x-y)^T B^T (x-y)} \quad (2.17)$$

όπου B είναι ένας συμμετρικός θετικά ορισμένος πίνακας. Όταν B=Σ-1, όπου Σ είναι ο εντός των ομάδων πίνακας διασποράς – συνδιασποράς (pooled within – groups variance - covariance), τότε το μέτρο αυτό καλείται Mahalanobis μετρική. Το μέτρο αυτό έχει το πλεονέκτημα έναντι των άλλων μέτρων, ότι επιτρέπει τις συσχετίσεις μεταξύ μεταβλητών. Όταν οι συσχετίσεις αυτές είναι μηδενικές τότε το μέτρο αυτό είναι ισοδύναμο με την τετραγωνική ευκλείδεια απόσταση.

- Όταν το p λάβει την τιμή 1 τότε λαμβάνεται η l_1 διατιμημένη μετρική ή η Manhattan (City - Block) μετρική.

$$d_1(x,y) = \left(\sum_{i=1}^{l} w_i |x_i - y_i| \right) \quad (2.18)$$

- Όταν το p λάβει την τιμή ∞ τότε λαμβάνεται η l_∞ διατιμημένη μετρική:

$$d_\infty(x,y) = \max_{1 \le i \le l} w_i |x_i - y_i| \quad (2.19)$$

- Δυο άλλα μέτρα ανομοιότητας πραγματικών τιμών είναι τα ακόλουθα [93, 90]:

$$d_G(x,y) = -\log_{10}\left(1 - \frac{1}{l} \sum_{j=1}^{l} \frac{|x_j - y_j|}{b_j - a_j} \right) \quad (2.20)$$

όπου bj και aj είναι η μέγιστη και η ελάχιστη τιμή του j – οστού χαρακτηριστικού στα N διανύσματα του X, αντίστοιχα. Αποδεικνύεται ότι το μέτρο αυτό είναι μια μετρική. Αξίζει να σημειωθεί πως η τιμή του $d_G(x,y)$ δεν εξαρτάται μόνο από τα δυο διανύσματα x και y αλλά από όλο το X. Έτσι αν $d_G(x,y)$ είναι η απόσταση μεταξύ των δύο διανυσμάτων x και y στο X και αν $d'_G(x,y)$ είναι η απόσταση μεταξύ αυτών των ίδιων δυο διανυσμάτων στο σύνολο X΄, τότε εν γένει $d_G(x,y) \neq d'_G(x,y)$.

Ένα άλλο μέτρο ανομοιότητας είναι το

$$d_Q(x,y) = \sqrt{\frac{1}{l}\sum_{j=1}^{l}\left(\frac{x_j - y_j}{x_j + y_j}\right)^2} \quad (2.21)$$

Στο σημείο αυτό θα πρέπει ίσως να κάνουμε ένα μικρό σχόλιο για το πώς η Ευκλείδεια απόσταση επηρεάζεται αρνητικά από την αλλαγή στην κλίμακα των μεταβλητών [42, 121]. Ας δώσουμε ένα μικρό παράδειγμα. Έστω ότι οι τρεις περιπτώσεις A, B, C μετριούνται σε δυο μεταβλητές το βάρος σε λίβρες και το ύψος σε πόδια με τα ακόλουθα αποτελέσματα [42]:

	Βάρος σε λίβρες	Ύψος σε πόδια
A	60	3.0
B	65	3.5
C	63	4.0

Οι ευκλείδειες τους αποστάσεις είναι

$$D^2_{AB} = 25.25$$

$$D^2_{AC} = 10.00$$

$$D^2_{BC} = 04.25$$

Ωστόσο αν το ύψος μετρηθεί σε ίντσες οι αποστάσεις αυτές γίνονται:

$$D^2_{AB} = 50$$

$$D^2_{AC} = 153$$

$$D^2_{BC} = 53$$

Συνεπώς η περίπτωση A είναι στην δεύτερη φορά πιο κοντά στην περίπτωση B πάρά στην C. Επιπλέον η ευκλείδεια απόσταση δεν διατηρεί ούτε και τις ταξινομήσεις απόστασης. Εξαιτίας αυτού, οι μεταβλητές συχνά τυποποιούνται πριν εφαρμοστεί η ευκλείδεια απόσταση, δηλαδή $Z_{ik} = \frac{x_{ik}}{\sigma_k}$, όπου σ_κ είναι η τυπική απόκλιση της κ – οστής μεταβλητής. Με τον τρόπο αυτό η ευκλείδεια μετρική διατηρεί τις σχετικές αποστάσεις [42].

2.4.3.2 Μέτρα ομοιότητας Πραγματικών τιμών.

Θα ασχοληθούμε πολύ σύντομα, στο όνομα της πληρότητας, με μερικά μέτρα ομοιότητας μεταξύ πραγματικών τιμών.

- Εσωτερικό γινόμενο. Ορίζεται ως

$$s_{inner}(x, y) = x^T y = \sum_{i=1}^{l} x_i y_i \quad (2.22)$$

Στις περισσότερες περιπτώσεις το εσωτερικό γινόμενο χρησιμοποιείται όταν τα διανύσματα x και y είναι κανονικοποιημένα, έτσι ώστε να έχουν το ίδιο μήκος α. Στις περιπτώσεις αυτές το άνω και το κάτω φράγμα του εσωτερικού γινομένου είναι $+\alpha^2$ και $-\alpha^2$ αντίστοιχα και το εσωτερικό γινόμενο εξαρτάται αποκλειστικά από τη γωνία μεταξύ των διανυσμάτων.

- Μέτρα Tanimoto. Αυτό το μέτρο μπορεί να χρησιμοποιηθεί τόσο σε πραγματικά όσο και σε διακριτά διανύσματα. Ορίζεται ως εξής:

$$s_T(x, y) = \frac{x^T y}{\|x\|^2 + \|y\|^2 - x^T y} \quad (2.23)$$

Προσθέτοντας και αφαιρώντας τον όρο $x^T y$ από τον παρονομαστή της (2.23) και μετά από μερικές αλγεβρικές πράξεις παίρνουμε

$$s_T(x, y) = \frac{1}{1 + \dfrac{(x - y)^T (x - y)}{x^T y}} \quad (2.24)$$

- Ένα άλλο μέτρο ομοιότητας χρήσιμο σε συγκεκριμένες εφαρμογές δίνεται από τον τύπο

$$s_c(x, y) = 1 - \frac{s_2(x, y)}{\|x\| + \|y\|} \quad (2.25)$$

όπου το $s_c(x, y)$ παίρνει την μέγιστη τιμή του 1 όταν x = y και την ελάχιστη 0 όταν x = -y.

2.4.3.3 Μέτρα ανομοιότητας Διακριτών τιμών.

Έστω ότι οι συντεταγμένες των διανυσμάτων x ανήκουν σε ένα πεπερασμένο σύνολο F = {0, 1, …, k – 1}, όπου k είναι ένα θετικός ακέραιος. Ας υποθέσουμε τώρα ότι η διάσταση των δεδομένων μας είναι l, τότε είναι φανερό πως υπάρχουν k^l διανύσματα $x \in F^l$. Θα μπορούσε κάποιος να θεωρήσει ότι τα διανύσματα αυτά είναι κορυφές ενός l – διάστατου πλέγματος.

Έστω ότι $x, y \in F^l$ τότε ο πίνακας

$$A(x,y) = [a_{ij}] \; i,j = 0,1,...,k-1 \quad (2.26)$$

είναι ένας k×k πίνακας όπου το στοιχείο a_{ij} είναι το πλήθος των θέσεων όπου το πρώτο διάνυσμα έχει το σύμβολο i και το αντίστοιχο στοιχείο του δεύτερου διανύσματος έχει το σύμβολο j, όπου $i, j \in F$ τότε ο πίνακας αυτός καλείται contingency matrix. Για την καλύτερη κατανόηση του παραπάνω ορισμού ας δώσουμε ένα παράδειγμα.

Έστω l = 6, k = 3, και $x = [0,1,2,1,2,1]^T$ $y = [1,0,2,1,0,1]^T$ τότε ο πίνακας A(x,y) είναι ο:

$$A(x,y) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad (2.27)$$

διότι F = {0,1,2} και το στοιχείο a_{ij} του πίνακα ισούται με το πόσες φορές εμφανίζεται σε αντίστοιχες θέσεις στα διανύσματα x και y κάποιο συγκεκριμένο ζεύγος τιμών του συνόλου F. Παραδείγματος χάριν αν i = 1 και j = 1 τότε το a_{11} θα ισούται με 2 γιατί το ζεύγος τιμών (1,1) εμφανίζεται σε αντίστοιχες θέσεις στα διανύσματα x και y δύο φορές. Μια φορά στην θέση 4 και μια φορά στη θέση 6. Ας σημειωθεί εδώ ότι οι δείκτες του πίνακα A(x,y) αρχίζουν από το 0,0. το Πρώτο στοιχείο του πίνακα είναι δηλαδή το $a_{00} = 0$.

- Απόσταση Hamming. Η απόσταση Hamming ορίζεται ως το πλήθος τον θέσεων όπου τα δυο διανύσματα διαφέρουν. Χρησιμοποιώντας τον πίνακα A που ορίσαμε λίγο παραπάνω μπορούμε να ορίσουμε την απόσταση Hamming πιο τυπικά ως εξής:

$$d_H(x,y) = \sum_{i=0}^{k-1} \sum_{\substack{j=0 \\ j \neq i}}^{k-1} a_{ij} \quad (2.28)$$

  Δηλαδή η απόσταση Hamming παριστάνει το άθροισμα όλων των στοιχείων του πίνακα A, πλην αυτών της κύριας διαγωνίου. Αυτό προφανώς το άθροισμα παριστάνει σε πόσα σημεία διαφέρουν τα δυο διανύσματα. Πράγματι το άθροισμα των στοιχείων της κύριας διαγωνίου του πίνακα A μας δίνει το πλήθος των αντίστοιχων θέσεων που τα δυο διανύσματα συμφωνούν. Συνεπώς το άθροισμα όλων των άλλων στοιχείων μας δίνει το πλήθος των θέσεων που υπάρχουν μη κοινά στοιχεία στα δυο διανύσματα.

- Η απόσταση l_1. Μια άλλη απόσταση η οποία μπορεί να χρησιμοποιηθεί σε διακριτά διανύσματα είναι η απόσταση l_1 η οποία έχει οριστεί παραπάνω. Παραθέτουμε ξανά τον ορισμό της.

$$d_1(x,y) = \left(\sum_{i=1}^{l} w_i |x_i - y_i| \right) \quad (2.29)$$

Η απόσταση l_1 και η απόσταση Hamming ταυτίζονται όταν τα δεδομένα έχουν δυαδικές τιμές.

Είναι προφανές ότι τα μέτρα ανομοιότητας διακριτών τιμών μπορούν να χρησιμοποιηθούν άμεσα σαν μέτρα κατηγορικών δεδομένων. Δίνοντας μια διακριτή τιμή από το 0 ως το k-1 σε καθεμιά από τις διαφορετικές τιμές κάθε μεταβλητής μας, μπορούμε να χρησιμοποιήσουμε άμεσα τις παραπάνω αποστάσεις. Η απόσταση Hamming είναι πιο κοντά στην φιλοσοφία των κατηγορικών δεδομένων αφού λαμβάνει υπ' όψιν της μόνο τα σημεία στα οποία διαφέρουν τα δυο διανύσματα χωρίς να ασχολείται με την αριθμητική της αντιστοίχησης των κατηγορικών δεδομένων σε φυσικούς αριθμούς. Αντίθετα η απόσταση l_1 το κάνει αυτό και είναι ίσως λιγότερο κοντά στην φιλοσοφία των κατηγορικών δεδομένων.

2.4.4 Συντελεστές σχέσης.

Οι συντελεστές Σχέσης (Association Coefficients) χρησιμοποιούνται για να υπολογιστεί η ομοιότητα μεταξύ περιπτώσεων που περιγράφονται από δυαδικές μεταβλητές. Οι συντελεστές σχέσης παίρνουν τιμές μεταξύ 0 και 1. Οι συντελεστές αυτοί μπορούν να περιγραφούν μέσω ενός 2×2 πίνακα σχέσης (association table), στον οποίο το 1 ή το + αναφέρεται στην παρουσία μιας μεταβλητής ενώ το 0 ή το – αναφέρεται στην απουσία μιας μεταβλητής όπως φαίνεται στον παρακάτω πίνακα [42]:

		Περίπτωση i		
		+	–	
Περίπτωση j	+	a	c	a + b
	–	b	d	c + d
		a + c	b +d	p

Έστω δυο l – διάστατα διανύσματα x και y. Τα a, b , c και d αναφέρονται στο πόσες φορές εμφανίζεται το αντίστοιχο πρότυπο των δυαδικών μεταβλητών σ' αυτές τις δυο περιπτώσεις. Παραδείγματος χάριν το a μας δίνει το πλήθος των φορών εμφάνισης στις δυο περιπτώσεις δύο θετικών τιμών. Αντίστοιχα ισχύουν και για τα b, c, και d.

Έχουν προταθεί πολλοί συντελεστές σχέσης κυρίως εξαιτίας της αβεβαιότητας στον τρόπο ενσωμάτωσης των αρνητικών ταιριασμάτων (δηλαδή των d του παραπάνω πίνακα) στον συντελεστή, καθώς επίσης και εξαιτίας του τρόπου αξιολόγησης των ταιριασμένων ζευγαριών των μεταβλητών. Είναι δυνατόν αυτά τα ζεύγη να είναι ισοδύναμα διατιμημένα ή και να έχουν διπλάσιο βάρος από ότι τα μη ταιριασμένα ζευγάρια ή και το αντίστροφο. Μερικοί συντελεστές αγνοούν τα αρνητικά ταιριάσματα όπως παραδείγματος χάριν ο συντελεστής (ii) στον παρακάτω πίνακα, ενώ άλλοι δίνουν περισσότερο βάρος στα ταιριασμένα ζευγάρια παρά στα μη

ταιριασμένα όπως για παράδειγμα οι συντελεστές (iii) και (iv). Στον ακόλουθο πίνακα δίνουμε μερικούς συντελεστές συσχέτισης για δυαδικά δεδομένα [42, 34]:

(i) $\frac{a+d}{p}$ (ii) $\frac{a}{a+b+c}$

(iii) $\frac{2a}{2a+b+c}$ (iv) $\frac{2(a+d)}{2(a+d)+b+c}$

(v) $\frac{a}{a+2(b+c)}$ (vi) $\frac{a}{p}$

Ο συντελεστής (i) ονομάζεται Συντελεστής Απλού ταιριάσματος (Simple Matching Coefficient) και η ομοιότητα κυμαίνεται από 0 ως 1. Ο συντελεστής αυτός δεν μετασχηματίζεται εύκολα σε μετρική και λαμβάνει υπ' όψιν του την αρθρωτή απουσία μιας μεταβλητής. Το γεγονός αυτό οδηγεί κάποιες περιπτώσεις να εμφανίζονται πολύ όμοιες εξαιτίας κυρίως της ταυτόχρονης έλλειψης των ίδιων χαρακτηριστικών παρά εξαιτίας των όμοιων χαρακτηριστικών [1].

Ο συντελεστής (ii) καλείται συντελεστής Jaccard και η ομοιότητα που απορρέει απ' αυτόν τον συντελεστή κυμαίνεται από 0 ως 1. Ο συντελεστής αυτός αποφεύγει τη χρήση της αρθρωτής απουσίας μιας μεταβλητής στον υπολογισμό της ομοιότητας [1, 42].

Ας σημειώσουμε εδώ ότι διαφορετικοί συντελεστές σχέσης όταν εφαρμοστούν στο ίδιο σύνολο δεδομένων μπορεί να λάβουν εντελώς διαφορετικές τιμές. Ας δούμε ένα παράδειγμα όπου δυο περιπτώσεις χαρακτηρίζονται από την παρουσία ή την απουσία 10 μεταβλητών, όπως φαίνεται παρακάτω:

Μεταβλητές	1 2 3 4 5 6 7 8 9 10
Περίπτωση 1	1 0 0 0 1 1 0 0 1 0
Περίπτωση 2	0 0 0 0 1 0 0 1 1 0

Ο 2×2 πίνακας σχέσης για αυτές τις δύο περιπτώσεις είναι ο εξής:

		Περίπτωση 1		
		1	0	
Περίπτωση 2	1	2	1	3
	0	2	5	7
		4	6	10

Οι τιμές ομοιότητας που θα πάρουμε αν εφαρμόσουμε με την βοήθεια του παραπάνω πίνακα τους συντελεστές σχέσης θα είναι:

(i) 0.70	(ii) 0.40
(iii) 0.50	(iv) 0.82
(v) 0.25	(vi) 0.20

Παρατηρούμε πάρα πολύ καθαρά την μεγάλη απόκλιση που έχουμε μεταξύ των διαφορών συντελεστών σχέσης όταν εφαρμόζονται πάνω στο ίδιο ζεύγος δεδομένων. Ο συντελεστής (iv) δίνει ομοιότητα 0.82 ενώ ο συντελεστής (vi) μας δίνει μια ομοιότητα 0.20.

Το γεγονός αυτό δεν θα ήταν τόσο σημαντικό αν όλοι οι συντελεστές ήταν αρθρωτά μονότονοι (jointly monotonic) δηλαδή, αν όλες οι τιμές των ζευγαριών των περιπτώσεων ενός συντελεστή μπορούσαν να ταξινομηθούν κατά τέτοιο τρόπο ώστε να σχηματίζουν μια μονότονη ακολουθία (μια ακολουθία δηλαδή που οι τιμές της είτε θα αυξάνουν είτε θα φθίνουν σ' όλο το μήκος της). Σ' αυτήν την περίπτωση τότε οι τιμές για τα ζεύγη που θα λαμβάνονταν μέσω ενός άλλου συντελεστή θα ήταν επίσης μονότονες. Αυτό όμως δεν ισχύει, γιατί αν υποθέσουμε ότι μια τρίτη περίπτωση έχει την ακόλουθη μορφή για τις 10 δυαδικές μεταβλητές:

Μεταβλητές	1 2 3 4 5 6 7 8 9 10
Περίπτωση 1	0 0 0 0 0 0 0 1 0 0

τότε εφαρμόζοντας τους συντελεστές (i) και (ii) θα λάβουμε:

Συντελεστής (i)	Συντελεστής (ii)
$S_{12} = 0.7$	$S_{12} = 0.4$
$S_{13} = 0.5$	$S_{13} = 0.0$
$S_{23} = 0.8$	$S_{23} = 0.1$

Όπως φαίνεται καθαρά οι συντελεστές δεν είναι αρθρωτά μονότονοι [42, 34].

Ο Gower (1971) [54] όρισε έναν γενικό συντελεστή ομοιότητας, ο οποίος μπορεί να χρησιμοποιηθεί για διανύσματα ανάμεικτων τιμών, για διανύσματα δηλαδή που περιέχουν πραγματικές αλλά και διακριτές τιμές. Ο συντελεστής αυτός ορίζεται ως εξής:

$$S_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}} \quad (2.30)$$

Το βάρος w_{ijk} μπορεί να είναι ίσο με 1 ή με 0 ανάλογα με το αν η σύγκριση θεωρηθεί έγκυρη (valid) για την μεταβλητή k. Το βάρος μπορεί να είναι 0 όταν η μεταβλητή k είναι άγνωστη για τη μία ή και για τις δυο περιπτώσεις. Όταν το w_{ijk} είναι ίσο με το 0 τότε το s_{ijk} τίθεται ίσο με το 0 και αν το w_{ijk} είναι ίσο με το 0 για όλες τις μεταβλητές τότε το S_{ij} δεν ορίζεται.

Για δυαδικά δεδομένα οι τιμές και τα βάρη δίνονται από τον ακόλουθο πίνακα:

	Τιμές της Μεταβλητής k			
Περίπτωση i	+	+	–	–
Περίπτωση j	+	–	+	–
Τιμή s_{ijk}	1	0	0	0
Βάρος w_{ijk}	1	1	1	0

Για κατηγορικά δεδομένα η τιμή του s_{ijk} ισούται με τη μονάδα εάν οι δυο περιπτώσεις i και j είναι ίδιες στην k – οστή μεταβλητή, αλλιώς ισούται με το μηδέν.

Για ποσοτικά δεδομένα ισχύει ο τύπος:

$$s_{ijk} = 1 - \frac{\left|X_{ik} - X_{jk}\right|}{R_k} \quad (2.31)$$

όπου X_{ik} είναι η τιμή της περίπτωσης i στην μεταβλητή k και R_k είναι το εύρος της μεταβλητής k [1, 42, 136].

Ο συντελεστής του Gower έχει την ικανότητα να φιλοξενεί ανάμεικτους τύπους δεδομένων. Ο τελεστής αυτός μπορεί να χειριστεί δηλαδή δεδομένα που περιέχουν δυαδικές, ποσοτικές και κατηγορικές μεταβλητές.

Ας δώσουμε ένα παράδειγμα. Υποθέτουμε ότι το ύψος, το βάρος, το χρώμα των ματιών και το χρώμα των μαλλιών είναι οι μεταβλητές που περιγράφουν τρεις περιπτώσεις ανθρώπων, οι οποίοι ταυτόχρονα χαρακτηρίζονται και ως καπνιστές ή μη καπνιστές. Τα δεδομένα μας είναι τα ακόλουθα:

	Ύψος (ίντσες)	Βάρος (λίβρες)	Μάτια (χρώμα)	Μαλλιά (χρώμα)	Καπνιστής/μη Καπνιστής
Περίπτωση 1	66	120	Μπλε	Ξανθά	Καπνιστής
Περίπτωση 2	72	130	Πράσινα	Μαύρα	Καπνιστής
Περίπτωση 3	70	150	Μπλε	Ξανθά	Μη Καπνιστής

Οι τιμές του συντελεστή Gower για τα δεδομένα αυτά υπολογίζεται ως εξής:

$$s_{12} = \frac{s_{121} + s_{122} + s_{123} + s_{124} + s_{125}}{w_{121} + w_{122} + w_{123} + w_{124} + w_{125}} = \frac{\left(1 - \frac{|66-72|}{72-66}\right) + \left(1 - \frac{|120-130|}{150-120}\right) + 0 + 0 + 1}{1+1+1+1} = 0.334$$

$$s_{13} = \frac{s_{131} + s_{132} + s_{133} + s_{134} + s_{135}}{w_{131} + w_{132} + w_{133} + w_{134} + w_{135}} = \frac{\left(1 - \frac{|66-70|}{72-66}\right) + \left(1 - \frac{|120-150|}{150-120}\right) + 1 + 1 + 0}{1+1+1+1} = 0.466$$

$$s_{23} = \frac{s_{231} + s_{232} + s_{233} + s_{234} + s_{235}}{w_{231} + w_{232} + w_{233} + w_{234} + w_{235}} = \frac{\left(1 - \frac{|72-70|}{72-66}\right) + \left(1 - \frac{|130-150|}{150-120}\right) + 0 + 0 + 0}{1+1+1+1} = 0.2$$

2.4.5 Συντελεστές συσχέτισης

Οι συντελεστές αυτοί καλούνται μερικές φορές και μέτρα γωνιών (angular measures) εξαιτίας της γεωμετρικής του ερμηνείας. Χρησιμοποιούνται πολύ συχνά στις κοινωνιολογικές επιστήμες. Ο πιο διαδεδομένος συντελεστής είναι ο product – moment συντελεστής συσχέτισης που προτάθηκε από τον Karl Pearson. Αρχικά, χρησιμοποιήθηκε σαν μέθοδος για την συσχέτιση μεταβλητών, αλλά έχει χρησιμοποιηθεί και στην ποσοτική κατηγοριοποίηση για συσχέτιση περιπτώσεων. Ο συντελεστής ορίζεται ως εξής:

$$r_{jk} = \frac{\sum \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{\sqrt{\sum \left(x_{ij} - \bar{x}_j\right)^2 \left(x_{ik} - \bar{x}_k\right)^2}} \quad (2.32)$$

όπου x_{ij} είναι η τιμή της μεταβλητής i για την περίπτωση j, $\bar{x}_j$ η μέση τιμή όλων των τιμών των μεταβλητών για την περίπτωση j [1, 121].

Η τιμή του συντελεστή κυμαίνεται από -1 ως 1. Το πρόσημο υποδηλώνει την κατεύθυνση της σχέσης. Το θετικό πρόσημο δείχνει πως η σχέση είναι θετική ενώ το αρνητικό πρόσημο μας δείχνει ότι η σχέση είναι αρνητική. Η απόλυτη τιμή του συντελεστή δηλώνει την σημαντικότητα (magnitude) της σχέσης.

Ο συντελεστής του Pearson είναι ένας δείκτης για την γραμμική σχέση μεταξύ δύο περιπτώσεων. Η υψηλή συσχέτιση μπορεί να λάβει χώρα μεταξύ των περιπτώσεων εφόσον οι μετρήσεις μιας περίπτωσης είναι σε γραμμική σχέση με την άλλη. Γραμμική σχέση (linear relationship) δεν σημαίνει ότι τα σημεία πέφτουν ακριβώς πάνω σε μα ευθεία γραμμή, αλλά ότι τα σημεία είναι τοποθετημένα γενικά κατά μήκος μιας γραμμής.

Ένα μειονέκτημα του συντελεστή του Pearson, είναι ότι η χρήση του για τον καθορισμό της συσχέτισης μεταξύ των περιπτώσεων δεν έχει κανένα στατιστικό νόημα, διότι απαιτεί τον υπολογισμό της μέσης τιμής μεταξύ διαφορετικών τύπων

μεταβλητών και όχι τον υπολογισμό της μέσης τιμής κάθε μεταβλητής ανά περίπτωση. Ένας άλλος περιορισμός του συντελεστή είναι ότι συχνά αποτυγχάνει να ικανοποιήσει την τριγωνική ανισότητα [1, 121].

2.4.6 Πιθανοτικοί Συντελεστές Ομοιότητας.

Θα αναφερθούμε τώρα πολύ σύντομα στους πιθανοτικούς συντελεστές ομοιότητας. Οι πιθανοτικοί συντελεστές ομοιότητας (Probabilistic Similarity Coefficients) χρησιμοποιούνται μόνο για δυαδικά δεδομένα. Για δυο διανύσματα δυαδικών τιμών x και y ένα μέτρο του είδους αυτού, s βασίζεται στο πλήθος των θέσεων που τα διανύσματα x και y συμφωνούν. Η τιμή του s(x,y) συγκρίνεται με ζεύγη τυχαία επιλεγμένων διανυσμάτων, προκειμένου να εξεταστεί αν τα διανύσματα x και y είναι κοντά το ένα με το άλλο ή όχι. Αυτή η εργασία πραγματοποιείται με την εκτέλεση στατιστικών ελέγχων [1, 136].

2.5 Ελλιπή Δεδομένα

Ένα πρόβλημα που συναντάται πολύ συχνά στις εφαρμογές του πραγματικού κόσμου είναι το πρόβλημα των ελλιπών δεδομένων (missing data). Πιο συγκεκριμένα, το πρόβλημα υφίσταται από την στιγμή που για ορισμένα διανύσματα περιπτώσεων είναι άγνωστες οι τιμές κάποιων χαρακτηριστικών τους. Αυτό μπορεί να οδηγήσει σε εσφαλμένες μετρήσεις εγγύτητας των περιπτώσεων. Επομένως έχουν αναπτυχθεί ορισμένες τεχνικές που αντιμετωπίζουν το πρόβλημα αυτό [62, 136]:

1. Απορρίπτονται όλα τα διανύσματα των περιπτώσεων που περιλαμβάνουν ελλιπή χαρακτηριστικά. Η προσέγγιση αυτή μπορεί να χρησιμοποιηθεί μόνο όταν το πλήθος των περιπτώσεων με ελλιπή χαρακτηριστικά είναι μικρό συγκρινόμενο με το συνολικό πλήθος των διαθέσιμων περιπτώσεων. Αν δεν πληρείται αυτή η προϋπόθεση, τότε μια τέτοια αντιμετώπιση μπορεί να επηρεάσει την φύση του προβλήματος.

2. Για το i – οστό χαρακτηριστικό, υπολογίζεται η μέση τιμή του βασιζόμενη στις αντίστοιχες διαθέσιμες τιμές όλων των περιπτώσεων του συνόλου X. Εν συνεχεία, αντικαθίσταται η τιμή αυτή σε εκείνα τα διανύσματα, στα οποία η τιμή του i – οστού χαρακτηριστικού λείπει.

3. Για όλα τα ζεύγη των συστατικών x_i και y_i των διανυσμάτων x και y ορίζεται το b_i ως εξής:

$$b_i = \begin{cases} 0, x_i \text{ και } y_i \text{ είναι διαθέσημα} \\ 1, \text{αλλιώς} \end{cases}$$

Τότε, η εγγύτητα μεταξύ των x και y ορίζεται από τη σχέση:

$$P(x,y) = \frac{l}{1 - \sum_{i=1}^{l} b_i} \sum_{\text{κάθε } i: b_i = 0} \phi(x_i, y_i) \quad (2.32)$$

όπου $\phi(x_i, y_i)$ συμβολίζει την εγγύτητα μεταξύ δυο βαθμωτών x_i και y_i. Μια επιλογή του φ όταν εμπλέκεται ένα μέτρο ανομοιότητας, είναι το $\phi(x_i, y_i) = |x_i - y_i|$. Η λογική της προσέγγισης αυτής είναι σχετικά απλή. Έστω $[a,b]$ ένα διάστημα επιτρεπτών τιμών του P(x,y). Η παραπάνω σχέση εξασφαλίζει πως το μέτρο εγγύτητας των διανυσμάτων x και y θα καλύψει όλο το διάστημα $[a,b]$, αδιαφορώντας για τα μη διαθέσιμα χαρακτηριστικά των διανυσμάτων.

4. Υπολογίζεται ο μέσος όρος των ομοιοτήτων (proximities) $\phi_{aug}(i)$ μεταξύ όλων των διανυσμάτων του X κατά μήκος όλων των χαρακτηριστικών i = 1, …, l. Είναι φανερό πως για κάποια διανύσματα x η τιμή του i - οστού χαρακτηριστικού δεν είναι διαθέσιμη. Στην περίπτωση αυτή οι ομοιότητες που περιλαμβάνουν τα x_i αποκλείονται από τον υπολογισμό του $\phi_{aug}(i)$. Ορίζεται η εγγύτητα $\psi(x_i, y_i)$ μεταξύ του i – οστού χαρακτηριστικού των x και y ως $\phi_{aug}(i)$ αν τουλάχιστον ένα από τα x_i και y_i δεν είναι διαθέσιμα και ως $\phi(x_i, y_i)$ αν και τα δυο διατίθενται. Το $\phi(x_i, y_i)$ μπορεί να οριστεί όπως και στην προηγούμενη περίπτωση. Τότε

$$P(x,y) = \sum_{i=1}^{l} \psi(x_i, y_i) \quad (2.33).$$

3 Μέθοδοι ομαδοποίησης

3.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο εξετάσαμε κάποια μέτρα ομαδοποίησης. Είδαμε ότι κάθε ένα από τα μέτρα ομοιότητας δίνει διαφορετική ερμηνεία στις έννοιες της ομοιότητας και ανομοιότητας, οι οποίες έννοιες συνδέονται με τους τύπους των ομάδων που πρέπει να ανακαλυφθούν από τα δεδομένα. Είναι αξιοσημείωτο ότι διαφορετικοί συνδυασμοί μέτρων εγγύτητας και αλγορίθμων ομαδοποίησης, μας δίνουν διαφορετικά αποτελέσματα.

Σ΄ αυτό το κεφάλαιο θα αναφερθούμε γενικά στους αλγόριθμους ομαδοποίησης. Θα αναφέρουμε κατ' αρχάς δυο λόγια για τις διαφορετικές τεχνικές ομαδοποίησης και έπειτα θα επικεντρωθούμε σε κάποιους αλγόριθμούς ομαδοποίησης κατηγορικών δεδομένων καθώς και σε αλγόριθμούς εννοιολογικής ομαδοποίησης.

3.2 Πλήθος πιθανών ομαδοποιήσεων

Ας υποθέσουμε ότι $x_i, i = 1,..., N$ είναι τα χαρακτηριστικά διανύσματα του συνόλου δεδομένων X. Το πρόβλημα τώρα της ομαδοποίησης έγκειται στην τμηματοποίηση σε m ομάδες του συνόλου αυτού των N διανυσμάτων με βάση κάποιο κριτήριο. Η ιδανική λύση του προβλήματος αυτού θα ήταν ο υπολογισμός όλων των δυνατών ομαδοποιήσεων του συνόλου αυτού σε m ομάδες και η επιλογή της «καλύτερης», της πιο λογικής τμηματοποίησης. Αυτή η προφανής λύση του προβλήματος είναι, όπως είναι λογικό και υπολογιστικά μη εφικτή, ακόμα και αν το σύνολο των N αντικειμένων είναι σχετικά μικρό. Πράγματι, έστω ότι S(N, m) είναι το πλήθος των πιθανών ομαδοποιήσεων των N αντικειμένων σε m ομάδες. Τότε προφανώς θα ισχύουν τα εξής [62, 136, 130]:

1. S(N, m) = 1.
2. S(N, N) = 1.
3. S(N, m) = , όταν m< N.

Έστω τώρα ότι L_{N-1}^k είναι μια λίστα που περιέχει όλες τις δυνατές ομαδοποιήσεις των N – 1 παρατηρήσεων σε k ομάδες για k = m, m – 1. Τότε η N – οστή παρατήρηση

- Είτε θα προστεθεί σε κάποια από τις ομάδες που ήδη υπάρχουν σε κάθε μέλος της λίστας L_{N-1}^k
- Είτε θα σχηματιστεί μια νέα ομάδα σε κάθε ένα μέλος της λίστας L_{N-1}^k.

Συνεπώς θα πρέπει να ισχύει η σχέση:

$$S(N,m) = mS(N-1,m) + S(N-1,m-1) \quad (3.1)$$

Οι λύσεις της εξίσωσης (3.1) είναι οι αριθμοί που καλούνται αριθμοί Stirling δεύτερου είδους:

$$S(N,m) = \frac{1}{m!}\sum_{i=0}^{m}(-1)^{m-i}\binom{m}{i}i^N \quad (3.2)$$

Τώρα για N = 2 το πλήθος των πιθανών ομαδοποιήσεων σύμφωνα με την παραπάνω εξίσωση (3.2) δίνεται από την σχέση:

$$S(N,2) = 2^{N-1} - 1 \quad (3.3).$$

Ας δούμε τώρα ορισμένες αριθμητικές τιμές της παραπάνω εξίσωσης [90]:

- S(15, 3) = 2 375 101
- S(20, 4) = 45 232 115 901
- S(25, 8) = 690 223 721 118 368 580
- S(100, 5) = 10^{68}.

Από τα παραπάνω αποτελέσματα γίνεται φανερό ότι η εύρεση όλων των δυνατών ομαδοποιήσεων των N αντικείμενων σε m ομάδες με σκοπό να επιλεγεί η καλύτερη ομαδοποίηση, δεν λύνει ικανοποιητικά το πρόβλημα της ομαδοποίησης. Για παράδειγμα, ας αναφέρουμε ότι θέλουμε να ομαδοποιήσουμε 100 αντικείμενα σε 5 ομάδες. Το πλήθος των διαφορετικών ομαδοποιήσεων απ' την οποία θα πρέπει να επιλεγεί η καλύτερη είναι, όπως αναφέραμε και λίγο παραπάνω, 10^{68}.Αν υποθέσουμε πως για την εκτίμηση κάθε ομαδοποίησης ένας υπολογιστής απαιτεί χρόνο 10^{-12} δευτερολέπτων, που είναι ένας σχετικά λογικός χρόνος με τα σημερινά δεδομένα, τότε θα έχουμε την πιο λογική ομαδοποίηση των 100 αντικειμένων σε 5 ομάδες μετά από περίπου 10^{48} χρόνια, μια και όχι τόσο εφικτή περίοδος μια και ξεπερνάει κατά πολλές τάξεις μεγέθους την διάρκεια του ανθρώπινου πολιτισμού.

3.3 Κατηγορίες μεθόδων ομαδοποίησης.

Όπως έχουμε ήδη αναφέρει οι μέθοδοι ομαδοποίησης μπορούν να χωριστούν σε δυο μεγάλες κατηγορίες : τις παραμετρικές (parametric) μεθόδους ομαδοποίησης και τις μη παραμετρικές (non parametric) μεθόδους ομαδοποίησης. Η θεμελιώδης διαφορά των δυο αυτών μεθόδων ομαδοποίησης είναι ότι οι παραμετρικές μέθοδοι απαιτούν τον καθορισμό εξ αρχής ορισμένων παραμέτρων, όπως το πλήθος των ομάδων, η κατανομή των δεδομένων, το κριτήριο ομαδοποίησης και άλλα, σε αντίθεση με τις μη παραμετρικές μεθόδους, οι οποίες εκτελούν την ομαδοποίηση χωρίς τον εκ των προτέρων καθορισμό τέτοιων παραμέτρων, χωρίς ουσιαστικά εκ των προτέρων γνώση για την δομή των δεδομένων.

Μια διαφορετική κατηγοριοποίηση των δεδομένων μπορεί να γίνει σύμφωνα με τον τρόπου που κάθε αλγόριθμος ομαδοποιεί τα δεδομένα. Έτσι μπορούμε να κατατάξουμε τις μεθόδους ομαδοποίησης στις ακόλουθες κατηγορίες:

1. **Ιεραρχικοί (Hierarchical) Αλγόριθμοι Ομαδοποίησης.** Χωρίζονται σε συσσωρευτικούς (agglomerative) και σε διαιρετικούς (divisive) οι οποίοι διασπούν την βάση των N δεδομένων σε πολλά επίπεδα από φωλιασμένες ομάδες που αναπαριστώνται μέσω ενός δέντρου που λέγεται δεντρόγραμμα (dentrogramm).

2. **Αλγόριθμοι Ομαδοποίησης που Βασίζονται στην Βελτιστοποίηση μιας Συνάρτησης Κόστους.** Στις τεχνικές αυτές οι ομάδες σχηματίζονται βελτιστοποιώντας κάποιο κριτήριο ομαδοποίησης με αποτέλεσμα μη τεμνόμενες ομάδες όμοιων αντικειμένων. Συνήθως το πλήθος των ομάδων στους αλγόριθμούς αυτής της κατηγορίας διατηρείται σταθερό κατά την εκτέλεση του αλγορίθμου. Οι αλγόριθμοι της κατηγορίας αυτής χωρίζονται σε δυο ομάδες:

   - Επαναληπτικές τεχνικές τμηματοποίησης (Iterative Partitioning).

   - Πιθανοθεωρητικοί (Probabilistic) αλγόριθμοι Ομαδοποίησης.

3. **Density ή Mode Seeking Τεχνικές.** Οι ομάδες σ' αυτήν την κατηγορία αλγορίθμων θεωρούνται ως περιοχές υψηλής πυκνότητας του χώρου των δεδομένων οι οποίες χωρίζονται μεταξύ τους με περιοχές χαμηλής πυκνότητας. Οι τεχνικές αυτές αναζητούν περιοχές υψηλής συχνότητας χρησιμοποιώντας κάποιο κριτήριο ομαδοποίησης.

4. **Ανταγωνιστικοί αλγόριθμοι Μάθησης.** Τα επαναληπτικά σχήματα αυτής της κατηγορίας δεν χρησιμοποιούν συναρτήσεις κόστους, αντίθετα παράγουν ομαδοποιήσεις και συγκλίνουν στην πιο λογική από αυτές, σύμφωνα με κάποιο μέτρο απόστασης. Τυπικοί εκπρόσωποι της κατηγορίας αυτής είναι το βασικό ανταγωνιστικό σχήμα μάθησης (basic competitive learning scheme), ο διαρρέων (leaky) αλγόριθμος μάθησης και οι αυτοοργανωτικές απεικονίσεις (Self – Organizing Maps).

5. **Branch and Bound Αλγόριθμοι Ομαδοποίησης.** Οι αλγόριθμοι αυτής της οικογένειας παρέχουν ολικά βέλτιστες ομαδοποιήσεις, χωρίς την εξέταση όλων των πιθανών ομαδοποιήσεων, για ένα συγκεκριμένο πλήθος M ομάδων και για κάποιο προκαθορισμένο κριτήριο ομαδοποίησης. Ωστόσο, η υπολογιστική πολυπλοκότητα των αλγορίθμων αυτών είναι υπερβολική.

Πριν περάσουμε στην ανάλυση κάποιων αλγορίθμων ομαδοποίησης κατηγορικών που θα μας απασχολήσουν και αργότερα, θα δώσουμε καταρχάς με σχετική συντομία και την πληρότητα που μας επιτρέπει αυτή η συντομία κάποιες γενικές γραμμές για τις προαναφερθείσες μεθόδους ομαδοποίησης.

3.3.1 Ιεραρχικοί αλγόριθμοι

Ας αρχίσουμε με την πρώτη κατηγόρια αλγορίθμων τους Ιεραρχικούς αλγορίθμους. Αυτοί οι αλγόριθμοι εκτελούν συγχωνεύσεις ή διαχωρίσεις των δεδομένων για να πετύχουν την επιθυμητή ομαδοποίηση για αυτό και διακρίνονται σε δυο ομάδες. Τους ιεραρχικά συσσωρευτικούς (hierarchical agglomerative) και τους ιεραρχικά διαιρετικούς (hierarchical divisive) αλγορίθμους.

Ένα από τα χαρακτηριστικά των ιεραρχικών μεθόδων ομαδοποίησης είναι ότι η ανάθεση ενός αντικείμενου σε μια ομάδα είναι αμετάκλητη. Όταν δηλαδή έναν αντικείμενο αποδοθεί σε μια ομάδα ποτέ δεν απομακρύνεται από αυτήν και ποτέ δεν θα ενωθεί με αντικείμενα που ανήκουν σε άλλη ομάδα.

Οι συσσωρευτικοί ιεραρχικοί αλγόριθμοι προχωρούν σχηματίζοντας μια σειρά από συγχωνεύσεις των N αντικειμένων σε ομάδες, καταλήγοντας σε μια ομάδα που περιέχει όλα τα αντικείμενα, ενώ αντίθετα οι διαιρετικοί ιεραρχικοί αλγόριθμοι ομαδοποίησης χωρίζουν το σύνολο των N αντικείμενων διαδοχικά μέχρι που να καταλήξουν σε N ομάδες, που η κάθε μια περιέχει από ένα αντικείμενο.

Τα αποτελέσματα των ιεραρχικών αυτών μεθόδων αναπαριστούνται σε έναν δισδιάστατο διάγραμμα που είναι γνωστό ως δεντρόγραμμα. Όταν ένα ζεύγος περιπτώσεων συγχωνευτεί αναπαρίσταται ως ένας κλάδος στο δέντρο. Αντίθετα, όταν μια ομάδα διαιρεθεί, αναπαρίσταται ως δυο παιδιά – κλάδοι οι οποίοι εκπορεύονται από τον κόμβο που παρίστανε την ομάδα ολόκληρη.

Οι ομάδες που παράγονται από τους ιεραρχικούς αλγόριθμους είναι φωλιασμένες και κάθε μια μπορεί να θεωρηθεί τμήμα μιας μεγαλύτερης, περισσότερο περιεκτικής ομάδας, η οποία βρίσκεται σε κάποιο υψηλότερο επίπεδο του δεντρογράμματος [1, 42, 34, 115, 121].

3.3.1.1 Ιεραρχικές συσσωρευτικές (agglomerative) μέθοδοι.

Ας δούμε τώρα τα γενικά βήματα ενός ιεραρχικού συσσωρευτικού αλγορίθμου για την ομαδοποίηση N αντικειμένων.

Βήμα 1ο. Αρχικά, τα N αντικείμενα χωρίζονται σε N ομάδες. Έστω ο N × N συμμετρικός πίνακας εγγύτητας $D = d_{ij}$.

Βήμα 2ο. Αναζητείται στον πίνακα το κοντινότερο, το πιο όμοιο, ζευγάρι ομάδων. Έστω ότι η απόσταση των πιο όμοιων ομάδων U και V είναι d_{UV}.

Βήμα 3ο. Συγχωνεύονται οι ομάδες U και V. Στην καινούργια ομάδα δίνεται η ετικέτα (UV) Ενημερώνονται τα πεδία του πίνακα εγγύτητας:

Βήμα 4ο. Διαγράφονται οι γραμμές και οι στήλες που αναφέρονται στις ομάδες U και V.

Βήμα 5ο. Προστίθεται μια γραμμή και μια στήλη που δείχνει την απόσταση της ομάδας (UV) από τις υπόλοιπες ομάδες.

Βήμα 6ο. Επαναλαμβάνονται τα βήματα 2 και 3 συνολικά N – 1 φορές. Όταν ο αλγόριθμος τερματιστεί όλα τα αντικείμενα θα βρίσκονται σε μια ομάδα. Εγγράφονται οι ομάδες που έχουν σχηματιστεί και το επίπεδο ομοιότητας στο οποίο έχουν σχηματιστεί.

Οι ιεραρχικοί συσσωρευτικοί αλγόριθμοι ομαδοποίησης διακρίνονται από τους διαφορετικούς τύπους σύνδεσης των ομάδων. Οι πιο σημαντικοί κανόνες είναι οι παρακάτω[1, 12, 34, 36, 42, 64, 115, 121, 136]:

- Απλή σύνδεση ή ο κοντινότερος γείτονας (Single Linkage or The nearest neighbor)
- Πλήρης σύνδεση ή ο μακρινότερος γείτονας (Complete Linkage or The Furthest Neighbor).
- Σύνδεση Μέσου Όρου (Average Liknkage).
- Μέθοδος του Ward.
- Centroid ομαδοποίηση.
- Median ομαδοποίηση.

Ας δούμε τώρα πιο αναλυτικά τους τύπους αυτούς των ιεραρχικών συσσωρευτικών αλγορίθμων. Προτού όμως προχωρήσουμε στην περιγραφή των κανόνων είναι χρήσιμο για λόγους καλύτερης κατανόησης, να υποθέσουμε ότι το μέτρο εγγύτητας που υιοθετείται είναι ένα μέτρο ανομοιότητας. Αντί του μέτρου ανομοιότητας θα μπορούσε κάλλιστα να χρησιμοποιηθεί κάποιο μέτρο ομοιότητας, θα έπρεπε όμως να γίνουν οι κατάλληλες αλλαγές στις παρακάτω περιγραφές. Τα βήματα και ο τρόπος λειτουργίας των ιεραρχικών μεθόδων ομαδοποίησης είναι ανεξάρτητος από το μέτρο εγγύτητας που θα χρησιμοποιηθεί.

3.3.1.1.1 Απλή Σύνδεση ή ο Κοντινότερος Γείτονας.

> ΚΑΝΟΝΑΣ: Μια περίπτωση θα ενωθεί με κάποια από τις υπάρχουσες ομάδες, αν τουλάχιστον ένα από τα μέλη μιας υπάρχουσας ομάδας είναι στο ίδιο επίπεδο ομοιότητας με την υπό εξέταση περίπτωση.

Ο αλγόριθμος της απλής σύνδεσης ξεκινά βρίσκοντας το ελάχιστο στοιχείο του πίνακα ανομοιότητας D. Δηλαδή βρίσκει τα δυο εκείνα αντικείμενα που έχουν την μικρότερη απόσταση. Έστω C_i και C_j τα αντικείμενα αυτά. Αυτά θα αποτελέσουν την πρώτη ομάδα $C_q = C_i \cup C_j$. Στο επόμενο βήμα ένα από τα δύο μπορεί να συμβεί:

- Είτε ένα τρίτο αντικείμενο θα έρθει να προστεθεί στην ήδη σχηματισμένη ομάδα C_q
- Είτε τα δυο πιο κοντινά μη ομαδοποιημένα (unclustered) αντικείμενα θα ομαδοποιηθούν σε μια καινούργια ομάδα.

Η απόφαση για το ποια ενέργεια θα εκτελεστεί, θα στηριχθεί στο αν η απόσταση ενός μη ομαδοποιημένου αντικειμένου από την σχηματισμένη ομάδα C_q είναι μικρότερη από την απόσταση των δυο πιο κοντινών μη ομαδοποιημένων αντικειμένων. Η διαδικασία συνεχίζεται μέχρι να σχηματιστεί μια ομάδα που να υπάρχει όλα τα αντικείμενα της βάσης.

Η απόσταση μεταξύ των ομάδων C_q και C_s υπολογίζεται από τον εξής τύπο

$$d(C_q, C_s) = \min\{d(C_i, C_s), d(C_j, C_s)\} \text{ (3.4).}$$

Η μέθοδος της απλής σύνδεσης είναι αμετάβλητη σε μετασχηματισμούς του πίνακα ομοιότητας. Δεν επηρεάζεται δηλαδή από κανέναν μετασχηματισμό στα δεδομένα. Έχει όμως την τάση να σχηματίζει αλυσίδες ή μακριές επιμήκεις ομάδες, όταν χρησιμοποιείται σε αριθμητικά δεδομένα. Ακόμα, από την παρατήρηση του δενδρικού διαγράμματος της μεθόδου δεν μπορεί να υποθέσει κανείς πόσες ομάδες υπάρχουν στα δεδομένα [1].

3.3.1.1.2 Πλήρης σύνδεση ή ο απώτατος γείτονας.

> ΚΑΝΟΝΑΣ: Μια περίπτωση για να συμπεριληφθεί σε μια υπάρχουσα ομάδα πρέπει να είναι μέσα σε ένα συγκεκριμένο επίπεδο ομοιότητας με όλα τα μέλη της συγκεκριμένης ομάδας.

Η απόσταση μεταξύ δυο ομάδων καθορίζεται από την απόσταση δυο στοιχείων (ένα από κάθε ομάδα), των οποίων η απόσταση είναι μέγιστη.

Ο αλγόριθμος της πλήρους σύνδεσης ξεκινά βρίσκοντας το ελάχιστο στοιχείο του πίνακα D και μετά συγχωνεύει τα αντίστοιχα αντικείμενα. Έστω C_i και C_j τα αντικείμενα που συγχωνεύονται σχηματίζοντας την ομάδα C_q. Έπειτα υπολογίζονται οι αποστάσεις της ομάδας C_q από κάθε άλλη υπάρχουσα ομάδα C_s από τον τύπο:

$$d(C_q, C_s) = \max\{d(C_i, C_s), d(C_j, C_s)\} \text{ (3.5).}$$

Η πλήρης σύνδεση έχει την τάση να βρίσκει σχετικά συμπαγείς υπερσφαιρικές ομάδες οι οποίες αποτελούνται από περιπτώσεις σε πολύ μεγάλο βαθμό όμοιες μεταξύ τους. Το μειονέκτημα της μεθόδου αυτής είναι ότι οι ομάδες που θα σχηματιστούν πιθανώς να υπερκαλύπτονται μεταξύ τους [1].

3.3.1.1.3 Σύνδεση μέσου όρου

> ΚΑΝΟΝΑΣ: Υπολογίζεται ο μέσος όρος της ομοιότητας της υπό εξέταση περίπτωσης με όλες τις άλλες περιπτώσεις στην ήδη υπάρχουσα ομάδα και επακόλουθα ενώνεται η περίπτωση αυτή με την υπάρχουσα ομάδα, εάν επιτυγχάνεται ένα συγκεκριμένο επίπεδο ομοιότητας.

Η απόσταση μεταξύ των ομάδων ορίζεται ως ο μέσος όρος της απόστασης μεταξύ όλων των ζευγαριών των στοιχείων, με το ένα στοιχείο να ανήκει στη μία ομάδα και το άλλο στοιχείο στην άλλη ομάδα [42].

Ο αλγόριθμος του μέσου όρου ξεκινά αναζητώντας την μικρότερη απόσταση στον πίνακα D για να βρει τα κοντινότερα (ομοιότερα) αντικείμενα. Έστω C_i και C_j τα πιο όμοια αντικείμενα. Η συγχώνευση αυτών των δυο ομάδων σχηματίζει την ομάδα C_q. Στη συνέχεια υπολογίζονται οι αποστάσεις μεταξύ της ομάδας C_q και κάθε άλλης ομάδας C_s από τον τύπο:

$$d(C_q, C_s) = \frac{n_i}{n_i + n_j} d(C_i, C_s) + \frac{n_j}{n_i + n_j} d(C_j, C_s) \quad (3.6)$$

με n_i και n_j να είναι ο πληθυσμός των ομάδων C_i και C_j αντίστοιχα.

3.3.1.1.4 Μέθοδος του Warp

> ΚΑΝΟΝΑΣ: Ο σχηματισμός των ομάδων βασίζεται στην απώλεια πληροφορίας, που προκαλείται από την ομαδοποίηση των ξεχωριστών περιπτώσεων σε ομάδες, η οποία μετριέται μέσω του ολικού αθροίσματος των τετραγώνων των αποκλίσεων κάθε παρατήρησης από το κέντρο (μέσο της ομάδας στην οποία ανήκει. Ο κανόνας ανάθεσης βασίζεται στην αύξηση του λάθους του αθροίσματος τετραγώνων, που προέρχεται από το συνδυασμό κάθε δυνατού ζευγαριού από ομάδες. Αυτή η τιμή χρησιμοποιείται ως αντικειμενική συνάρτηση και συμβολίζεται ως ESS [1, 42, 34, 136].

Το ολικό άθροισμα των τετραγωνικών αποκλίσεων (Error Sum of Squares ESS) υπολογίζεται από τον τύπο:

$$ESS = \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} X_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_j} X_{ij} \right)^2 \right) \quad (3.7)$$

όπου X_{ij} είναι η τιμή της i – σοστής περίπτωσης στην j – οστή ομάδα, κ είναι το πλήθος των ομάδων και n_j ν είναι το πλήθος των περιπτώσεων στης j – οστής ομάδας[1, 42, 34, 121].

Στο πρώτο βήμα της διαδικασίας της ομαδοποίησης κάθε περίπτωση θεωρείται ως μια ομάδα. Στην αρχή έχουμε λοιπόν N διαφορετικές ομάδες του ενός στοιχείου και ισχύει $ESS_j = 0$ για j=0, 1, …, N. Επειδή η αντικειμενική συνάρτηση ESS ορίζεται ως το άθροισμα όλων των ESS_j για j=0, 1, …, N θα έχουμε ESS = 0.

Η πρώτη ομάδα σχηματίζεται από την επιλογή των δύο εκείνων ομάδων, οι οποίες όταν ενωθούν θα δώσουν την ελάχιστη αύξηση στην αντικειμενική συνάρτηση. Με τον τρόπο αυτόν οι N ομάδες μειώνονται διαδοχικά σε N-1, N-2, …,1. Σε κάθε επίπεδο της διαδικασίας η τιμή της αντικειμενικής συνάρτησης εκτιμάται [1, 42, 34, 121].

3.3.1.1.5 Centroid Linkage

> ΚΑΝΟΝΑΣ: Οι ομάδες (ή οι περιπτώσεις) συγχωνεύονται σύμφωνα με την απόσταση μεταξύ των κέντρων τους. Οι ομάδες με τα πιο κοντινά κέντρα ομαδοποιούνται πρώτες.

Αν $\bar{x}_i = \frac{1}{n_i}\sum_{x \in C_i} x$ είναι το κέντρο των n_i μελών της ομάδας C_i και ομοίως $\bar{x}_j$ είναι το κέντρο των n_j μελών της ομάδας C_j τότε οι αποστάσεις μεταξύ της ομάδας C_q και κάθε άλλης ομάδας C_s μπορούν να υπολογιστούν από τον τύπο:

$$d(C_q, C_s) = \frac{n_i}{n_i + n_j} d(C_i, C_s) + \frac{n_j}{n_i + n_j} d(C_j, C_s) - \frac{n_j}{\left(n_i + n_j\right)^2} d(C_i, C_j) \quad (3.8)$$

όπου d είναι ένα μέτρο ομοιότητας. Το κέντρο της νέας ομάδας C_q, που προκύπτει από την συνένωση των ομάδων C_i και C_j είναι ο διατιμημένος μέσος όρος

$$\bar{x} = \frac{n_i \bar{x}_i + n_j \bar{x}_j}{n_i + n_j} \quad (3.9).$$

3.3.1.1.6 Median Linkage

Ένα μειονέκτημα της μεθόδου Centroid Linkage είναι ότι όταν τα μεγέθη των δυο ομάδων είναι πολύ διαφορετικά μεταξύ τους, τότε η νέα ομάδα θα είναι πολύ πιο κοντά στην μεγάλη ομάδα απ' ότι θα είναι στην μικρή και πιθανόν να παραμείνει μέσα της. Συνέπεια αυτού του γεγονότος θα είναι να χαθούν ουσιαστικά τα χαρακτηριστικά της μικρότερης ομάδας. Για να αποφύγουμε αυτό το μειονέκτημα θα πρέπει να χρησιμοποιήσουμε μια στρατηγική που να είναι ανεξάρτητη από το μέγεθος της ομάδας. Μια τέτοια στρατηγική υποθέτει λοιπόν , πως οι ομάδες που συγχωνεύονται είναι του ίδιου μεγέθους και ως συνέπεια η φαινομενική θέση της καινούργιας ομάδας θα είναι ανάμεσα στις δυο υπάρχουσες ομάδες. Αν αναπαραστήσουμε τα κέντρα των ομάδων που θα συγχωνευτούν με $\bar{x}_i$ και $\bar{x}_j$ αντίστοιχα τότε το κέντρο της τρίτης ομάδας $\bar{x}_s$ θα βρίσκεται κατά μήκος της διαμέσου του τριγώνου που ορίζεται από τα $\bar{x}_i$, $\bar{x}_j$ και $\bar{x}_s$. [42, 121, 136].

Συνεπώς το κέντρο της καινούργιας ομάδας θα δίνεται από τον τύπο

$$\bar{x} = \frac{1}{2}\left(\bar{x}_i + \bar{x}_j\right) \quad (3.10).$$

Όλες οι παραπάνω τεχνικές συσσωρευτικής ιεραρχικής ομαδοποίησης είναι φτιαγμένες για αριθμητικά δεδομένα. Θα δούμε όμως παρακάτω ότι οι ίδιες αυτές τεχνικές θα μπορούσαν να χρησιμοποιηθούν και σε κατηγορικά δεδομένα, εκτός ίσως από την μέθοδο του Warp, με την απόσταση που θα προτείνουμε και με τις επεκτάσεις των εννοιών του μέσου όρου που θα δώσουμε αργότερα. Αυτές οι επεκτάσεις θα μπορούν να χρησιμοποιηθούν άμεσα και χωρίς καμία περαιτέρω αλλαγή του συμβολισμού.

3.3.1.2 Ιεραρχικές Διαιρετικές Μέθοδοι

Θα δώσουμε τώρα μια μικρή επισκόπηση των ιεραρχικών διαιρετικών μεθόδων ομαδοποίησης. Στην αρχή μιας τέτοιας μεθόδου όλες οι οντότητες ανήκουν σε μία και μοναδική ομάδα. Στη συνέχεια αυτή η ομάδα χωρίζεται σε μικρότερα κομμάτια μέχρι να καταλήξουμε σε N ομάδες που περιέχουν από ένα αντικείμενο της βάσης. Μόλις πραγματοποιηθεί ο αρχικός διαχωρισμός, τα αντικείμενα μετακινούνται από την μια ομάδα στην άλλη ή εκτελούνται πιο εκλεπτυσμένες υποδιαιρέσεις των ήδη σχηματισμένων ομάδων [42, 34, 121].

Υπάρχουν δυο στρατηγικές διαίρεσης:

1. Μονοθετικές (Monothetic): Μονοθετική είναι μια ομάδα στην οποία όλες οι οντότητες έχουν προσεγγιστικά την ίδια τιμή για μια συγκεκριμένη μεταβλητή. Οι μονοθετικές ομάδες, δηλαδή χαρακτηρίζονται από οντότητες με συγκεκριμένες μεταβλητές στις οποίες συγκεκριμένες τιμές είναι απαραίτητες για να γίνουν οι οντότητες δεκτές ως μέλη αυτών των ομάδων.

   Οι μονοθετικές τεχνικές συνήθως χρησιμοποιούνται σε περιπτώσεις δυαδικών δεδομένων. Αρχικά, το σύνολο των δεδομένων χωρίζεται σε εκείνες τις περιπτώσεις που κατέχουν και σε εκείνες που δεν κατέχουν κάποιο χαρακτηριστικό. Αν οι διαιρέσεις είναι αυτού του απλού τύπου, τότε για δεδομένα που έχουν m χαρακτηριστικά υπάρχουν m ενδεχόμενες αρχικές διαιρέσεις του αρχικού χώρου και έτσι σχηματίζονται m – 1 ενδεχόμενες διαιρέσεις σε δυο υποσύνολα και ούτω καθεξής. Μια τέτοια διαίρεση ορίζεται ως μονοθετική και μια ιεραρχία από τέτοιες διαιρέσεις καλείται μονοθετική ομαδοποίηση [42].

2. Πολυθετικές (Polothetic): Μια πολυθετική ομάδα είναι μια ομάδα στην οποία όλες οι οντότητες έχουν προσεγγιστικά τις ίδιες τιμές σε ένα υποσύνολο συγκεκριμένων μεταβλητών. Οι πολυθετικές ομάδες, χαρακτηρίζονται από οντότητες που διαθέτουν ένα υποσύνολο συγκεκριμένων μεταβλητών που θα πρέπει να πάρουν συγκεκριμένη τιμή για να γίνουν δεκτές ως μέλη της ομάδας.

3.3.2 Αλγόριθμοι Βελτιστοποίησης συνάρτησης κόστους

3.3.2.1 Εισαγωγή

Μια από τις πιο διαδεδομένες οικογένειες ομαδοποίησης είναι αυτή που στηρίζεται στην βελτιστοποίηση μιας συνάρτησης κόστους F. Η F είναι μια συνάρτηση των διανυσμάτων του συνόλου δεδομένων X και παραμετρικοποιείται σε σχέση με το διάνυσμα των αγνώστων παραμέτρων θ, έτσι ώστε να περιγράφονται όσο το δυνατόν καλύτερα οι ομάδες που περιέχονται στα δεδομένα. Το διάνυσμα παραμέτρων θ εξαρτάται ισχυρά από το σχήμα των ομάδων.

3.3.2.2 Επαναληπτικές τεχνικές τμηματοποίησης

Η πλειονότητα των επαναληπτικών μεθόδων τμηματοποίησης (Iterative partitioning) διαχωρίζει ένα σύνολο οντοτήτων έτσι ώστε να βελτιστοποιείται κάποιο προκαθορισμένο κριτήριο, όπως παραδείγματος χάριν μια συνάρτηση κόστους. Οι μέθοδοι αυτής της οικογένειας προϋποθέτουν εκ των προτέρων το πλήθος των ομάδων στις οποίες θα χωριστούν οι οντότητες των δεδομένων μας [42, 34].

Ας δούμε όμως τα βασικά βήματα των επαναληπτικών μεθόδων [1]:

Βήμα 1ο. Γίνεται ένας αρχικός διαχωρισμός του συνόλου των δεδομένων σε κάποιο προκαθορισμένο πλήθος ομάδων και υπολογίζονται τα κέντρα των ομάδων αυτών.

Βήμα 2ο. Τοποθετείται κάθε σημείο του συνόλου δεδομένων στην ομάδα το κέντρο της οποίας είναι πιο κοντά στη συγκεκριμένη οντότητα.

Βήμα 3ο. Υπολογίζονται τα καινούργια κέντρα των ομάδων. Ας σημειώσουμε εδώ ότι οι ομάδες δεν ενημερώνονται μέχρι να ολοκληρωθεί ένα πέρασμα στα δεδομένα.

Βήμα 4ο. Επαναλαμβάνονται τα βήματα 2 και 3 μέχρις ότου κανένα σημείο δεδομένων να μην αλλάξει ομάδα.

Ας αναφέρουμε τώρα κάποια πλεονεκτήματα των επαναληπτικών μεθόδων ομαδοποίησης σε σχέση με τους ιεραρχικούς αλγόριθμους. Καταρχάς σε αντίθεση με τις ιεραρχικές μεθόδούς ομαδοποίησης οι επαναληπτικές μέθοδοι δουλεύουν απευθείας πάνω στον πίνακα δεδομένων (raw data) με αποτέλεσμα να μην απαιτείται ο υπολογισμός και η αποθήκευση ενός $N \times N$ πίνακα, του πίνακα ομοιότητας δηλαδή μεταξύ των δεδομένων. Αυτό έχει ως άμεση συνέπεια ότι αυτές οι μέθοδοι μπορούν να επεξεργαστούν μεγαλύτερα σύνολα δεδομένων. Κατά δεύτερο λόγο, οι αλγόριθμοι αυτοί εκτελούν περισσότερα από ένα περάσματα στο σύνολο των περιπτώσεων με αποτέλεσμα να μπορούν να διορθώσουν έναν μη ακριβή και φτωχό αρχικό διαχωρισμό των δεδομένων. Οι ομάδες που παράγονται είναι ομάδες single – rank και μη φωλιασμένες, πράγμα που σημαίνει ότι δεν είναι μέρος μιας ιεραρχίας. Αυτό έχει σαν συνέπεια ότι αν έναν αντικείμενο αποδοθεί σε μια ομάδα τότε αυτό μπορεί να αποδοθεί σε κάποιο άλλο βήμα σε κάποια άλλη ομάδα. Τέλος θα πρέπει να αναφερθεί ότι οι περισσότερες επαναληπτικές μέθοδοι δεν επιτρέπουν την παραγωγή επικαλυπτόμενων ομάδων [1, 34].

Παρά τα παραπάνω ελκυστικά χαρακτηριστικά των επαναληπτικών μεθόδων ομαδοποίησης, οι τεχνικές αυτές έχουν και σημαντικές αδυναμίες. Για να βρεθεί ένας βέλτιστος διαχωρισμός του συνόλου των δεδομένων θα πρέπει να σχηματιστούν όλες οι δυνατές τμηματοποιήσεις του και έπειτα να επιλεχθεί η καλύτερη. Αυτό όμως είναι υπολογιστικά αδύνατο. Αντί για αυτό όμως έχουν αναπτυχθεί πολλές ευρεστικές μέθοδοι οι οποίες εφαρμόζονται σε ένα δείγμα ενός μικρού υποσυνόλου των δεδομένων με την ελπίδα να βρεθεί ή τουλάχιστον να προσεγγιστεί ο βέλτιστος διαχωρισμός του συνόλου των δεδομένων [1].

Μπορούμε να διακρίνουμε τρεις μεγάλες κατηγορίες στις διαδικασίες που ακολουθούνται κατά την διάρκεια της εκτέλεσης των επαναληπτικών μεθόδων ομαδοποίησης:

1. Επιλογή της αρχικής διαμέρισης.
2. Τύπος περάσματος
3. Κριτήρια ομαδοποίησης

Υπάρχουν διαφορετικές πρακτικές οι οποίες μπορούν να υλοποιήσουν τις τρεις αυτές διαδικασίες, οι συνδυασμοί των οποίων μπορούν να οδηγήσουν σε διαφορετικούς αλγόριθμους ομαδοποίησης. Είναι αξιοσημείωτο ότι αν διαφορετικοί συνδυασμοί μεθόδων που υλοποιούν τις παραπάνω διαδικασίες χρησιμοποιηθούν για την δημιουργία αλγορίθμων και αυτοί χρησιμοποιηθούν στο ίδιο σύνολο δεδομένων, τότε οι ομάδες που θα πάρουμε ως εξαγόμενα αυτών των αλγορίθμων θα ενδέχεται να είναι πολύ διαφορετικά.[1]

Ας δούμε όμως από λίγο πιο κοντά αυτές τις διαδικασίες:

1. **Αρχική Διαμέριση.** Το πρόβλημα που πρέπει να αντιμετωπιστεί σ' αυτήν την φάση είναι η τμηματοποίηση του συνόλου των δεδομένων μας σε ένα συγκεκριμένο πλήθος ομάδων, έστω k. Συγκεκριμένα το πρόβλημα έγκειται στην επιλογή της κατάλληλης τιμής του k για το συγκεκριμένο σύνολο δεδομένων.

   Υπάρχουν δυο βασικοί τρόποι για να ξεκινήσει μια επαναληπτική μέθοδος [1, 42, 121]:

   - Ορισμός k αρχικών σημείων τα οποία είναι εκτιμήσεις των κέντρων των ομάδων. Για παράδειγμα ο MacQueen το 1967, επέλεξε τα k πρώτα σημεία του συνόλου των δεδομένων ως τις k αρχικές εκτιμήσεις (seeds) των κέντρων των ομάδων. Μετά από αυτήν την επιλογή κάθε σημείο της βάσης ανατίθεται στην ομάδα με το πλησιέστερο σ' αυτό κέντρο κατά το πρώτο πέρασμα.

   - Κατάλληλη αρχική διαμέριση, η οποία να εμπλέκει τον καθορισμό των k ομάδων. Τα κέντρα κάθε ομάδας υπολογίζονται ως ο πολυδιάστατος μέσος των περιπτώσεων σε κάθε ομάδα. Η αρχική διαμέριση είτε επιλέγεται τυχαία είτε επιλέγεται με κάποιον τρόπο από τον χρήστη.

2. **Τύπος περάσματος.** Ο παράγοντας αυτός αναφέρεται στον τρόπο με τον οποίο οι περιπτώσεις ανατίθενται στις ομάδες. Η ανάθεση των περιπτώσεων στις ομάδες γίνεται με σκοπό να βελτιστοποιηθεί κάποιο κριτήριο ομαδοποίησης. Υπάρχουν δυο βασικοί τρόποι ανάθεσης μιας περίπτωσης σε μια ομάδα [1, 12]:

   - **k – means περάσματα.** Αυτά αναφέρονται και ως περάσματα κοντινότερου κέντρου (nearest centroid sorting pass) ή και ως περάσματα επανάθεσης (reassignment pass). Το κέντρο κάθε ομάδας μπορεί να δοθεί

είτε από τον αριθμητικό μέσο όρο των στοιχείων της ομάδας εάν πρόκειται για αριθμητικά δεδομένα, είτε με κάποιον άλλον υπολογισμό του μέσου εάν πρόκειται για κατηγορικά δεδομένα. Η διαδικασία της επανάθεσης προχωρά μελετώντας κάθε περίπτωση, η οποία επανατοποθετείται στην ομάδα με το κοντινότερο κέντρο. Τα k – means περάσματα διακρίνονται στα συνδυαστικά και στα μη συνδυαστικά περάσματα. Στα συνδυαστικά περάσματα το κέντρο κάθε ομάδας επαναϋπολογίζεται κάθε φορά δηλαδή που υπάρχει αλλαγή στην σύνθεση της ομάδας, κάθε φορά που ένα στοιχείο προστίθεται στην ομάδα. Στα μη συνδυαστικά περάσματα αντίθετα το κέντρο κάθε ομάδας υπολογίζεται μόνο μετά από ένα ολοκληρωμένο πέρασμα των δεδομένων.

- **k – medoids περάσματα.** Σ' αυτά τα περάσματα μια ομάδα παριστάνεται από μια περίπτωσή της. Ως κέντρο σ΄ αυτήν την περίπτωση θεωρείται το στοιχείο εκείνο που χωρίζει την ομάδα σε δύο ίσα μέρη.

Εν γένει, η διαδικασία επανάθεσης προχωρά μελετώντας κάθε περίπτωση ξεχωριστά και η επανάθεση κάθε περίπτωσης γίνεται ανάλογα με το αν προκαλεί μείωση ή αύξηση στην τιμή κάποιου κριτηρίου. Η διαδικασία συνεχίζεται μέχρι του σημείου που καμιά πλέον μετακίνηση δεν προκαλεί βελτίωση του επιλεγμένου κριτηρίου ομαδοποίησης και η λύση είτε γίνεται δεκτή είτε επιχειρείται μια βελτίωσής της χρησιμοποιώντας μια διαφορετική αρχική διαμέριση [42, 121].

3. **Κριτήριο ομαδοποίησης.** Ένα μεγάλο πλήθος επαναληπτικών μεθόδων τμηματοποίησης βασίζεται στην μεγιστοποίηση (ελαχιστοποίηση) των μεταξύ των ομάδων (between – groups) ή της εντός των ομάδων (within – groups) διασκόρπισης [34].

Αν συμβολίσουμε με Τ τον ολικό πίνακα διασκόρπισης, με Β τον πίνακα της μεταξύ των ομάδων διασκόρπισης και με W τον πίνακα της εντός των ομάδων διασκόρπισης , δηλαδή $W = \sum_{i=1}^{g} W_i$ τότε ισχύει ότι:

$$T = W + B. \quad (3.11)$$

Τα κριτήρια που χρησιμοποιούνται συνήθως είναι τα ακόλουθα [1, 42, 34, 121]:

a) **Ίχνος του πίνακα W tr(W).** Το κριτήριο αυτό προσπαθεί να ελαχιστοποιήσει το ίχνος του πίνακα W. Παραλλαγές του έχουν προταθεί από τους Singelton & Kautz (1966), Friedman & Rubin (1967) και άλλους. Σε πολλές μεθόδους όπου το κριτήριο ομαδοποίησης δεν είναι σαφώς ορισμένο, αλλά η επανοτοποθέτηση εκτελείται μέχρις ότου όλες οι περιπτώσεις σε μια ομάδα να είναι πιο κοντά στο κέντρο της ομάδας στην οποία ανήκουν, μπορεί επίσης να θεωρηθεί ως μια προσπάθεια ελαχιστοποίησης του ίχνους του W.

Η ελαχιστοποίηση του ίχνους του πίνακα W είναι ισοδύναμη με την μεγιστοποίηση του ίχνους του πίνακα Β εφ' όσον ισχύει:

$$trace(T) = trace(W) + trace(B) \quad (3.12).$$

Οι ομάδες που σχηματίζονται με αυτό το κριτήριο είναι υπερσφαιρικές και πολύ ομογενείς. Το κριτήριο αυτό όμως επηρεάζεται από απλούς μετασχηματισμούς του πίνακα δεδομένων.

b) **Ορίζουσα του πίνακα W, det(W).** Αυτό το κριτήριο είναι ανεξάρτητο από μετασχηματισμούς στον πίνακα δεδομένων ή στην κλίμακα των δεδομένων. Οι ομάδες που δημιουργούνται χρησιμοποιώντας αυτό το κριτήριο δεν είναι υπερσφαιρικές, αλλά έχουν το ίδιο σχήμα. Και έχουν την τάση να σχηματίζουν ομάδες ίσου μεγέθους ακόμα και αν τέτοιες ομάδες δεν εμφανίζονται στα δεδομένα μας. Το κριτήριο αυτό προτάθηκε από τους Friedman & Rubin το (1967).

Η ελαχιστοποίηση της ορίζουσας του W ισοδυναμεί με την μεγιστοποίηση της του λόγου |T|/|W|.

c) **Ίχνος του πίνακα** BW^{-1}, $tr(BW^{-1})$. Το κριτήριο αυτό κριτήριο αυτό προσπαθεί να μεγιστοποιήσει το ίχνος του πίνακα BW^{-1}, ο οποίος λαμβάνεται από το γινόμενο του μεταξύ των ομάδων πίνακα διασκόρπισης και του αντίστροφου εντός των ομάδων πίνακα διασκόρπισης W. Το κριτήριο αυτό προτάθηκε από τους Friedman & Rubin το 1967.

Το $tr(BW^{-1})$ καθώς και το |T|/|W| μπορούν αν εκφραστούν με όρους των ιδιοτιμών του πίνακα BW^{-1} με τον παρακάτω τρόπο:

$$tr(BW^{-1}) = \sum_{i=1}^{p} \lambda_i \quad (3.13)$$

$$\frac{T}{W} = \prod_{i=1}^{p} (1 + \lambda_i) \quad (3.14)$$

όπου λ_i είναι οι ρίζες της χαρακτηριστικής εξίσωσης $|B - \lambda W| = 0$.

d) **Μέσος όρος Ευστάθειας Οντότητας (Average entity stability).** Ο Rubin το 1967 πρότεινε μια διαδικασία ομαδοποίησης που βασίζεται στην βελτιστοποίηση μιας ποσότητας που ονομάζεται όρος ευστάθειας οντότητας (average entity stability). Το μέτρο αυτό βασίζεται στην «έλξη» μιας οντότητας από μια ομάδα. Η «έλξη» μιας οντότητας από μια ομάδα είναι ο μέσος όρος ομοιότητας μεταξύ της οντότητας και των μελών της ομάδας. Αν μια οντότητα έλκεται περισσότερο από μια εξωτερική ομάδα παρά από την ομάδα στην οποία βρίσκεται τότε αυτή η οντότητα λέγεται ασταθής (unstable). Χρησιμοποιώντας την προσέγγιση αυτή ο Rubin προσπάθησε να ορίσει ένα μέτρο ευστάθειας της οντότητας (entity stability) για το αντικείμενο i στην ομάδα G χρησιμοποιώντας του εξής όρους:

- M = έλξη της i από την G [G – {i}], (δηλαδή από τις άλλες οντότητες της ομάδας).

- $\overline{M}$ = η μέγιστη έλξη της i από οποιαδήποτε ομάδα εκτός της G.

Ένας προφανής ορισμός της ευστάθειας ενός αντικειμένου είναι ο $M - \overline{M}$. Ωστόσο ο Rubin δεν χρησιμοποίησε αυτό το μέτρο άλλα όρισε την ευστάθεια μιας οντότητας, o_i ως εξής:

$$0_i = S^* M - (1 - S^*)\overline{M}, i = 1,..., N \quad (3.15)$$

με S* να είναι η έλξη ενός αντικειμένου από ένα άδειο σύνολο (μιας ομάδας δηλαδή που δεν έχει μέλη). Αυτός ο όρος χρειάζεται, ειδάλλως η τιμή M δεν θα έχει νόημα όταν η οντότητα i αποτελεί από μόνη της μια ομάδα. Αρχίζοντας από μια αρχική τμηματοποίηση βελτιστοποιείται ο μέσος όρος ευστάθειας, δηλαδή ο

$$\sum_{i=1}^{N} \frac{0_i}{N} \quad (3.16).$$

3.3.2.2.1 Ο αλγόριθμος k – means.

Σ' αυτήν την παράγραφο θα εξετάσουμε τον πολύ δημοφιλή αλγόριθμο k – means ως έναν από τους βασικούς επαναληπτικούς αλγόριθμους ο οποίος αποτελεί την βάση για πολλούς άλλους αλγορίθμους αυτού του τύπου καθώς και για τον αλγόριθμο που θα προτείνουμε παρακάτω. Εφαρμόζεται για την ομαδοποίηση μεγάλων συνόλων δεδομένων σε ομοιογενείς ομάδες [3, 12, 34, 36, 145, 146, 64, 121, 113, 115, 137].

Όπως σε όλες τις επαναληπτικές μεθόδους ομαδοποίησης, έτσι και ο k – means βασίζεται στην ιδέα της βελτιστοποίησης κάποιας συνάρτησης. Αυτή η συνάρτηση, αποτελεί μια προσπάθεια να αποδοθεί η έννοια της ομάδας μαθηματικά. Όπως είδαμε και παραπάνω η συνάρτηση αυτή αναφέρεται είτε ως κριτήριο ομαδοποίησης (clustering criterion) είτε ως συνάρτηση κόστους E. Η τιμή αυτής της συνάρτησης εξαρτάται συνήθως από την τρέχουσα ομαδοποίηση $\{C_1, C_2, ..., C_k\}$ του συνόλου των δεδομένων. Δηλαδή ισχύει

$$E : \wp_k(\Omega) \to \Re$$

όπου $\wp_k(\Omega)$ είναι το σύνολο όλων των τμηματοποιήσεων της βάσης δεδομένων $\Omega = \{w_1, w_2, ..., w_N\}$ σε k μη κενές ομάδες. Κάθε μια από τις N περιπτώσεις $w_i i = 1,2,..., N$ είναι ένα d – διάστατο διάνυσμα. Πιο συγκεκριμένα ο k – means βρίσκει τοπικά βέλτιστες λύσεις χρησιμοποιώντας για κριτήριο ομαδοποίησης E το άθροισμα των L^2 αποστάσεων μεταξύ κάθε στοιχείου και του κοντινότερου κέντρου της ομάδας. Το κριτήριο αυτό αναφέρεται και ως κριτήριο τετραγωνικού σφάλματος (square – error criterion). Επομένως προκύπτει ότι το κριτήριο ομαδοποίησης του αλγορίθμου k – means είναι το:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d\left(w_{ij}, \overline{w}_i\right) \text{ (3.17).}$$

όπου

k	είναι το πλήθος των ομάδων
n_i	είναι το πλήθος των στοιχείων της ομάδας C_i
w_{ij}	είναι η j – οστή περίπτωση της i – οστής ομάδας
$\overline{w}_i = \frac{1}{n_i} E = \sum_{j=1}^{n_i} w_{ij}$	Είναι το κέντρο (centroid) της ομάδας C_i
$d(x, y) = \|x - y\|^2$	Είναι η τετραγωνική Ευκλείδεια απόσταση

Ο αλγόριθμός k – means, αποτελείται από δύο φάσεις. Στην πρώτη φάση εκτελείται μια τμηματοποίηση των δεδομένων σε k ομάδες, ενώ κατά την διάρκεια της δεύτερης φάσης καθορίζεται η ποιότητα αυτής της τμηματοποίησης.

Ο k – means είναι μια διαδικασία τεσσάρων βημάτων η οποία αρχίζει από μια τυχαία τμηματοποίηση των δεδομένων στις k ομάδες. Ας δούμε όμως τα βήματα του αλγορίθμού πιο αναλυτικά:

1. Επιλογή των k αρχικών κέντρων για τις k ομάδες.
2. Υπολογισμός της ανομοιότητας μεταξύ ενός αντικειμένου και των κέντρων των k ομάδων.
3. Τοποθέτηση του αντικειμένου σε εκείνη την ομάδα της οποίας το κέντρο είναι πιο κοντά στο αντικείμενο αυτό.
4. Ενημέρωση του κέντρου της ομάδας έτσι ώστε να ελαχιστοποιηθεί η ανομοιότητα εντός της ομάδας.

Εκτός από το πρώτο βήμα, όλα τα άλλα βήματα του αλγορίθμου εκτελούνται επαναληπτικά μέχρις ότου ο αλγόριθμος συγκλίνει, μέχρις ότου δηλαδή να μην υπάρχει βελτίωση του κριτηρίου ομαδοποίησης.

Θα μπορούσαμε να τυποποιήσουμε την παραπάνω διαδικασία με το ακόλουθο πρόβλημα μαθηματικού προγραμματισμού P.

$$\text{Minimize} \quad P(W, Q) = \sum_{l=1}^{k} \sum_{i=1}^{N} w_{i,l} d\left(X_i, Q_l\right)$$

$$\text{subject to} \quad P(W,Q)=\sum_{l=1}^{k} w_{i,l}=1, w_{i,l}\in\{0,1\}, i=1,...,n, l=1,...,N$$

όπου W είναι ένας N×k πίνακας τμηματοποίησης, $Q=\{Q_1,Q_2,...,Q_k\}$ είναι ένα σύνολο αντικειμένων από τον χώρο αντικειμένων και $d(\cdot,\cdot)$ είναι η τετραγωνική Ευκλείδεια απόσταση μεταξύ δυο αντικειμένων.

Ο αλγόριθμος k – means έχει κάποιες σημαντικές ιδιότητες:

1. Η υπολογιστική πολυπλοκότητα του αλγορίθμου αυτού είναι της τάξης O(tkmn), όπου m είναι το πλήθος των χαρακτηριστικών, n είναι το πλήθος των αντικειμένων, k είναι το πλήθος των ομάδων και t είναι το πλήθος των επαναλήψεων πάνω σε όλο το σύνολο των δεδομένων. Στις περισσότερες περιπτώσεις το πλήθος των αντικειμένων είναι πολύ μεγαλύτερο από το πλήθος των χαρακτηριστικών, από το πλήθος των ομάδων και από το πλήθος των επαναλήψεων., ισχύει δηλαδή k, m, t << n. Στην ομαδοποίηση μεγάλων δεδομένων ο αλγόριθμος k – means είναι πολύ πιο γρήγορος από τους ιεραρχικούς αλγόριθμους ομαδοποίησης οι οποίοι έχουν γενικά πολυπλοκότητα $O(n^2)$.

2. Πολύ συχνά ο αλγόριθμος k – means τερματίζει σε ένα τοπικό βέλτιστο. Για να βρεθεί το ολικό βέλτιστο υιοθετούνται τεχνικές όπως οι γενετικοί αλγόριθμοι ή το deterministic annealing, οι οποίες σχετικά εύκολα μπορούν να ενσωματωθούν στον αλγόριθμο.

3. Τα δεδομένα που εμπλέκονται στον αλγόριθμο θα πρέπει να λαμβάνουν μόνο αριθμητικές τιμές, αφού ο αλγόριθμος προσπαθεί να ελαχιστοποιήσει την συνάρτηση κόστους υπολογίζοντας τους μέσους όρους των ομάδων.

4. Το γεγονός ότι οι ομάδες που ανακαλύπτει ο αλγόριθμος k – means έχουν κυρτά σχήματα καθιστά δυσχερή την χρήση του αλγορίθμου αυτού για τον εντοπισμό μη κυρτών ομάδων.

Παρ' όλη την αποτελεσματικότητα του και τη διαδεδομένη χρήση του ο αλγόριθμος K – means έχει και αρκετά μειονεκτήματα, τα σημαντικότερα από τα οποία αναλύονται παρακάτω:

1. Υποθέτει ότι το πλήθος των ομάδων k σε μια βάση δεδομένων είναι εκ των προτέρων γνωστό, κάτι το οποίο δεν είναι απαραίτητα σωστό στις περισσότερες εφαρμογές του πραγματικού κόσμου.

2. Ως μια επαναληπτική μέθοδος, ο αλγόριθμος k – means είναι αρκετά ευαίσθητος στην επιλογή των αρχικών ομάδων. Επίσης επηρεάζεται σημαντικά από την παρουσία θορύβου και απομακρυσμένων τιμών (outliers).

3. Συγκλίνει σε ένα τοπικό ελάχιστο συνήθως κακής ποιότητας, δηλαδή η λύση που προτείνει δεν είναι πάντα ικανοποιητική.

4. Ο αλγόριθμος θεωρεί τα k κέντρα ως αντιπροσώπους των δεδομένων κάθε ομάδας. Ωστόσο μερικές φορές είναι δυνατόν ο αριθμητικός μέσος να μην έχει καμία έγκυρη ερμηνεία.

5. Εξαιτίας του χρόνου που χρειάζεται για να ολοκληρωθεί μια επανάληψη δεν μπορεί να χειριστεί μεγάλες βάσεις δεδομένων γρήγορα.

6. Τα δεδομένα που μπορεί να επεξεργαστεί είναι μόνο αριθμητικά. Ο αλγόριθμος k – means δεν μπορεί να χειριστεί κατηγορικά ή μεικτά δεδομένα.

3.3.2.2.2 K – modes

Το 1998 ο Zhexue Huang πρότεινε έναν γρήγορο αλγόριθμο ομαδοποίησης, για την ομαδοποίηση κατηγορικών δεδομένων. Ο αλγόριθμος αυτός ονομάζεται k – modes και είναι μια επέκταση του αλγορίθμου k – means που έχει αναλυθεί παραπάνω.

Όπως έχουμε αναφέρει και παραπάνω ο αλγόριθμος k – means μπορεί να χειριστεί και να ομαδοποιήσει μόνο αριθμητικά δεδομένα. Στις εφαρμογές όμως συχνά της εξόρυξης δεδομένων εμπλέκονται πολλές φορές κατηγορικά δεδομένα. Η αντιστοίχιση των κατηγορικών αυτών δεδομένων σε αριθμητικές τιμές και η χρησιμοποίηση ενός αλγορίθμου που χειρίζεται μόνο αριθμητικά δεδομένα όπως παραδείγματος χάριν ο k – means δεν είναι απαραίτητο να μας δώσει ερμηνεύσιμα αποτελέσματα.

Ο k – modes δεν διαφέρει από τον k – means μόνο στο ότι χρησιμοποιεί κατηγορικά δεδομένα αλλά και στο ότι

a. Χρησιμοποιεί ένα απλό μέτρο ανομοιότητας για τα κατηγορικά αντικείμενα

b. Χρησιμοποιεί τον μέσο (mode) μιας ομάδας αντί για το κέντρο της

c. Χρησιμοποιεί μια μέθοδο που βασίζεται στην συχνότητα εμφάνισης (frequency – based) των περιπτώσεων, για την ενημέρωση των μέσων και συνεπώς για την ελαχιστοποίηση του κριτηρίου ομαδοποίησης.

Παρακάτω θα δώσουμε κάποιους ορισμούς και κάποιους συμβολισμού για την καλύτερη κατανόηση του αλγορίθμου.

> **Κατηγορικά (Categorical) δεδομένα** θεωρούνται εκείνα τα δεδομένα που περιγράφουν αντικείμενα τα οποία αποτελούνται μόνο από κατηγορικά χαρακτηριστικά.

Έστω $A_1, A_2, ..., A_m$ είναι m χαρακτηριστικά, που το καθένα από αυτά ορίζεται σε μια περιοχή $DOM(A_i)$.

> Η περιοχή $DOM(A_i)$ καλείται **κατηγορική περιοχή**, αν είναι πεπερασμένη και μη διατεταγμένη. Για παράδειγμα για κάθε $a, b \in DOM(A_i)$ θα ισχύει είτε $a = b$ είτε $a \neq b$.

Το A_i καλείται **κατηγορικό χαρακτηριστικό** και ο χώρος των χαρακτηριστικών Ω καλείται **κατηγορικός χώρος** εφ' όσον όλα τα $A_1, A_2, ..., A_m$ είναι κατηγορικά.

Μια ειδική τιμή που συμβολίζεται με ε, ορίζεται σε όλες τις κατηγορικές περιοχές και χρησιμοποιείται για να περιγράψει τις τιμές που λείπουν (missing values).

Ένα κατηγορικό αντικείμενο αναπαρίσταται στην γλώσσα της λογικής ως μια τομή ζευγαριών χαρακτηριστικού τιμής:

$$[A_1 = x_1] \wedge [A_2 = x_2] \wedge ... \wedge [A_m = x_m] \quad (3.18)$$

όπου $x_j \in DOM(A_j), j = 1,2,...,m$.

Ένα ζεύγος χαρακτηριστικού τιμής $[A_j = x_j]$ καλείται **επιλογέας (selector).**

Για λόγους απλότητας θα συμβολίζουμε το κατηγορικό αντικείμενο X ως ένα διάνυσμα

$$X = [x_1^r, x_2^r, ..., x_p^r, x_p^r + 1^c, ..., x_m^c] \quad (3.19)$$

όπου τα πρώτα p στοιχεία είναι αριθμητικές τιμές και τα υπόλοιπα είναι οι κατηγορικές τιμές.

Αν τα στοιχεία του X παίρνουν τιμές από έναν μόνο τύπο τότε το διάνυσμα X θα συμβολίζεται με

$$X = [x_1, x_2, ...x_m] \quad (3.20).$$

Κάθε αντικείμενο του Ω θα έχει m χαρακτηριστικές τιμές. Αν η τιμή του χαρακτηριστικού A_j λείπει τότε θα έχουμε $A_j = \varepsilon$.

Έστω $X = \{X_1, X_2, ..., X_m\}$ ένα σύνολο από n κατηγορικά αντικείμενα και $X \subseteq \Omega$. Το αντικείμενο X_i παριστάνεται ως $[x_{i1}, x_{i2}, ...x_{im}]$. Αν $x_{i,j} = x_{k,j}, j = 1,2,...,m$ τότε $X_i = X_k$. Αυτή η σχέση δεν σημαίνει πως τα αντικείμενα X_i και X_k ταυτίζονται στην πραγματική βάση δεδομένων, αλλά πως τα δυο αυτά αντικείμενα έχουν ίσες κατηγορικές τιμές για τα χαρακτηριστικά $A_1, A_2, ..., A_m$.

Το μέτρο ανομοιότητας δυο κατηγορικών αντικειμένων X και Y που περιγράφονται από m χαρακτηριστικά, ορίζεται ως οι συνολικές διαφορές (mismatches) των αντίστοιχων κατηγορικών χαρακτηριστικών των δυο αντικειμένων. Όσο μικρότερο είναι το πλήθος των διαφορών τόσο περισσότερα όμοια είναι τα αντικείμενα. Τυπικά έχουμε:

$$d(X,Y) = \sum_{i=1}^{m} \delta(x_i, y_i) \quad (3.21)$$

όπου

$$\delta(x_i, y_i) = \begin{cases} 0, x_i = y_i \\ 1, x_i \neq y_i \end{cases} \quad (3.22).$$

Το μέτρο αυτό δίνει ίση σημασία σε κάθε κατηγορία ενός χαρακτηριστικού. Αν ληφθούν υπ' όψιν οι συχνότητες εμφάνισης των κατηγοριών στο σύνολο των δεδομένων μπορεί να οριστεί το ακόλουθο μέτρο ανομοιότητας:

$$d_{X^2}(X,Y) = \sum_{i=1}^{m} \frac{n_{x_i} + n_{y_j}}{n_{x_i} n_{y_i}} \delta(x_i, y_i) \quad (3.23)$$

όπου n_{x_i}, n_{y_i} είναι το πλήθος των αντικειμένων στο σύνολο δεδομένων που έχουν τις κατηγορίες x_i, y_i για το χαρακτηριστικό i. Αυτή η απόσταση καλείται x^2 απόσταση. Αυτό το μέτρο ανομοιότητας δίνει μεγαλύτερη έμφαση σε όχι συχνά εμφανιζόμενες κατηγορίες.

Έστω ένα σύνολο από κατηγορικά αντικείμενα που περιγράφονται από τα χαρακτηριστικά $A_1, A_2, ..., A_m$.

> Ένας **μέσος (mode)** του X είναι ένα διάνυσμα $Q = [q_1, q_2, ...q_m] \in \Omega$ ελαχιστοποιεί το:
>
> $$d(Q,X) = \sum_{i=1}^{n} d(X_i, Q) \quad (3.24)$$
>
> όπου $X = \{X_1, X_2, ..., X_m\}$ και d μπορεί να οριστεί είτε από τη σχέση (3.21) είτε από την σχέση (3.23). Το Q δεν είναι απαραίτητο να είναι στοιχείο του X.

Έστω ότι $n_{c_{k,j}}$ είναι το πλήθος των αντικειμένων που έχουν ως τιμή του χαρακτηριστικού A_j την κατηγορία $c_{k,j}$ και $f_r(A_j = c_{k,j} \mid X) = \frac{n_{c_{k,j}}}{n}$ είναι η σχετική συχνότητα της κατηγορίας $c_{k,j}$ στο X. Ισχύει το ακόλουθο θεώρημα:

> **Θεώρημα.** Η συνάρτηση D(Q,X) ελαχιστοποιείται αν και μόνο αν $f_r(A_j = q_j \mid X) \geq f_r(A_j = c_{k,j} \mid X)$ για $q_j \neq c_{k,j}$ και για όλα τα j = 1, 2, …, m.

Το θεώρημα καθορίζει έναν τρόπο εύρεσης του μέσου Q ενός δοσμένου συνόλου X και επομένως είναι σημαντικό διότι επιτρέπει την χρήση του k – means για ομαδοποίηση κατηγορικών δεδομένων χωρίς μείωση της απόδοσης του αλγορίθμου.

Επίσης υπονοεί ότι ο μέσος του συνόλου δεδομένων δεν είναι μοναδικός. Για παράδειγμα ο μέσος του συνόλου

{[a,b], [a,c], [c,d], [b,c]}

μπορεί να είναι είτε ο [a,b] είτε ο [a,c].

Έστω $\{S_1, S_2, ..., S_k\}$ μια τμηματοποίηση του X, όπου $S_l \neq \text{ø}$ για $1 \neq l \neq k$ και $\{Q_1, Q_2, ..., Q_k\}$ είναι οι μέσοι των ομάδων $\{S_1, S_2, ..., S_k\}$. Το ολικό κόστος της τμηματοποίησης ορίζεται ως εξής:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{i,l} d(X_i, Q_l) \quad (3.25)$$

όπου $y_{i,l}$ είναι ένα στοιχείο του πίνακα τμηματοποίησης $Y_{n \times l}$ και d μπορεί να οριστεί από την εξίσωση (3.21) ή από την εξίσωση (3.23).

Ο σκοπός του αλγορίθμου k – modes είναι η εύρεση ενός συνόλου $\{Q_1, Q_2, ..., Q_k\}$ που να μπορεί να ελαχιστοποιεί την συνάρτηση E. Ο αλγόριθμος k – modes αποτελείται από τα ακόλουθα βήματα:

1. Επιλογή των k αρχικών μέσων, ένας μέσος για κάθε ομάδα.

2. Τοποθέτηση ενός αντικειμένου στην ομάδα της οποίας ο μέσος είναι ο κοντινότερος σε αυτό σύμφωνα με το μέτρο απόστασης d που θα χρησιμοποιηθεί. Ενημέρωση του μέσου της ομάδας μετά από κάθε ανάθεση σύμφωνα με το παραπάνω θεώρημα.

3. Όταν όλα τα αντικείμενα έχουν τοποθετηθεί στις ομάδες ελέγχεται ξανά η ανομοιότητα των αντικειμένων από τους μέσους των ομάδων. Αν βρεθεί αντικείμενο τέτοιο ώστε ο κοντινότερος μέσος προς αυτό να ανήκει σε άλλη ομάδα από αυτήν που το αντικείμενο βρίσκεται, τότε τοποθετείται στην νέα ομάδα και ενημερώνονται οι μέσοι των δυο ομάδων.

4. Επαναλαμβάνεται το τρίτο βήμα έως ότου κανένα αντικείμενο να μην αλλάζει ομάδα μετά από τον έλεγχο σε όλο το σύνολο των αντικειμένων.

$$\begin{Bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & c_{1,4} \\ c_{2,1} & c_{2,2} & c_{2,3} & c_{2,4} \\ c_{3,1} & c_{3,1} & c_{3,3} & c_{3,4} \\ c_{4,1} & & c_{4,3} & \\ & & c_{5,3} & \end{Bmatrix}$$

Εικόνα 3.1: Το διάνυσμα των κατηγοριών του συνόλου δεδομένων με 4 χαρακτηριστικά έχοντας 4, 2, 5 και 3 κατηγορίες αντίστοιχα.

Όπως και ο αλγόριθμος k – means έτσι και ο k – modes παράγει τοπικά βέλτιστες λύσεις που εξαρτώνται από την επιλογή των αρχικών μέσων και τη διάταξη των αντικειμένων στο σύνολο των δεδομένων.

Ο Huang πρότεινε δυο μεθόδους επιλογής των αρχικών k μέσων. Η πρώτη μέθοδος επιλέγει τις πρώτες k διακριτές εγγραφές από το σύνολο των δεδομένων ως τους k αρχικούς μέσους. Η δεύτερη μέθοδος αποτελείται από τα ακόλουθα βήματα:

1. Υπολογίζονται οι συχνότητες όλων των κατηγοριών για όλα τα χαρακτηριστικά και αποθηκεύονται σε ένα διάνυσμα με φθίνουσα σειρά συχνοτήτων, όπως φαίνεται στο παρακάτω σχήμα. Το $c_{i,j}$ συμβολίζει την κατηγορία i του χαρακτηριστικού j και $f(c_{i,j}) \geq f(c_{i+1,j})$ όπου $f(c_{i,j})$ είναι η συχνότητα της κατηγορίας $c_{i,j}$.

2. Ανατίθενται οι πιο συχνές συχνότητες κατηγοριών ισοδύναμα στους k αρχικούς μέσους. Για παράδειγμα αν υποτεθεί ότι k = 3 τότε

$$Q_1 = [q_{1,1} = c_{1,1}, q_{1,2} = c_{2,2}, q_{1,3} = c_{3,3}, q_{1,4} = c_{1,4}]$$

$$Q_2 = [q_{2,1} = c_{2,1}, q_{2,2} = c_{1,2}, q_{2,3} = c_{4,3}, q_{2,4} = c_{2,4}]$$

$$Q_3 = [q_{3,1} = c_{3,1}, q_{3,2} = c_{2,2}, q_{3,3} = c_{1,3}, q_{31,4} = c_{3,4}]$$

3. Αρχίζοντας με τον μέσο Q_1, επιλέγεται η εγγραφή που είναι πιο όμοιο με τον Q_1 και αντικαθίσταται το Q_1 με την εγγραφή αυτή. Έπειτα, επιλέγεται η πιο όμοια εγγραφή ως προς τον Q_2 και αντικαθίσταται ο μέσος αυτός από την νέα εγγραφή. Η διαδικασία αυτή συνεχίζεται έως ότου αντικατασταθεί και ο Q_k. Για την επιλογή της πιο όμοιας εγγραφής θα πρέπει να ισχύει ότι $Q_l \neq Q_t, \gamma\iota\alpha l \neq t$.

Ο σκοπός της μεθόδου επιλογής είναι η δημιουργία όσο το δυνατόν πιο διαφορετικών μέσων, οι οποίοι θα δώσουν καλύτερα αποτελέσματα ομαδοποίησης.

3.3.2.2.3 Ο αλγόριθμος k – windows

Θα αναφερθούμε τώρα στον αλγόριθμο k – windows ο οποίος αποτελεί μια βελτίωση του αλγορίθμου k – means, τόσο στον τομέα της χρονικής πολυπλοκότητας όσο και στην καλύτερη ποιότητα των ομάδων που παράγει [137]. Ο Αλγόριθμος αυτός όπως θα δούμε παρακάτω επεξεργάζεται μόνο αριθμητικές τιμές. Παρόλα αυτά όμως η απόσταση μεταξύ κατηγορικών δεδομένων που θα προτείνουμε αργότερα θα μπορεί να χρησιμοποιηθεί και σε αυτόν τον αλγόριθμο έτσι ώστε να τον κάνει να χειρίζεται και κατηγορικά δεδομένα. Αλλά ας εξετάσουμε κατ' αρχάς με κάποια λεπτομέρεια τον αλγόριθμο αυτόν.

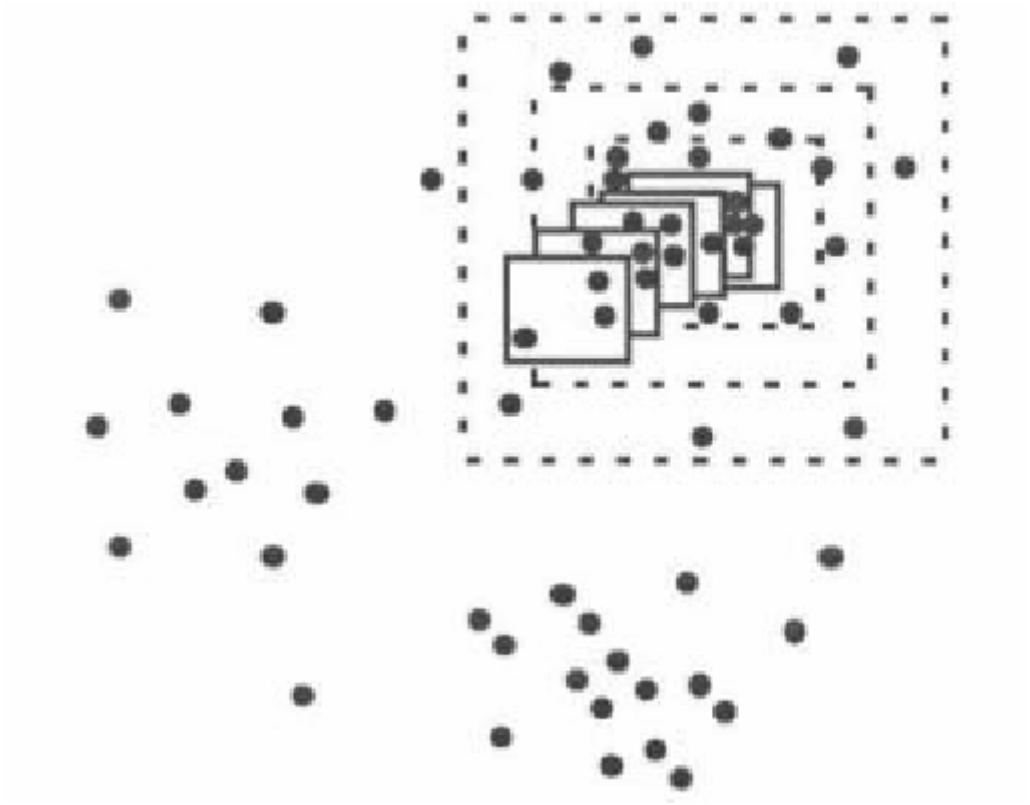
Ας δούμε όμως πρώτα σε τι υστερεί ο k – means έναντι του k – windows. Ο k – means καταναλώνει κυρίως χρόνο στο βήμα της ανάθεσης κάθε προτύπου στην

ομάδα με τον κοντινότερο μέσο. Επίσης μια βασική χρονοβόρα λειτουργία του αλγορίθμου αυτού είναι ο υπολογισμός της τετραγωνικής ευκλείδειας απόστασης. Από την άλλη μεριά ο αλγόριθμος k – windows καταφέρνει να μειώσει σημαντικά το πλήθος των υπό εξέταση περιπτώσεων χρησιμοποιώντας μια παραθυρική τεχνική. Επιπλέον, χρόνος κερδίζεται και κατά την βασική λειτουργία στο πρώτο βήμα όπου η ανάθεση των προτύπων στις ομάδες δεν είναι τίποτα άλλο από μια απλή αριθμητική σύγκριση.

Η βασική ιδέα αυτής της τεχνικής είναι η χρήση ενός παραθύρου για τον καθορισμό της ομάδας. Ένα παράθυρο είναι μια ορθογώνια περιοχή σε έναν d – διάστατο ευκλείδειο χώρο, με το d να συμβολίζει το πλήθος των αριθμητικών χαρακτηριστικών που περιγράφουν τις περιπτώσεις μας. Κάθε παράθυρο δηλαδή είναι μια d – περιοχή αρχικά σταθερού εμβαδού α. Η σημασία του αριθμού α εξαρτάται από την πυκνότητα των δεδομένων μας. Μια πρόταση για τον ορισμό του α είναι να ορίζεται κατά μήκος κάθε κατεύθυνσης i ως εξής :

$$a_i = \frac{\text{μέση απόσταση των προτύπων στη } \nu i}{\text{πλήθος παραθύρων}} \times 0.5 \quad (3.26)$$

διαισθητικά αυτό που προσπαθείται είναι να γεμίσει ο χώρος των προτύπων με μη επικαλυπτόμενα παράθυρα. Κάθε πρότυπο που ανήκει σε ένα παράθυρο θεωρείται ότι ανήκει σε μια συγκεκριμένη ομάδα. Επαναληπτικά κάθε παράθυρο μετακινείται στον ευκλείδειο χώρο έχοντας ως κέντρο του πάντα το κέντρο το μέσο των προτύπων που περιέχει. Αυτή η διαδικασία συνεχίζεται μέχρις εκείνο το σημείο που καμία πλέον μετακίνηση δεν θα επιφέρει αύξηση στο πλήθος των προτύπων που βρίσκονται μέσα στο παράθυρο (Εικόνα 3.2). Έπειτα καθορίζονται τα κέντρα των ομάδων ως τα κέντρα των αντίστοιχων παραθύρων. Επειδή λίγα μόνο πρότυπα θα περιέχονται μέσα στο παράθυρο, αυτό μεγεθύνεται με σκοπό να περιλάβει όσο το δυνατόν περισσότερα πρότυπα (διακεκομμένες γραμμές στην Εικόνα 3.2).



Εικόνα 3.2: Μετακινήσεις και μεγεθύνσεις ενός παραθύρου.

Στο πρώτο βήμα του αλγορίθμου επιλέγονται τυχαία k μέσοι. Οι d – περιοχές (παράθυρα) έχουν μέσα αυτούς τους μέσους και έχουν εμβαδόν α. Στο δεύτερο βήμα μπορούν να βρεθούν τα πρότυπα που βρίσκονται μέσα σ' αυτές τις περιοχές. Για τον σκοπό αυτό χρησιμοποιείται η τεχνική της ορθογώνιας αναζήτησης περιοχής (orthogonal range search), η οποία βασίζεται σε μια φάση προεργασίας όπου κατασκευάζεται ένα δέντρο περιοχής (range tree). Τα πρότυπα που βρίσκονται μέσα σε μια d – περιοχή μπορούν να βρεθούν διανύοντας το δέντρο της περιοχής σε πολυλογαριθμικό χρόνο. Στο τρίτο βήμα υπολογίζονται οι μέσοι των προτύπων κάθε περιοχής. Ο μέσος των ομάδων καθορίζεται ως ο μέσος των αντίστοιχων παραθύρων. Έπειτα κάθε παράθυρο μετακινείται στον ευκλείδειο χώρο έτσι ώστε ο νέος μέσος των προτύπων του παραθύρου να βρίσκεται στο κέντρο του παραθύρου. Τα δύο τελευταία βήματα επαναλαμβάνονται μέχρι που καμία μετακίνηση να μην επιφέρει αύξηση στον αριθμό των προτύπων που βρίσκονται μέσα στο παράθυρο.

Ωστόσο αφού περιορισμένος αριθμός περιπτώσεων μελετάται σε κάθε μετακίνηση, η ποιότητα της ομαδοποίησης ίσως δεν είναι η βέλτιστη. Για τον λόγο αυτό η ποιότητα της τμηματοποίησης βελτιώνεται στο δεύτερο στάδιο του αλγορίθμου. Η ποιότητα της ομαδοποίησης καθορίζεται από το πόσα πρότυπα βρίσκονται σε κάθε παράθυρο σε σχέση με όλα τα πρότυπα. Στην αρχή μεγεθύνεται το παράθυρο έτσι ώστε να περιλάβει όσο το δυνατόν περισσότερα πρότυπα από την αντίστοιχη ομάδα. Αυτό επιτυγχάνεται αναγκάζοντας τους μέσους κάθε παραθύρου να παραμένουν σταθεροί και αυξάνοντας τον όγκο τους. Στην συνέχεια υπολογίζονται η σχετική συχνότητα των προτύπων που ανατίθενται σε κάθε d – περιοχή ως προς όλο το σύνολο των προτύπων. Αν η συγκεκριμένη είναι μικρότερη από ένα κατώφλι υ (ορισμένο από τον χρήστη), τότε είναι πιθανόν να υπάρχει ομάδα (ή ομάδες) της οποίας η παράλειψη να είναι δυνατή. Την περίπτωση αυτή η διαδικασία επαναλαμβάνεται.

Η νέα επανάληψη μπορεί να αρχίσει με τους ίδιους αρχικούς μέσους με μεγαλύτερο όμως εμβαδόν $a' > a$ για τις d – περιοχές. Μια άλλη προσέγγιση είναι να ξαναρχίσει ο αλγόριθμος διαλέγοντας καινούργιους μέσους τοποθετημένους όμως στην μεγαλύτερη δυνατή απόσταση από τους προηγούμενους.

Η τιμή του α καθορίζεται από τον χρήστη. Το παραπάνω κατώφλι χρησιμοποιείται ως κριτήριο τερματισμού του αλγορίθμου. Η τιμή του δεν παίζει κάποιον ρόλο στην επιτυχή επιλογή των αρχικών d – περιοχών.

Τέλος ας σημειώσουμε ότι η βασική λειτουργία του αλγορίθμου k – windows είναι η σύγκριση μεταξύ δυο αριθμών. Τέτοιες συγκρίσεις οδηγούν την διαδρομή στο δέντρο περιοχής από τη ρίζα προς τα φύλλα.

Στο δεύτερο βήμα του αλγορίθμου χρησιμοποιείται η ορθογώνια αναζήτηση περιοχής (orthogonal range search) που είναι μια έννοια της υπολογιστικής γεωμετρίας. Το πρόβλημα της ορθογώνιας αναζήτησης περιοχής μπορεί να τεθεί ως εξής:

Είσοδος:

1. $V = \{p_{1,} p_{2,}..., p_n\}$ είναι ένα σύνολο σημείων του $\Re^d$, d – διάστατου ευκλείδειου χώρου, με συντεταγμένες αξόνων $\{Ox_{1,} Ox_{2,}..., Ox_n\}$.
2. Ένα ερώτημα d – περιοχής $Q = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_n, b_n]$ που ορίζεται μεταξύ δυο σημείων $(a_1, a_2, ..., a_n)$ και $(b_1, b_2, ..., b_n)$ με $a_j \le b_j \, j = 1,2,...,n$.

Έξοδος:

Αναφέρονται όλα τα σημεία του V που βρίσκονται μέσα στην d – περιοχή του X.

Η γνωστή μέθοδος του δέντρου περιοχής (range tree) επιτρέπει την επίλυση των ορθογωνίων ερωτημάτων περιοχής σε $O(n \log^d n)$ χρόνο και χώρο προεπεξεργασίας και σε $O(s + \log^d n)$ χρόνο αναζήτησης για s σημεία. Ένα d – διάστατο δέντρο περιοχής αποτελείται από ένα ισορροπημένο δυαδικό δέντρο αναζήτησης με T φύλλα, το οποίο αποθηκεύει στα φύλλα του τα σημεία του δοσμένου συνόλου V κατά αύξουσα σειρά σε σχέση με την πρώτη τους συντεταγμένη. Σε κάθε εσωτερικό κόμβο p_i του δέντρου αποθηκεύεται το σημείο που εμφανίζεται στο δεξιότερο φύλλο του αριστερού υποδέντρου του p_i. Κάθε εσωτερικός κόμβος του p_s ενός (d – 1) – διάστατου δέντρου περιοχής είναι συνδεδεμένο με ένα (d – i – 1) – διάστατο δέντρο περιοχής $T_{p_s} \quad \forall i \in \{1,2,...,d-1\}$, το οποίο αποθηκεύει στα φύλλα του όλα τα σημεία του υποδέντρου, που έχει ρίζα το p_s κατά αύξουσα σειρά σε σχέση με την i + 1 συντεταγμένη των σημείων αυτών.

Για να εκτελεστεί μια αναζήτηση περιοχής αρχικά αναζητούνται τα a_1 και b_1 με $a_1 < b_1$ σε ένα d – διάστατο δέντρο περιοχής T με σκοπό να βρεθούν δυο φύλλα, έστω τα $p_a = \left(x_1^a, x_2^a, ..., x_d^a\right)$ και $p_b = \left(x_1^b, x_2^b, ..., x_d^b\right)$ στο T, έτσι ώστε p_a να είναι το

κοντινότερο πριν το a_1 ($x_1^a < a_1$) και το p_b να είναι το κοντινότερο μετά το b_1 ($x_1^b < b_1$). Για ορισμένο χρόνο η αναζήτηση για τα a_1 και b_1 ίσως ακολουθήσει το ίδιο μονοπάτι, αλλά σε κάποιον κόμβο p_t το a_1 θα βρίσκεται από το αριστερότερο παιδί του p_t, ενώ το b_1 θα βρίσκεται πάνω από το δεξιότερο παιδί του p_t.

Έστω ότι οι κόμβοι $p_w \in T$ είναι είτε δεξιά παιδιά ενός κόμβου στο μονοπάτι $p_t,...,p_a$ είτε αριστερά παιδιά στο μονοπάτι $p_t,...,p_b$. Επειδή το δέντρο είναι ισορροπημένο, υπάρχουν το πολύ $O(\log n)$ τέτοιοι κόμβοι p_w. Στην συνέχεια αναζητούνται τα a_2 και b_2 σε κάθε δέντρο περιοχής T_{p_w}. Στην αναζήτηση αυτή, η ένωση των απαντήσεων των $O(\log n)$ (d – 1) – διάστατων δέντρων περιοχής T_{p_w} αποτελείται το πολύ από $O(\log^2 n)$. Στο τέλος του αλγορίθμου η αναζήτηση των a_d και b_d λαμβάνει χώρα το πολύ σε $O(\log^{d-1} n)$ δισδιάστατα δέντρα περιοχής και θα βρεθούν το πολύ $O(\log^d n)$ μονοδιάστατα δέντρα περιοχής. Προφανώς όλα τα σημεία που θα αποθηκευτούν ως φύλλα στα τελευταία δέντρα θα βρίσκονται μέσα στο Q.

Η ανάθεση των προτύπων σε μια d – περιοχή χρειάζεται $O(s + \log^{d-2} n)$ χρόνο όπου s είναι το πλήθος των προτύπων που βρίσκονται μέσα στην d – περιοχή. Σημειώνεται πως οι d – περιοχές είναι αρκετά μικρές έτσι ώστε s << n. Επομένως σε κάθε μετακίνηση (επανάληψη), το βήμα της ανάθεσης προτύπων στις d – περιοχές έχει χρονική πολυπλοκότητα $O(k(s + \log^{d-2} n))$. Το βήμα όπου υπολογίζονται οι μέσοι των d – περιοχών χρειάζεται $O(sdk)$ χρόνο. Είναι φανερό πως sk << n. Τέλος η ποιότητα της συνάρτησης επίσης υπολογίζεται σε $O(sdk)$ χρόνο. Έτσι η πολυπλοκότητα του αλγορίθμου k – windows είναι $O(dkqr(s + \frac{\log^{d-2} n}{d}))$, όπου q είναι το πλήθος των μετακινήσεων (επαναλήψεων) και r είναι το πλήθος των επαναλήψεων που οφείλεται στις ομάδες που παραλείπονται.

3.3.2.2.4 Αλγόριθμος Z – Windows

Στην συνέχεια θα εξετάσουμε άλλον έναν αλγόριθμο ομαδοποίησης ο οποίος βασίζεται στον k – windows αλγόριθμο και ο οποίος ομαδοποιεί αριθμητικά δεδομένα. Όπως και ο k – windows θα δείξουμε σε λίγο πολύ εύκολά μπορεί να μετατραπεί χρησιμοποιώντας την απόσταση που θα προτείνουμε έτσι ώστε να χειρίζεται και κατηγορικά δεδομένα. Θα επανέλθουμε όμως λίγο αργότερα σ' αυτό το ζήτημα. Ας δώσουμε προς το παρόν, για χάρη της πληρότητας, μια κάπως σε βάθος εξέταση αυτού του αλγορίθμου.

Στην διαδικασία της ομαδοποίησης υπάρχει ένα θεμελιώδες θέμα, ανεξάρτητο από τις συγκεκριμένες τεχνικές ομαδοποίησης. Αυτό δεν είναι άλλο από τον προσδιορισμό του πλήθους των ομάδων που υπάρχουν στα δεδομένα. Αυτό το ουσιαστικό ερώτημα παραμένει ένα άλυτο πρόβλημα της ομαδοποίησης [17]. (Παραδείγματος χάριν όπως είδαμε και παραπάνω ένα από τα δεδομένα εισόδου του αλγορίθμου k – means είναι και το πλήθος των ομάδων που θέλουμε να ανακαλύψουμε στα δεδομένα).

Ο αλγόριθμος Z – windows λοιπόν επιχειρεί να λύσει το πρόβλημα αυτό. Ο Αλγόριθμος αυτός βασίζεται στον αλγόριθμο k – windows, ο οποίος όπως είπαμε και παραπάνω βασίζεται σε μια παραθυρική τεχνική για την μείωση των προτύπων που θα πρέπει να εξεταστούν ως προς την ομοιότητα σε κάθε επανάληψη. Η ιδέα κλειδί πίσω απ΄ τον αλγόριθμο είναι η χρησιμοποίηση ενός επαρκούς πλήθους αρχικών παραθύρων τα οποία κατάλληλα θα συγχωνευτούν κατά την διάρκεια του αλγορίθμου ώστε να παραχθούν οι επιθυμητές ομάδες. Η παραθυρική τεχνική του k – windows μας επιτρέπει την εξέταση ενός μεγάλου πλήθους αρχικών ομάδων (παραθύρων), χωρίς κάποια σημαντική χρονική επιβάρυνση. Η διαδικασία της συγχώνεύσης οδηγείται έπειτα μέσω συγκεκριμένων κατωφλιών που καθορίζονται από τον χρήστη. Τα παράθυρα που απομένουν, αξιολογούνται σύμφωνα με το κριτήριο της τμηματοποίησης και έτσι καθορίζεται το τελικό σύνολο των ομάδων.

Κατ' αρχάς τα αρχικά παράθυρα καθορίζονται χρησιμοποιώντας το range tree που αναλύσαμε πιο πάνω. Οι εσωτερικοί κόμβοι του ίδιου επιπέδου σε ένα range tree μπορούν να χρησιμοποιηθούν ως αντιπρόσωποι διαφορετικών υποσυνόλων των προτύπων. Αυτοί οι κόμβοι χρησιμοποιούνται ως κέντρα των αρχικών παραθύρων. Οι δύο πιο αποτελεσματικές ευρετικές μέθοδοι αρχικοποίησης που στηρίζονται στο range tree είναι η ακόλουθες:

1. Προσαυξητική (incremental) μέθοδος. Αρχίζει από το πρώτο επίπεδο και επαναληπτικά αναθέτει ένα παράθυρο για κάθε διαφορετικό κόμβο σε επίπεδο. Η διαδικασία αρχικοποίησης τελειώνει αν το πλήθος των παραθύρων που αναθέτονται σε ένα συγκεκριμένο επίπεδο είναι ίδιο με το πλήθος των παραθύρων που έχουν ανατεθεί σε προηγούμενο επίπεδο.

2. Ευθεία (direct) μέθοδος. Αναθέτει ένα παράθυρο σε κάθε κόμβο του τελευταίου εσωτερικού επιπέδου του range tree.

Τα αρχικά παράθυρα καλύπτουν ένα αρχικό εμβαδόν α, το οποίο εξαρτάται από την πυκνότητα του συνόλου των δεδομένων. Παρόμοια με τον k – windows το εμβαδόν α καθορίζεται κατά μήκος κάθε κατεύθυνσης i, δηλαδή:

$$a_i = \frac{\text{μέση απόσταση των προτύπων στην } vi}{\text{πλήθος παραθύρων}} \times 0.5 \quad (3.27)$$

Διαισθητικά γίνεται προσπάθεια να καλυφθεί ο ενδιάμεσος χώρος μεταξύ δυο παραθύρων χωρίς να επικαλύπτονται τα παράθυρα. Γι αυτό πολλαπλασιάζουμε με 0.5 στον παραπάνω τύπο.

1. Είσοδος a, u, th_e, th_m, th_c, th_v
2. x = DetermineInitialWindows()
3. Αρχικοποίηση των x – ranges $w_{m1}, w_{m2}, ..., w_{mx}$ εμβαδού α η κάθε μία κατά μήκος των μέσων $i_{m1}, i_{m2}, ..., i_{mx}$ τους.
4. Για κάθε διάνυσμα εισόδου $i_l, i = 1, ..., n$
   κάνε
   ανάθεσε το i_l στην w_j
   επανέλαβε

5. Για κάθε d – range w_j

Κάνε

Υπολόγισε το μέσο της $i_{mj} = \frac{1}{|w_j|}\sum_{i_l \in w_j} i_l$

Και ενημέρωσε τις d – ranges

6. Για κάθε d – range w_j

Κάνε

Επανέλαβε

Για κάθε διάσταση d_i

Κάνε

Επανέλαβε

Μεγέθυνε τις w_j κατά μήκος της διάστασης d_i κατά th_e%

Μέχρι η αύξηση του πλήθους των προτύπων κατά μήκος της διάστασης d_i να είναι μικρότερη του ποσοστού th_c%

Μέχρι η καμία σημαντική αλλαγή (<th_v) των μέσων των d – ranges να λάβει χώρα

7. Για κάθε d – range w_j που δεν έχει σημειωθεί με ετικέτα w_j

Κάνε

Σημείωσε την w_j με ετικέτα w_j

Αν $\exists w_i \neq w_j$ που επικαλύπτεται με την w_j

Τότε σημείωσε την w_i με ετικέτα w_j

8. υπολόγισε τον λόγο $r = \frac{1}{n}\sum_{j=1}^{x} i_l \in |w_j|$

9. αν r < u

κάνε

επανεκτέλεσε τον αλγόριθμο

10. Για κάθε πρότυπο εισόδου $i_l, i = 1,...,n$

Κάνε

Ανέθεσε το i_l στην w_j με τον κοντινότερο μέσο i_{mj} έτσι που $\|i_l - i_{mj}\|^2 \leq \|i_l - i_{mu}\|^2$

11. Έξοδος ομάδες $c_{l1}, c_{l2},...$ έτσι που $\{c_{li} = i \mid i \in w_j, label(w_j) = l_i\}$

Αλγόριθμος 3.1: Ο αλγόριθμος Z - Windows.

Έπειτα, κάθε παράθυρο επαναληπτικά μετακινείται στον ευκλείδειο χώρο έτσι ώστε το κέντρο των προτύπων που αυτό περιλαμβάνει να βρίσκεται στο κέντρο αυτού του παραθύρου (βήματα 4,5). Αυτή η διεργασία εκτελείται μέχρι εκείνο το σημείο όπου καμία πλέον μετακίνηση δεν προκαλεί αύξηση των προτύπων που βρίσκονται μέσα στο παράθυρο. Το κριτήριο αυτό προσδιορίζεται μέσω του κατωφλιού μεταβλητότητας (variability threshold), th_v που καθορίζει την ελάχιστη αλλαγή στο κέντρο ενός παραθύρου που μπορεί να θεωρηθεί ως κίνηση.

Στην συνέχεια λαμβάνει χώρα η φάση της μεγέθυνσης (enlargement). Τα παράθυρα μεγεθύνονται για να περιβάλλουν όσο το δυνατόν περισσότερα πρότυπα από την συγκεκριμένη ομάδα (βήμα 6). Κάθε παράθυρο μεγεθύνεται βαθμιαία μέσω του th_e ποσοστού μεγέθυνσης κατά μήκος κάθε διάστασης. Το th_e είναι ένα κατώφλι

μεγέθυνσης που καθορίζεται από τον χρήστη στο βήμα 1. Η φάση αυτή συνεχίζεται μέχρις ότου καμιά πλέον μεγέθυνση να μην καταλήγει στην αύξηση των προτύπων που βρίσκονται μέσα στο παράθυρο. Το κριτήριο αυτό καθορίζεται μέσω του κατωφλιού κάλυψης th_c (coverage threshold) το οποίο επίσης καθορίζεται από τον χρήστη στο βήμα 1. Το κατώφλι κάλυψης καθορίζει την ελάχιστη αύξηση προτύπων στο παράθυρο που μπορεί να θεωρηθεί σημαντική.

Οι φάσεις της μετακίνησης και της μεγέθυνσης εκτελούνται η μία μετά την άλλη επαναληπτικά μέχρις ότου καμία ουσιαστική αλλαγή των μέσων d – ranges σε σχέση με την προηγούμενη κατάσταση δεν συμβεί, σύμφωνα πάντα με το κατώφλι μεταβλητότητας th_v.

Μετά από την φάση της μετακίνησης και της μεγέθυνσης εκτελείται μια αναζήτηση για επικαλυπτόμενα παράθυρα (βήμα 7). Κάθε ζευγάρι επικαλυπτόμενων παραθύρων αντικαθίσταται από ένα νέο παράθυρο. Ένα άμεσο κριτήριο για την αναγνώριση παραθύρων που θα συγχωνευτούν είναι απλώς η επικάλυψη μεταξύ τους. Ένας άλλος αποτελεσματικός ευρεστικός έλεγχος είναι η διαβάθμιση της επικάλυψης σύμφωνα με το πλήθος των προτύπων που βρίσκονται στην τομή αυτών των παραθύρων. Ειδικότερα η λειτουργία της συγχώνευσης μπορεί να κατευθυνθεί μέσω του κατωφλιού συγχώνευσης (merge threshold) th_m που καθορίζεται από τον χρήστη έτσι που να ισχύει:

$$average\left(\frac{\left|i_l \in w_i \wedge i_l \in w_j\right|}{\left|w_i\right|}, \frac{\left|i_l \in w_i \wedge i_l \in w_j\right|}{\left|w_j\right|}\right) > th_m \quad (3.28)$$

τότε τα w_i και w_j πρέπει να συγχωνευτούν. Διαισθητικά, θα μπορούσαμε να πούμε ότι η συγχώνευση καθοδηγείται από τον βαθμό που ένα παράθυρο ενσωματώνεται σε ένα άλλο.

Η τελευταία φάση περιέχει τον έλεγχο της ποιότητας των ομάδων που έχουν προκύψει. Όπως και στην περίπτωση του k – windows η ποιότητα καθορίζεται μέσω του πλήθους των προτύπων που περιέχονται σε ένα παράθυρο σε σχέση με όλα τα πρότυπα στην βάση (βήματα 8,9). Αν η επιθυμητή ποιότητα που προσδιορίζεται από τον χρήστη μέσω της παραμέτρου u δεν πετυχαίνεται τότε ο αλγόριθμος ξαναεκτελείται. Η καινούργια εκτέλεση μπορεί να αρχίσει με τα ίδια αρχικά κέντρα αλλά με μεγαλύτερο εμβαδόν d – περιοχών, $a' > a$. Μια άλλη λύση είναι η επιλογή νέων αρχικών μέσων τοποθετημένων όμως στη μέγιστη δυνατή απόσταση από τα προηγούμενα. Ωστόσο τα πειράματα έδειξαν ότι τις περισσότερες φορές δεν υπάρχει η ανάγκη για επανεκτέλεση του αλγορίθμου. Ας σημειώσουμε επίσης ότι το πολύ μεγάλο πλήθος αρχικών παραθύρων αποτελεί σχεδόν εγγύηση για την εύρεση όλων των ομάδων που υπάρχουν στα δεδομένα.

Τέλος καθορίζονται οι ομάδες εξόδου. Πρώτα κάθε πρότυπο ανατίθεται στην ομάδα της οποίας το κέντρο είναι πιο κοντά στο πρότυπο (βήμα 10). Ο αλγόριθμος δίνει ως έξοδο τις τελικές ομάδες που αποτελούνται από τα πρότυπα που βρίσκονται εντός των παραθύρων που έχουν τις ίδιες ετικέτες μ' αυτές (βήμα 11). Μετά τον προσδιορισμό των ομάδων εύκολα καθορίζονται εκείνες οι ομάδες που περιέχουν πρότυπα θορύβου, καθώς η πυκνότητα των προτύπων μέσα στις ομάδες θα πρέπει να

είναι σημαντικά μεγαλύτερη από την πυκνότητα των προτύπων στο εξωτερικό των ομάδων.

Ο αλγόριθμος Z – Windows είναι πολύ γρηγορότερος από άλλους αλγορίθμους ομαδοποίησης. Πιο συγκεκριμένα ο πρώτος βρόχος στο βήμα 4 όπου τα πρότυπα ανατίθενται στις d – περιοχές έχει χρονική πολυπλοκότητα $O(x(s+\log^{d-2} n))$. Το βήμα 5 όπου τα κέντρα των d – περιοχών υπολογίζονται απαιτεί $O(sdx)$ χρόνο, ενώ το βήμα 6 όπου κατά την διάρκεια της μεγέθυνσης των d – περιοχών, λαμβάνουν χώρα οι αναθέσεις των προτύπων στις d – περιοχές, έχει χρονική πολυπλοκότητα $O(x(s+\log^{d-2} n))$. Τέλος η ποιότητα των ομάδων υπολογίζεται σε χρόνο $O(sdx)$. Συνεπώς η ολική πολυπλοκότητα του αλγορίθμου είναι $O(dxqr(s+\frac{\log^{d-2} n}{d}))$, όπου q είναι το πλήθος των μετακινήσεων και r είναι το πλήθος των επαναλήψεων εξαιτίας των ελλιπών ομάδων (βήμα 9).

Από τα πειράματα που έγιναν πάνω σ' αυτόν τον αλγόριθμο εξάγονται μερικά σημαντικά συμπεράσματα όσον αφορά την αποτελεσματικότητα του αλγορίθμου σε σχέση με τις κρίσιμες παραμέτρους, οι οποίες καθορίζονται από τον χρήστη στο πρώτο βήμα (a, u, th_e, th_m, th_c, th_v). Έτσι λοιπόν η παράμετρος α υπολογίζεται αυτόματα κατά μήκος κάθε διάστασης μέσω του τύπου (3.27):

$$a_i = \frac{\text{μέση απόσταση των προτύπων στη } i}{\text{πλήθος παραθύρων}} \times 0.5 \quad (3.27).$$

Στα πειράματα που έγιναν [148] δεν χρειάστηκε ποτέ επανεκτέλεση του αλγορίθμου συνεπώς η παράμετρος u δεν φαίνεται να παίζει ουσιαστικό ρόλο. Επίσης τα κατώφλια της μεταβλητότητας, της κάλυψης και της μεγέθυνσης δεν φαίνεται να συνεισφέρουν σημαντικά στην ποιότητα της ομαδοποίησης. Το πλήθος των ομάδων φαίνεται εξαρτάται μόνο από το κατώφλι συγχώνευσης το οποίο πραγματικά φαίνεται να προσδιορίζει την ποιότητα της ομαδοποίησης.

4 Εξόρυξη κειμένου

Σε αυτό το κεφάλαιο θα ασχοληθούμε με τον τομέα της εξόρυξης δεδομένων που ονομάζεται εξόρυξη κειμένου. Όπως είδαμε και παραπάνω πολλές από τις τεχνικές που εφαρμόζονται στο κομμάτι της ομαδοποίησης αριθμητικών δεδομένων μπορούν με κάποιες αλλαγές να χρησιμοποιηθούν και στην ομαδοποίηση κατηγορικών δεδομένων. Σε αυτά που θα ακολουθήσουν θα αναφερθούμε αναλυτικότερα τους σημαντικότερους τομείς της εξόρυξης κειμένου.

4.1 Εισαγωγή

Ο τομέας της Ανακάλυψης Γνώσης από Κείμενο (Knowledge Discovery in Text KDT) καθώς και ο τομέας της εξόρυξης κειμένου (Text Mining) είναι δυο ταχέως αναπτυσσόμενοι τομείς κυρίως λόγω της μεγάλης ανάγκης για ανάλυση τεράστιων ποσοτήτων δεδομένων κειμένων, το οποίο σε μεγάλο βαθμό οφείλεται στην αλματώδη ανάπτυξη του διαδικτύου και κυρίως του WWW.

Ζούμε στην «εποχή της πληροφορίας». Το κυριότερο χαρακτηριστικό αυτής της εποχής είναι το τεράστιο πλήθος πληροφοριών οι οποίες παράγονται και αποθηκεύονται πράγμα που κάνει πολύ δύσκολο στα ανθρώπινα όντα να τις κατανοήσουν στο σύνολό τους. Στις περισσότερες εταιρίες και οργανισμούς μεγάλος εργατοχρόνος και προσπάθεια σπαταλείται σε μη αποτελεσματικές έρευνες σε κείμενα του διαδικτύου καθώς και κείμενα άλλων συμβατικών πηγών. Το πρόβλημα αυτό της υπερβολικής ύπαρξης πληροφορίας οξύνεται ακόμα περισσότερο λόγο της έλλειψης κάποιας κοινής αποδεκτής δομής στην πλειοψηφία των κειμένων. Το μεγαλύτερο ποσό της πληροφορίας που μπορεί να βρει κανείς σε μια εταιρία είναι κειμενικού τύπου. Κάποιες εκτιμήσεις ανεβάζουν το ποσοστό των κειμενικών δεδομένων στο 80% του συνολικό ποσού πληροφορίας των εταιριών, το οποίο βέβαια περιλαμβάνει αναφορές, ηλεκτρονικό ταχυδρομείο κ.α. [88]. Για αυτού του τύπου τα δεδομένα δεν υπάρχουν συνήθως μεταδεδομένα (δεδομένα που να αναφέρονται στα δεδομένα) και σαν συνέπεια δεν υπάρχουν πρότυπα μέσα που να διευκολύνουν την αναζήτηση (search), την αναζήτηση μέσω ερωτήσεων (query) και την ανάλυση αυτών των δεδομένων. Από την άλλη πλευρά το διαδίκτυο έχει εξελιχθεί σε μία τεραστίων διαστάσεων συλλογή κειμένων τα οποία όμως γράφονται για να διαβαστούν και να κατανοηθούν από τους ανθρώπους και όχι από προγράμματα υπολογιστών. Προς το παρόν, ο παγκόσμιος ιστός του διαδικτύου έχει αναπτύξει ένα μέσο κειμένων για χρήση από τους ανθρώπους και όχι για χρήση ως δεδομένα τα οποία να μπορούν να μπορούν να επεξεργαστούν αυτόματα. [13].

Αν και η ποσότητα των δεδομένων κειμένου που είναι διαθέσιμες σε εμάς μεγαλώνει σταθερά μέρα με τη μέρα, η ικανότητά μας να τα καταλάβουμε και να τα επεξεργαστούμε αυτήν την πληροφορία παραμένει σταθερή. Ένας άνθρωπος – συντάκτης μπορεί να αναγνωρίσει ότι ένα καινούργιο γεγονός εμφανίστηκε μόνο με το να ακολουθήσει και να διαβάσει προσεκτικά όλες της δημοσιευμένες σελίδες στο διαδίκτυο και στο σε άλλες πηγές. Αυτό είναι πρακτικά αδύνατο αν αναλογιστούμε τον όγκο και την πολυπλοκότητα των δεδομένων που υπάρχουν διαθέσιμα. Γίνεται, λοιπόν προφανής η ανάγκη για αυτόματη εξαγωγή χρήσιμης πληροφορίας από

τεράστια ποσά κειμενικών δεδομένων με σκοπό να βοηθήσουν την ανάλυσή τους από ανθρώπους. Η γρήγορη υιοθέτηση από τις επιχειρήσεις του μοντέλου του ηλεκτρονικού εμπορείου (e – commerce) οδηγεί την ζήτηση για λογισμικό που να βοηθάει τις εταιρίες να αναλύουνε, να διαχειριστούνε, να ελέγχουνε και να ελέγχουν αποτελεσματικά πως οι επιχειρήσεις διαδίδουν την πληροφορία για ανταγωνιστικά ανάπτυξη [88]. Η ανακάλυψη γνώσης και η εξόρυξη κειμένου είναι αυτοματοποιημένες τεχνικές που σαν σκοπό έχουν την ανακάλυψη υψηλού επιπέδου πληροφοριών σε τεράστια ποσά κειμενικών δεδομένων και να τα παρουσιάσουν στον εν γένει χρήστη (τον αναλυτή, τον αποφασίζοντα κ.τ.λ.).

4.2 Τι είναι KDT και TM

Η ανακάλυψη γνώσης σε κείμενο (KDT) καθώς και η εξόρυξη κειμένου (TM) είναι ένα σχετικά νέο πεδίο έρευνας που προσπαθεί να λύσει το πρόβλημα της υπερ – ύπαρξης πληροφοριών χρησιμοποιώντας τεχνικές από την εξόρυξη δεδομένων (data mining), την μηχανική μάθηση (machine learning), την επεξεργασία φυσικής γλώσσας (Natural Language Processing NLP), την ανάκτηση πληροφοριών (Information Retrieval), εξαγωγή πληροφοριών (Information extraction) και την διαχείριση πληροφορίας (Information Management). Όπως και σε κάθε γρήγορα αναπτυσσόμενο τομέα της έρευνας, δεν υπάρχει κάποιο καθιερωμένο λεξιλόγιο για την KDT και το TM, γεγονός που μπορεί να οδηγήσει σε συγχύσεις όταν προσπαθήσει κάποιος να συγκρίνει τεχνικές και αποτελέσματα. Συχνά χρησιμοποιούνται διαφορετικοί όροι για να δηλώσουν το ίδιο πράγμα. Ανακάλυψη γνώσης σε κείμενο (Knowledge Discovery in Text) [44], [72], κειμενική εξόρυξη δεδομένων (Text Data Mining) [57] και εξόρυξη κειμένου (Text mining) [117], [108], [133], [58], [132] είναι μερικοί από τους όρους που μπορούν να βρεθούν στην βιβλιογραφία.

Χρησιμοποιούμε συνήθως τον όρο KDT για να δηλώσουμε την διαδικασία της μετατροπής μη δομημένων κειμενικών δεδομένων σε υψηλής στάθμης πληροφοριών και γνώσης, ενώ ο όρος TM αναφέρεται κυρίως για ένα από τα βήματα της διαδικασίας Ανακάλυψης Γνώσης σε Κείμενο (KDT) το οποίο έχει να κάνει με την εξαγωγή προτύπων από κειμενικά δεδομένα. Αν επεκτείνουμε τον ορισμό για την ανακάλυψη γνώσης σε δεδομένα των Fayyad και Piatetsky – Shapiro [43] μπορούμε να δώσουμε τον εξής απλό ορισμό:

> Η ανακάλυψη γνώσης σε Κείμενο (KTD) είναι μια μη τετριμμένη διαδικασία ανακάλυψης έγκυρων, καινούργιων, δυνητικά χρήσιμων και τελικά κατανοητών προτύπων σε δεδομένα κειμένου.

Ας δούμε από λίγο πιο κοντά τις βασικές έννοιες που εμφανίζονται στον παραπάνω ορισμό:

- Μη δομημένα δεδομένα κειμένου (Unstructured textual Data) είναι μια συλλογή κειμένων. Χρησιμοποιούμε τον όρο έγγραφο (document) για να αναφερθούμε σε μια λογική μονάδα κειμένου. Αυτό θα μπορούσε να είναι ένα status memo, μια σελίδα στο παγκόσμιο ιστό (Webpage), ένα τιμολόγιο, ένα ηλεκτρονικό ταχυδρομείο (e - mail) κ.α. Μια μονάδα κείμενου θα μπορούσε να είναι μακριά και πολύπλοκή [132] ακόμα και να περιέχει κάτι παραπάνω

από απλώς κείμενο όπως γραφικά και Multimedia. Εμείς εδώ θα ασχοληθούμε μόνο με έγγραφα που περιέχουν μόνο κείμενο. Τα έγγραφα που χρησιμοποιούνε την extensible markup language (XML) ή την standard generalized markup language (SGML) και παρόμοιες συμβάσεις λέγονται ημιδομημένα κειμενικά δεδομένα.

- Πρότυπο (pattern). Εάν θεωρήσουμε τα δεδομένα μας ως ένα σύνολο γεγονότων F (π.χ. περιπτώσεις σε μια βάση δεδομένων) ένα πρότυπο είναι ένας κανόνας E ο οποίος περιγράφει γεγονότα σε ένα υποσύνολο F_E του F [43]. Γενικότερα θα μπορούσαμε να πούμε ότι υπάρχουν δύο τύποι προτύπων: τα πρότυπα πρόβλεψης (predictive pattern) και τα πρότυπα ενημέρωσης (informative pattern). Χρησιμοποιούμε τα πρότυπα πρόβλεψης για να προβλέψουν ένα ή περισσότερα γνωρίσματα (attributes) από αυτά που υπάρχουν στη βάση. Αυτό το είδος των προτύπων κάνουν μια εικασία για την τιμή ενός άγνωστου γνωρίσματος δεδομένων των τιμών των γνωρισμάτων των άλλων δεδομένων. Από την άλλη μεριά τα ενημερωτικά πρότυπα δεν επιλύουν ένα συγκεκριμένο πρόβλημα αλλά παρουσιάζουν στον χρήστη ενδιαφέροντα πρότυπα που θα έπρεπε να γνωρίζει.

Η ανακάλυψη γνώσης σε Κείμενο είναι μια διαδικασία πολλών βημάτων (multi – step procedure) που περιλαμβάνει όλες τις διαδικασίες από την συλλογή των εγγράφων μέχρι την οπτικοποίηση τις γνώσης που έχει προκύψει. Η διαδικασία θα πρέπει να είναι μη τετριμμένη, δηλαδή το αποτέλεσμα θα πρέπει μπορεί να αξιολογηθεί ως ανακάλυψη. Τα πρότυπα που ανακαλύφθηκαν θα πρέπει να είναι έγκυρα σε καινούργια δεδομένα με κάποιον βαθμό βεβαιότητας. Τα πρότυπα θα πρέπει να είναι καινούργια, τουλάχιστον για το σύστημα, και θα πρέπει να οδηγούν σε χρήσιμες δράσεις, όπως προσδιορίζονται από μια συνάρτηση χρησιμότητας (utility function). Ο κύριος σκοπός της KDT είναι να κάνει τα πρότυπα κατανοητά στους ανθρώπους για τη διευκόλυνση της κατανόησης των παρόντων δεδομένων [43].

Η εξόρυξη κειμένου (Text mining) τώρα είναι ένα βήμα στην διαδικασία KDT που αποτελείται από συγκεκριμένους NLP αλγορίθμους της εξόρυξης δεδομένων οι οποίοι κάτω από υπολογιστικά παραδεχτούς περιορισμούς παράγει μια συγκεκριμένη ποσότητα προτύπων από ένα σύνολο μη δομημένων δεδομένων κειμένου.

Η εξόρυξη κειμένου χρησιμοποιεί μη δομημένα κείμενα που τα εξετάζει με σκοπό να ανακαλύψει δομή και αυτονόητα «νοήματα» που βρίσκονται κρυμμένα μες στο κείμενο [68]. Ο κύριος στόχος της εξόρυξης κειμένου είναι η υποστήριξη της ανακάλυψης γνώσης σε τεράστιες συλλογές κειμένων (που αποθηκεύονται είτε στο διαδύκτιο είτε συμβατικά.). Η εξόρυξη κειμένου χρησιμοποιεί ειδικές NLP τεχνικές εξόρυξης δεδομένων που εφαρμόζονται σε δεδομένα κειμένου με σκοπό την εξαγωγή χρήσιμων πληροφοριών. Οι εφαρμογές εξόρυξης κειμένου επιβάλουν αυστηρούς περιορισμούς στις συνήθεις NLP τεχνικές [117]. Για παράδειγμα αφού οι τεχνικές αυτές εφαρμόζονται σε τεράστια ποσά πληροφορίας δεν μπορούν να περιέχουν πολύπλοκες διαδικασίες. Είναι αυτή η κοντινή συγγένεια που κάνει την εξόρυξη κειμένου έναν καινούργιο τομέα της έρευνας που προέρχεται από την εξόρυξη δεδομένων και από το NLP.

Η φύση της εξόρυξης δεδομένων γίνεται ακόμα πιο κατανοητή αν συγκριθεί με την ήδη υπάρχουσα τεχνολογία: την εξόρυξη δεδομένων. Η εξόρυξη δεδομένων είναι η

εφαρμογή αλγορίθμων σε ένα σύνολο για να αποκαλυφθούν συνδέσεις και συσχετίσεις που δεν ήταν από πριν προφανείς. Η εξόρυξη δεδομένων δουλεύει με δομημένα δεδομένα, τα οποία είναι τις πιο πολλές φορές αριθμητικά. Η εξόρυξη κειμένου είναι ανάλογη με την εξόρυξη δεδομένων στο γεγονός ότι αποκαλύπτει σχέσεις που ενυπάρχουν στην πληροφορία μας. Από την άλλη πλευρά όμως η εξόρυξη κειμένου, διαφέρει από την εξόρυξη δεδομένων, στο ότι αυτή δουλεύει με πληροφορίες που βρίσκονται αποθηκευμένες ως μια μη δομημένη συλλογή εγγράφων κειμένου.

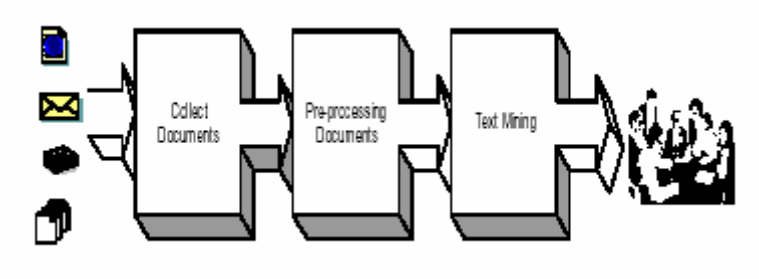
Οι τεχνικές της εξόρυξης κειμένου δεν είναι τα μόνα εργαλεία που μπορούν να χρησιμοποιηθούν για την λύση του προβλήματος της ύπαρξης υπερβολικής πληροφορίας. Μπορούμε να αναφέρουμε άλλες τεχνικές από διαφορετικούς τομείς της έρευνας όπως την ανεύρεση πληροφοριών (Information Retrieval), την επεξεργασία φυσικής γλώσσας (Natural Language Processing), κ.α.. η εξόρυξη κειμένου θα μπορούσε να χρησιμοποιηθεί είτε άμεσα είτε έμμεσα (όπως ακριβώς και το Web Mining [76]). Με την άμεση προσέγγιση εννοούμε την εφαρμογή των τεχνικών της εξόρυξης κειμένου κατευθείαν σε προβλήματα που παράγονται από την ύπαρξη μεγάλων ποσών πληροφορίας. Για παράδειγμα η εύρεση παρόμοιας πληροφορίας σε μία τεράστια βάση δεδομένων. Με την έμμεση εφαρμογή των τεχνικών της εξόρυξης κειμένου εννοούμε ότι οι τεχνικές αυτές θα είναι μόνο κάποιο μέρος μιας μεγαλύτερης διαδικασίας που προσπαθεί να λύσει το πρόβλημα της υπερχείλισης της πληροφορίας (information overload).

Η διαδικασία της ανακάλυψης γνώσης σε κείμενο περιέχει τρία βασικά στάδια [132], [76], Εικόνα 4.1:

1. Συλλογή των σχετικών εγγράφων: Στο πρώτο βήμα προσπαθούμε να αναγνωρίσουμε ποια κείμενα θα βρούμε. Αφού πρώτα βρούμε την πηγή από όπου θα ανακτήσουμε τα κείμενά μας (διαδύκτιο ή άλλες συμβατικές πηγές), θα πρέπει στη συνέχεια να τα συλλέξουμε.

2. Προ – επεξεργασία των εγγράφων: Αυτό το στάδιο περιλαμβάνει κάθε είδους μετατροπής των αρχικών δεδομένων. Αυτές οι μετατροπές θα μπορούσαν να σκοπεύουν στο να αποκτήσουμε μια επιθυμητή αναπαράσταση των δεδομένων όπως π.χ. XML, SGML. Τα έγγραφα που θα προκύψουν θα μπορούν πλέον να υποστούν επεξεργασίες για να παραχθούν βασικές λεκτικές πληροφορίες πάνω στο περιεχόμενο κάθε εγγράφου.

3. Διαδικασίες εξόρυξης κειμένου: Σε αυτό το στάδιο υψηλής ποιότητας πληροφορίες εξάγονται (δημιουργία μεταδεδομένων (metadata creation)), αναγνωρίζονται πρότυπα και συσχετίσεις των δεδομένων.



Εικόνα 4.1: Τα βασικά στάδια της ανακάλυψης γνώσης σε κείμενο.

4.3 Βασικές τεχνικές της εξόρυξης κειμένου

4.3.1 Εισαγωγή

Όπως είπαμε και παραπάνω ο κύριος σκοπός της εξόρυξης δεδομένων είναι η εξαγωγή πληροφοριών από τεράστια ποσά δεδομένων κειμένου. Τεχνικές επεξεργασίας φυσικής γλώσσας, εξόρυξης δεδομένων και μηχανικής μάθησης συνεργάζονται για την αυτόματη ανακάλυψη προτύπων και σχέσεων μέσα στα δεδομένα.

Οι κυριότερες κατηγορίες των αντικειμένων που χειρίζεται η εξόρυξη κειμένου είναι:

- Εξαγωγή χαρακτηριστικών (Feature Extraction).
- Πλοήγηση με βάση το κείμενο (Text Based Navigation).
- Αναζήτηση και Ανάκτηση (Search and Retrieval).
- Κατηγοριοποίηση, κατάταξη με επίβλεψη (Categorization, Supervised Classification).
- Ομαδοποίηση, μη επιβλεπόμενη κατάταξη (Clustering, Unsupervised Classification).
- Εξαγωγή περίληψης (Summarization).
- Ανάλυση τάσεων (Trends Analysis).
- Συσχετίσεις (Associations).
- Οπτικοποίηση (Visualisation).

Ας δούμε τώρα λίγο πιο αναλυτικά κάθε ένα από αυτά τα σημεία.

4.3.2 Εξαγωγή Χαρακτηριστικών.

Αντικειμενικός στόχος της διαδικασίας εξαγωγής χαρακτηριστικών είναι να ανακαλύψει στο κείμενο γεγονότα και σχέσεις. Αυτή η διεργασία περιλαμβάνει μερικές φορές την διάκριση αν το ουσιαστικό μιας φράσης αναφέρεται σε πρόσωπο, σε τόπο, σε οργανισμό ή σε κάποιο άλλο διακριτό αντικείμενο.

Οι αλγόριθμοι εξαγωγής χαρακτηριστικών χρησιμοποιούν μερικές φορές λεξικά για να αναγνωρίσουν κάποιους όρους. Άλλες πάλι φορές χρησιμοποιούν λεκτικά πρότυπα (linguistic patterns) για να αναγνωρίσουν άλλους όρους. Για παράδειγμα το όνομα ενός οργανισμού όπως «ITE» μπορεί να μην υπάρχει σε ένα λεξικό αλλά ένας αλγόριθμος εξαγωγής χαρακτηριστικών θα μπορούσε να το αναγνωρίσει ως

ουσιαστικό και μάλιστα ως έναν σημαντικό όρο. Αλγόριθμοι αναγνώρισης προτύπων (όπως παραδείγματος χάριν Hidden Markov Models HMM), θα μπορούσαν να εκπαιδευτούν για να αναγνωρίζουν λεκτικά πρότυπα, όπως μια φράση ουσιαστικού ακολουθείται από μια φράση ρήματος και αυτή συνήθως ακολουθείται από μια φράση ουσιαστικού όπως παραδείγματος χάριν στο «το ITE προσλαμβάνει ερευνητές». Φυσικά σχεδόν πάντα απαιτείται μια προ – και μια μετά – επεξεργασία για να καθοριστούν οι σημαντικοί όροι και άλλοι που εξάχθηκαν ως σημαντικοί και τελικά δεν ήταν.

Επιπλέον οι όροι θα πρέπει να είναι βρίσκονται σε μια κανονική ή καθιερωμένη μορφή. Αυτό κάνει την ανάκτηση ευρετηρίου (indexing retrieval) και άλλες διεργασίες που θα ακολουθήσουν πιο ακριβείς. Για παράδειγμα οι λέξεις «διαβάζοντας» και «διαβάζω» θα πρέπει να ανιχνευθούν ως η ίδια λέξη.

Η διεργασία της εξαγωγής χαρακτηριστικών θα πρέπει επίσης να μας παρουσιάζει και τον αριθμό των εμφανίσεων κάθε όρου (word frequency). Αυτό κυρίως υποστηρίζει τις διεργασίες κατάταξης κειμένου.

4.3.3 Πλοήγηση με βάση το κείμενο

Η πλοήγηση με βάση το κείμενο επιτρέπει τους χρήστες να πλοηγούνται μέσα σε μια συλλογή εγγράφων με βάση σχετικά θέματα ή σημαντικούς όρους. Αυτό βοηθάει στον να αναγνωριστούν έννοιες κλειδιά και επιπλέον να αναγνωριστούν και συσχετίσεις μεταξύ σημαντικών όρων. Για παράδειγμα, όταν αναζητούμε έγγραφα με τον όρο «ITE» θα πρέπει να έχουμε γρήγορη πρόσβαση σε και να μετακινούμαστε σε έννοιες όπως «ερευνητές του ITE» ή «εργαστήρια του ITE» και σε άλλους παρεμφερείς όρους που να περιέχουν το «ITE». Τα σημαντικά σημεία σε αυτήν την διεργασία είναι δύο. Πρώτον η ικανότητα να μπορούμε να βλέπουμε και άλλους σχετικούς όρους. Παραδείγματος χάριν αν αναγνωρίσουμε ότι δύο όροι εφανίζονται συχνά μαζί τότε θα μπορούμε να υποθέσουμε ότι αυτοί οι δυο όροι έχουν πιθανόν κάποια σχέση μαζί μεταξύ τους. Το δεύτερο σημαντικό σημείο είναι ότι μπορούμε να κινηθούμε από ένα ζευγάρι εννοιών σε ένα άλλο ζευγάρι εννοιών, όπως παραδείγματος χάριν αν δεν μας ικανοποιεί το ζευγάρι «ITE» και «ερευνητές» τότε ίσως να μας ικανοποιήσει το ζευγάρι «ITE» και «εργαστήρια». [139] [135].

4.3.4 Αναζήτηση και Επανάκτηση

Αυτό χρησιμοποιείται για την αναζήτηση σε εσωτερικές συλλογές εγγράφων ή σε συλλογές που βρίσκονται στο διαδύκτιο. Το βασικό του χαρακτηριστικό είναι οι διάφορες επιλογές αναζήτησης κειμένου. Μετά τη δημιουργία ευρετηρίου που είναι το πρώτο βήμα, μια αρκετά μεγάλη γκάμα επιλογών αναζήτησης κειμένου μπορούν να προσφερθούν. Αυτές μπορούν να περιλαμβάνουν από απλές επιλογές αναζήτησης όπως επιλογές Bool (and/or/not), εγγύτητα (proximity), τμηματική αναζήτηση (segment), αριθμητικό εύρος (numeric range) ως και πιο σύνθετες επιλογές αναζήτησης όπως (relevancy – ranked natural language searching) ή ασαφής αναζήτηση (fuzzy searching) κ.α. [35], [110], [134].

4.3.5 Κατηγοριοποίηση

Η κατηγοριοποίηση είναι η διαδικασία που χρησιμοποιούμε για να κατατάξουμε έγγραφα σε προκαθορισμένες κατηγορίες. Η κατηγοριοποίηση, λοιπόν μας βοηθάει να ανακαλύψουμε τα κύρια θέματα μιας συλλογής εγγράφων.

Οι κατηγορίες μπορούν να είναι προκαθορισμένες (π.χ. από τον προγραμματιστή) ή μπορεί να αφήνεται να προσδιοριστούν από τον χρήστη. Υπάρχουν δυο τρόποι για να δημιουργήσουμε κατηγορίες. Στην πρώτη περίπτωση δημιουργείται ένας θησαυρός ο οποίος να ορίζει ένα σύνολο από όρους σχετικά με το θέμα καθώς και σχέσεις μεταξύ τους (οι πιο κοινές σχέσεις είναι οι διευρυμένος όρος, κοντινότερος όρος, συνώνυμο και σχετικός όρος). Ο κατηγοριοποιητής μπορεί τότε να ορίσει το αντικείμενο του κειμένου σύμφωνα με την συχνότητα των όρων σχετικά με το πεδίο που υπάρχουν στο έγγραφο. Στην δεύτερη περίπτωση ο κατηγοριοποιητής εκπαιδεύεται με πρότυπα έγγραφα. Αναλύονται στατιστικά λεκτικά πρότυπα (linguistic patterns) όπως λεκτικές συγγένειες (lexical affinities), συχνότητες λέξεων από πρότυπα κείμενα τα οποία είναι προ – κατηγοριοποιημένα και τα οποία ανήκουν σε κάθε ξεχωριστή κατηγορία με σκοπό να δημιουργηθεί μία στατιστική υπογραφή για κάθε κατηγορία. Μετά χρησιμοποιείται αυτή η στατιστική υπογραφή για να καταταχθούν τα καινούργια έγγραφα στις κατηγορίες. Το πλεονέκτημα της δεύτερης προσέγγισης είναι ότι δεν χρειάζεται να δημιουργηθεί κάποιος θησαυρός εννοιών, πράγμα εξαιρετικά δύσκολο για μεγάλους τομείς.

Για να αποφύγουμε την λάθος κατάταξη ενός εγγράφου είναι μερικές φορές απαραίτητο να εφοδιάζουμε με διάφορες κατηγορίες κάθε έγγραφο.

4.3.6 Ομαδοποίηση

Μια ομάδα είναι μια συλλογή από σχετικά κείμενα, και ομαδοποίηση είναι η διαδικασία της δημιουργίας ομάδων εγγράφων βάσει κάποιο μέτρου ομοιότητας, αυτόματα χωρίς να έχουμε από πριν προσδιορίσει τις κατηγορίες.

Οι πιο γνωστοί αλγόριθμοι κατάταξης που υπάρχουν είναι ιεραρχικοί (hierarchical), δυαδικοί σχεσιακοί (binary relational) και ασαφής (fuzzy). Οι ιεραρχικοί αλγόριθμοι ομαδοποίησης όπως έχουν πει πιο αναλυτικά και παραπάνω κατασκευάζουν ένα δέντρο με όλα τα έγγραφα στη ρίζα του και ένα έγγραφο σε κάθε φύλλο του. Οι ενδιάμεσοι κόμβοι είναι περιέχουν κάποιον αριθμό εγγράφων και γίνονται όλο και πιο ειδικοί όσο πλησιάζουν προς τα φύλλα. Αυτό είναι εξαιρετικά χρήσιμοι όταν εξερευνούμε μια καινούργια συλλογή κειμένων και θέλουμε να έχουμε μια γενική επισκόπηση της συλλογής. Η δυαδική σχεσιακή ομαδοποίηση χωρίζει τα έγγραφα σε μια οριζόντια δομή όπου κάθε ομάδα τοποθετείται μόνο σε ένα σύνολο. Στην ασαφή κατηγοριοποίηση όλα τα έγγραφα περιέχονται σε όλες τις ομάδες αλλά με διαφορετικούς βαθμούς κυριότητας.

Ο πιο σημαντικός παράγοντας στην ομαδοποίηση είναι το μέτρο ομοιότητας. Όλοι οι αλγόριθμοι ομαδοποίησης, όπως έχουμε πει και παραπάνω, βασίζονται σε μέτρα ομοιότητας και υπάρχουν διάφορα είδη μέτρων ομοιότητας. Ένας τύπος χρησιμοποιεί λέξεις οι οποίες εμφανίζονται συχνά μαζί (λεκτικές συγγένειες π.χ. Τμήμα Πληροφορικής) σαν κοινά χαρακτηριστικά για να κατατάξει τα έγγραφα στις ομάδες.

Ένας άλλος τύπος χρησιμοποιεί χαρακτηριστικά που έχουν εξαχθεί όπως παραδείγματος χάριν κ. Θωμάς Παπαστεργιού [13], [135], [61].

4.3.7 Εξαγωγή περίληψης

Η εξαγωγή περίληψης είναι η διαδικασία κατά την οποία μειώνεται η ποσότητα του κειμένου διατηρώντας όμως πάντα το βασικό του νόημα.

Σε αυτού του είδους τις διεργασίες ο χρήστης καλείται τις πιο πολλές φορές να καθορίσει ένα πλήθός παραμέτρων όπως παραδείγματος χάριν το πλήθος των λέξεων που θα εξαχθούν ή το ποσοστό επί του συνολικού κειμένου το οποίο θα αποτελεί την περίληψη. Το αποτέλεσμα περιλαμβάνει συνήθως τις πιο σημαντικές προτάσεις του κειμένου [135], [67], [8].

4.3.8 Ανάλυση τάσεων

Αυτή η διεργασία χρησιμοποιείται για την αναγνώριση τάσεων σε έγγραφα που συλλέγονται σε κάποια χρονική περίοδο. Οι τάσεις χρησιμοποιούνται για παράδειγμα για να ανακαλύψουμε ότι μια εταιρία μετακινεί τα ενδιαφέροντά της από τον έναν τομέα στον άλλον [83].

4.3.9 Ανάλυση συσχετίσεων

Η ανάλυση συσχετίσεων προσπαθεί να λύσει το εξής πρόβλημα: Δεδομένης μιας συλλογής εγγράφων, αναγνώρισε σχέσεις μεταξύ γνωρισμάτων (χαρακτηριστικά που έχουν εξαχθεί από τα έγγραφα) τέτοιες ώστε η εμφάνιση ενός χαρακτηριστικού να συνεπάγεται την εμφάνιση ενός άλλου προτύπου. Παραδείγματος χάριν Neurosoft SA, Intrasoft -> Take over, θα μπορούσε να είναι ένας κανόνας που έχει ανακαλυφθεί. Μια εφαρμογή που βασίζεται σε αυτήν την διαδικασία έχει παρουσιαστεί από τον Feldman [45], [26].

4.3.10 Οπτικοποίηση

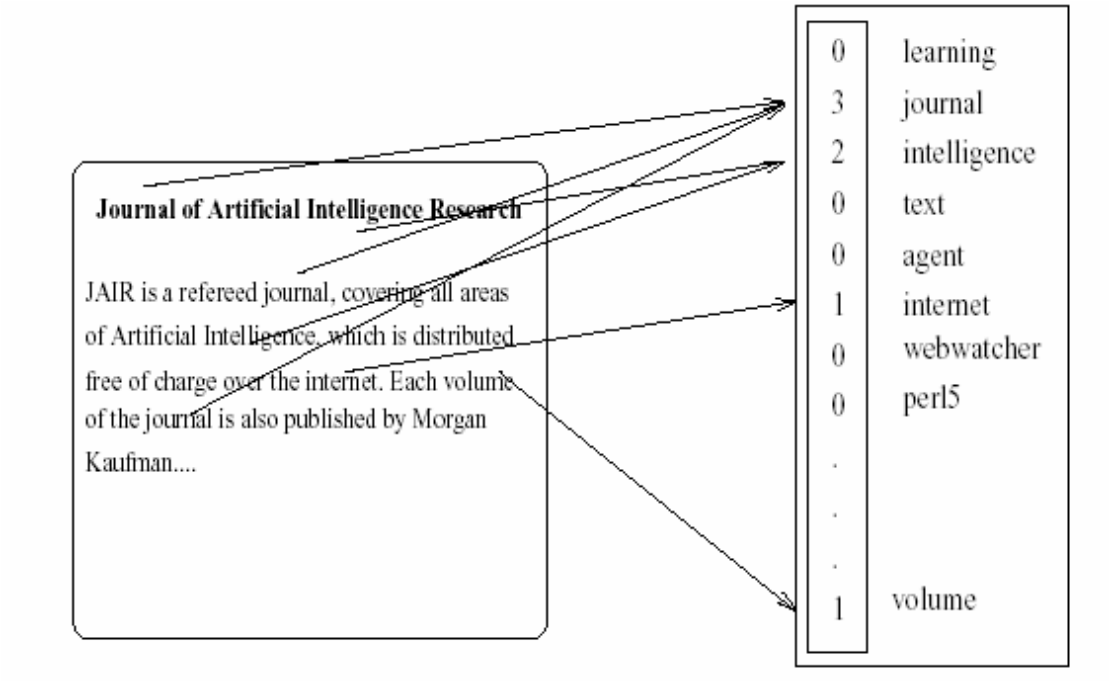
Η οπτικοποίηση χρησιμοποιεί την εξαγωγή χαρακτηριστικών και το ευρετήριο βασικών όρων για να κατασκευάσει μια γραφική αναπαράσταση μιας συλλογής εγγράφων. Αυτή η προσέγγιση βοηθάει τον χρήστη να αναγνωρίζει γρήγορα τα βασικά θέματα ή τις βασικές έννοιες μιας συλλογής εγγράφων βάσει της σπουδαιότητας τους στην αναπαράσταση. Επιπλέον είναι σχετικά εύκολο να ανακαλύψει κανείς την θέση κάποιων εγγράφων σε μια γραφική αναπαράσταση της συλλογής [26], [123].

4.4 Αναπαράσταση Κειμένου στην Εξόρυξη Κειμένου

Αφού λοιπόν πήραμε μια γεύση για το τι είναι Ανακάλυψη γνώσης σε κείμενο, τι είναι εξόρυξη κειμένου και αφού είδαμε ποιες είναι οι κυριότερες διεργασίες και τα κυριότερα προβλήματα που τίθενται σε αυτόν το τομέα της έρευνας εδώ θα

ασχοληθούμε με το πώς αναπαριστάται ένα κείμενο με σκοπό να λάβει μέρος στις διαδικασίες εξόρυξης κειμένου και κυρίως στην ομαδοποίηση κειμένου.

Τρία βασικά ερωτήματα ανακύπτουν κατά τη διαδικασία αναπαράστασης ενός κειμένου με σκοπό να πάρει μια μορφή κατάλληλη ώστε να μπορέσει να υποστεί επεξεργασία από αλγορίθμους ομαδοποίησης (ή άλλων τεχνικών εξόρυξης κειμένου). Θα δώσουμε εδώ μια γενική επισκόπηση της δουλειάς που έχει γίνει πάνω σε αυτόν το τομέα υπό το πρίσμα αυτών των τριών θεμελιωδών ερωτημάτων.



Εικόνα 4.2: Απεικόνιση της bag – of – words αναπαράστασης ενός κειμένου χρησιμοποιώντας ένα διάνυσμα συχνοτήτων.

(1) Η πιο συχνά χρησιμοποιούμενη αναπαράσταση εγγράφων στην ανάκτηση πληροφοριών και στην μάθηση – κειμένου είναι η λεγόμενη αναπαράσταση διανύσματος (vector representation). Αυτή η αναπαράσταση δεν είναι τίποτα άλλο από μια bag – of – words [100] αναπαράσταση: όλες οι λέξεις του κειμένου χρησιμοποιούνται, χωρίς να λαμβάνεται υπ' όψιν η σειρά των λέξεων ή κάποιου άλλου είδους δομής του κειμένου. Όταν έχουμε μια συλλογή κειμένων τότε κάθε κείμενο αναπαρίσταται με μια σακούλα λέξεων (bag – of – words), η οποία περιλαμβάνει όλες τις λέξεις που εμφανίζονται στο κείμενο (Εικόνα 4.2). Μερικές φορές μπορεί να χρησιμοποιηθούν και κάποιες επιπλέον πληροφορίες γύρω από το κείμενο που αναπαριστάται όπως παραδείγματος χάριν η δομή των προτάσεων, οι θέσεις των λέξεων ή γειτονικές λέξεις. Το ερώτημα που ανακύπτει είναι κατά πόσο κερδίζουμε εάν θεωρήσουμε επιπλέον πληροφορίες από το κείμενο (και ποιες είναι αυτές οι πληροφορίες που θα επιλέξουμε) και ποιο θα είναι το κόστος που πληρώσουμε θεωρώντας αυτές τις παραπάνω πληροφορίες. Δεν υπάρχει κάποια πρόσφατη σύγκριση ή κάποιες κατευθύνσεις για την αναπαράσταση κειμένου. Υπάρχουν κάποιες ενδείξεις στην ανάκτηση πληροφοριών που μας λένε ότι για μεγάλα κείμενα η επιπλέον θεώρηση άλλων πληροφοριών περάν της bag – of – words αναπαράσταση συχνοτήτων δεν αξίζει πραγματικά τον κόπο. Έχει γίνει επίσης δουλειά στην ομαδοποίηση κειμένων η οποία επεκτείνει την bag – of – words

αναπαράσταση συχνοτήτων που χρησιμοποιεί ακολουθίες λέξεων που ονομάζονται (n – grams), στην θέση απλών λέξεων [99], [98]. Σ' αυτές τις εργασίες υποστηρίζεται ότι η χρησιμοποίηση λέξεων κατά μονάδες και κατά ζευγάρια βελτιώνει την επίδοση της κατάταξης σχετικά μικρών κειμένων.

Πολλά από τα συστήματα που μαθαίνουν από κείμενο χρησιμοποιούν την bag – of – words αναπαράσταση χρησιμοποιώντας είτε λογικά (Boolean) χαρακτηριστικά για να αναπαραστήσουν το γεγονός αν κάποια λέξη εμφανίζεται στο κείμενο ή όχι παραδείγματος χάριν στα [7, 31, 27, 53, 84, 85, 86, 90, 107, 109, 111, 124, 126, 141], είτε χρησιμοποιώντας την συχνότητα εμφάνισης κάθε λέξης παραδείγματος χάριν στα [6, 5, 10, 11, 15, 65, 66, 82, 81, 102, 99, 141, 140]. Υπάρχει ακόμα και κάποιες εργασίες που χρησιμοποιούν κάποιες επιπλέον πληροφορίες όπως την θέση των λέξεων [27, 124], ή χρησιμοποιούν n – άδες λέξεων (τα λεγόμενα n - gramms) [93, 99, 98, 129], (παραδείγματος χάριν το «μηχανική μάθηση» είναι ένα 2 – γραμμα ενώ το «World Wide Web» είναι ένα 3 - γραμμα). Πιο πρόσφατες δουλειές [126] μας δείχνουν ότι η χρησιμοποίηση της δομής υπερκειμένου (hypertext structure) καθώς και την γραφική οργάνωση του ιστοσελίδων (graph organization of Web pages) βελτιώνει τα αποτελέσματα της κατάταξης.

(2) Μια από τις πιο συχνά χρησιμοποιούμενες μεθόδους για την μείωση του αριθμού των διαφορετικών λέξεων που εμφανίζονται σε ένα κείμενο είναι το να διαγράψουμε τις λέξεις που περιέχονται στην λεγόμενη «stop – list» λίστα που περιέχει πάρα πολύ κοινές στην χρήση λέξεις μιας συγκεκριμένης γλώσσας όπως παραδείγματος χάριν «μια», «το», «με» [6, 10, 81, 85, 99, 112, 124]. Μια άλλη προσέγγιση είναι να παραλείπουμε τις λέξεις που δεν εμφανίζονται συχνά σε ένα κείμενο (συχνότητα εμφάνισης < min. συχνότητα) [27, 65, 66, 99]. Μια άλλη πρακτική είναι και ο περιορισμός των λέξεων (word stemming), πράγμα που είναι συνδεδεμένο με την γλώσσα στο οποίο είναι γραμμένο το κείμενο, που χρησιμοποιείται στα [5, 10, 124] και το οποίο συνίσταται στον περιορισμό των λέξεων χρησιμοποιώντας έναν αλγόριθμο περιορισμού ο οποίος παραδείγματος χάριν αντικαθιστά της λέξεις «δουλεύει», «δουλεύοντας», «δουλευτής» με την λέξη «δουλειά». Κάποιες προσεγγίσεις χρησιμοποιούν κάποιες ανεξάρτητες της γλώσσας τεχνικές και εισάγουν κάποιου είδους βαθμολογία λέξεων (word scoring), με σκοπό να διαλέξουν μόνο τις καλύτερες λέξεις [6, 7, 10, 65, 82, 84, 85, 99, 107, 111, 126]. Άλλες πάλι προσεγγίσεις χρησιμοποιούν το λεγόμενο latent semantic indexing with singular value representation (LSI) [11, 15].

Πειράματα με διαφορετικό πλήθος επιλεγμένων χαρακτηριστικών στην ομαδοποίηση κειμένων δείχνουν ότι καλύτερα αποτελέσματα επιτυγχάνονται όταν είτε χρησιμοποιείται μόνο ένα μικρό μέρος από τα χαρακτηριστικά των κειμένων (10% περίπου) είτε όταν χρησιμοποιούνται όλα τα χαρακτηριστικά σε κάποιες από τις περιπτώσεις. Μια σύγκριση των διαφορετικών μέτρων βαθμολόγησης των λέξεων που χρησιμοποιούνται στην επιλογή των ενός υποσυνόλου χαρακτηριστικών από τα συνολικά χαρακτηριστικά ενός κειμένου δείχνει ότι τα πιο υποσχόμενα χαρακτηριστικά λαμβάνουν υπ' όψιν τους την φύση του τομέα του προβλήματος και τον χρησιμοποιούμενο αλγόριθμο κατάταξης [111]. καλά αποτελέσματα έδωσε και η χρησιμοποίηση ενός απλού μέτρου συχνότητας (frequency measure) σε συνδυασμό με μια «stop – list" λίστα [11, 99].

Μια από τις πιο πλατιά χρησιμοποιούμενες μεθόδους αναπαράστασης κειμένου στην ανάκτηση πληροφοριών είναι η παράσταση κάθε εγγράφου με ένα bag – of – words, σαν TFIDF διάνυσμα δηλαδή, στον χώρο των λέξεων που εμφανίζονται σε όλη τη συλλογή κειμένων. Έπειτα αθροίζονται όλα τα ενδιαφέροντα διανύσματα κειμένου. Κάθε συστατικό ενός διανύσματος κειμένου $d^{(i)} = TF(w_i, d)IDF(w_i)$ υπολογίζεται ως γινόμενο των συχνοτήτων των όρων TF (Term Frequency) (το πλήθος των φορών που εμφανίζεται μία λέξη w_i στο κείμενο) με το IDF, όπου $IDF = \log\frac{D}{DF(w_i)}$ (Ανάστροφη συχνότητα (Inverse Frequency)) όπου D είναι ο αριθμός των εγγράφων και $DF(w_i)$ είναι ο αριθμός των εγγράφων στα οποία η λέξη w_i εμφανίζεται τουλάχιστον μία φορά. Οι ακριβείς τύποι που εμφανίζονται στις διάφορες εφαρμογές και προσεγγίσεις μπορεί να διαφέρουν ελαφρώς (μπορούν να προσθέτονται κάποιοι παράγοντες ή να εκτελείται κάποια κανονικοποίηση [120]), αλλά η βασική ιδέα παραμένει η ίδια. Έτσι ένα καινούργιο έγγραφο παριστάνεται τότε σαν διάνυσμα στον ίδιο χώρο και μπορεί να μετρηθεί η απόστασή του από άλλα έγγραφα χρησιμοποιώντας κάποιο μέτρο ανομοιότητας, συνήθως το μέτρο ομοιότητας συνημιτόνου (cosine similarity measure).

Μια επέκταση της TFIDF αναπαράστασης εγγράφων που προτάθηκε από τον Joacims [65] ονομάζεται πιθανοτηκή TFIDF αναπαράσταση και λαμβάνει υπό όψιν της την αναπαράσταση του κειμένου. Σε πειράματα που έχουν γίνει φαίνεται να δίνει καλύτερα αποτελέσματα από την TFIDF αναπαράσταση.

4.5 Ομαδοποίηση Κειμένου

4.5.1 Εισαγωγή

Η ομαδοποίηση είναι μια σημαντική διεργασία, όπως έχουμε αναφέρει και παραπάνω, η οποία χρησιμοποιείται κατά κόρον σαν συστατικό στοιχείο σε πολλά συστήματα εξόρυξης κειμένου και ανάκτησης πληροφορίας. Η ομαδοποίηση κειμένου μπορεί να χρησιμοποιηθεί για να βρεθούν οι κοντινότεροι γείτονες (κοντινότερα κείμενα) ενός εγγράφου [21], για να βελτιωθεί η ακρίβεια ή η ανάκτηση στον τομέα της ανάκτησης πληροφορίας [118, 77], για βοήθεια σε μια συλλογή κειμένων [32], για την οργάνωση των αποτελεσμάτων από μηχανές αναζήτησης [143] και τελευταία για την προσωπικοποίηση (personalization) των αποτελεσμάτων μηχανών αναζήτησης [101].

Οι πιο πολλές προσεγγίσεις ομαδοποίησης κείμενου λειτουργούν με το λεγόμενο μοντέλο διανύσματος χώρου (vector space model), όπου κάθε έγγραφο αντιπροσωπεύεται όπως είδαμε και παραπάνω από ένα διάνυσμα στον χώρο των όρων της συλλογής κειμένων (term – space). Όπως αναφέραμε και προηγουμένως ο χώρος των όρων αποτελείται από όλους εκείνους τους όρους που είναι σημαντικοί σε μια συλλογή κειμένων. Για την ακρίβεια οι αντίστοιχες συχνότητες όρων (Term Frequencies TF) [75] σε ένα δεδομένο έγγραφο μπορούν αν χρησιμοποιηθούν δε ένα διάνυσμα για να αποτελέσουν ένα μοντέλο του εγγράφου. Για να αγνοήσουμε συχνά εμφανιζόμενες λέξεις με μικρή διακριτική ικανότητα, σε κάθε όρο – λέξη μπορεί να αποδοθεί ένα βάρος που να βασίζεται στο Ανάστροφη Συχνότητα του Κειμένου (Inverse Document Frequency IDF) [75, 101] της υποκείμενης συλλογής εγγράφων.

Είναι όμως προφανές ότι η κατανομή των λέξεων στις πιο πολλές συλλογές εγγράφων της πραγματικής ζωής, μπορεί να διαφέρουν δραστικά από ένα σύνολο κειμένων σε ένα άλλο. Έτσι το να βασιζόμαστε μόνο στο IDF για την επιλογή λέξεων κλειδιών (key words) μπορεί μερικές φορές να μην είναι και εντελώς αποδεκτό και θα μπορούσε να αλλοιώσει πολύ τις διαδικασίες ομαδοποίησης καθώς και άλλες διαδικασίες εξόρυξης κειμένου που θα μπορούσαν να ακολουθήσουν. Παραδείγματος χάριν ένα σύνολο κειμένων «Νέων» και ένα σύνολο κειμένων «επιχειρήσεων» θα περίμενε κανείς να έχει διαφορετικό σύνολο σημαντικών όρων. Τώρα αν τα έγγραφα αυτά είχαν εκ των προτέρων προκαταταχεί χειροκίνητα σε ομάδες τότε θα ήταν μια τετριμμένη διαδικασία να διαλέξει κανείς εν γένει διαφορετικό σύνολο σημαντικών λέξεων σε κάθε κατηγορία που να βασίζεται στο IDF. Αλλά για μεγάλες δυναμικές συλλογές κειμένων όπως είναι παραδείγματος χάριν το World Wide Web είναι μη πραγματοποιήσιμη μια τέτοια χειροκίνητη προκατάταξη των εγγράφων. Για αυτόν τον λόγο χρειαζόμαστε από μια αυτόματη ή μη επιβλεπόμενη διαδικασία κατάταξης – ομαδοποίησης η οποία να μπορέι να χειριστεί κατηγορίες που να διαφέρουν ριζικά στο σύνολο των πιο σημαντικών τους όρων. Δυστυχώς δεν είναι δυνατόν να μπορούμε να ξεχωρίσουμε διαφορετικά σύνολα σημαντικών όρων, εκτός και αν τα σύνολά μας είναι ομαδοποιημένα. Αυτό σημαίνει ότι σε μια μη επιβλεπόμενη διαδικασία οι κατηγορίες και οι αντίστοιχοι σημαντικοί όροι θα πρέπει να ανακαλυφθούν ταυτόχρονα. Το να διαλέγουμε και να αποδίδουμε βάρη σε υποσύνολα λέξεων κλειδιά σε έγγραφα κειμένου είναι μια διαδικασία παρόμοιας φύσης με το πρόβλημα της επιλογής χαρακτηριστικών και της απονομής βαρών στην αναγνώριση προτύπων και στην εξόρυξη δεδομένων. Το πρόβλημα της επιλογής ενός καλού υποσυνόλου γνωρισμάτων αποτελεί ένα σημαντικό μέρος του σχεδιασμού ενός καλού αλγορίθμου μάθησης που να τα καταφέρνει αρκετά καλά στις εφαρμογές πραγματικού κόσμου. Μη σχετικά χαρακτηριστικά θα μπορούσαν να μειώσουν σημαντικά την ικανότητα γενίκευσης των αλγορίθμων αυτών. Στην πραγματικότητα ακόμα και αν πρότυπα δεδομένων έχουν ήδη καταταχθεί σε γνωστές κλάσεις, είναι γενικά πιο επιθυμητό να μοντελοποιήσουμε τις πολύπλοκες αυτές κλάσεις με ένα σύνολο απλών υποκλάσεων ή ομάδων και να χρησιμοποιήσουμε διαφορετικό σύνολο βαρών για κάθε διαφορετική κλάση. Αυτό βοηθάει στο να κατατάξουμε νέα αντικείμενα στις ήδη υπάρχουσες κλάσεις. Μέχρι τώρα το πρόβλημα της ομαδοποίησης και της επιλογής χαρακτηριστικών έχει μάλλον προσεγγιστεί ανεξάρτητα [2, 71, 79, 74].

4.5.2 Ένας αλγόριθμος για ομαδοποίηση και ταυτόχρονη απόδοση βαρών σε όρους κλειδιά.

4.5.2.1 Εισαγωγή

Παραπάνω αναφερθήκαμε στους κυριότερους τρόπους αναπαράστασης ενός εγγράφου προκειμένου να λάβει μέρος σε μια διαδικασία επιβλεπόμενης (κατάταξη) ή μη επιβλεπόμενης (ομαδοποίηση) κατάταξη. Εδώ θα αναφερθούμε σε έναν και μόνο αλγόριθμο ομαδοποίηση ο οποίος ταυτόχρονα αποδίδει και βάρη στους όρους κλειδιά κάθε ομάδας.

Στο [48] παρουσιάζεται ένας αλγόριθμος που λέγεται ταυτόχρονη ομαδοποίηση και διαχώριση χαρακτηριστικών (Simultaneous Clustering and Attribute Discrimination SCAD) ο οποίος εκτελεί τις διεργασίες της ομαδοποίησης και της απόδοσης βάρους

στους όρους ταυτόχρονα. Όταν ο SCAD χρησιμοποιείται σαν μέρος μιας επιβλεπόμενης ή μη επιβλεπόμενης διαδικασίας μάθησης προσφέρει αρκετά θετικά στοιχεία. Κατ' αρχάς η συνεχής απόδοση βαρών στους όρους προσφέρει πολύ πιο πλούσια αναπαράσταση συνάφειας χαρακτηριστικών (feature relevance representation) από ότι η δυαδική εκλογή χαρακτηριστικών (feature selection). Δεύτερον ο SCAD μαθαίνει μια διαφορετική αναπαράσταση συνάφειας χαρακτηριστικών (feature relevance representation) για κάθε διαφορετική ομάδα (cluster) με έναν μη επιβλεπόμενο τρόπο. Από την άλλη μεριά όμως ο SCAD είχε δημιουργηθεί έτσι ώστε να λειτουργεί με δεδομένα που κείτονται σε κάποιον Ευκλείδειο χώρο και το μέτρο απόστασης που χρησιμοποιήθηκε ήταν η Ευκλείδεια απόσταση. Για ειδική περίπτωση των εγγράφων κειμένων, είναι πολύ καλά γνωστό ότι η ευκλείδεια απόσταση δεν είναι η κατάλληλη επιλογή. Είναι επίσης γνωστό ότι αποστάσεις σαν τις αποστάσεις Jaccard ή τις αποστάσεις συνημιτόνου (cosine similarity) είναι πολύ πιο αποτελεσματικές στο να εκφράσουν την ομοιότητα ή την ανομοιότητα μεταξύ κειμένων.

Παρακάτω θα δούμε μια επέκταση του αλγορίθμου SCAD σε μια διαδικασία ταυτόχρονης ομαδοποίησης εγγράφων κειμένων και δυναμική εξαρτημένη από τις κατηγορίες απόδοση βαρών σε όρους κλειδιά (simultaneous text document clustering and dynamic category – dependent keyword set weighting). Αυτή η τεχνική προσέγγισης της ομαδοποίησης κειμένου ονομάστηκε «Simultaneous KeyWord Identification and Clustering – SKWIC» και είναι απλή στην υλοποίηση αλλά και στην σύλληψη. Τα πλεονεκτήματα, σε σχέση με άλλες μεθόδους είναι: Πρώτον η απόδοση βαρών στους όρους γίνεται με συνεχή τρόπο πράγμα που μας δίνει μια πλούσια αναπαράσταση συνάφειας χαρακτηριστικών (feature relevance representation) σε σχέση με την δυαδική επιλογή χαρακτηριστικών: Στην πρώτη περίπτωση δεν θεωρούνται όλοι οι όροι εξίσου συναφείς σε κάθε διαφορετική κατηγορία εγγράφων κειμένου. Αυτό είναι πολύ κοντά στην πραγματικότητα κυρίως στις περιπτώσεις που το πλήθος των λέξεων κλειδιών (key words) είναι σχετικά μεγάλο. Για παράδειγμα θα μπορούσε κάποιος να περιμένει ότι η λέξη «playoff» θα ήταν πολύ πιο σημαντική από τη λέξη «πρόγραμμα» στο να ξεχωρίσουμε ένα σύνολο εγγράφων που αναφέρονται στα αθλητικά. Δεύτερον, δεδομένου ενός όρου δεν θεωρείται εξίσου συναφής σε όλες τις κατηγορίες: Για παράδειγμα ο όρος «film» θα μπορούσε να είναι πολύ πιο σημαντικός σε μια κατηγορία κειμένων που αναφέρονται στη «διασκέδαση» από ότι σε μια κατηγορία «αθλητικών». Τέλος, ακόμα ένα πλεονέκτημα της μεθόδου SKWIC είναι ότι μαθαίνει διαφορετικό σύνολο βαρών των όρων για κάθε ξεχωριστή ομάδα (cluster) και μάλιστα με έναν μη επιβλεπόμενο τρόπο.

Υπάρχει ακόμα και μια άλλη επέκταση του SKWIC η οποία έχει να κάνει με την κληρονομουμένη ασάφεια σε έγγραφα κειμένων, παράγοντας αυτόματα ασαφείς ετικέτες (fuzzy or soft labels) στη θέση της κατηγοριοποίησης και μοναδικές ταμπέλες (single – label categorization). Αυτό σημαίνει ότι ένα έγγραφο μπορεί να ανήκει σε πολλές κατηγορίες αλλά με κάποιο βαθμό συμμετοχής σε κάθε κατηγορία [49].

Με την αποτελεσματικότητα της δυναμικής απόδοσης βαρών και της συνεχής αλληλεπίδρασης με υπολογισμούς απόστασης και συναρτήσεις μέλους, η ασαφής αυτή μέθοδος μπορεί να χειριστεί και κείμενα που περιέχουν θόρυβο, με την

αυτόματη σχεδίαση ενός ή δυο μαγνητών θορύβου (noisy magnets) οι οποίες θα αποσπούν όλες σχεδόν τις απομακρυσμένες τιμές από τις άλλες ομάδες [49].

4.5.2.2 Ο αλγόριθμος SKWIC.

Όπως αναφέραμε και παραπάνω ο SCAD [48] βασίστηκε στην ευκλείδεια απόσταση. Είναι όμως γνωστό ότι για πολλές εφαρμογές της εξόρυξης δεδομένων όπως η ομαδοποίηση κειμένων ή άλλες πολυδιάστατες διεργασίες η ευκλείδεια απόσταση δεν είναι κατάλληλη επιλογή. Γενικά θα μπορούσε να πει κάποιος ότι η ευκλείδεια απόσταση δεν είναι ένα καλό μέτρο ανομοιότητας για κατηγοριοποίηση εγγράφων. Αυτό οφείλεται κατά πρώτο λόγο στην μεγάλη διάσταση του προβλήματος και κατά δεύτερο λόγο στο γεγονός ότι δυο κείμενα δεν μπορούν αν θεωρηθούν όμοια αν λείπουν και από τα δύο οι ίδιοι όροι κλειδιά. Πιο κατάλληλη για αυτήν την εφαρμογή είναι το μέτρο ομοιότητας συνημιτόνου (cosine similarity measure) [75]. Το μέτρο αυτό ορίζεται μεταξύ των διανυσμάτων συχνότητας των εγγράφων x_i και y_i που ορίζονται σε ένα λεξιλόγιο (vocabulary) n όρων ως εξής:

$$S(x_i, y_i) = \frac{\sum_{k=1}^{n} x_{ik} \times x_{jk}}{\sqrt{\sum_{k=1}^{n} x_{ik}^2} \sqrt{\sum_{k=1}^{n} x_{jk}^2}} \quad (4.1).$$

Για να είμαστε σε θέση να επεκτείνουμε την συνάρτηση κριτηρίου του SCAD στην περίπτωση που κάποιο άλλο μέτρο ανομοιότητας χρησιμοποιηθεί, θα πρέπει να είμαστε σε θέση να αποσυνθέσουμε το μέτρο ανομοιότητας κατά μήκος κάθε διαφορετικής κατεύθυνσης χαρακτηριστικού. Η προσπάθεια εδώ είναι να δημιουργηθεί ένα μέτρο ανομοιότητας που να βασίζεται στο μέτρο συνημιτόνου. Αυτό γίνεται ορίζοντας την ανομοιότητα μεταξύ του κειμένου y_i και του κέντρου της i – οστής ομάδας ως εξής:

$$\tilde{D}_{wc_{ij}} = \sum_{k=1}^{n} v_{ik} D_{wc_{ij}}^{k} \quad (4.2),$$

το οποίο είναι το συνολικό βεβαρημένο συνολικό άθροισμα των αποστάσεων που βασίζονται στο μέτρο του συνημιτόνου κατά μήκος των διαφορετικών διαστάσεων, όπου

$$D_{wc_{ij}}^{k} = \frac{1}{n} - \left(x_{jk} \cdot c_{ik}\right) \quad (4.3)$$

με n να είναι ο συνολικός αριθμός των όρων της συλλογής των N εγγράφων, c_{ik} να είναι το k – οστό στοιχείο του διανύσματος του κέντρου της i – οστής ομάδας και $V = \left[v_{ik}\right]$ είναι το βάρος της συνάφειας (relevance) του όρου k στην ομάδα i. Ας σημειώσουμε ότι στην εξίσωση (4.2) τα μεμονωμένα γινόμενα δεν κανονικοποιούνται γιατί έχει υποτεθεί ότι διανύσματα των δεδομένων έχουν κανονικοποιηθεί σε

μοναδιαίο μήκος πριν από την διαδικασία ομαδοποίησης και ότι τα κέντρα των ομάδων κανονικοποιούνται πριν από κάθε επανάληψη.

Ο SKWIC είναι έτσι σχεδιασμένος ώστε να ψάχνει για τα βέλτιστα κέντρα των ομάδων C και το βέλτιστο σύνολο των βαρών των όρων V ταυτόχρονα. Κάθε ομάδα έχει το δικό της σύνολο βαρών των όρων (feature weights) $V_i = [v_{i1},...,v_{in}]$. Ορίζουμε την παρακάτω αντικειμενική συνάρτηση

$$J(C,V;\aleph) = \sum_{i=1}^{C}\sum_{x_j \in \aleph_i}\sum_{k=1}^{n} v_{ik} D_{wc_{ij}}^{k} + \sum_{i=1}^{C}\delta_i \sum_{k=1}^{n} v_{ik}^2 \quad (4.4)$$

με τους περιορισμούς:

$$v_{ik} \in [0,1], \forall i,k, \kappa\alpha\iota \sum_{k=1}^{n} v_{ik} = 1, \forall i \quad (4.5).$$

Η αντικειμενική συνάρτηση της εξίσωσης (4.4) αποτελείται από τα εξής στοιχεία: Το πρώτο συστατικό είναι το άθροισμα των αποστάσεων ή των λαθών από τα κέντρα των ομάδων. Αυτό το συστατικό μας βοηθάει στο να αποκτάμε συμπαγείς ομάδες. Αυτός ο όρος ελαχιστοποιείται όταν μονάχα ένας όρος σε κάθε ομάδα είναι εντελώς συναφής (relevant) ενώ όλοι οι άλλοι όροι κλειδιά είναι εντελώς μη συναφής (irrelevant). Το δεύτερο συστατικό της εξίσωσης (4.4) είναι το άθροισμα των τετραγώνων των βαρών των όρων κλειδιών. Το ολικό ελάχιστο αυτού του συστατικού αποκτάται όταν όλοι οι όροι κλειδιά έχουν το ίδιο βάρος. Όταν και τα δυο συστατικά σχετιστούν και τα δ_t επιλεχθούν κατάλληλα, η τελική διαμέριση θα ελαχιστοποιήσει το άθροισμα των ενδοομαδικών βεβαρημένων αποστάσεων με τα βάρη των όρων κλειδιών να βελτιστοποιούνται για κάθε ομάδα.

Για να βελτιστοποιήσουμε το J υπό τον περιορισμό V, χρησιμοποιούμε την πολλαπλασιαστική τεχνική του Lagrange και λαμβάνουμε:

$$J(\Lambda,V) = \sum_{i=1}^{C}\sum_{x_j \in \aleph_i}\sum_{k=1}^{n} v_{ik} D_{wc_{ij}}^{k} + \sum_{i=1}^{C}\delta_i \sum_{k=1}^{n} v_{ik}^2 - \sum_{i=1}^{C}\lambda_i \left(\sum_{k=1}^{n} v_{i\kappa} - 1 \right) \quad (4.6),$$

όπου $\Lambda = [\lambda_1,...,\lambda_n]^T$. Αφού οι στήλες του V είναι ανεξάρτητες μεταξύ τους το παραπάνω πρόβλημα μπορεί να μετασχηματιστεί στο ακόλουθο σύνολο των C ανεξάρτητων προβλημάτων:

$$J_i(\lambda_i, V_i) = \sum_{x_j \in \aleph_i}\sum_{k=1}^{n} v_{ik} D_{wc_{ij}}^{k} + \delta_i \sum_{k=1}^{n} v_{ik}^2 - \lambda_i \left(\sum_{k=1}^{n} v_{i\kappa} - 1 \right), i = 1,...,C, \quad (4.7)$$

όπου V_i είναι η i – οστή γραμμή του V. Θέτοντας την κλίση (gradient) του J_i μηδέν παίρνουμε:

$$\frac{\partial J_i(\lambda_i, V_i)}{\partial \lambda_i} = \sum_{k=1}^{n} (v_{\iota\kappa} - 1) = 0 \quad (4.8)$$

και

$$\frac{\partial J_i(\lambda_i, V_i)}{\partial v_{\iota\kappa}} = \sum_{x_j \in \aleph_i} D^k_{wc_{ij}} + 2\delta_t v_{ik} - \lambda_i = 0 \quad (4.9).$$

Λύνοντας τις εξισώσεις (4.8) και (4.9) ως προς v_{ik} παίρνουμε

$$v_{ik} = \frac{1}{n} + \frac{1}{2\delta_t} \sum_{x_j \in \aleph_i} \left[\frac{1}{n} \sum_{x_j \in \aleph_i} D^k_{wc_{ij}} - D^k_{wc_{ij}} \right] \quad (4.10).$$

Ο πρώτος όρος στην εξίσωση (4.10) (1/n) είναι η προκαθορισμένη τιμή αν σε όλα τα χαρακτηριστικά / όροι αποδοθεί η ίδια τιμή και δεν καμία διάκριση δεν επιχειρηθεί. Ο δεύτερος όρος είναι ένας πολωτικός όρος ο οποίος μπορεί να είναι είτε θετικός είτε αρνητικός. Αυτός ο όρος είναι θετικός για συμπαγή χαρακτηριστικά όταν η απόσταση κατά αυτήν την κατεύθυνση είναι κατά μέσο όρο μικρότερη από την συνολική απόσταση όλων των διαστάσεων. Αν ένα χαρακτηριστικό είναι πολύ συμπαγές σε σχέση με τα άλλα χαρακτηριστικά, για τα περισσότερα σημεία που ανήκουν σε μια δεδομένη ομάδα, τότε αυτό είναι πολύ συναφές για τη συγκεκριμένη ομάδα. Ας σημειωθεί ότι είναι πιθανό για την κάθε ανομοιότητα όρο προς όρο στην εξίσωση (4.3) να γίνει αρνητικό. Αυτό θα δώσει επιπλέον έμφαση σε αυτήν τη διάσταση και θα έχει ως αποτέλεσμα σχετικά μεγαλύτερα βάρη χαρακτηριστικών v_{ik} (δες εξίσωση 4.10). Επιπλέον η συνολική συγκεντρωτική ανομοιότητα στην εξίσωση (4.2) μπορεί να γίνει και αρνητική. Αυτό δεν θα αποτελέσει κάποιο πρόβλημα αφού διαμερίζουμε τα δεδομένα μας βασιζόμενη στην ελάχιστη απόσταση.

Η επιλογή των δ_t στην εξίσωση (4.4) είναι σημαντική στον αλγόριθμο SKWIC αφού αυτή η επιλογή αντανακλά τη σημαντικότητα μεταξύ του δεύτερου όρου σε σχέση με τον πρώτο. Εάν τα δ_t είναι πολύ μικρά τότε μόνο ένας όρος κλειδί θα είναι συναφής στην ομάδα i και σ' αυτόν θα αποδοθεί το βάρος 1. σε όλες τις άλλες λέξεις θα αποδοθούν μηδενικά βάρη. Από την άλλη πλευρά αν τα δ_t είναι πολύ μεγάλα τότε όλες οι λέξεις στην ομάδα i θα είναι συναφείς και θα τους αποδοθούν ίσα βάρη 1/n. Οι τιμές των δ_t θα πρέπει να διαλέγονται έτσι ώστε και οι δυο όροι να έχουν την ίδια τάξη μεγέθους. Μια επιλογή των δ_t που δίνει ικανοποιητικά αποτελέσματα [49] μπορεί να υπολογιστεί αναδρομικά ως εξής:

$$\delta_i^{(t)} = K_\delta \frac{\sum_{x_j \in \aleph_i} \sum_{k=1}^{n} v_{ik}^{(t-1)} \left(D^{k^{(t-1)}}_{wc_{ij}} \right)}{\sum_{k=1}^{n} \left(v_{ik}^{(t-1)} \right)^2} \quad (4.11).$$

Στην εξίσωση (4.11) το K_δ είναι μια σταθερά και ο υπερδείκτης (t – 1) χρησιμοποιείται για να δηλώσει τις τιμές των μεταβλητών στην επανάληψη (t – 1).

Θα πρέπει επίσης να σημειωθεί ότι σε σχέση με τις τιμές των δ_t οι τιμές της συνάφειας των χαρακτηριστικών v_{ik} μπορεί να μην περιορίζονται στο διάστημα [0,1]. Αν αυτό εμφανίζεται πολύ συχνά τότε αυτό είναι μια ένδειξη ότι η τιμή των δ_t είναι πολύ μικρές και θα πρέπει να μεγαλώσουν (θα πρέπει να αυξήσουμε το K_δ). Αντίθετα αν αυτό εμφανιστεί μόνο σε λίγες ομάδες και μόνο σε κάποιες επαναλήψεις τότε οι αρνητικές τιμές της συνάφειας των χαρακτηριστικών μπορούν να προσαρμοστούν ως εξής:

$$v_{ik} \leftarrow v_{ik} + \left| \min_{k=1}^{n} v_{ik} \right| \alpha w_{ik} < 0 \quad (4.12).$$

Μπορεί ακόμα να δειχθεί ότι ο διαμερισμός σε ομάδες ο οποίος ελαχιστοποιεί το J είναι αυτός που αναθέτει κάθε δεδομένο στην ομάδα με το κοντινότερο κέντρο, το οποίο σημαίνει:

$$\aleph_i = \left\{ x_j \mid \tilde{D}_{wc_{ij}} \le \tilde{D}_{wc_{kj}} \forall k \ne i \right\} \quad (4.13),$$

όπου $\tilde{D}_{wc_{kj}}$ είναι η βεβαρημένη συνολική απόσταση που βασίζεται στην απόσταση του συνημιτόνου της εξίσωσης (4.2).

Δεν είναι δυνατόν να ελαχιστοποιηθεί το J σε σχέση με τα κέντρα. Έτσι αυτό που γίνεται είναι να υπολογιστούν να νέα κέντρα των ομάδων κανονικά όπως και στον αλγόριθμο SCAD [48] και στη συνέχεια να τα κανονικοποιήσουμε και έτσι να αποκτήσουμε τα καινούργια κέντρα (centroids) των ομάδων. Υπάρχουν δυο περιπτώσεις σε σχέση με την τιμή της v_{ik}.

<u>Περίπτωση 1:</u> $v_{ik} = 0$

> Σε αυτήν την περίπτωση το k – οστό χαρακτηριστικό είναι εντελώς μη σχετικό με την ομάδα i. Έτσι ανεξάρτητα από την τιμή των c_{ik}, οι τιμές αυτού του χαρακτηριστικού δεν θα συνεισφέρουν στον υπολογισμό της συνολικής απόστασης. Έτσι σε αυτήν την περίπτωση κάθε τυχαία τιμή θα μπορεί να διαλεχτεί για το c_{ik}. Συνήθως στις εφαρμογές χρησιμοποιείται για το c_{ik} σε αυτήν την περίπτωση η τιμή 0 [49].

<u>Περίπτωση 2:</u> $v_{ik} \ne 0$

> Σε αυτήν την περίπτωση που το k – οστό χαρακτηριστικό έχει κάποια συνάφεια με την i – οστή ομάδα τότε το κέντρο της περιορίζεται στο:

$$c_{ik} = \frac{\sum_{x_j \in \aleph_i} x_{ij}}{\sum_{x_j \in \aleph_i} 1} \quad (4.14).$$

Για να συνοψίσουμε η ολοκληρωμένη συνάρτηση για τα κέντρα είναι:

$$c_{ik} = \begin{cases} 0, \alpha\nu_{ik} = 0 \\ \dfrac{\sum_{x_j \in \aleph_i} x_{ij}}{|\aleph_i|} \end{cases} \quad (4.15).$$

Συνοψίζουμε τώρα τον αλγόριθμο SKWIC για την ομαδοποίηση μιας συλλογής N κανονικοποιημένων διανυσμάτων κειμένου πάνω σε ένα λεξιλόγιο n λέξεων.

Δώσε τον αριθμό των ομάδων C.
Αρχικοποίησε τα κέντρα των ομάδων διαλέγοντας τυχαία C κείμενα.
Αρχικοποίησε τις διαμερίσεις $\aleph_i$, χρησιμοποιώντας την εξίσωση (4.13) και απόδωσε ίσα βάρη στους όρους ίσα με 1/n.
ΕΠΑΝΕΛΑΒΕ
Υπολόγισε το $D^k_{wc_{ij}} = \frac{1}{n} - \left(x_{jk} \cdot c_{ik}\right)$ για i = 1, …, C, j = 1, …, N και k = 1, …, n.
Ενημέρωσε τα βάρη συνάφειας ν_{ik} χρησιμοποιώντας την εξίσωση (4.10).
Υπολόγισε τα $\widetilde{D}_{wc_{ij}}$ για i = 1, …, C και j = 1, …, N χρησιμοποιώντας την εξίσωση (4.2).
Ενημέρωσε την διαμέριση των ομάδων $\aleph_i$ χρησιμοποιώντας την (4.13).
Ενημέρωσε τα κέντρα χρησιμοποιώντας την (4.15).
Ενημέρωσε τα δ_t χρησιμοποιώντας την (4.11).
ΜΕΧΡΙ (να σταθεροποιηθούν τα κέντρα**).**

Αλγόριθμος 4.1: Αλγόριθμος SKWIC.

5 Το προτεινόμενο μέτρο ανομοιότητας

5.1 Εισαγωγή

Όπως είπαμε και παραπάνω οι μέθοδοι ομαδοποίησης έχουν εφαρμοστεί σε πάρα πολλά και διαφορετικά εφαρμοσμένα πεδία της επιστήμης (στο marketing, στην μηχανική, στην ιατρική, στην βιολογία, στην ψυχολογία, στην αστρονομία καθώς και σε στρατιωτικές εφαρμογές). Επίσης η ομαδοποίηση έχει χρησιμοποιηθεί και σε πεδία όπως στην εξόρυξη δεδομένων, στην ταξινομία και στην στατιστική ανάλυση δεδομένων. Όπως είπαμε αναλυτικότερα παραπάνω η ομαδοποίηση είναι ο χωρισμός ενός συνόλου αντικειμένων σε διακριτές και ομογενείς ομάδες, έτσι ώστε τα αντικείμενα που ανήκουν σε μία ομάδα να είναι πιο όμοια από κάθε άλλο αντικείμενο που δεν ανήκει στην συγκεκριμένη ομάδα. Τα αντικείμενα που θα ομαδοποιηθούν θεωρούνται ως ένα σύνολο από χαρακτηριστικά (attributes), κάθε αντικείμενο δηλαδή είναι μια σύζευξη από τιμές χαρακτηριστικών.

Οι μέθοδοι της ομαδοποίησης έχουν χρησιμοποιηθεί σε επιστημονικά πεδία όπως η Μηχανική Μάθηση, τα Νευρωνικά δίκτυα και η στατιστική. Όπως αναφέραμε πιο αναλυτικά παραπάνω οι αλγόριθμοι ομαδοποίησης μπορούν να ταξινομηθούν [1] σε ιεραρχικούς (hierarchical) και αναδρομικούς (iterative) (διαμεριστικούς (partitional), αναζήτησης πυκνότητας (density search), αναλυτικού παράγοντα (factor analytic) και γραφοθεωρητικοί (graph theoretic)). Μερικοί από τους πιο δημοφιλείς ιεραρχικούς αλγορίθμους είναι οι Complete – link, average – link και single – link αλγόριθμοι [37,67,128]. Από την άλλη μεριά μερικοί από τους πιο δημοφιλείς διαμεριστικούς αλγόριθμους είναι ο K – means [62,89], με τις παραλλαγές του π.χ. [146,119,137] καθώς και οι αλγόριθμοι hill - climbing [4].

Όπως είδαμε και πιο αναλυτικά παραπάνω ο αλγόριθμος k – means καθώς και οι παραλλαγές του αποτελούνται από δυο φάσεις. Στην πρώτη φάση υπολογίζεται μια διαμέριση των δεδομένων σε k διαφορετικές ομάδες, ενώ στην δεύτερη φάση υπολογίζονται βελτιώσεις αυτής της διαμέρισης, για να βελτιωθεί η ποιότητά της. Ο k – means ξεκινάει από μια τυχαία αρχική διαμέριση. Ύστερα οι υπόλοιπες διαμερίσεις ξαναϋπολογίζονται επαναληπτικά μέχρι η συνάρτηση ποιότητας της διαμέρισης (quality function) να φτάσει σε κάποιο βέλτιστο. Τα βήματα που ακολουθούνται στον k – means και στις παραλλαγές του, όπως αναφέραμε και παραπάνω είναι τα ακόλουθα:

1. επιλογή των k αρχικών μέσων
2. απόδοση κάθε αντικειμένου στην ομάδα που διαθέτει το πιο κοντινό μέσο (mean) στο αντικείμενο
3. επαναϋπολογισμός των k καινούργιων μέσων για τις ομάδες
4. υπολογισμός της συνάρτησης ποιότητας

με τα τελευταία 3 βήματα να εκτελούνται επαναληπτικά μέχρι που ο αλγόριθμος να συγκλίνει. Οι περισσότεροι αλγόριθμοι που είναι παραλλαγές του k – means έχουν αποδειχθεί ότι συγκλίνουν [122]. Από την άλλη πλευρά οι αλγόριθμοι τύπου k – means συγκλίνουν τις περισσότερες φορές σε ένα τοπικό βέλτιστο. Στην

πραγματικότητα μπορούμε να μετασχηματίσουμε το πρόβλημα της ομαδοποίησης χρησιμοποιώντας τον k – means σε ένα μη κυρτό πρόβλημα ομαδοποίησης [70]:

Το πρόβλημα θα μπορούσε να τεθεί πιο αυστηρά ως εξής:

> Έστω $i_1, i_2, ...i_N$ είναι τα αντικείμενα εισόδου και κάθε ένα από αυτά μπορεί να αναπαρασταθεί από μια d – άδα, $(au_1, au_2, ...au_d)$ όπου $au_i, i = 1,...,d$ συμβολίζει την τιμή του i – οστού χαρακτηριστικού με πεδίο $dom(a_i)$ να είναι το κάποιο υποσύνολο των πραγματικών αριθμών, εάν πρόκειται για αριθμητικά δεδομένα ή ένα σύνολο ετικετών εάν πρόκειται για κατηγορικά (categorical) δεδομένα και τα k αρχικά μέσα (means (στην περίπτωση των αριθμητικών δεδομένων) / medoids (στην περίπτωση των κατηγορικών δεδομένων)) αρχικοποιούνται σε k από τα N αντικείμενα εισόδου $i_{m1}, i_{m2}, ...i_{mk}$, τότε ζητείται η ελαχιστοποίηση της συνάρτησης ποιότητας (quality function):
>
> $$E = \sum_{j=1}^{k} \sum_{i_l \in C_j} q\left(i_l, i_{mj}\right) \quad (5.1),$$
>
> όπου $C = \left\{C_j \mid j = 1,...,k\right\}$ είναι το σύνολο των k διαφορετικών ομάδων και q είναι ένα μέτρο ομοιότητας (ανομοιότητας) που ορίζεται διαφορετικά σε κάθε παραλλαγή του k – means. Στον αρχικό k – means το q ορίζεται ως το τετράγωνο της ευκλείδειας απόστασης, δηλαδή: $q(x, y) = \|x - y\|^2$.

Σε όλες τις διαφορετικές παραλλαγές του k – means οι ομάδες αναπαρίστανται είτε από την βαρύτητα τους, δηλαδή το μέσον τους (ο μέσος όρος των στοιχείων της ομάδας) ή από ένα αντιπροσωπευτικό στοιχείο της ομάδας (medoid).

Παραδοσιακά οι μέθοδοι ομαδοποίησης μπορούν χρησιμοποιούν κυρίως αριθμητικά δεδομένα (ποσοτικά ή μετρικά δεδομένα (quantitative or metric data)), δηλαδή αντικείμενα που παριστάνονται από μια σύζευξη αριθμητικών τιμών. Η κριτική εναντίων των αριθμητικών μεθόδων ομαδοποίησης είναι βασικά ότι στερούνται κατανοητότητας. Αυτές οι μέθοδοι βασίζονται σε κάποιο κατά ζεύγη μέτρα ομοιότητας, όπως το εσωτερικό γινόμενο ή κάποιο μέτρο ανομοιότητας, όπως η ευκλείδεια απόσταση, μεταξύ των δεδομένων. Οι τελικές ομάδες αναπαρίστανται από ένα σύνολο συζευγμένων αριθμών οι οποίοι είναι σχετικά δύσκολο να ερμηνευτούν από ανθρώπους. Από την άλλη πλευρά οι σύγχρονες βάσεις δεδομένων περιέχουν πολλές φορές και δεδομένα τα οποία είναι κατηγορικά ή αλλιώς ονομαστικά (nominal), μη μετρικά (non metric) ή συμβολικά (symbolic). Θα πρέπει να επισημάνουμε ότι μερικές φορές ο όρος «συμβολικός» χρησιμοποιείται επίσης για να δηλώσει πολύπλοκα συγκεντρωτικά αντικείμενα (complex aggregated objects) τα οποία όμως περιέχουν μια εσωτερική διακύμανση και είναι δομημένα (π.χ. στο [16]). Τα κατηγορικά δεδομένα αναπαρίστανται από μια σύζευξη κατηγορικών τιμών. Ένα αντικείμενο είναι κατηγορικό εάν το σύνολο των τιμών που μπορεί να λάβει είναι πεπερασμένο (π.χ. (επάγγελμα, φύλο και σύμπτωμα).

5.2 Υπάρχουσες λύσεις στο πρόβλημα

Μια προφανής λύση στο πρόβλημα της ομαδοποίησης κατηγορικών δεδομένων είναι η κωδικοποίηση των κατηγορικών δεδομένων σε αριθμητικά και η μετέπειτα εφαρμογή κάποιου γνωστού αλγορίθμου ομαδοποίησης αριθμητικών δεδομένων. Το σημαντικό μειονέκτημα αυτής της μεθόδου είναι ότι η κωδικοποίηση των κατηγορικών δεδομένων σε αριθμούς δεν μπορεί να διατηρήσει την σημαντική των κατηγορικών δεδομένων.

Η πιο συνηθισμένη αντιμετώπιση του προβλήματος είναι η χρησιμοποίηση του μέτρου επικάλυψης (overlap measure), το οποίο ορίζει την ανομοιότητα μεταξύ δυο κατηγορικών δεδομένων βασιζόμενο στο μη – ταιριάσματα των χαρακτηριστικών. Αυτό το απλό αλλά κατά τα άλλα συχνά χρησιμοποιούμενο μέτρο ανομοιότητας για δυο απλές τιμές χαρακτηριστικού a_i και a_j η ανομοιότητα ορίζεται ως μηδέν όταν τα a_i και a_j είναι όμοια και ένα όταν είναι διαφορετικά. Η αδυναμία αυτής της μεθόδου είναι ότι δίνει την ίδια ακριβώς σημαντικότητα σε κάθε τιμή του χαρακτηριστικού και σε κάθε χαρακτηριστικό των αντικειμένων. Επίσης ένα ακόμα μειονέκτημα αυτής της μεθόδου είναι ότι κάθε τιμή ενός χαρακτηριστικού απέχει ακριβώς την ίδια απόσταση από κάθε άλλη. Έτσι δεν μπορεί να αναπαραστήσει ζευγάρια τιμών με διαφορετικούς βαθμούς ανομοιότητας όπως παραδείγματος χάριν στο αντικείμενο γεύση η τιμή ξινό θα πρέπει να είναι πιο κοντά στο γλυκό παρά στο πικρό.

Πολλές παραλλαγές υπάρχουν στην βιβλιογραφία που προσπαθούν να αντιμετωπίσουν αυτό το πρόβλημα. Παραδείγματος χάριν στο [94] όπου η ανομοιότητα παριστάνεται μέσω κάποιων παραμέτρων, που αντιστοιχούν σε αυτά τα χαρακτηριστικά στα οποία διαφέρουν τα δυο αντικείμενα. Αυτές οι παράμετροι παίρνουν τιμές στο διάστημα (0,1]. Ας σημειωθεί ότι στην περίπτωση που αυτές οι παράμετροι τεθούν μηδέν τότε το μέτρο ανομοιότητας περιορίζεται στο παραπάνω γνωστό μέτρο επικάλυψης. Ως ένα άλλο παράδειγμα μπορούμε να αναφέρουμε ότι στον αλγόριθμο k – modes [146] έχει χρησιμοποιηθεί μια απόσταση chi – square [55], η οποία λαμβάνει επίσης υπ' όψιν την συχνότητα εμφάνισης των τιμών των χαρακτηριστικών μέσα σε μια ομάδα με σεβασμό στη συνολική βάση δεδομένων. Πιο αυστηρά αν A και B είναι δυο κατηγορικά δεδομένα που περιγράφονται από m διαφορετικά χαρακτηριστικά τότε το μέτρο ανομοιότητας εκφράζεται ως εξής:

$$d_{x^2}(A,B) = \sum_{j=1}^{m} \frac{\left(n_{a_j} + n_{b_j}\right)}{\left(n_{a_j} n_{b_j}\right)} \delta\left(a_j, b_j\right) \quad (5.2)$$

$$\text{όπου } \delta\left(a_j, b_j\right) = \begin{cases} 0, \alpha\nu a_j = b_j \\ 1, \alpha\nu a_j \neq b_j \end{cases}$$

και n_{a_j} και n_{b_j} είναι ο αριθμός των αντικειμένων εισόδου που έχουν τιμές a_j και b_j για το χαρακτηριστικό j.

Μια άλλη δημοφιλής προσέγγιση στο πρόβλημα της ομαδοποίησης κατηγορικών δεδομένων, που είναι παρόμοια με την παραπάνω, συνίσταται στην μετατροπή των κατηγορικών χαρακτηριστικών σε μια μακριά λίστα δυαδικών (binary) χαρακτηριστικών και στην εφαρμογή μετέπειτα συγκεκριμένων μέτρων ανομοιότητας δυαδικών δεδομένων. Παραδείγματος χάριν το χαρακτηριστικό Επάγγελμα που έχει πεδίο ορισμού το D = {τεχνικός, εργάτης, …} μπορεί να μετασχηματιστεί στην ακόλουθη λίστα χαρακτηριστικών {Επάγγελμα$_{τεχνικός}$, Επάγγελμα$_{εργάτης}$,…} με κάθε ένα χαρακτηριστικό να ορίζεται στο πεδίο D_i={0,1}. Τώρα μπορεί να εφαρμοστεί το μέτρο επικάλυψης. Ας σημειώσουμε εδώ ότι για δυαδικά χαρακτηριστικά το μέτρο ανομοιότητας περιορίζεται στο γνωστό μέτρο του Hamming, στο οποίο αναφερθήκαμε πιο αναλυτικά παραπάνω. Υπάρχουν αρκετά μέτρα ανομοιότητας στην βιβλιογραφία που ορίζονται σε δυαδικά χαρακτηριστικά [144].

Μια διαφορετική προσέγγιση είναι να προϋπολογιστούν οι εννοιολογικές αποστάσεις μεταξύ των τιμών των χαρακτηριστικών, αριθμητικά. Η πιο τετριμμένη περίπτωση είναι αυτή όπου ο πίνακας των αποστάσεων δίνεται κατ' ευθείαν από τον χρήστη. Ένας τέτοιος πίνακας εγγύτητας μπορεί να κατασκευαστεί ζητώντας από τον χρήστη να κρίνει την εγγύτητα διάφορων τιμών. Υπάρχουν διάφοροι τρόποι για να εξαχθούν αυτές οι κρίσεις [40]. Επιπλέον στο [25] προτείνεται μια αυτόματη μέθοδος εξαγωγής αυτών των πινάκων ανομοιότητας από τα δεδομένα. Μια άλλη περίπτωση της προσέγγισης προϋπολογισμού των αποστάσεων είναι η μέθοδος του multidimensional scaling [30, 78]. Η μέθοδος του Multidimensional Scaling μετατρέπει το πρόβλημα σε ένα πρόβλημα μικρότερης διάστασης συνήθως στον ευκλείδειο χώρο. Οι αποστάσεις μεταξύ σημείων σε αυτόν τον χώρο αντιστοιχούν στις ανομοιότητες οι οποίες ορίζονται από πίνακες ανομοιότητας που δίνονται από τους χρήστες. Υπάρχουν ακόμα και περιπτώσεις που η προσέγγιση του προϋπολογισμού αναφέρεται μόνο σε αντικείμενα συγκεκριμένου τύπου, όπως παραδείγματος χάριν τα συμβολικά αντικείμενα [91]. Τέλος υπάρχουν και πιο εξεζητημένες προσεγγίσεις όπου πραγματοποιείται ένας έμμεσος προϋπολογισμός των ανομοιοτήτων ο οποίος βασίζεται στη δομή και στην γνώση πάνω στο πεδίο των δεδομένων, όπως για παράδειγμα [41, 46, 73]. Τέτοια μέτρα ομοιότητας χρησιμοποιούνται στο πεδίο που ονομάζεται εννοιολογική ομαδοποίηση (conceptual clustering) [97]. Η εννοιολογική ομαδοποίηση αφορά το πρόβλημα της διάκρισης των κλάσεων στις οποίες αντικείμενα χωρίς ετικέτα κλάσης θα μπορούσαν να ομαδοποιηθούν. Μια εξαγόμενη ομάδα χαρακτηρίζεται από εκτατική (extesional) περιγραφή, δηλαδή τα αντικείμενα τις ομάδας που υπάρχουν στην ομάδα, και μια εννοιολογική (intensional, conceptual) περιγραφή. Επιπλέον οι παραγόμενες ομάδες ταξινομούνται σε μία ιεραρχία ανάλογα με την γενικότητα ή ειδικότητα κάθε ομάδας. Για παράδειγμα στο [73] η ανομοιότητα μεταξύ δυο διαφορετικών τιμών ενός χαρακτηριστικού βασίζεται στην δομή του πεδίου ορισμού του χαρακτηριστικού το οποίο συνήθως αναπαριστάται με μια δομή δέντρου. Για την ακρίβεια υπάρχει ένας βαθμός γενικότητας (degree of generality) $g(a_i)$ που ορίζεται για κάθε τιμή a_i κάθε χαρακτηριστικού i. Τότε αν a_i και b_i είναι δυο διαφορετικές τιμές του χαρακτηριστικού i τότε η ανομοιότητα ορίζεται ως εξής:

$$\delta(a_i, b_i, c_i) = 0.5 \frac{g(c_i)}{2} \quad (5.3),$$

όπου c_i είναι η λιγότερο γενική γενίκευση (least general generalization) των a_i και b_i βάση της δομής του πεδίου του χαρακτηριστικού i. Τότε αν A και B είναι δυο

κατηγορικά αντικείμενα που περιγράφονται με m χαρακτηριστικά το κάθε ένα, τότε το μέτρο ανομοιότητας ορίζεται ως εξής:

$$d(A,B) = \sum_{i=1}^{m} \frac{\delta(a_i, b_i, c_i)}{1 - \delta(a_i, b_i, c_i)} \quad (5.4).$$

5.3 Το προτεινόμενο μέτρο ανομοιότητας

5.3.1 Εισαγωγή

Αφού περιγράψαμε σχετικά γρήγορα ποιες είναι οι κυριότερες δουλειές που έχουν γίνει στον τομέα της ομαδοποίησης κατηγορικών δεδομένων, θα παρουσιάσουμε το προτεινόμενο μέτρο ανομοιότητας.

Θα παρουσιάσουμε παρακάτω ένα μέτρο ανομοιότητας το οποίο βασίζεται σε οντολογίες ορισμένες από τον χρήστη και αναπαριστούν τη σχετική με το πεδίο γνώση. Η ανομοιότητα ορίζεται με μια διαδικασία κοινή για όλες τις οντολογίες η οποία όμως λαμβάνει υπ' όψιν της την δομή κάθε οντολογίας ξεχωριστά. Το μέτρο ανομοιότητας που θα παρουσιάσουμε μπορεί να ενσωματωθεί σε διάφορες παραλλαγές του k – means αλγορίθμου ομαδοποίησης και να τις μετασχηματίσει σε αλγορίθμους ομαδοποίησης κατηγορικών δεδομένων. Θα αποδείξουμε εμπειρικά ότι το προτεινόμενο μέτρο ανομοιότητας είναι ακριβές σε σύγκριση με άλλα είδη υπάρχοντα και ευρέως χρησιμοποιούμενα μέτρα ανομοιότητας.

Στην συνέχεια θα περιγράψουμε κατ' αρχάς το προτεινόμενο μέτρο ανομοιότητας και θα δούμε πως μπορούμε να μετασχηματίσουμε τον k – means καθώς και κάποιες παραλλαγές του, με σκοπό να μπορούν να ομαδοποιούν και κατηγορικά δεδομένα. Θα δούμε επίσης πως το προτεινόμενο μέτρο ανομοιότητας μπορεί να χρησιμοποιηθεί για πιο πολύπλοκους υπολογισμούς, όπως αυτούς του αλγορίθμου k – windows. Τέλος θα εξετάσουμε και θα παρουσιάσουμε την ακρίβεια του μέτρου ανομοιότητας παρουσιάζοντας τα αποτελέσματα τις σύγκρισής του με είδη γνωστά μέτρα ανομοιότητας. Μετά θα συζητήσουμε κατά πόσο είναι δυνατόν να χρησιμοποιηθεί και για αριθμητικά δεδομένα.

5.3.2 Περιγραφή

Η βασική ιδέα πίσω από το προτεινόμενο μέτρο ανομοιότητας είναι η χρησιμοποίηση μίας οντολογίας με την μορφή ενός δέντρου, για κάθε διαφορετικό χαρακτηριστικό. Κάθε τέτοια οντολογία θα περιγράφει με έναν ιεραρχικό τρόπο τις έννοιες του πεδίου του χαρακτηριστικού. Το γεγονός αυτό, αφ' ενός δίνει μεγάλη σημασία στην εννοιολογική δομή του πεδίου του κάθε χαρακτηριστικού αφ' εταίρου δίνει την δυνατότητα στον χρήστη να εισάγει την δικιά του διαισθητική γνώση πάνω στο πεδίο. Αυτό το τελευταίο είναι αρκετά σημαντικό αν λάβουμε υπ' όψιν κατ' αρχήν ότι κατά ένα μεγάλο μέρος η εμπειρία κάποιου ανθρώπου σε έναν τομέα είναι διαισθητική και πολύ δύσκολα μεταφράζεται με τη βοήθεια μαθηματικών και κατά δεύτερο λόγο ο τρόπος σκέψης, μάθησης και γενικά αντίληψης του περιβάλλοντα χώρου ενός ανθρώπου, γίνεται τις πιο πολλές φορές με έναν ιεραρχικό τρόπο.

Έτσι ο χρήστης, όπως είπαμε και παραπάνω δίνει μια οντολογία για κάθε χαρακτηριστικό, η οποία αφορά τις σχέσεις μεταξύ των τιμών του χαρακτηριστικού αυτού. Μια οντολογία ενός χαρακτηριστικού a_i είναι ένας κατευθυνόμενος ακυκλικός γράφος (directed acyclic graph), ένα δέντρο δηλαδή $T_{a_i} = (N, E)$, όπου N είναι το σύνολο των κόμβων και E είναι το σύνολο των ακμών του, έτσι ώστε κάθε ακμή $e_i \in E$ να είναι ένα διατεταγμένο ζεύγος $(n_k, n_l) \in E$ με $n_k, n_l \in N$. Κάθε κόμβος n_i παριστάνει είτε μια τιμή $a_{i_j} \in dom(a_i)$ του χαρακτηριστικού i είτε μια γενίκευση μιας τιμής του χαρακτηριστικού i η οποία να μην ανήκει στο πεδίο ορισμού του a_i. Δηλαδή ισχύει ότι $N \supseteq dom(a_i)$. Αυτό σημαίνει ότι υπάρχει περίπτωση κάποιοι κόμβοι της οντολογίας να μην ανήκουν στο πεδίο ορισμού του χαρακτηριστικού. Με αυτόν τον τρόπο ο χρήστης που θα φτιάξει την οντολογία ενός χαρακτηριστικού, έχει την ελευθερία να παρεμβάλει κόμβους και γενικεύσεις οι οποίες δεν περιέχονται στο πεδίο ορισμού του χαρακτηριστικού. Έτσι ο έμπειρος στο πεδίο γνώσης του χρήστης μπορεί να δημιουργήσει μια πλουσιότερη και πιο ακριβή αναπαράσταση του πεδίου, από ότι θα μπορούσε να δημιουργήσει, αν είχε την δυνατότητα να χρησιμοποιήσει μόνο κόμβους, που να αντιστοιχούν στο πεδίο ορισμού του συγκεκριμένου χαρακτηριστικού.

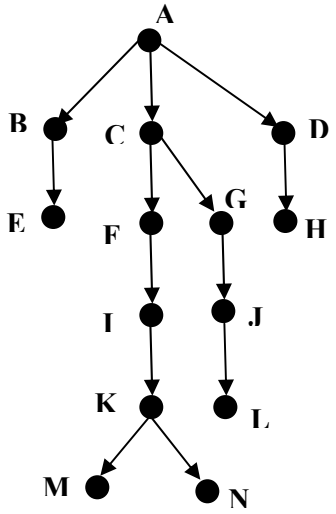
Ας ορίσουμε κατ' αρχάς την ανομοιότητα μεταξύ δυο τιμών ενός χαρακτηριστικού και ύστερα ας γενικεύσουμε την ανομοιότητα στην περίπτωση που συγκρίνουμε δύο αντικείμενα που περιγράφονται με m χαρακτηριστικά το καθένα.

Η ανομοιότητα μεταξύ των τιμών X και Y του ενός χαρακτηριστικού δίνεται από τον τύπο:

$$d(X,Y) = \frac{1}{fl(X,Y)+1} \cdot Average\left(\frac{l(X) - fl(X,Y)}{\max(P(X))}, \frac{l(Y) - fl(X,Y)}{\max(P(Y))} \right) \cdot \frac{P(X,Y)}{\max(P(X)) + \max(P(Y))} \quad (5.5)$$

όπου τα X και Y παριστάνουν δυο οποιουσδήποτε κόμβους της οντολογίας του συγκεκριμένου χαρακτηριστικού (τιμές του χαρακτηριστικού). Με fl(X,Y) συμβολίζουμε το επίπεδο (level) του κοινού προγόνου (πατέρα) των κόμβων X και Y. Με l(X) συμβολίζουμε το επίπεδο (δηλαδή το βάθος) του κόμβου X στην δενδρική δομή. Με max(p(X)) συμβολίζουμε το μήκος του μέγιστου κατευθυνόμενου μονοπατιού που ξεκινάει από τη ρίζα και περιέχει τον κόμβο X. Με p(X,Y) συμβολίζουμε το μήκος του κατευθυνόμενου μονοπατιού που ενώνει τα X και Y. Αν δεν υπάρχει τέτοιο μονοπάτι τότε θέτουμε p(X,Y) = p(X, fl(X,Y)) + p(Y, fl(X,Y)).

Ας εξετάσουμε τώρα όρο προς όρο τον παραπάνω τύπο προσπαθώντας να εξηγήσουμε την αναγκαιότητά τους.



Εικόνα 5.1: Ένα παράδειγμα μιας οντολογίας

Η απόσταση που προτείνουμε, αποτελείται ουσιαστικά από τρεις διαφορετικούς όρους οι οποίοι πολλαπλασιάζονται για να μας δώσουν την τελική απόσταση μεταξύ δυο τιμών κάποιου χαρακτηριστικού. Κάθε όρος της απόστασης συνεισφέρει με κάποιον ιδιαίτερο τρόπο στο να μειώσει ή να ενισχύσει το τελικό αποτέλεσμα.

Ο κύριος όρος αυτού του γινομένου είναι ο $\frac{P(X,Y)}{\max(P(X))+\max(P(Y))}$, ο οποίος μετράει ουσιαστικά την απόσταση, το μήκος του μονοπατιού μεταξύ των κόμβων X και Y στην οντολογία. Το μήκος αυτό δεν συνεισφέρει με την πραγματική του τιμή στην απόσταση των τιμών, αλλά με μια κανονικοποιημένη μορφή του. Το μήκος του μονοπατιού διαιρείται με το άθροισμα των μέγιστων μονοπατιών που ξεκινάνε από τη ρίζα και περιέχουν το X και το Y αντίστοιχα. Ο όρος αυτός παίρνει πάντα τιμές στο διάστημα (0,1].

Με αυτήν την κανονικοποίηση αποφεύγουμε το πρόβλημα της άνισης συνεισφοράς μικρών και μεγάλων ιεραρχιών. Μπορούμε να θεωρήσουμε ως μια ολόκληρη ιεραρχία κάθε μονοπάτι που ξεκινάει από τη ρίζα και καταλήγει σε κάποιο φύλλο. Διαισθητικά κάθε μονοπάτι από την ρίζα ως κάποιο φύλλο είναι ένα ολοκληρωμένο ιεραρχημένο κομμάτι του πεδίου του χαρακτηριστικού μας. Αυτό σημαίνει ότι ανεξάρτητα από το αν κάποιο μονοπάτι από τη ρίζα στα φύλλα είναι κοντό ή μακρύ θα πρέπει η ρίζα από τα φύλλα να απέχει σχεδόν την ίδια απόσταση αφού κάθε φορά πρόκειται για μια ολοκληρωμένη ιεραρχία. Ας δούμε για παράδειγμα τα μονοπάτια από το A στο E και από το A στο M. Η απόσταση από το A στο E θα πρέπει να είναι περίπου η ίδια με την απόσταση από το A στο M. Στην πρώτη περίπτωση η ιεραρχία δεν έχει μεγάλη διακρητότητα, με αποτέλεσμα πρώτον να μην υπάρχουν πολλοί κόμβοι σε αυτή και δεύτερον οι αποστάσεις του A με το B και του B από το E να είναι κατά πολύ μεγαλύτερες από τις αντίστοιχες αποστάσεις μεταξύ των γειτονικών κόμβων του μονοπατιού από το A στο M (παραδείγματος χάριν AC, CF …).

Έτσι ο όρος $\frac{P(X,Y)}{\max(P(X))+\max(P(Y))}$ αφ' ενός υπολογίζει το μονοπάτι μεταξύ των X και Y ως προς τα μέγιστα μονοπάτια που περνάνε από τα X και Y αφετέρου όμως υποστηρίζει και το γεγονός ότι όσο τα μέγιστα μονοπάτια που περνούν από τα X και Y μεγαλώνουν τότε η απόσταση μεταξύ X και Y μικραίνει. Έτσι η απόσταση μεταξύ των κόμβων C και F θα πρέπει να είναι μικρότερη από την απόσταση μεταξύ των κόμβων B και E αφού το πρώτο μέγιστο μονοπάτι είναι μεγαλύτερο από το δεύτερο. Ένα άλλο παράδειγμα που μας δείχνει λίγο πιο ξεκάθαρα τα όσα είπαμε παραπάνω, αναφέρεται στην ιεραρχία του χαρακτηριστικού «Επάγγελμα» που θα χρησιμοποιηθεί αργότερα στην εφαρμογή του αλγόριθμου και που φαίνεται στην Εικόνα 5.2. η απόσταση μεταξύ «Γενικών Γιατρών» και «Καρδιολόγων» θα πρέπει, σύμφωνα με αυτά που είπαμε παραπάνω να είναι μικρότερη από την απόσταση μεταξύ των «Διαφόρων» και «Συνταξιούχων».

Οι άλλοι όροι τώρα του γινομένου ενισχύουν ή αποδυναμώνουν τον βασικό αυτόν όρο. Θα μπορούσε να πει κανείς ότι δρουν ως παράμετροι που ρυθμίζουν την απόσταση μεταξύ των κόμβων του δέντρου.

Έτσι ο όρος $\frac{1}{fl(X,Y)+1}$ γίνεται μικρότερος όσο μεγαλύτερο γίνεται το fl(X,Y). Θα πρέπει κατ' αρχάς να αναφέρουμε ότι ο όρος αυτός παίρνει τιμές στο διάστημα (0,1]. Έτσι όσο βαθύτερα βρίσκετε ο κοινός πρόγονος τόσο πιο μεγάλο θα είναι το fl(X,Y) και άρα τόσο πιο μικρός θα γίνεται ο όρος $\frac{1}{fl(X,Y)+1}$ και συνεπώς και η απόσταση μεταξύ των X και Y. Ο κοινός πρόγονος δυο κόμβων ουσιαστικά είναι η ελάχιστη γενική γενίκευση (least general generalization) των δυο αυτών όρων. Έτσι όσο πιο βαθιά είναι η ελάχιστη γενική γενίκευση δυο όρων τόσο πιο σχετικές μεταξύ τους θα πρέπει να είναι οι έννοιες αυτές, αφού όσο προχωράμε πιο κάτω στην ιεραρχία τόσο πιο ειδικές γίνονται οι έννοιες που αναπαρίστανται και συνεπώς τόσο πιο μικρή θα πρέπει να είναι και η απόσταση μεταξύ τους. Ας δούμε όμως ένα παράδειγμα βασισμένο στην οντολογία της Εικόνας 5.1. Σύμφωνα με τα παραπάνω η απόσταση μεταξύ των κόμβων M και N θα πρέπει να είναι μικρότερη από την απόσταση μεταξύ των κόμβων F και G. Και αυτό γιατί ο κοινός πατέρας K των M και N είναι πολύ πιο ειδικός, αφού βρίσκεται βαθιά μέσα στην ιεραρχία απ' ότι ο κοινός πατέρας τωνF και G ο οποίος είναι πολύ πιο γενικός. Ας επανέλθουμε όμως στο παράδειγμα τις οντολογίας του χαρακτηριστικού «Επάγγελμα». Εδώ η απόσταση μεταξύ της τιμής «τεχνικοί – εργάτες γεωργοκτηνοτροφικών μονάδων» και «κτηνοτρόφοι – ψαράδες – κυνηγοί» θα πρέπει να είναι μικρότερη από την απόσταση μεταξύ των τιμών «πρωτογενή τομέα» και «κατασκευαστικού τομέα», αφού οι πρώτες έννοιες είναι πιο ειδικές αρά και πιο κοντά από τις δεύτερες.

Τέλος ο όρος $Average\left(\frac{l(X)-fl(X,Y)}{\max(P(X))},\frac{l(Y)-fl(X,Y)}{\max(P(Y))}\right)$ υπολογίζει την μέση απόσταση των κόμβων των τιμών X, Y του χαρακτηριστικού από τον κοινό τους πρόγονο ως προς τα μέγιστα μονοπάτια max(p(X)) και max(p(Y)). Όσο πιο κοντά είναι οι έννοιες στον κοινό τους πρόγονο τόσο πιο μικρός γίνεται ο όρος αυτός με συνέπεια να μικραίνει και η απόσταση μεταξύ των εννοιών. Έτσι όσο πιο κοντά στον κοινό πατέρα, στην ελάχιστη γενική γενίκευση δηλαδή, βρίσκονται οι έννοιες τόσο

πιο σχετικές θα είναι με τον κοινό πρόγονο άρα τόσο πιο σχετικές θα είναι και μεταξύ τους πράγμα που σημαίνει ότι η απόστασή τους θα πρέπει να είναι μικρότερη. Έτσι στην Εικόνα 5.1 βλέπουμε ότι η απόσταση μεταξύ των κόμβων (τιμών του χαρακτηριστικού) F και G θα πρέπει να είναι μικρότερη, οι έννοιες αυτές θα πρέπει να είναι πιο σχετικές, απ' ότι οι έννοιες που αναπαριστώνται στην δεντρική δομή του χαρακτηριστικού με του ς κόμβους K και L αφού αυτοί οι δυο απέχουν περισσότερο από τον κοινό τους πρόγονο απ' ότι οι άλλες δυο. Στο παράδειγμα της οντολογίας «Επάγγελμα» η απόσταση μεταξύ του κόμβου «Πρωτογενής Τομέας» και του κόμβου «Κατασκευαστικός τομέας» θα πρέπει να είναι μικρότερη από την απόσταση μεταξύ των κόμβων «Αγρότης» και «Αρχιτέκτονας».

Αφού αναλύσαμε έναν προς έναν τους όρους της απόστασης και είδαμε την χρησιμότητα του καθενός, στον υπολογισμό της απόστασης μεταξύ τιμών ενός χαρακτηριστικού ας δούμε τώρα πως θα υπολογίσουμε την απόσταση μεταξύ δυο περιπτώσεων της βάσης μας A και B οι οποίες περιγράφονται με m χαρακτηριστικά.

Για να υπολογίσουμε την συνολική απόσταση των περιπτώσεων A και B το μόνο που έχουμε να κάνουμε είναι να υπολογίσουμε τις επιμέρους αποστάσεις των m χαρακτηριστικών, που βασίζονται σε m διαφορετικές οντολογίες, ξεχωριστά και μετά να προσθέσουμε τις επιμέρους αποστάσεις. Έτσι έχουμε:

$$D(A,B)=\sum_{i=1}^{m} d(X_i,Y_i)=$$

$$=\sum_{i=1}^{m}\frac{1}{fl(X_i,Y_i)+1}\cdot Average\left(\frac{l(X_i)-fl(X_i,Y_i)}{\max(P(X_i))},\frac{l(Y_i)-fl(X_i,Y_i)}{\max(P(Y_i))}\right)\cdot\frac{P(X_i,Y_i)}{\max(P(X_i))+\max(P(Y_i))}$$

5.3.3 Ιδιότητες της απόστασης

Ας δούμε τώρα κάποιες από τις ιδιότητες της απόστασης για να μπορέσουμε να έχουμε μια καλύτερη εικόνα της απόστασης.

1. Κάθε μονοπάτι από την ρίζα μέχρι κάποιο φύλλο έχει σχεδόν την ίδια απόσταση. Αυτό γιατί όπως εξηγήσαμε και παραπάνω κάθε τέτοιο μονοπάτι μπορεί να θεωρηθεί ως μια ολοκληρωμένη ιεραρχία και κάθε ολοκληρωμένη ιεραρχία θα πρέπει να έχει σχεδόν το ίδιο μήκος.

2. Όσο προχωράμε από την ρίζα προς τα φύλλα πάνω σε ένα μονοπάτι οι αποστάσεις μεταξύ των διαδοχικών κόμβων του μονοπατιού αυτού βρίσκονται όλο και πιο κοντά. Όσο προχωράμε προς τα κάτω σε ένα μονοπάτι δηλαδή οι αποστάσεις των κόμβων πυκνώνουν. Στο παράδειγμα της Εικόνας 5.1 παραδείγματος χάριν βλέπουμε ότι d(A, C) > d(C, F) > d(F, I) > d(I, K) > d(K, M). Αυτό οφείλεται κυρίως στη συνεισφορά του πρώτου όρου του γινομένου της απόστασης. Όσο προχωράμε προς τα κάτω ο κοινός πατέρας των κόμβων του μονοπατιού συνεχώς βυθίζεται στην ιεραρχία με αποτέλεσμα ο όρος $\frac{1}{fl(X,Y)+1}$ να μικραίνει. Αν απουσίαζε ο όρος αυτός τότε όλες οι αποστάσεις των διαδοχικών κόμβων του μονοπατιού θα ήταν ίσες.

3. Όσο πιο βαθιά στην ιεραρχία βρίσκετε ο κοινός πατέρας των δύο κόμβων τόσο πιο μικρή θα είναι η απόσταση των δυο αυτών κόμβων. Τόσο πιο κοντά θα είναι δηλαδή αυτοί οι δυο κόμβοι. Αυτήν την ιδιότητα την αναλύσαμε εκτενώς παραπάνω.

4. Όσο πιο κοντά στον κοινό πρόγονο βρίσκονται οι δύο κόμβοι τόσο πιο κοντά θα βρίσκονται μεταξύ τους, τόσο πιο μικρή θα είναι δηλαδή η απόσταση. Και αυτήν την ιδιότητα την έχουμε αναλύσει παραπάνω.

5. Όσο πιο μακριά στο δέντρο βρίσκονται οι δυο κόμβοι των εννοιών τόσο πιο μεγάλη θα είναι η απόστασή μεταξύ τους. Αυτό οφείλεται κυρίως στην συνεισφορά του τρίτου όρου του γινομένου της απόστασης: $\frac{P(X,Y)}{\max(P(X))+\max(P(Y))}$.

5.4 Αλγόριθμος ομαδοποίησης κατηγορικών δεδομένων

Έχουμε περιγράψει ήδη την απόσταση μεταξύ κατηγορικών δεδομένων που προτείνουμε. Όπως έχουμε αναφέρει και παραπάνω ο αλγόριθμος ομαδοποίησης κατηγορικών δεδομένων δεν είναι τίποτα άλλο από μια επέκταση του γνωστού αλγόριθμου k – means, έτσι ώστε να μπορεί να χειριστεί κατηγορικά δεδομένα. Η επέκταση αυτή θα γίνει αντικαθιστώντας την ευκλείδεια απόσταση του k – means με την απόσταση που προτείναμε λίγο παραπάνω. Η μετατροπή όμως αυτή δεν είναι αρκετή για να μετασχηματίσουμε τον k – means έτσι ώστε να ομαδοποιεί κατηγορικά δεδομένα. Θα πρέπει επίσης να επεκτείνουνε την έννοια του μέσου μιας ομάδας κατηγορικών αντικειμένων όπως επίσης θα πρέπει να περιγράψουμε και μια αποδοτική τεχνική επιλογής των αρχικών μέσων έτσι ώστε να μπορέσει να ξεκινήσει ο αλγόριθμος. Η αποδοτική αυτή επιλογή των αρχικών μέσων είναι αρκετά σημαντική αφού όπως ξέρουμε ο αλγόριθμος k – means από την φύση του είναι αφ' ενός ευαίσθητος στην επιλογή των αρχικών μέσων των ομάδων και αφ' εταίρου τις πιο πολλές φορές συγκλίνει σε ένα τοπικό βέλτιστο της συνάρτησης ποιότητας.

Ας περιγράψουμε όμως πρώτα, διαισθητικά, τον ορισμό της έννοιας του μέσου μιας ομάδας κατηγορικών χαρακτηριστικών. Θα ορίσουμε την έννοια αυτή χρησιμοποιώντας την δεντρική δομή των χαρακτηριστικών. Θα ορίσουμε την έννοια του μέσου για κάθε χαρακτηριστικό ξεχωριστά και συνεπώς ο συνολικός μέσος θα αποτελείται από τους επιμέρους μέσους των χαρακτηριστικών.

Για να υπολογίσουμε λοιπόν τον μέσο ενός χαρακτηριστικού μιας ομάδας υπολογίζουμε τον μέσο όρο όλων των επιπέδων (levels) των κόμβων του χαρακτηριστικού που συμμετέχουν στην ομάδα. Αυτός ο μέσος όρος, στρογγυλεμένος θα αποτελέσει το επίπεδο στο οποίο θα βρίσκεται ο μέσος όρος. Έπειτα υπολογίζουμε πιο από τα μονοπάτια, από την ρίζα μέχρι κάποιο φύλλο, που περιέχουν κόμβους που συμμετέχουν στην ομάδα διαθέτει το μεγαλύτερο πλήθος κόμβων απ' αυτούς που περιέχονται στην ομάδα. Ουσιαστικά πρέπει να υπολογίσουμε το μονοπάτι με την μέγιστη συνεισφορά στην ομάδα. Αφού λοιπόν το έχουμε υπολογίσει και αυτό, διαλέγουμε ως μέσον της ομάδας εκείνον τον κόμβο που ανήκει στο μονοπάτι με την μέγιστη συνεισφορά σε κόμβους στην ομάδα και που βρίσκεται στο μέσο επίπεδο των κόμβων που αποτελούν την ομάδα.

Υπολογίζοντας τον μέσο όρο των επιπέδων των κόμβων που αποτελούν την ομάδα, αλλά και του μέγιστου σε πλήθος τιμών μονοπατιού, ουσιαστικά λαμβάνουμε υπ' όψιν το πλήθος της εμφάνισης κάθε τιμής του χαρακτηριστικού. Η απόσταση αυτή δηλαδή δεν αδιαφορεί, για το πόσες φορές έχει εμφανιστεί κάποια τιμή στην ομάδα.

Ας δώσουμε τώρα έναν τυπικό ορισμό για τον μέσο (medoid) μια ομάδας κατηγορικών δεδομένων:

Έστω $O = \{o_r \mid o_r = (a_{1_x},...,a_{d_y}), 1 \le r \le t\}$ ένα σύνολο κατηγορικών αντικειμένων και $T_{a_1}, T_{a_2},..., T_{a_{d1}}$ οι αντίστοιχες οντολογίες. Έστω $l(a_{i_j})$ το επίπεδο (level) του κόμβου που αναπαριστά την τιμή του χαρακτηριστικού a_{i_j} στην οντολογία $T_{a_i} = (N_{a_i}, E_{a_i})$ και έστω $n_{a_{i_j}}$ το πλήθος των αντικειμένων εισόδου που έχουν την τιμή a_{i_j} για το χαρακτηριστικό a_{i_j}. Έστω επίσης $\max\_p(a_i) = (a_{iq_1}, a_{iq_2},..., a_{iq_m})$ με $a_{iq_1}, a_{iq_2},..., a_{iq_m} \in N_{a_i}$ το μονοπάτι της οντολογίας T_{a_i} έτσι ώστε $\sum_{j=1}^{m} n_{a_{iqj}}$ να είναι μέγιστο πάνω σε όλα τα μονοπάτια του T_{a_i}. Τότε ο μέσος (medoid) της ομάδας Ο θα είναι $Medoid(O) = (a_{1_{ml1}}, a_{1_{ml2}},..., a_{1_{mld}})$ με $a_{1_{mli}}$ να είναι η τιμή του χαρακτηριστικού του μονοπατιού $\max\_p(a_i)$ το οποίο όμως είναι στο επίπεδο $\text{int}\left(\frac{\sum_{q=1}^{t} l(a_{i_q})}{t}\right)$.

5.4.1 Περιγραφή του αλγορίθμου ομαδοποίησης

Ο αλγόριθμος ομαδοποίησης που θα χρησιμοποιήσουμε βασίζεται όπως έχουμε ήδη αναφέρει στον πολύ δημοφιλή αλγόριθμο k – means. Έχουμε αναφέρει και πιο πάνω όταν περιγράφαμε αναλυτικά τον k – means τα βασικά βήματά του:

1. Επιλογή των k αρχικών κέντρων για τις k ομάδες.
2. Υπολογισμός της ανομοιότητας μεταξύ ενός αντικειμένου και των κέντρων των k ομάδων.
3. Τοποθέτηση το αντικειμένου σε εκείνη την ομάδα της οποίας το κέντρο είναι πιο κοντά στο αντικείμενο αυτό.
4. Ενημέρωση του κέντρου της ομάδας έτσι ώστε να ελαχιστοποιηθεί η ανομοιότητα εντός της ομάδας.

Εκτός από το πρώτο βήμα, όλα τα άλλα βήματα του αλγορίθμου εκτελούνται επαναληπτικά μέχρις ότου ο αλγόριθμος συγκλίνει, μέχρις ότου δηλαδή να μην υπάρχει κάποια σημαντική μετακίνηση αντικειμένων από μια ομάδα σε μια άλλη.

Παραπάνω ορίσαμε την έννοια του μέσου μιας ομάδας (medoid). Αυτή η έννοια είναι θεμελιώδης όπως εξηγήσαμε για την εφαρμογή του μέτρου ανομοιότητας στον k –

means. Αυτό που μας μένει τώρα πλέον να ορίσουμε είναι η διαδικασία αρχικοποίησης, η οποία αποτελεί το πρώτο βήμα του K – means.

Η διαδικασία της αρχικοποίησης είναι ο ορισμός των k αρχικών κέντρων των ομάδων. Ο αρχικός κλασικός k – means ορίζει ως τα k πρώτα κέντρα των ομάδων τα k πρώτα αντικείμενα του συνόλου, των προς ομαδοποίηση δεδομένων. Αυτή η επιλογή είναι μια λύση του προβλήματος της αρχικοποίησης, αλλά αν λάβουμε υπ' όψιν μας πρώτων, την ευαισθησία του k – means από τα αρχικά μέσα και δεύτερον, ότι ο αλγόριθμος αυτός συνήθως συγκλίνει σε ένα τοπικό βέλτιστο και όχι στο ολικό βέλτιστο μας κάνει να αναζητήσουμε μια πιο αποδοτική τεχνική επιλογής των αρχικών μέσων.

Η προσπάθεια που κάνουμε, είναι να επιλέξουμε κατά τέτοιον τρόπο τα αρχικά μέσα έτσι ώστε να βρούμε μια όσο το δυνατόν κοντά στο βέλτιστο ομαδοποίηση. Προσπαθούμε να πάρουμε εκείνα τα k αρχικά μέσα τα οποία εφ' ενός θα είναι στις πιο k πιο πυκνές περιοχές των δεδομένων μας και αφ' εταίρου θα ανήκουν στην βάση των αντικειμένων προς κατάταξη. Αυτή η δεύτερη απαίτηση είναι κάτι που προέκυψε από την εμπειρία διάφορων σετ δεδομένων. Έτσι η εμπειρία έδειξε ότι η ομαδοποίηση ήταν πολύ πιο κοντά στην εννοιολογικά βέλτιστη όταν οι αρχικοί μέσοι διαλέγονταν από τα δεδομένα εισόδου του αλγορίθμου.

Η διαδικασία επιλογής χωρίζεται σε δύο φάσεις. Στην πρώτη φάση επιλέγουμε k τιμές – κόμβους, για κάθε χαρακτηριστικό, από κάθε οντολογία. Ας σημειώσουμε εδώ πως οι τιμές – κόμβοι που θα επιλεγούν σε αυτήν τη φάση δεν χρειάζεται να εμφανίζονται στη βάση των προς ομαδοποίησης δεδομένων. Στην δεύτερη φάση οι διάφοροι συνδυασμοί των επιλεγμένων τιμών συνδυάζονται και επιλέγονται, για τους k αρχικούς μέσους, εκείνα τα αντικείμενα από τα προς κατάταξη δεδομένα τα οποία βρίσκονται πιο κοντά στους k καλύτερους συνδυασμούς των επιλεγμένων, από την πρώτη φάση, τιμών των χαρακτηριστικών. Ας εξετάσουμε όμως πιο αναλυτικά τις δύο αυτές φάσεις.

$1^{η}$ φάση

Για κάθε ένα χαρακτηριστικό ξεχωριστά υπολογίζονται οι μέσοι όροι των επιπέδων (levels) των δεδομένων $I = \left\{ o_r \mid o_r = \left(a_{1_x}, ..., a_{d_y} \right), 1 \le r \le t \right\}$ σε κάθε οντολογία T_{a_i} ως εξής : $\text{int}\left(\frac{\sum_{q=1}^{t} l\left(a_{i_q} \right)}{t} \right)$. Έπειτα, για κάθε χαρακτηριστικό, δηλαδή για κάθε οντολογία ξεχωριστά, υπολογίζονται τα k μέγιστα από την ρίζα στα φύλλα μονοπάτια. Αυτό σημαίνει ότι υπολογίζονται τα k μονοπάτια τα οποία περιέχουν το μεγαλύτερο πλήθος από αντικείμενα της αρχικής βάσης δεδομένων I. Έπειτα για κάθε χαρακτηριστικό επιλέγονται οι k τιμές – κόμβοι οι οποίοι βρίσκονται στα παραπάνω k μέγιστα μονοπάτια και στον μέσο όρο των επιπέδων που έχουμε ήδη υπολογίσει. Στο τέλος της πρώτης φάσης δηλαδή διαθέτουμε k τιμές από κάθε χαρακτηριστικό.

$2^{η}$ φάση

Στην συνέχεια εξετάζονται οι διάφοροι συνδυασμοί αυτών των χαρακτηριστικών. Συνδυάζονται δηλαδή οι k διαφορετικές τιμές που έχουμε από κάθε χαρακτηριστικό έτσι ώστε να προκύψουν αντικείμενα. Από αυτά τα αντικείμενα που σχηματίζονται επιλέγονται τα k που βρίσκονται πιο κοντά σε κάποιο από αντικείμενα του συνόλου I. Τα k αυτά αντικείμενα του συνόλου I θα πρέπει να είναι διαφορετικά μεταξύ τους. Στην περίπτωση που ο αλγόριθμος πάει να επιλέξει κάποιο από τα ήδη επιλεγμένα αντικείμενα της βάσης, αυτό αγνοείται και γίνεται προσπάθεια να βρεθεί το αμέσως επόμενο καλύτερο αντικείμενο. Στο τέλος της δεύτερης αυτής φάσης επιλέγουμε για τα αρχικά κέντρα τα k αντικείμενα από τη βάση δεδομένων που έχουν επιλεγεί.

Ας δούμε όμως λίγο πιο τυπικά τον αλγόριθμο με τον οποίο επιλέγονται τα k αρχικά κέντρα των ομάδων, έτσι ώστε να μπορέσει να ξεκινήσει η ομαδοποίηση. Ας σημειώσουμε εδώ ότι στον αλγόριθμο αρχικοποίησης που παρουσιάζεται παρακάτω δεν εξετάζονται όλοι οι διαφορετικοί συνδυασμοί μεταξύ των χαρακτηριστικών. Γεγονός που κάνει την διαδικασία λιγότερο δαπανηρή όσον αφορά τον υπολογιστικό χρόνο χωρίς μειώνει καθόλου την ακρίβεια του αλγορίθμου αρχικοποίησης.

επιλογή των αρχικών μέσων (medoids)

input $I = \{o_r \mid o_r = (a_{1_x},...,a_{d_y}), 1 \le r \le t\}$

input $comb_1,...,comb_d$ το σύνολο των k τιμών των χαρακτηριστικών (attribute values) για κάθε χαρακτηριστικό

for $i_1 = 1$ **to** k
  if $comb_1[i_1]$ είναι σημειωμένο **then next** i_1
  ...
    for $i_d = 1$ **to** k
      if $comb_d[i_d]$ είναι σημειωμένο **then next** i_d
      find $cm_{i_d} \in I$, το πιο όμοιο αντικείμενο στο $(comb_1[i_1],...,comb_d[i_d])$.
    end for
    choose μέσο (medoid) αυτό για το οποίο $d(cm_{i_d},(comb_1[i_1],...,comb_d[i_d]))$ είναι μέγιστο.
    if $m = (a_{1_x},...,a_{d_y})$ είναι κάποιο μέσο **then**
      delete $(a_{1_x},...,a_{d_y})$ από $comb_1,...,comb_d$
  …
end for

Αλγόριθμος 5.1: Ο αλγόριθμος αρχικοποίησης των κέντρων

Ας δώσουμε ένα παράδειγμα για το πώς δουλεύει ο αλγόριθμος αυτός. Το παράδειγμά μας θα βασιστεί στις οντολογίες που θα χρησιμοποιήσουμε για την σύγκριση του μέτρου ανομοιότητας που προτείνεται με άλλα γνωστά μέτρα ανομοιότητας, σε επόμενη παράγραφο. Οι δυο αυτές οντολογίες φαίνονται στα

αντίστοιχα σχήματα (Εικόνα 5.3 και Εικόνα 5.4) και είναι οι οντολογίες με ρίζες τα: «Ελλάδα» και «Επαγγέλματα» αντίστοιχα.

Όπως είπαμε και παραπάνω στην πρώτη φάση του αλγορίθμου επιλέγονται k (στην περίπτωσή μας 3) τιμές για κάθε χαρακτηριστικό. Αυτές οι τιμές είναι οι κόμβοι που σημειώνονται με κόκκινο στα παρακάτω σχήματα και είναι οι εξείς:

Οντολογία «Ελλάδα»	«Αττική»	«Μακεδονία»	«Θεσσαλία»
Οντολογία «Επαγγέλματα»	«Οικιακά»	«Ιδιωτικά Νοσοκομεία»	«Κατασκευών»

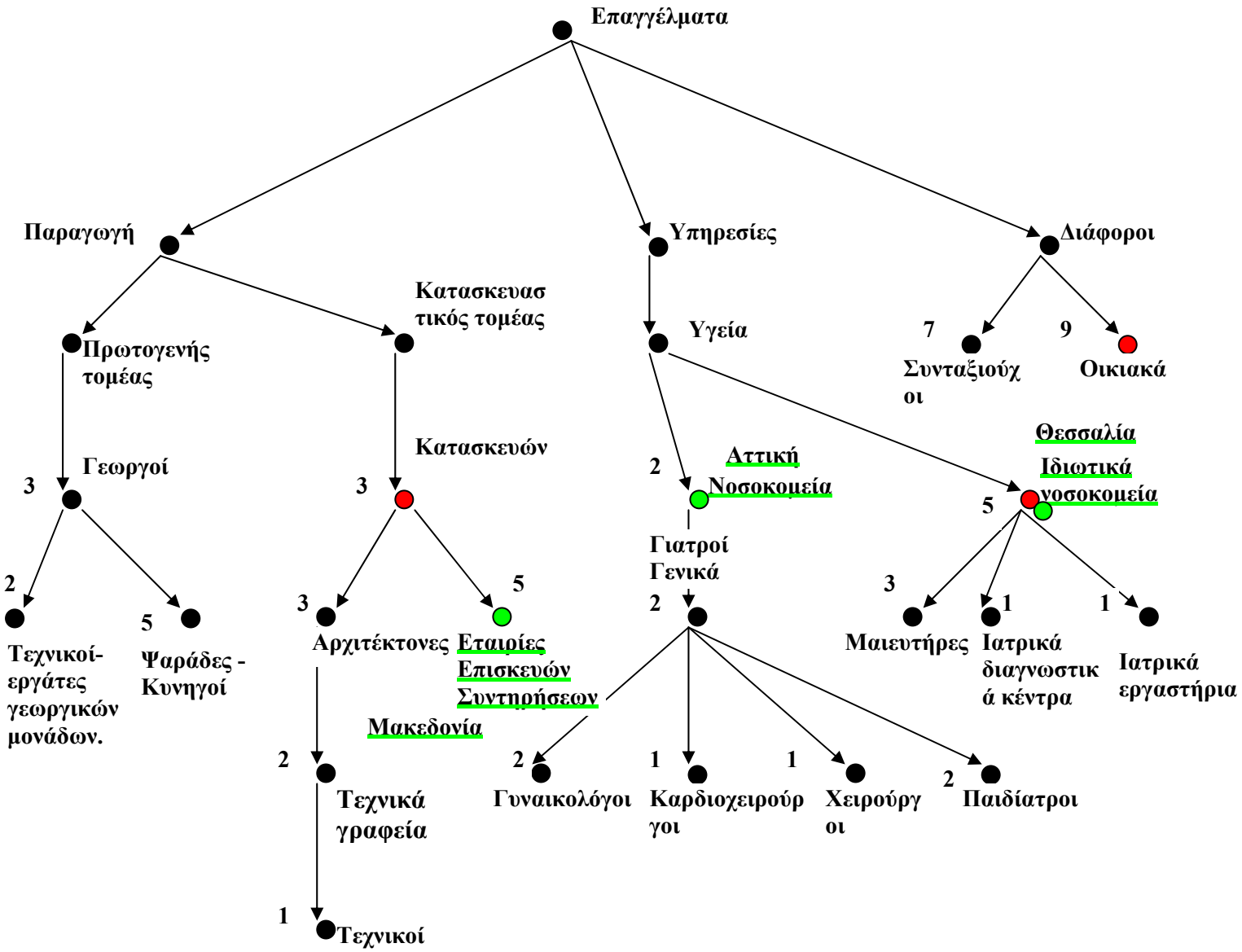
Στη συνέχεια υπολογίζονται οι συνδυασμοί αυτών των χαρακτηριστικών και επιλέγονται εκείνοι οι τρεις συνδυασμοί που είναι περισσότερο όμοιοι με κάποια (διαφορετικά μεταξύ τους) αντικείμενα της αρχικής προς ομαδοποίηση βάσης δεδομένων. Έτσι οι συνδυασμοί που επιλέγονται είναι οι:

Οντολογία «Ελλάδα»	«Αττική» «Οικιακά»	«Μακεδονία» «Κατασκευών»	«Θεσσαλία» «Ιδιωτικά Νοσοκομεία»
Οντολογία «Επαγγέλματα»			

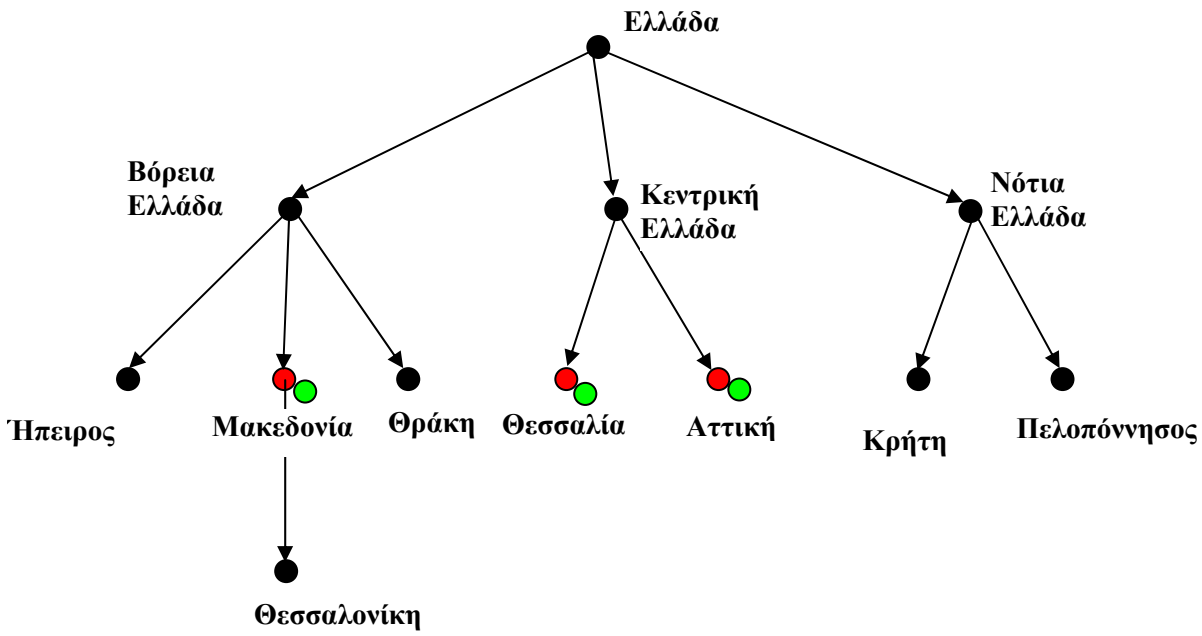
Οι οποίοι βρίσκονται τελικά πιο κοντά στα αντικείμενα της βάσης μας τα οποία σημειώνονται με πράσινο χρώμα στις Εικόνες 5.3 και 5.4.

Οντολογία «Ελλάδα»	«Αττική» «Νοσοκομεία»	«Μακεδονία» «Εταιρίες επισκευών συντηρήσεων»	«Θεσσαλία» «Ιδιωτικά Νοσοκομεία»
Οντολογία «Επαγγέλματα»			

Έτσι λοιπόν αφού αποσαφηνίσαμε την διαδικασία επιλογής των αρχικών κέντρων μπορούμε να υλοποιήσουμε έναν k – means like αλγόριθμο. Ας σημειώσουμε μόνο ότι τα κέντρα κάθε ομάδας ενημερώνονται κάθε φορά που γίνεται κάποια αλλαγή σε μια ομάδα και όχι στο τέλος ενός πλήρους βήματος του k – means.



Εικόνα 5.2: Οι κόκκινοι κόμβοι παριστάνουν τις επιλογές της πρώτης φάσης του αλγόριθμου αρχικοποίησης, ενώ οι πράσινοι κόμβοι παρουσιάζουν τις επιλογές της δεύτερης φάσης και συνεπώς και τους αρχικούς κόμβους του αλγόριθμου.



Εικόνα 5.3: Η οντολογία «Ελλάδα». Οι κόκκινοι κόμβοι επιλέχθηκαν κατά την πρώτη φάση του αλγόριθμου αρχικοποίησης ενώ οι πράσινοι κόμβοι κατά την δεύτερη φάση.

5.4.2 Επέκταση του αλγόριθμου k – windows ώστε να μπορεί να δέχεται κατηγορικά δεδομένα.

Υπάρχουν κάποιες παραλλαγές του k – means αλγορίθμου των οποίων η επέκτασή τους έτσι ώστε να ομαδοποιούν κατηγορικά δεδομένα χρησιμοποιώντας το παραπάνω μέτρο ανομοιότητας δεν είναι τόσο προφανείς. Σε αυτές τις περιπτώσεις θα πρέπει να οριστούν οι υπολογισμοί με βάση την δεντρική δομή ώστε να μετασχηματιστούν αυτοί οι αλγόριθμοι. Φαίνεται πάντως ότι το προτεινόμενο μέτρο ανομοιότητας μπορεί να ενσωματωθεί με μεγάλη αποδοτικότητα στις περισσότερες από τις παραλλαγές του k – means αλγορίθμου.

Ας δώσουμε όμως ένα παράδειγμα θεωρώντας την περίπτωση του k – windows [137] αλγορίθμου, τον οποίο περιγράψαμε αναλυτικά πιο πάνω. Η βασική ιδέα στον k – windows είναι η χρήση ενός παραθύρου για τον ορισμό μιας ομάδας (cluster). Ως παράθυρο ορίζεται ένα ορθογώνιο d – διάστημα (orthogonal d – range) στον d – διάστατο ευκλείδειο χώρο, όπου d είναι το πλήθος των αριθμητικών χαρακτηριστικών των περιπτώσεων. Έτσι κάθε παράθυρο είναι ένα ορθογώνιο d – διάστημα και έχει σταθερό μέγεθος. Κάθε αντικείμενο που βρίσκεται μέσα σε ένα παράθυρο θεωρείται ότι ανήκει στην αντίστοιχη ομάδα. Ας σημειώσουμε εδώ ότι στην αριθμητική μορφή του k – windows τα αντικείμενα που βρίσκονται μέσα σε ένα d – διάστημα μπορούν να βρεθούν σε πολυλογαριθμικό χρόνο χρησιμοποιώντας την τεχνική της ορθογώνιας αναζήτησης διαστήματος (orthogonal range search technique) της υπολογιστικής γεωμετρίας.

Επαναληπτικά, κάθε παράθυρο μετακινείται στον ευκλείδειο χώρο κεντράροντάς το στον μέσο των αντικειμένων που περιέχονται σε αυτό. Αυτό λαμβάνει χώρα μέχρις ότου καμία επιπλέον κίνηση δεν αυξάνει τον αριθμό των αντικειμένων που βρίσκονται εντός του παραθύρου. Μετά από αυτό το βήμα είμαστε έτοιμοι να καθορίσουμε τα μέσα κάθε ομάδας ως τα μέσα των αντίστοιχων παραθύρων. Τα τελευταία δυο βήματα επαναλαμβάνονται εξακολουθητικά μέχρις ότου δεν υπάρχει κανένα d – διάστημα στο οποίο να παρατηρείται κάποια σημαντική αύξηση των αντικειμένων μετά από αυτό το βήμα.

Όπως και να έχει, αφού μόνο περιορισμένος αριθμός αντικειμένων θεωρούνται σε κάθε κίνηση, ο διαμερισμός μπορεί να μην είναι ο βέλτιστος. Για αυτόν τον λόγο η ποιότητα του χωρισμού υπολογίζεται σε μια δεύτερη φάση. Κατ' αρχάς τα παράθυρα μεγεθύνονται για να περιλάβουν όσο το δυνατόν περισσότερα αντικείμενα στην ομάδα. Αυτό γίνεται αναγκάζοντας τα d – διαστήματα να διατηρούν το κέντρο τους κατά την διάρκεια της μεγέθυνσης. Έπειτα υπολογίζεται η σχετική συχνότητα των περιπτώσεων που περιέχονται στην d – περιοχή σε σχέση με το σύνολο όλων των περιπτώσεων. Αν η σχετική συχνότητα είναι μικρή τότε πιθανόν να υπάρχουν κάποιες ομάδες που δεν εντοπίσθηκαν. Σε αυτήν την περίπτωση, η όλη διαδικασία επαναλαμβάνεται από την αρχή.

Παραπάνω παρουσιάσαμε και τον αλγόριθμο z – windows ο οποίος βασίζεται στην παραθυρική τεχνική του k – windows. Το πρόβλημα που επιχειρεί να λύσει ο αλγόριθμος z – windows είναι ο προσδιορισμός του των αριθμού των ομάδων που υπάρχουν μέσα στα δεδομένα, το οποίο παραμένει ένα άλυτο πρόβλημα στο πεδίο της ανάλυσης ομάδων. Ας σημειωθεί εδώ ότι ο δημοφιλής αλγόριθμος ομαδοποίησης k – means καθώς και οι παραλλαγές του απαιτούν από τον χρήστη τον καθορισμό του πλήθους των ομάδων εκ των προτέρων. Η βασική ιδέα του αλγορίθμου z – windows όπως είδαμε και παραπάνω είναι η χρησιμοποίηση ενός αναγκαίου αριθμού αρχικών παραθύρων τα οποία κατά την διάρκεια του αλγορίθμου θα συγχωνευτούν για να δημιουργήσουν τις τελικές ομάδες. Η παραθυρική τεχνική του k – windows επιτρέπει την εξέταση ενός μεγάλου αριθμού παραθύρων χωρίς μεγάλη υπολογιστική επιβάρυνση. Η διαδικασία της συγχώνευσης οδηγείται από κάποια συγκεκριμένα κατώφλια τα οποία ορίζονται από τον χρήστη.

Έτσι για να επεκτείνουμε τους αλγόριθμους k – windows και z – windows, έτσι ώστε να μπορούν να ομαδοποιούν κατηγορικά δεδομένα θα πρέπει να ορίσουμε την έννοια του d – διαστήματος, την έννοια της μετακίνησης των d – διαστημάτων, την έννοια της μεγέθυνσης καθώς και την έννοια της επικάλυψης d – διαστημάτων για να μπορέσουμε να ενσωματώσουμε την απόσταση που αναπτύξαμε σ' αυτούς τους αλγορίθμους. Θα παρουσιάσουμε τον ορισμό αυτών των εννοιών χρησιμοποιώντας την δεντρική δομή των οντολογιών κάθε χαρακτηριστικού.

Το διάστημα κατηγορικών τιμών ενός χαρακτηριστικού που παριστάνονται από κόμβους μιας δεντρικής δομής μπορεί να οριστεί μέσω του ελάχιστου και του μέγιστου επιπέδου των κόμβων αυτών των κόμβων. Τυπικά θα μπορούσαμε να γράψουμε:

$$[\min, \max] = \left\{ a_{i_j} \mid a_{i_j} \in N_{a_i}, \min \le l\left(a_{i_j}\right) \le \max \right\} \text{ (5.6).}$$

Ένα ανοιχτό διάστημα (min, max) θα μπορούσε να οριστεί αντίστοιχα. Έτσι ένα d – διάστημα, ένα d- διάστατο παράθυρο δηλαδή είναι ένα διάνυσμα από d διαστήματα, που κάθε ένα από αυτά ορίζεται σε ένα χαρακτηριστικό.

Ως ένα παράδειγμα ας δώσουμε τον αλγόριθμο που ορίζει τα αρχικά d – παράθυρα στον αλγόριθμο z – windows.

covering input space with d – windows

do while $\frac{i1 < (\max\_level\_of\_(Ta_{i1}))}{a}$

…

do while $\frac{id < (\max\_level\_of\_(Ta_{id}))}{a}$

output the d – range ([i1, i1 + a), …,[id, id + a))

end do

…

i1+=a

end do

Αλγόριθμος 5.2: ο αλγόριθμος αρχικοποίησης των d – παραθύρων του αλγόριθμου z – windows.

Μία κίνηση ενός d – διαστήματος ([i1, i1 + a), …,[id, id + a)) καταμήκος ενός συγκεκριμένου αντικειμένου ($\alpha 1_x$, …, αd_y) ορίζει ένα καινούργιο d – διάστημα $\left(\left[l(a1_x) - a/2, l(a1_x) + a/2\right),...,\left[l(ad_y) - a/2, l(ad_y) + a/2\right)\right)$.

Μια μεγέθυνση του d – διαστήματος ([i1, i1 + a),…,[id, id + a)) ορίζεται από το d – διάστημα $\left(\left[i1 - a*r, i1 + a + a*r\right),...,\left[id - a*r, id + a + a*r\right)\right)$.

Τέλος δυο d – περιοχές ([i1, i1 + a), …,[id, id + a)) και ([j1, j1 + a), …,[jd, jd + a)) αλληλοκαλύπτονται όταν i1<j1<i1+a<j1+a,…, id<jd<id+a<jd+a. Στην τελευταία περίπτωση η επικάλυψη είναι $\frac{\left((i1 + a - j1)\cdot...\cdot(id + a - jd)\cdot 100\right)}{\left(a^d\right)}\%$.

5.5 Περαιτέρω δυνατότητες του μέτρου ανομοιότητας

5.5.1 Γενικά

Ο ορισμός ενός αποδοτικού μέτρου ανομοιότητας μεταξύ κατηγορικών δεδομένων είναι μεγάλης σημασίας όχι μόνο σε εφαρμογές ομαδοποίησης (clustering) αλλά και σε πολλούς αλγόριθμους ταξινόμησης (classification). Για παράδειγμα οι αλγόριθμοι Μάθησης από Παραδείγματα (Instance Based Learning IBL – algorithms) χρησιμοποιούν κάποιο μέτρο ανομοιότητας για να κατατάξουν ένα άγνωστο αντικείμενο, με βάση το λιγότερο ανόμοιο (δηλαδή το περισσότερο όμοιο) από τα παραδείγματα εκπαίδευσης. Οι IBL – αλγόριθμοι ή αλλιώς αλγόριθμοι νωθρής μάθησης (lazy learning algorithms) ή απλώς αλγόριθμοι μάθησης βασισμένοι σε παραδείγματα (exemplar – based algorithms) αυτό που κάνουν είναι να αποθηκεύουν ολόκληρο το σύνολο εκπαίδευσης και να μεταθέτουν την όλη διαδικασία μάθησης μέσω επαγωγικής γενίκευσης (inductive generalization), μέχρι την ώρα της

ταξινόμησης. Η γενίκευση στους IBL – αλγόριθμους γίνεται με τον εξής τρόπο: Πρώτα βρίσκονται τα k λιγότερο ανόμοια (δηλαδή τα k περισσότερο όμοια) αντικείμενα, στο αντικείμενο κατάταξης. Έπειτα, το προς κατάταξη αντικείμενο ανατίθεται στην κλάση που περιέχει την πλειοψηφία των k αυτών αντικειμένων. Μερικές φορές τίθενται και κάποια βάρη στα k αυτά αντικείμενα. Είναι λοιπόν φανερό ότι η ποιότητα των IBL αλγορίθμων εξαρτάται σε πολύ μεγάλο βαθμό από τα αντικείμενα που θα βρεθούν να είναι πιο όμοια ως προς το αντικείμενο κατάταξης, το οποίο πάλι με τη σειρά του καθορίζεται από το χρησιμοποιούμενο μέτρο ανομοιότητας.

Ο ταξινομητής των k – κοντινότερων γειτόνων [29] (k nearest neighbor (k – NN) classifier) είναι η βάση πολλών IBL αλγορίθμων. Η είσοδος σε έναν k – NN ταξινομητή είναι ένα άγνωστο αντικείμενο και η έξοδος του είναι μια πρόβλεψη για την κλάση του. Πιο τυπικά αν A και B είναι δυο κατηγορικά αντικείμενα που περιγράφονται από m χαρακτηριστικά, το μέτρο ανομοιότητας ορίζεται ως εξής:

$$d(A,B) = \left(\sum_{j=1}^{m} w(j) \cdot \delta(a_j, b_j)^r \right)^{\frac{1}{r}} \quad (5.7),$$

όπου $w(j)$ είναι μια συνάρτηση βάρους χαρακτηριστικού (attribute weighting function). Ο k – NN ταξινομητής, ορίζει το $\delta(a_j, b_j)$ ως το μέτρο επικάλυψης το οποίο έχουμε αναφέρει πιο πάνω, το r = 2 και το $w(j)$ = s για κάποια σταθερά s. Είναι φανερό ότι το προτεινόμενο μέτρο ανομοιότητας θα μπορούσε να χρησιμοποιηθεί αντί για το $\delta(a_j, b_j)$.

Η επίδοση του k – NN είναι παρά πολύ ευαίσθητη στον ορισμό του μέτρου ανομοιότητας. Για αυτόν το λόγο πολλές από τις παραλλαγές του k – NN παραμετροποιούν το μέτρο ανομοιότητας για να μπορέσουν να απαλλαγούν από αυτήν την ευαισθησία. Για παράδειγμα ας θεωρήσουμε το πολύ δημοφιλές μέτρο ανομοιότητας MVDM [28] το οποίο είναι μια παραλλαγή του VDM [131] μέτρου ανομοιότητας, απαλλαγμένο όμως από τις παραμέτρους. Το μέτρο αυτό λοιπόν, βασίζεται στις συχνότητες των τιμών των χαρακτηριστικών (attribute values) ως προς όλες τις διαφορετικές κλάσεις. Πιο τυπικά έχουμε:

αν a_j και b_j είναι δυο τιμές ενός χαρακτηριστικού τότε το μέτρο ανομοιότητας μπορεί να οριστεί ως εξής:

$$\delta(a_j, b_j) = \sum_{c \in C} \left| P(c \mid a_j) - P(c \mid b_j) \right|^k \quad (5.8)$$

όπου το k είναι μια σταθερά που τις περισσότερες φορές τίθεται ίση με ένα και $P(c \mid a_j)$ και $P(c \mid b_j)$ είναι οι κατά συνθήκη πιθανότητες της κλάσης με τιμή c δεδομένων των a_j και b_j αντίστοιχα. Στο [138] παρουσιάζεται μια επισκόπηση των μεθόδων απόδοσης βαρών στους IBL αλγορίθμους.

Από την άλλη μεριά η εννοιολογική ομαδοποίηση [97] (conceptual clustering) βασίζεται επίσης στο μέτρο ανομοιότητας. Η εννοιολογική ομαδοποίηση προσπαθεί να λύσει δύο προβλήματα:

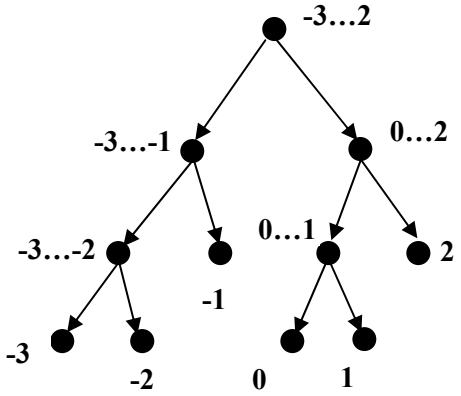
1. δεδομένου ενός συνόλου αντικειμένων θα πρέπει να τοποθετηθούν σε διακριτές (disjoint) ομάδες. Αυτό αποτελεί το λεγόμενο πρόβλημα συγκέντρωσης (aggregation problem) και
2. να διατυπώσει περιγραφές για κάθε μία από αυτές τις ομάδες. Αυτό αποτελεί το λεγόμενο πρόβλημα του χαρακτηρισμού (characterization problem).

Το πρόβλημα της συγκέντρωσης (aggregation problem) έχει να κάνει με μια εκτατική (extensional) περιγραφή της ομάδας. Μια τέτοια περιγραφή είναι η λίστα όλων των αντικειμένων της ομάδας. Ενώ το πρόβλημα του χαρακτηρισμού έχει να κάνει με μια λογική (extensional) εννοιολογική περιγραφή. Επιπλέον θα πρέπει να λαμβάνεται υπ' όψιν η ποιότητα των περιγραφών των εννοιών όταν αξιολογούνται εναλλακτικές ομάδες αντικειμένων. Και τα δυο προβλήματα είναι πολύ ευαίσθητα στην επιλογή του μέτρου ανομοιότητας που θα χρησιμοποιηθεί. Προφανώς το προτεινόμενο μέτρο ανομοιότητας μπορεί να χρησιμοποιηθεί στο πρόβλημα της συγκέντρωσης (aggregation problem). Θα μπορούσε επίσης το προτεινόμενο μέτρο ανομοιότητας να χρησιμοποιηθεί και στο πρόβλημα της περιγραφής. Η ιδέα είναι ότι το προτεινόμενο μέτρο ανομοιότητας βασίζεται σε γνώση πάνω στον τομέα κάθε συγκεκριμένου προβλήματος (domain – specific knowledge). Η γνώση αυτή, αναπαρίσταται με οντολογίες που ορίζονται από τον χρήστη. Έτσι θα μπορούσε κανείς να χρησιμοποιήσει το παραπάνω μέτρο ανομοιότητας για να πάρει μια ιεραρχική εννοιολογική περιγραφή των ομάδων [20].

5.5.2 Αριθμητικά δεδομένα

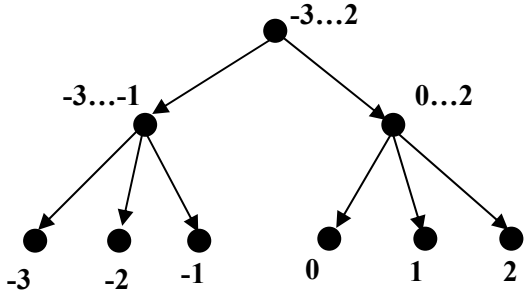
Θα εξετάσουμε εδώ, αν και πως μπορεί να χρησιμοποιηθεί η παραπάνω απόσταση σε αριθμητικά δεδομένα. Αν μπορεί να χρησιμοποιηθεί και σε αριθμητικά δεδομένα τότε η χρήση της, θα είναι ανεξάρτητη από το είδος των δεδομένων και θα μπορεί να χρησιμοποιηθεί και σε μεικτά δεδομένα.

Η βασική ιδέα είναι να αναπαραστήσουμε χαρακτηριστικά κάθε τύπου (συνεχούς ή διακριτού τύπου) με μία οντολογία. Για παράδειγμα, ας θεωρήσουμε το $S = \{u_1,...,u_n\}$ ένα σύνολο διακριτών αριθμητικών τιμών σε αύξουσα σειρά. Θα μπορούσαμε να αναπαραστήσουμε το σύνολο αυτό με την ρίζα κάποιας οντολογίας. Αν διαιρέσουμε το σύνολο αυτό σε δύο υποσύνολα τα $S_1 = \{u_i \in S \mid u_i \le u_1 + (u_n - u_1)/2\}$ και $S_2 = \{u_i \in S \mid u_i > u_1 + (u_n - u_1)/2\}$ τότε θα μπορέσουμε να βάλουμε στους κόμβους παιδιά της ρίζας τα σύνολα $[u_1, u_1 + (u_n - u_1)/2]$ και $[u_1 + (u_n - u_1)/2, u_n]$. Αναδρομικά θα μπορούσε κανείς να χωρίζει τα σύνολα μέχρι που να φτάσει σε σύνολα που δεν περιέχουν παρά μόνο μία τιμή. Αυτό που περιγράψαμε είναι λοιπόν μια από πάνω προς τα κάτω (top – down) τεχνική κατασκευής μιας οντολογίας. Για παράδειγμα αν S = {-3, …, 2}τότε η οντολογία που κατασκευάζεται με τον παραπάνω τρόπο φαίνεται στην Εικόνα 5.4. Ας σημειωθεί ότι θα μπορούμε να παραστήσουμε τιμές που λείπουν από το σύνολό μας με κάποιους από τους εσωτερικούς κόμβους της οντολογίας.



Εικόνα 5.4: Ένα παράδειγμα οντολογίας με αριθμητικά δεδομένα.

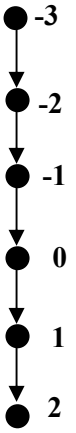
Μια οντολογία αριθμητικών τιμών θα πρέπει να κατασκευάζεται διατηρώντας τις αριθμητικές σχέσης ανισότητας των τιμών, παραδείγματος χάριν αν $x_1 - x_2 = y$, $x_1 - x_3 = z$ και y<z τότε $d(x_1, x_2) < d(x_1, x_3)$. Ας σημειώσουμε ότι αυτό δεν ισχύει στην οντολογία της Εικόνας 5.4 αφού, d(-1,0) > d(-1,2). Από πειράματα που έχουμε κάνει εξάγεται το συμπέρασμα ότι αυτήν την ιδιότητα την έχουν ζυγιασμένες οντολογίες όπως αυτή της παρακάτω Εικόνας 5.5.



Εικόνα 5.5: Μια ζυγιασμένη οντολογία αριθμών που διατηρεί τις σχέσεις μεταξύ τους.

Η οντολογία αυτή είναι μια από κάτω προς τα πάνω κατασκευή (bottom – up construction). Ξεκινάμε δηλαδή από τα φύλλα για να καταλήξουμε στην κορυφή. Κάθε φύλλο αναπαριστά ένα $u_i \in S$ και κάθε μονοπάτι από την ρίζα στα φύλλα έχει το ίδιο μήκος. Η κατασκευή της οντολογίας ικανοποιεί το γεγονός ότι αν $x_1 - x_2 = y$ και $x_1 - x_3 = z$ με y<z τότε $d(x_1, x_2) \leq d(x_1, x_3)$. Για παράδειγμα d(0,1) = d(0,2), δηλαδή δεν διατηρεί την αριθμητική σχέση μεταξύ των τιμών.

Μια προφανής λύση σ' αυτό το πρόβλημα είναι η κατασκευή μιας οντολογίας ενός μονοπατιού όπως φαίνεται στην Εικόνα 5.6. Κάτι τέτοιο όμως δεν αποτελεί μια αποδοτική λύση και για αυτό εργαζόμαστε ακόμα πάνω σε μια αποδοτική κατασκευή μιας οντολογίας αριθμητικών δεδομένων.



Εικόνα 5.6: Μια οντολογία αριθμητικών δεδομένων ενός μονοπατιού

5.6 Εμπειρικοί έλεγχοι

5.6.1 Περιγραφή των ελέγχων

Με τους εμπειρικούς ελέγχους που παρουσιάζουμε παρακάτω θέλουμε να εξετάσουμε την απόδοση και την διαμεριστική ακρίβεια του προτεινόμενου μέτρου ανομοιότητας σε σύγκριση με άλλα γνωστά μέτρα ανομοιότητας κατηγορικών δεδομένων. Θα συγκρίνουμε το προτεινόμενο μέτρο ανομοιότητας με το γνωστό chi – square μέτρο [146] καθώς επίσης και με το μέτρο ανομοιότητας που χρησιμοποιείται σε εννοιολογική ομαδοποίηση και που παρουσιάζεται στο [73].

Αφού δεν υπάρχει ένα κοινώς αποδεκτό ποσοτικό μέτρο για την εκτίμηση της ακρίβειας της ομαδοποίησης για όλους τους αλγορίθμους ομαδοποίησης και κατά συνέπεια δεν υπάρχει και κάποια βάση δεδομένων που να χρησιμοποιείται για τον έλεγχο της ακρίβειας της ομαδοποίησης, είμαστε αναγκασμένοι να καταφύγουμε σε συνθετικές βάσεις δεδομένων. Για να κατασκευάσουμε μια κυρίαρχη αλήθεια και για να μπορέσουμε να συγκρίνουμε τις ομάδες που θα παραχθούν αυτόματα από τους αλγόριθμους ομαδοποίησης θα κατασκευάσουμε μια συνθετική βάση δεδομένων με προσχηματισμένες λογικές ομάδες εντός των δεδομένων.

Πιο ειδικά αυτό που κάναμε ήταν να διελέγξουμε εγγραφές από μια πραγματική βάση δεδομένων. Η βάση που χρησιμοποιήθηκε ήταν μια βάση 62500 εγγραφών μιας μεγάλης τηλεπικοινωνιακής εταιρίας που αφορούσε επιχειρηματικά δεδομένα. Από ένα πλήθος διαφορετικών χαρακτηριστικών που διέθεταν οι εγγραφές της βάσης και που περιέγραφαν την σχετική με το κέρδος συμπεριφορά των πελατώ,ν επιλέχτηκαν δύο χαρακτηριστικά. Το χαρακτηριστικό «Επάγγελμα» και το χαρακτηριστικό «Τόπος Διαμονής» τα οποία παριστάνονται από τις οντολογίες με ρίζες «Επαγγέλματα» και «Ελλάδα» αντίστοιχα. Μέρος τις οντολογίας «Επαγγέλματα» φαίνεται παρακάτω στην Εικόνα (5.7) όπως επίσης η οντολογία «Ελλάδα» φαίνεται στην Εικόνα (5.3)

Από το σύνολο των εγγραφών της βάσης διαλέγουμε 60 οι οποίες συνθέτουν διαφορετικές, πλήρως κατανοητές και εξηγήσιμες ομάδες. Οι ομάδες αυτές φαίνονται μέσα στους κύκλους στην Εικόνα (5.7). Βλέπουμε ότι κάποιες ομάδες περιέχουν και μια πιο λεπτομερή δομή. Θα τις εξετάσουμε όμως όλες μία προς μία.

Η πρώτη ομάδα περιέχει όλους εκείνους τους πελάτες που δουλεύουν στον τομέα της παραγωγής και μένουν στην Βόρεια Ελλάδα (Ήπειρος, Θράκη, Μακεδονία, Θεσσαλονίκη). Αυτή η ομάδα διαθέτει άλλες δυο υποομάδες μέσα της. Η πρώτη υποομάδα αποτελείται από όλους εκείνους που εργάζονται στον πρωτογενή τομέα της παραγωγής και διαμένουν στην Ήπειρο και στην Θράκη, σε σχετικά ακριτικές δηλαδή περιοχές, ενώ η άλλη υποομάδα περιέχει αυτούς που εργάζονται στον τομέα των κατασκευών και διαμένουν στην κεντρική βόρεια Ελλάδα (Μακεδονία και Θεσσαλονίκη).

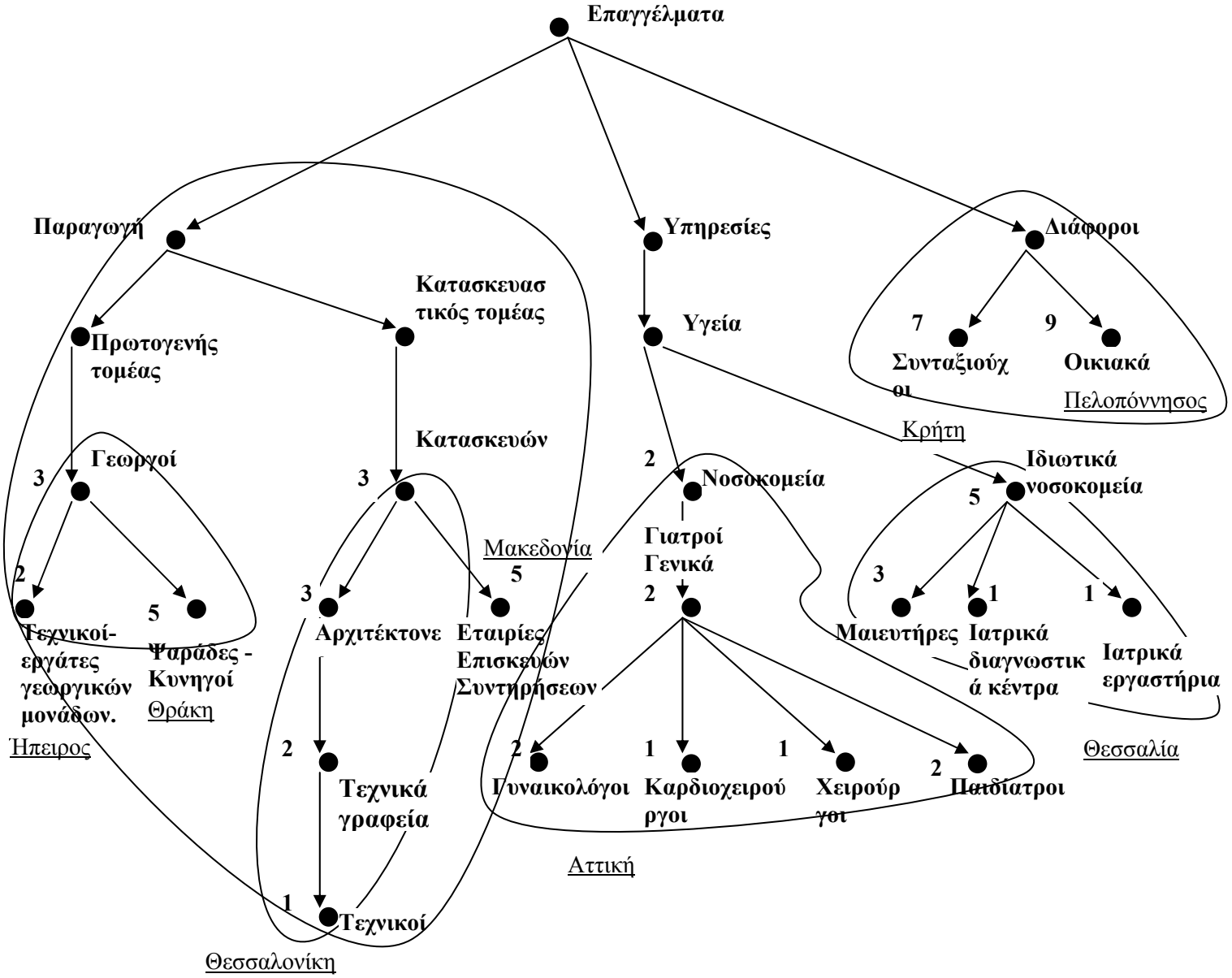
Η δεύτερη ομάδα αποτελείται από αυτούς που εργάζονται στον τομέα της υγείας και διαμένουν στην κεντρική Ελλάδα (Αττική και Θεσσαλία). Και αυτή η ομάδα περιέχει δυο μικρότερες υποομάδες μέσα της. Η πρώτη είναι αυτοί που διαμένουν στην Αττική και εργάζονται σε Δημόσια Νοσοκομεία, ενώ η δεύτερη είναι αυτοί που διαμένουν στην Θεσσαλία και εργάζονται σε ιδιωτικές επιχειρήσεις υγείας.

Τέλος μια τρίτη μεγάλη ομάδα αποτελείται από τους κατοίκους της νότιας Ελλάδας (Κρήτη, Πελοπόννησος) οι οποίοι μπορούν να μπουν κάτω από την ετικέτα «Διάφοροι» (συνταξιούχοι, νοικοκυρές κ.τ.λ.). Και αυτή η ομάδα περιέχει μια πιο λεπτομερή διαμέρηση: Αυτούς που είναι συνταξιούχους και κατοικούν στην Κρήτη και αυτούς που έχουν ως επάγγελμα τα οικιακά και κατοικούν στην Πελοπόννησο.

Ας δώσουμε λίγες ακόμα επεξηγήσεις για τον συμβολισμό που χρησιμοποιήσαμε στα σχήματα των Εικόνων (5.7) και (5.2). Οι αριθμοί δίπλα σε κάθε κόμβο, παριστάνουν το πλήθος των εγγραφών, που έχουν για τιμή του αντίστοιχου χαρακτηριστικού, την τιμή με την οποία επιγράφεται ο κόμβος. Επίσης οι υπογραμμισμένες ετικέτες αναφέρονται στην τιμή του χαρακτηριστικού «Τόπος Διαμονής» που έχουν όλα τα αντικείμενα της ομάδας.

Ο πίνακας με τα αντικείμενα που χρησιμοποιήθηκαν στα πειράματα για την αξιοπιστία του μέτρου ανομοιότητας, βρίσκεται στο Παράρτημα Α.

Τα αποτελέσματα των πειραμάτων που έγιναν έδειξαν ότι το προτεινόμενο μέτρο ανομοιότητας επέδειξε εξαιρετική συμπεριφορά ως προς την ανίχνευση των ομάδων που υπήρχαν στα δεδομένα. Ας σημειώσουμε ότι η μέθοδος ομαδοποίησης, ανίχνευσε τις ομάδες που υπήρχαν στα δεδομένα και που τις περιγράψαμε παραπάνω. Μπόρεσε δηλαδή, να ανιχνεύσει, ακόμα και τις μικρές υποομάδες που υπήρχαν μέσα στα δεδομένα με μεγάλη επιτυχία. Έτσι, στην τελική φάση των πειραμάτων μας, όπου ζητήσαμε από τη μέθοδο να μας επιστρέψει εφτά ομάδες, η μέθοδός μας επέστρεψε όλες τις υποομάδες που θα μπορούσε να φανταστεί λογικά κάποιος. Σε αντίθεση με τις άλλες μεθόδους ομαδοποίησης. Γενικά η μέθοδος που προτείνουμε έδωσε πολύ καλύτερα και πολύ ακριβέστερα αποτελέσματα από ότι οι άλλες δυο μέθοδοι που χρησιμοποιήθηκαν, δηλαδή την μέθοδο k – modes που βασίζεται στο γνωστό μέτρο ανομοιότητας chi – square και την μέθοδο τις εννοιολογικής ομαδοποίησης που χρησιμοποιεί ένα μέτρο ανομοιότητας που βασίζεται και αυτό σε μια δεντρική δομή.



Εικόνα 5.7: Η οντολογία «Επάγγελμα» με τις προϋπάρχουσες ομάδες σημειωμένες.

5.6.2 Σχολιασμός των αποτελεσμάτων

Ας δούμε και ας σχολιάσουμε όμως ξεχωριστά, ποια είναι τα αποτελέσματα που μας έδωσαν κάθε μία από τις παραπάνω μεθόδους. Ας σημειώσουμε εδώ, ότι στο παράρτημα Β παραθέτονται οι ομάδες που προέκυψαν από την εκτέλεση των τριών αλγορίθμων που χρησιμοποιήθηκαν: τον προτεινόμενο, τον k – modes και τον αλγόριθμο εννοιολογικής ομαδοποίησης του Kodratoff. Οι ομάδες φαίνονται πάνω στο μεγαλύτερο από τα δύο δέντρα των οντολογιών, στο οποίο όμως υπάρχουν πληροφορίες και για το δεύτερο δέντρο «Ελλάδα», όπως έχουμε εξηγήσει και παραπάνω. Επίσης στα σχήματα του παραρτήματος Β οι κόμβοι που είναι σημειωμένοι με κόκκινο είναι τα αρχικά κέντρα που προήλθαν από τον αλγόριθμο αρχικοποίησης που περιγράψαμε παραπάνω. Οι υπογραμμισμένες με διπλή κόκκινη γραμμή ετικέτες αναφέρονται στην αντίστοιχη τιμή που παίρνει το αρχικό αντικείμενο, στην άλλη οντολογία που δεν φαίνεται στο σχήμα. Τέλος ας σημειώσουμε ότι για αρχικά κέντρα και στους τρεις αλγορίθμους χρησιμοποιήθηκαν αυτά που προέκυψαν από τον αλγόριθμο αρχικοποίησης που προτάθηκε. Με αυτόν τον τρόπο μπορούμε να κάνουμε μια αντικειμενική σύγκριση των μέτρων ανομοιότητας.

Κατ’ αρχάς να δούμε πως μπορούν να ερμηνευτούν τα αποτελέσματα που πήραμε από το προτεινόμενο μέτρο ανομοιότητας.

2 ομάδες:

- Προτεινόμενος αλγόριθμος. Όταν ζητήσαμε από τον αλγόριθμο να μας επιστρέψει δυο ομάδες αυτός επέστρεψε τις ομάδες που φαίνονται στο αντίστοιχο σχήμα και μπορούν να ερμηνευτούν πάρα πολύ εύκολα με την διαίσθησή μας. Ουσιαστικά ο αλγόριθμος έβαλε σε μια ομάδα όλους αυτούς που κατοικούν στην Βόρεια Ελλάδα και εργάζονται στον τομέα της παραγωγής. Στην άλλη ομάδα έβαλε όλους εκείνους που κατοικούν στην Κεντρική ή Νότια Ελλάδα και εργάζονται στον τομέα της υγείας ή κάνουν διάφορες δουλειές (συνταξιούχοι, οικιακά). Η διαμέριση του συνόλου σε δύο ομάδες έγινε με τον βέλτιστο και πιο εύκολα επεξηγήσιμο τρόπο.
- Kodratoff. Όταν ζητήθηκε από τον αλγόριθμο που χρησιμοποιεί το μέτρο ανομοιότητας του Kodratoff να μας επιστρέψει δύο ομάδες, αυτός επέστρεψε σε μια ομάδα όλους αυτούς που εργάζονται στην Κεντρική Ελλάδα και εργάζονται στον τομέα της υγείας και σε μια άλλη ομάδα έβαλε αυτούς που κατοικούν στην Βόρεια Ελλάδα και αυτούς που κατοικούν στην Νότια Ελλάδα. Η ομαδοποίηση αυτή αν και δεν μπορεί να εξηγηθεί λογικά δεν αποτελεί την βέλτιστη γιατί βάζει αυτούς που κατοικούν στην Βόρεια Ελλάδα με αυτούς που κατοικούν στην νότια Ελλάδα στην ίδια ομάδα.
- K – modes. Όταν ζητήθηκε από τον k – modes να διαμερίσει το σύνολο σε δύο ομάδες ομαδοποίησε μόνα τους τα αντικείμενα που κατοικούν στην Αττική και εργάζονται σε νοσοκομείο και όλα τα άλλα αντικείμενα τα ομαδοποίησε μαζί Η ομαδοποίηση αυτή δεν μπορεί να εξηγηθεί με λογικό τρόπο. Ο αλγόριθμος αυτός δεν μπορεί να ανακαλύψει τις ομάδες που κρύβονται μέσα στα δεδομένα όπως τις περιγράψαμε διαισθητικά παραπάνω.

3 ομάδες

- Προτεινόμενος αλγόριθμος. Εδώ ο αλγόριθμός μας, μας δίνει τις τρεις ομάδες όπως ακριβώς τις περιγράψαμε παραπάνω. Στην μία ομάδα βρίσκονται αυτοί που κατοικούν στην Βόρεια Ελλάδα και δουλεύουν στην Παραγωγή, στην δεύτερη ομάδα βρίσκονται αυτοί που κατοικούν στην Νότια Ελλάδα και εργάζονται στον τομέα της Υγείας ενώ στην Τρίτη ομάδα είναι αυτοί που κατοικούν στην Νότια Ελλάδα. Εδώ πρέπει να σημειώσουμε ότι δυο από τα αρχικά κέντρα βρέθηκαν στο ίδιο υποδέντρο, αυτό που έχει ρίζα τον κόμβο Υγεία, και πάρ' όλα αυτά ο αλγόριθμος ομαδοποίησης δεν διαμέρισε σε δύο το υποδέντρο αυτό, αλλά σωστά ανακάλυψε και την τρίτη ανεξάρτητη ομάδα, αυτή που αποτελείται από αυτούς που κατοικούν στην Νότια Ελλάδα.
- Kodratoff. Ο αλγόριθμος αυτός «πέφτει στην παγίδα» που αναφέραμε παραπάνω. Χωρίζει σε δύο ομάδες τους εργαζόμενους στην Υγεία και ομαδοποιεί μαζί αυτούς που κατοικούν στην Βόρεια και αυτούς που κατοικούν στην Νότια Ελλάδα. Ουσιαστικά δεν μπορεί να ξεχωρίσει ακόμα την ομάδα αυτών που κατοικούν στην Νότια Ελλάδα.
- k – modes. Ο k – modes ξεχωρίζει μια μικρή ομάδα, αυτούς που κατοικούν στην Ήπειρο και που εργάζονται στον γεωργικό τομέα («γεωργός», «τεχνικοί εργάτες γεωργικών μονάδων»). Η δεύτερη ομάδα είναι αυτού που εργάζονται σε νοσοκομείο και κατοικούν στην Αττική και όλες οι άλλες εγγραφές συνωστίζονται σε μια μεγάλη ομάδα. Ας σημειώσουμε εδώ πως και αυτός ο αλγόριθμος άρχισε όπως φαίνεται στο αντίστοιχο σχήμα με δύο αρχικά μέσα που τελικά ανήκαν στην ίδια ομάδα, χωρίς να την χωρίζουν στα δυο.

4 ομάδες

- Προτεινόμενος αλγόριθμος. Εδώ ο αλγόριθμός μας χωρίζει την μεγάλη ομάδα που αποτελούνταν από αυτούς που κατοικούν στην Αττικοί και εργάζονται στον τομέα της Υγείας, ανακαλύπτοντας δυο υποομάδες. Αυτούς που εργάζονται στην δημόσια υγεία και αυτούς που εργάζονται στον τομέα της ιδιωτικής Υγείας («μαιευτήρες», «ιατρικά διαγνωστικά κέντρα», «ιατρικά εργαστήρια» κ.τ.λ.).
- Kodratoff. Σε αυτήν την περίπτωση ο αλγόριθμος αυτός δίνει ακριβώς τα ίδια αποτελέσματα με τον προτεινόμενο αλγόριθμο που περιγράψαμε παραπάνω. Σε αυτό το σημείο ο αλγόριθμος αυτός μπόρεσε να αναγνωρίσει και την ύπαρξη της ομάδας αυτών που κατοικούν στην Νότια Ελλάδα.
- k – modes. Σε αυτήν την περίπτωση ο k – modes αυτό που κάνει είναι να απομονώσει από την μεγάλη ομάδα που περιγράψαμε παραπάνω όλες εκείνες τις περιπτώσεις που κατοικούν στην Θεσσαλονίκη και εργάζονται στις κατασκευές. Ουσιαστικά δεν μπορεί να ανακαλύψει κάποια από τις ομάδες που υπάρχουν στο σύνολο των δεδομένων.

5 ομάδες

- προτεινόμενος Αλγόριθμος. Σε αυτήν τη περίπτωση ο αλγόριθμος διασπάει την ομάδα που αποτελείται από αυτούς που κατοικούν στην Νότια Ελλάδα και ανακαλύπτει δυο υποομάδες που περιέχονται μέσα σε αυτήν . Τους συνταξιούχους που κατοικούν στην Κρήτη και αυτούς που ασχολούνται με τα οικιακά και κατοικούν στην Πελοπόννησο.

- Kodratoff. Και σε αυτήν την περίπτωση ο αλγόριθμος αυτός δίνει ακριβώς τα ίδια αποτελέσματα με τον προτεινόμενο αλγόριθμο.
- k – modes. Ο k – modes σε αυτήν την περίπτωση αυτό που κάνει είναι να απομονώσει από την μεγάλη ομάδα ένα ακόμα σύνολο περιπτώσεων: αυτούς που κατοικούν στην Μακεδονία και δουλεύουν σε εταιρίες επισκευών και συντηρήσεων. Πάλι δεν ανακαλύπτει κάποια από τις προϋπάρχουσες ομάδες.

Ας σημειώσουμε εδώ πως, όπως φαίνεται καθαρά από τα σχήματα ο αλγόριθμος αρχικοποίησης που προτείνουμε διασπείρει τα αρχικά μέσα όσο το δυνατόν καλύτερα στην δενδρική δομή των οντολογιών. Αυτό όμως όπως θα δούμε και θα σχολιάσουμε παρακάτω δεν φτάνει για να αποκτήσουμε μια εκλεπτυσμένη και ακριβή ομαδοποίηση. Μέχρι στιγμής όμως φαίνεται ότι ο αλγόριθμος βασισμένος στο μέτρο ομοιότητας του Kodratoff εκμεταλλεύεται πάρα πολύ αποτελεσματικά αυτήν την ιδιότητα του αλγορίθμου αρχικοποίησης των μέσων, πράγμα που θα πάψει λίγο αργότερα όταν θα συζητάμε για περισσότερες ομάδες.

6 ομάδες

- Προτεινόμενος αλγόριθμος. Όταν ζητήσαμε από τον αλγόριθμό να μας επιστρέψει έξι ομάδες αυτό που έκανε ουσιαστικά είναι να ανακαλύψει δυο υποομάδες που βρίσκονταν μέσα στην μεγαλύτερη ομάδα που στο δέντρο της οντολογίας «Επαγγέλματα» είχε για ρίζα τον κόμβο «Παραγωγή». Αυτό που έκανε ουσιαστικά είναι να χωρίσει αυτούς που ασχολούνται στην πρωτογενή παραγωγή από αυτούς που απασχολούνται στον κατασκευαστικό τομέα. Οι μεν πρώτοι μένουν στην Ήπειρο ή στην Θράκη οι δε δεύτεροι βρίσκονται στην Μακεδονία ή στην Θεσσαλονίκη.
- Kodratoff. Ο αλγόριθμος αυτός χωρίζει την μεγάλη ομάδα αυτών που διαμένουν στην Βόρεια Ελλάδα. Ο διαχωρισμός αυτός έγινε με διαφορετικό τρόπο. Στην πρώτη υποομάδα βρέθηκαν αυτοί που διαμένουν στην Ήπειρο και στην Μακεδονία και οι μεν ασχολούνται με τον γεωργικό τομέα ενώ οι δε εργάζονται για εταιρίες επισκευών και συντηρήσεων. Στη δεύτερη υποομάδα βρέθηκαν αυτοί που κατοικούν στην Θράκη ή στην Μακεδονία και εργάζονται είτε στον κτηνοτροφικό τομέα είτε στον τομέα των κατασκευών εκτός του τομέα των συντηρήσεων. Θα μπορούσε κάποιος να πει ότι εδώ ο αλγόριθμος μπερδεύει κάπως τις αρχικές ομάδες που υπάρχουν στα δεδομένα χωρίς να δίνει από την άλλη κάποια καινούργια λογικά εξηγήσιμη υποδιαίρεση της μεγάλης ομάδας των ανθρώπων που μένουν στην Βόρειο Ελλάδα.
- k – means. Σε αυτήν την περίπτωση ο k – means αυτό που κάνει είναι να απομονώσει άλλο ένα σύνολο ίδιων αντικειμένων από το σώμα της μεγάλης ομάδας. Αυτήν την φορά όμως το αντικείμενο που απομονώνει αποτελεί μια μικρή ομάδα. Είναι όλοι εκείνοι που κατοικούν στην Θράκη και απασχολούνται στον κτηνοτροφικό τομέα. Αυτή η ανακάλυψη οφείλεται μάλλον στο γεγονός ότι η ομάδα αποτελείται από ένα και μόνο αντικείμενο.

7 ομάδες

- Προτεινόμενος αλγόριθμος. Σε αυτήν την περίπτωση ο αλγόριθμός μας ανακαλύπτει μια άλλη υποομάδα η οποία βρίσκετε στην ομάδα αυτών που ασχολούνται στον πρωτογενή τομέα της παραγωγής. Χωρίζει λοιπόν αυτούς

που μένουν στην Ήπειρο και που ασχολούνται με τον γεωργικό τομέα απ' αυτούς που μένουν στην Θράκη και ασχολούνται με την κτηνοτροφία. Μπορούμε να πούμε ότι ο αλγόριθμός ανακάλυψε την βέλτιστη τμηματοποίηση του συνόλου των παραδειγμάτων μας.

- Kodratoff. σε αυτήν την περίπτωση ο αλγόριθμος αυτός κάνει έναν περίεργο χωρισμό των υποομάδων που αποτελούνται απ' αυτούς που απασχολούνται στον τομέα της Παραγωγής. Οι ομάδες που ανακάλυψε ο αλγόριθμος φαίνονται στο αντίστοιχο σχήμα του παραρτήματος Β. ο χωρισμός αυτός δεν ανταποκρίνεται αφ' ενός στη δομή των ομάδων που υπάρχουν στο σύνολο των δεδομένων μας και αφ' εταίρου δεν μπορούμε να πούμε ότι οι ομάδες έτσι που σχηματίστικαν ότι περιγράφουν κάποια άλλη δομή των δεδομένων.
- k – modes. Σε αυτήν την περίπτωση αυτό που κάνει ο αλγόριθμος k – modes είναι να αναγνωρίσει την ομάδα αυτών που απασχολούνται στον τομέα της ιδιωτικής υγείας. Αυτή η ομάδα αποκόβεται από την μεγάλη ομάδα για την οποία μιλήσαμε παραπάνω και μάλιστα δεν αποτελείται από ένα μόνο αντικείμενο, αλλά περιέχει μια ποικιλία αντικειμένων.

Όπως προαναφέραμε, το σύνολο αυτό των περιπτώσεων αποτελούνταν από ομάδες λογικά αναγνωρίσιμες και διαχωρίσιμες μεταξύ τους που περιείχαν μέσα τους μικρότερες υποομάδες. Ο προτεινόμενος αλγόριθμος μπόρεσε και ανακάλυψε όχι μόνο τις διαφορετικές ομάδες, πράγμα που παραδείγματος χάριν ο k – means δεν μπόρεσε να κάνει καθόλου, αλλά και αναγνώρισε και τις πιο εκλεπτυσμένες υποομάδες που υπήρχαν, κάτι που ο αλγόριθμος που βασίστηκε στο μέτρο του Kodratoff απέτυχε να αναγνωρίσει με την ακρίβεια του προτεινόμενου αλγόριθμου. Από την άλλη πλευρά είδαμε ότι ο αλγόριθμος k – means δεν μπόρεσε να αναγνωρίσει καθόλου την δομή των ομάδων που υπήρχαν στα δεδομένα.

Έτσι συμπερασματικά, θα μπορούσε κανείς να πει ότι ο προτεινόμενος αλγόριθμος είναι ο μόνος από τους τρεις που πήραν μέρος στους ελέγχους που μπόρεσε να αναγνωρίσει την δομή των ομάδων που υπήρχαν στα δεδομένα. Η δομή αυτή ήταν αποτυπωμένη στα δέντρα των οντολογιών των χαρακτηριστικών, τα οποία περιγράφουν με έναν πολύ σαφή και κατανοητό τρόπο την γνώση γύρω από το πεδίο εφαρμογής. Ο k – modes που δεν χρησιμοποίησε κάποιον τρόπο αναπαράστασης γνώσης γύρω από το πεδίο εφαρμογής απέτυχε εντελώς να αναγνωρίσει την δομή των ομάδων των δεδομένων. Ο αλγόριθμος που βασίστηκε στο μέτρο ανομοιότητας του Kodratoff από την άλλη μεριά, χρησιμοποιεί κάποια αναπαράσταση γνώσης πάνω στο πεδίο εφαρμογής, η οποία μάλιστα είναι δεντρική. Παρ' όλα αυτά όμως αν και ο αλγόριθμος αυτός μπόρεσε να αναγνωρίσει κάποιες από τις ομάδες που προϋπήρχαν στα δεδομένα εν τούτοις δεν μπόρεσε να αναγνωρίσει με κάθε λεπτομέρεια και με την ακρίβεια που υπολόγισε ο προτεινόμενος αλγόριθμος τις ομάδες με την εσωτερική δομή.

Έτσι θα μπορούσαμε να πούμε ότι το προτεινόμενο μέτρο ανομοιότητας, πέρα από τις διάφορες εφαρμογές και τον εύκολο τρόπο με τον οποίο μπορεί να ενσωματωθεί σε διάφορες διαδικασίες ώστε να τις βελτιώσει, βλέπουμε ότι αποτελεί και έναν εξαιρετικό τρόπο επέκτασης του αλγορίθμου κατάταξης k – means έτσι ώστε να μπορεί να ομαδοποιεί και κατηγορικά δεδομένα, ενσωματώνοντας στον αλγόριθμο και γνώση γύρω από το πεδίο εφαρμογής.

Θα πρέπει να αναφέρουμε εδώ πως η ενσωμάτωση της γνώσης του πεδίου εφαρμογής δεν γίνεται με κάποιον αυτόματο τρόπο, αλλά απεναντίας αφήνεται στην διαίσθηση και την εμπειρία του χρήστη, κάτι που δίνει στην ομαδοποίηση την δύναμη που είδαμε παραπάνω.

6 Εφαρμογή του προτεινόμενου μέτρου στην εξόρυξη κειμένου

6.1 Εισαγωγή

Εδώ θα περιγράψουμε άλλη μία εφαρμογή της μεθόδου ομαδοποίησης που στηρίζεται στο μέτρο ανομοιότητας που προτείναμε παραπάνω. Αυτήν την φορά η εφαρμογή αφορά δεδομένα από λογοτεχνικά κείμενα. Για την ακρίβεια, ως δεδομένα του αλγορίθμου ομαδοποίησης θα χρησιμοποιήσουμε αυτήν την φορά προεπεξεργασμένα ποιήματα του Έλληνα ποιητή Χριστιανόπουλου καθώς και των πολύ γνωστών Ιαπώνων ποιητών Χαϊκού του Μπασό και του Μπουσόν.

Η μέθοδός μας χρησιμοποιείται για δυο κυρίως λόγους. Ο πρώτος αντιμετωπίζει το γνωστό πρόβλημα του συγγραφέα (authoring – attribution problem). Το πρόβλημα αυτό συνίσταται στην διάκριση κειμένων διαφορετικών συγγραφέων, χωρίς όμως την εκ των προτέρων γνώση της πατρότητάς τους. Με την μέθοδο αυτή που προτείνουμε μπορέσαμε να διακρίνουμε τα γιαπωνέζικα ποιήματα των Μασό και Μπουσόν από τα ελληνικά ποιήματα του Χριστιανόπουλου.

Στην συνέχεια η μέθοδος ομαδοποίησης χρησιμοποιείται για την ομαδοποίηση των ποιημάτων που εισάγονται στον αλγόριθμο, σε νοηματικές ομάδες. Αυτή η προσέγγιση μας χώρισε τα ποιήματα σε ξεχωριστές και πολύ καλά εξηγήσιμες ομάδες με βάση το νόημα των ποιημάτων.

Στη συνέχεια θα αναφερθούμε στο πρόβλημα του συγγραφέα, θα εκθέσουμε την μεθοδολογία που ακολουθήσαμε και θα αναφερθούμε στα αποτελέσματα.

6.2 Authoring Attribution Problem

Το πρόβλημα του συγγραφέα (Author attribution problem) συνίσταται στην ανίχνευση των στυλιστικών χαρακτηριστικών ενός κειμένου ή ενός σώματος κειμένων οι συγγραφείς των οποίων είτε είναι άγνωστοι είτε τίθενται υπό αμφισβήτηση, με σκοπό την απόδοση της πατρότητας του κειμένου σε κάποιον συγγραφέα. Ένα τυπικό παράδειγμα αυτού του τύπου είναι η περίπτωση των Federalist papers [105, 60]. Αυτά είναι δώδεκα εργασίες των οποίων αμφισβητούνται οι συγγραφείς.

Η έλλειψη ενός τυπικού ορισμού του ιδιοσυγκρασιακού στυλ ενός συγγραφέα οδήγησε στην ανάπτυξη στατιστικών μεθόδων για την αναγνώριση των «δακτυλικών αποτυπωμάτων» των συγγραφέων. Αυτό οδήγησε και στην ανάπτυξη ενός ολόκληρου κλάδου της στυλομετρίας (stylometry), της στατιστικής ανάλυσης του λογοτεχνικού στυλ των συγγραφέων. Η στυλομετρία συμπληρώνει την παραδοσιακή φιλολογική προσπάθεια επίλυσης του προβλήματος του συγγραφέα, αφού προσφέρει μέσα ικανά να ανιχνεύσουν κάποιες παραμέτρους του στυλ των συγγραφέων, που είναι συχνά πολύ ασαφείς καταφεύγοντας στην ποσοτικοποίηση κάποιων από τα χαρακτηριστικά του στυλ. Πολλές από τις στυλομετρικές προσεγγίσεις χρησιμοποιούν κάποια

γλωσσικά στοιχεία, με αποτέλεσμα πολλά από αυτά να βασίζονται σε λεξικά και συνεπώς να μην μπορούν να εφαρμοστούν σε διαφορετικές γλώσσες. Μία έγκυρη έκθεση της λογικής πίσω από τέτοιες μελέτες έχει προταθεί από τον Laan [80].

Η βασική υπόθεση πίσω από την στυλομετρία είναι ότι κάθε συγγραφέας έχει μία συνειδητή και μια ασυνείδητη πλευρά του στυλ του. Έτσι κάποια στοιχεία του στυλ κάθε συγγραφέα είναι ανεξάρτητα από τη θέλησή του και αφού αυτά τα στοιχεία δεν μπορούν να ελεγχθούν από τον συγγραφέα συνειδητά, θεωρείται ότι μπορούν να αποτελέσουν αξιόπιστα δεδομένα για μια στυλομετρική μελέτη.

Δυο είναι οι υποθέσεις που γίνονται πάνω σε αυτά τα ασυνείδητα στοιχεία του στυλ των συγγραφέων. Η πρώτη υπόθεση είναι ότι αυτά τα στοιχεία παραμένουν αναλλοίωτα μέσα στον χρόνο, με αποτέλεσμα να μπορεί να αναγνωριστεί ο συγγραφέας σε κείμενα που ανήκουν σε τελείως διαφορετικές περιόδους της ζωής του. Η δεύτερη υπόθεση αναφέρει ότι τα ασυνείδητα στοιχεία του στυλ του συγγραφέα διαφοροποιούνται και εξελίσσονται στην διάρκεια της δημιουργικής του πορείας, οπότε, αναλύοντάς τα να μπορεί να γίνει χρονολογική κατάταξη του συγκεκριμένου έργου.

Θα αναφερθούμε εδώ σε μια ιστορική αναδρομή της πορείας της στυλομετρίας από τα πρώτα βήματά της το $19^{ο}$ αιώνα, ως τις μέρες μας. Θα αναφερθούμε κυρίως σε στυλιστικά κριτήρια που χρησιμοποιήθηκαν ιστορικά στην στυλομετρία.

6.2.1 Μήκος λέξης και Μήκος πρότασης (word – length and sentence – length)

Οι ρίζες της στυλομετρίας θα πρέπει να αναζητηθούν στην δουλειά του Mendenhall (1887) [95] πάνω στην μελέτη του μήκους των λέξεων και στην επέκταση αυτής της ιδέας ώστε να περιλάβει και μήκη προτάσεων από τον Yule το 1938 [142]. Ο Morton το 1965 [104] χρησιμοποίησε τα μήκη των προτάσεων για να εξετάσει την πατρότητα αρχαιοελληνικών κειμένων, αλλά τώρα γνωρίζουμε ότι κανένα από αυτά τα μέτρα δεν αποτελεί αξιόπιστο δείκτη της πατρότητας των κειμένων.

6.2.2 Συναρτήσεις λέξεων (function words)

Η χρήση των λέξεων προσφέρει μεγάλες ευκαιρίες για διάκριση. Μερικές λέξεις διαφέρουν αρκετά όσο αναφορά την συχνότητα εμφάνισης τους από έργο σε έργο του ίδιου συγγραφέα ενώ άλλες επιδεικνύουν μια μεγάλη σταθερότητα στον ίδιο συγγραφέα. Για σκοπούς διάκρισης χρειαζόμαστε ελεύθερες από συμφραζόμενα λέξεις ή αλλιώς συναρτήσεις λέξεων. Μια δουλειά πάνω σε συναρτήσεις συχνότητας λέξεων είναι αυτή των Mosteller και Wallace toy 1964 [106]. Ο Morton το 1978 [103] ανάπτυξε τεχνικές για την μελέτη της θέσης και τα άμεσα συμφραζόμενα κάθε λέξης αλλά η μέθοδός του έτυχε μεγάλης κριτικής και ο Smith το 1985 στο [127] έδειξε ότι η μέθοδος αυτή δεν είναι ικανή να ξεχωρίσει έργα μεταξύ των Ιακωβιανών και των Ελισαβετιανών θεατρικών συγγραφέων.

Η ιδέα της χρήσης συνόλων από συνηθισμένες υψηλής συχνότητας λέξεις και η διεξαγωγή αυτού που λέγετε ανάλυση κυρίων στοιχείων (principal components

analysis) προτάθηκε από τον Burrows το 1987 [22] και αποτελεί έναν σημαντικό σταθμό στην ιστορία της στυλομετρίας. Η τεχνική αυτή είναι πολύ δημοφιλής σήμερα και θεωρείται ως μια αξιόπιστη στυλομετρική διαδικασία. Ο Holmes και ο Forsyth (1995) [60] χρησιμοποίησαν αυτήν την τεχνική με επιτυχία για να λύσουν το πρόβλημα των «Federalist papers».

6.2.3 Κατανομές λεξιλογίου (Vocabulary distributions)

Μία από τις θεμελιώδεις έννοιες στην στιλομετρία είναι η μέτρηση αυτού που λέγετε πλούτος «richness» ή ποικιλία «diversity» του λεξιλογίου ενός συγγραφέα. Εάν λάβουμε ως παράδειγμα το κείμενο ενός συγγραφέα τότε θα πρέπει να περιμένουμε ότι η έκταση του λεξιλογίου του θα αντανακλάται στην κατανομή των συχνοτήτων των λέξεων που χρησιμοποιεί.

Διάφορα μαθηματικά μοντέλα για την κατανομή των συχνοτήτων του αριθμού των λέξεων του λεξιλογίου που εμφανίζονται ακριβώς r φορές κίνησαν το ενδιαφέρων των στατιστικών μετά την εργασία του Zipf το 1932 [147]. Το καλύτερο από τα μοντέλα που προτάθηκαν είναι αυτό του Sichel (1975) [125] στο οποίο αναφέρονται αναλύονται ως χρήσιμα στιλομετικά εργαλεία και οι λέξεις που εμφανίζονται μόνο μία (άπαξ λεγόμενα – hapax legomena) και αυτές που εμφανίζονται μόνο δύο φορές (άπαξ δυσλεγόμενα – hapax dislegomena) στο σύνολο του κειμένου.

6.2.4 Ανάλυση περιεχομένου (Content Analysis)

Η ανάλυση περιεχομένου αναφέρεται στην πινακοποίηση των συχνοτήτων των τύπων των λέξεων που εμφανίζονται στο κείμενο με σκοπό να προσεγγιστεί το σημαινόμενο ή το υπονοούμενο νόημα του κειμένου. Αν και η ανάλυση περιεχομένου είναι πολύ χρήσιμη στην στυλομετρία εντούτοις δεν έχει χρησιμοποιηθεί ευρέως. Ένα παράδειγμα είναι η εφαρμογή της ανάλυσης περιεχομένου των Martindale και McKenzie το 1995 [92] στο πρόβλημα των «Federalist papers».

6.2.5 Νευρωνικά δίκτυα (Neural Networks)

Η στυλομετρία είναι ουσιαστικά ένα πρόβλημα αναγνώρισης προτύπων. Τα νευρωνικά δίκτυα έχουν την ικανότητα να αναγνωρίζουν την υποκείμενη οργάνωση των δεδομένων κάτι το οποίο είναι μείζονος σημασίας για κάθε πρόβλημα αναγνώρισης προτύπων, έτσι η χρησιμοποίησή τους στην στυλομετρία είναι και αναπόφευκτη αλλά και ευπρόσδεκτη. Κάποια αποτελέσματα από αυτόν τον τομέα μπορούν να βρεθούν στις εργασίες των Merriam και Matthews (1994)[96] και των Lowe και Matthews (1995) [87].

6.2.6 Εξόρυξη δεδομένων (Data Mining)

Μια άλλη προσέγγιση στο ζήτημα του προβλήματος της πατρότητας ενός κειμένου είναι και η εφαρμογή τεχνικών της εξόρυξης δεδομένων. Στην εργασία [114] προτείνεται μια μέθοδος που βασίζεται στην τεχνική της εξόρυξης κανόνων συσχέτισης (association rules) με τη βοήθεια του αλγόριθμου a priori για την

απάντηση του ομηρικού ερωτήματος, αν δηλαδή η Ιλιάδα και η Οδύσσεια είναι κείμενα του ίδιου συγγραφέα. Η προσέγγιση εδώ έχει να κάνει με το συσχετισμό εννοιών και στα δύο έπη.

Η μέθοδος αυτή αρχίζει με τον έλεγχο των δυο ποιημάτων για να διαπιστωθεί ότι δεν υπάρχουν διαφορές στην ορθογραφία. Σε ένα δεύτερο επίπεδο προεπεξεργασίας αφαιρούνται από τα δύο κείμενα όλα τα σημεία της στίξης (απόστροφοι, διακριτικά, τελείες, κόμματα κ.τ.λ.), όλα τα σημεία των πνευμάτων (δασείες, βαρείες) καθώς και όλοι οι τόνοι, κάτι που διευκολύνει κατά πολύ τους υπολογισμούς. Μετά από αυτήν την μορφολογική ανάλυση τα δεδομένα αποτελούν δυο καθαρισμένες οντότητες από κείμενο.

Στην συνέχεια κάθε ένα από τα κείμενα διαιρείται σε ενότητες. Η διαίρεση αυτή γίνεται είτε σε κομμάτια της μιας παραγράφου, ακολουθώντας την διαίρεση των αρχικών κειμένων σε παραγράφους είτε, για την απόκτηση ακριβέστερων αποτελεσμάτων προτιμάται ο διαχωρισμός των κειμένων σε τμήματα που ορίζονται από τις προτάσεις των δυο ποιημάτων. Το σύνολο από όλα τα τμήματα κάθε έπους ορίζουν το σύνολο των δεδομένων. Το επόμενο βήμα συνίσταται στον ορισμό των κεντρικών εννοιών των λέξεων κλειδιών. Αυτή η διαδικασία γίνεται με την χρήση του γλωσσολογικού εργαλείου wordnet με το οποίο ανατίθεται ή όχι μια έννοια σε ένα τμήμα κειμένου ανάλογα αν εμφανίζεται εκεί ή όχι. Το τελευταίο βήμα της διαδικασίας συνίσταται στην εφαρμογή του αλγορίθμου apriori για την εξαγωγή των κανόνων συσχέτισης. Για την εφαρμογή του αλγορίθμου αυτού ως «συναλλαγές» (transactions) θεωρούνται τα κομμάτια στα οποία έχει χωριστεί κάθε έπος.

Οι κανόνες που θεωρούνται είναι της μορφής «90% των «συναλλαγών» οι οποίες περιέχουν την έννοια x περιέχουν επίσης και την έννοια y». Στο παρόν πλαίσιο ένας κανόνας συσχέτισης μας πληροφορεί γύρω από τη συσχέτιση μεταξύ δύο ή περισσότερων εννοιών. Για παράδειγμα στο 80% των περιπτώσεων που ο Όμηρος χρησιμοποιεί την έννοια πατρίδα χρησιμοποιεί επίσης και την έννοια ηλικία, κάτι που μας πληροφορεί για τη συσχέτιση των εννοιών πατρίδα και ηλικία. Το γεγονός αυτό παριστάνεται ως Πατρίδα => Ηλικία 80%.

Η ανάλυση των δυο ποιημάτων δίνει έναν σχετικά μεγάλο αριθμό ισχυρών συσχετίσεων (strong associations) που είναι όμοιες και στα δύο ποιήματα όπως παραδείγματος χάριν οι συσχετίσεις μεταξύ των εννοιών «πατρίδα» και «άντρας». Υπάρχουν επίσης ισχυρές συσχετίσεις που εμφανίζονται μόνο σε ένα από τα δυο ποιήματα όπως παραδείγματος χάριν η συσχέτιση των εννοιών «μάχη» και «άντρας» που εμφανίζεται μόνο στην Ιλιάδα. Υπάρχουν επίσης και ισχυρές συσχετίσεις που διαφέρουν από ποίημα σε ποίημα όπως παραδείγματος χάριν οι συσχετίσεις μεταξύ των εννοιών «ήρωας» και «μάχη» ή «ήρωας» και «οικία». Τα αποτελέσματα της ανάλυσης δείχνουν ότι απουσιάζουν εντελώς οι αντιφάσεις στις συσχετίσεις των εννοιών στα δυο έπη κάτι που δίνει μια ισχυρή ένδειξη ότι η Ιλιάδα και η Οδύσσεια έχουν γραφτεί από τον ίδιο άνθρωπο, τον Όμηρο.

6.2.7 Μελλοντικές εξελίξεις

Όσο το ποσό των κειμένων που είναι διαθέσιμα προς επεξεργασία από ηλεκτρονικού υπολογιστές θα αυξάνει, τόσο θα πρέπει να περιμένει κανείς ότι οι τεχνικές της

αυτοματοποιημένης αναγνώρισης προτύπων όπως παραδείγματος χάριν τα νευρωνικά δίκτυα, θα χρησιμοποιούνται όλο και συχνότερα ως βοηθήματα για την λύση του προβλήματος της πατρότητας των κειμένων. Αυτοματοποιημένοι εξαγωγείς χαρακτηριστικών θα αναπτύσσονται όπως στο [47] για να εξάγονται εκείνα τα χαρακτηριστικά που θα μπορέσουν καλύτερα να συνεισφέρουν στην διακριτοποίηση μεταξύ των κειμένων δύο διαφορετικών συγγραφέων. Θα υπάρξουν επίσης και θεωρητικές πρόοδοι όπως παραδείγματος χάριν το πέρασμα από λεξικογραφικές τεχνικές σε τεχνικές συντακτικού σχολιασμού (syntactic annotation) όπως προτάθηκε από τους Baayen, Van Halteren και Tweedie το 1996 [9].

6.3 Εφαρμογή

Η προσέγγισή μας όσο αφορά το authoring attribution πρόβλημα θα είναι από την σκοπιά της εξόρυξης δεδομένων. Πιο συγκεκριμένα θα χρησιμοποιήσουμε την διαδικασία ομαδοποίησης κατηγορικών δεδομένων που αναπτύξαμε λίγο πιο πάνω για αυτόν τον σκοπό. Θα προσπαθήσουμε να τμηματοποιήσουμε κατά τέτοιον τρόπο τα δεδομένα μας ώστε κάθε ομάδα να περιέχει τα κείμενα μόνο ενός συγγραφέα.

6.3.1 Επιλογή των Δεδομένων

Για την εφαρμογή αυτή επιλέχθηκαν σύντομα ποιήματα από την διεθνή λογοτεχνία. Για όσα ποιήματα δεν ήταν ελληνικά επιλέχθηκαν μεταφράσεις, κάτι που όπως θα δούμε δεν αποτελεί μειονέκτημα της μεθόδου αφού η προσέγγιση είναι εννοιολογική και όχι λεξικογραφική.

Από την ελληνική λογοτεχνία επιλέχθηκαν ποιήματα του σύγχρονου Θεσσαλονικιού ποιητή Ντίνου Χριστιανόπουλου ενώ από την διεθνή λογοτεχνία επιλέχθηκαν σύντομα ποιήματα των Ιαπώνων ποιητών Μπασό και Μπουσόν.

Η επιλογή των ποιητών και των ποιημάτων δεν έγινε καθόλου τυχαία. Τόσο ο Ντίνος Χριστιανόπουλος όσο και οι Ιάπωνες ποιητές έχουν αναπτύξει σύντομες φόρμες ποιημάτων. Από την μια μεριά οι Ιάπωνες έχουν αναπτύξει την παραδοσιακή πλέον σύντομη φόρμα ποιήματος τα χαϊκού, τα οποία είναι σύντομα αυστηρά δομημένα ποιήματα τριών στίχων και από την άλλη ο Χριστιανόπουλος έχει αναπτύξει και αυτός μια σύντομη ποιητική φόρμα η οποία όμως δεν είναι τόσο αυστηρή όσο αυτή των ιαπώνων. Παρ' όλα αυτά όμως μπορεί κανείς να βρει υποδόριες συγγένειες στον Χριστιανόπουλο και στα Ιαπωνικά χαϊκού. Το στοιχείο της έκπληξης και του ξαφνιάσματος μέσα από σύντομα ποιήματα είναι μια από τις κοινές ποιότητες τους. Ένα δεύτερο κοινό στοιχείο είναι η θεματολογία. Τα ιαπωνικά ποιήματα που επιλέχθηκαν έχουν ως κύρια πηγή έμπνευσής τους τον άνθρωπο και το περιβάλλον κάτι που παρατηρείται και στα ποιήματα του Χριστιανόπουλου αν και από διαφορετική σκοπιά. Αυτό σημαίνει ότι ενώ τα ιαπωνικά ποιήματα έχουν μια άμεση σχέση με την φύση και το φυσικό περιβάλλον τα ποιήματα του Χριστιανόπουλου έχουν κατά κύριο λόγο σχέση με την πόλη και τους ανθρώπους της.

Συνοψίζοντας μπορούμε να πούμε ότι τα ποιήματα που επιλέχθηκαν, επιλέχθηκαν λόγω της μικρής τους έκτασης, του γεγονότος ότι η ανάπτυξη τους γίνεται κατά κύριο λόγο με εικόνες και λόγω της θεματολογίας που όπως παρατηρήσαμε και παραπάνω κινείται γύρω από δύο κοινούς άξονες: την φύση και τον άνθρωπο. Αυτό το τελευταίο

γεγονός αποτελεί και την πρόκληση του όλου εγχειρήματος. Το γεγονός ότι τα ποιήματα που επιλέχθηκαν κινούνται σε κοινούς θεματικούς άξονες σε συνδυασμό με το γεγονός ότι η τεχνική ομαδοποίησης βασίζεται σε εννοιολογικές οντολογίες καθιστά τον διαχωρισμό των ποιημάτων όχι μόνο ένα μη τετριμμένο πρόβλημα άλλα θα μπορούσαμε να πούμε ότι του προσδίδει και έναν σημαντικό βαθμό δυσκολίας.

Συνολικά λοιπόν επιλέχθηκαν 77 ποιήματα. 23 ποιήματα του ιάπωνα ποιητή Μπασό, 25 ποιήματα του ιάπωνα ποιητή Μπουσόν και 29 ποιήματα του έλληνα ποιητή Χριστιανόπουλου.

6.3.2 Επιλογή των χαρακτηριστικών από τα ποιήματα.

Αφού λοιπόν επιλέχθηκαν τα ποιήματα ακολούθησε μια προεπεξεργασία τους πριν εισαχθούν ως είσοδος στον αλγόριθμο ομαδοποίησης.

Η προεπεξεργασία αυτή είχε ως σκοπό την εξαγωγή χαρακτηριστικών από τα ποιήματα ώστε να έρθουν στην κατάλληλη μορφή, για να μπορούν να υποστούν επεξεργασία από τον αλγόριθμο ομαδοποίησης. Ας σημειώσουμε εδώ ότι η διαδικασία της εξαγωγής των χαρακτηριστικών δεν έγινε από μια αυτοματοποιημένη διαδικασία άλλα «χειροκίνητα», αφού ο σκοπός μας δεν ήταν να αναπτύξουμε μια πλήρως αυτοματοποιημένη διαδικασία για την λύση του authoring attribution problem αλλά μια πρόταση για την λύση του. Θα πρέπει δηλαδή να γίνει ακόμα έρευνα για μια αυτοματοποιημένη εξαγωγή χαρακτηριστικών.

Παρ' όλα αυτά όμως, ας περιγράψουμε τον τρόπο με τον οποίο εξαγάγαμε τα χαρακτηριστικά από τα ποιήματα. Κατ' αρχάς επιλέξαμε να περιγράψουμε κάθε ποίημα με δύο χαρακτηριστικά. Το μικρό μέγεθος των ποιημάτων απαίτησε αυτήν την επιλογή. Προσπάθειες να περιγράψουμε τα ποιήματα με τρία ή και παραπάνω χαρακτηριστικά κατέληγαν συνήθως σε ένα σύνολο δεδομένων με πολλές ελλιπείς τιμές (missing values).

Μετά από πολλές διαφορετικές προσεγγίσεις καταλήξαμε στα δύο χαρακτηριστικά που θα μπορούσαν να περιγραφούν με τους εξής όρους: «Υποκείμενο» και «Περιβάλλον». Με τον όρο υποκείμενο εννοούμε το υποκείμενο του ποιήματος, αυτό δηλαδή που έχει τη δράση του ποιήματος. Με άλλα λόγια θα μπορούσαμε να περιγράψουμε αυτό το χαρακτηριστικό ως τον πρωταγωνιστή του ποιήματος. Στο χαϊκού παραδείγματος χάριν του Μπασό:

Τι ωραίο
για μια φορά το Φούτζι
χαμένο στην ομίχλη.

το «υποκείμενο» του ποιήματος είναι το Φούτζι, το γνωστό Ιαπωνικό βουνό. Το άλλο χαρακτηριστικό που το ονομάζουμε «περιβάλλον», περιγράφει το γενικό πλαίσιο μέσα στο οποίο εκτελείται η δράση του ποιήματος. Αυτό μπορεί να είναι κάποιο καιρικό φαινόμενο όπως εδώ στο παράδειγμά μας η ομίχλη, ή κάποια κατάσταση όπως μοναξιά ή ερημιά και ούτω καθεξής. Έτσι το παραπάνω χαϊκού συμπυκνώνεται στα εξής χαρακτηριστικά:

"ΒΟΥΝΟ ΦΟΥΤΖΙ";"ΟΜΙΧΛΗ"

Το σύνολο όλων των ποιημάτων με την μορφή των χαρακτηριστικών φαίνονται στο παράρτημα Γ. Αυτό είναι και το αρχείο που εισήχθηκε στον αλγόριθμο ομαδοποίησης, εκτός των επεξηγηματικών σχολίων που μας ενημερώνουν για το ποιος είναι ο ποιητής του συγκεκριμένου ποιήματος.

Ας σημειώσουμε εδώ ότι η αρχική σκέψη ήταν να περιγράψουμε τα ποιήματα με τρία χαρακτηριστικά το «Υποκείμενο», το «Αντικείμενο» και το «Περιβάλλον» του ποιήματος. Το «αντικείμενο» του ποιήματος θα ήταν αυτό που δεχόταν την δράση του υποκειμένου. Δηλαδή η δράση θα παράγεται από το «υποκείμενο» και το αποτέλεσμά της θα ήταν φανερό πάνω στο «αντικείμενο». Αυτή η προσέγγιση όμως δεν υιοθετήθηκε, γιατί αφ' ενός όπως στο παραπάνω παράδειγμα δεν υπάρχει το «αντικείμενο» ή απλώς υπονοείται και αφ'ετέρου σε άλλες περιπτώσεις ήταν εξαιρετικά ασαφές ποιος εκτελούσε την δράση και ποιος δέχονταν το αποτέλεσμά της. Έτσι αποφασίστηκε να χρησιμοποιηθεί μόνο το χαρακτηριστικό «υποκείμενο» το οποίο είναι ο πρωταγωνιστής του ποιήματος.

Αφού επιλέξαμε λοιπόν τα χαρακτηριστικά που θα χρησιμοποιηθούν θα πρέπει να τα οργανώσουμε σε οντολογίες, για να μπορέσουμε να τα χρησιμοποιήσουμε στον αλγόριθμο ομαδοποίησης. Οι οντολογίες φαίνονται στις εικόνες του παραρτήματος Ε.

Αυτό που θα πρέπει να παρατηρήσουμε σε αυτήν την διαδικασία κατασκευής των οντολογιών, είναι ότι ουσιαστικά έχει μικρή σημασία η εισαγωγή καινούργιων επεξηγηματικών κόμβων στην οντολογία. Αυτό το γεγονός οφείλεται στην ιδιότητα της κανονικοποίησης της απόστασης. Όλες οι πλήρης ιεραρχίες από την ρίζα ως κάποιο φύλλο, έχουν περίπου το ίδιο μήκος, με αποτέλεσμα και τα μεγάλα μονοπάτια και τα μικρά μονοπάτια (λεπτομερείς και λιγότερο λεπτομερείς ιεραρχίες) να είναι ουσιαστικά ισοδύναμα. Έτσι οι επεξηγηματικοί κόμβοι καλό είναι να μπαίνουν για την καλύτερη παρουσίαση των οντολογιών, αλλά και για την διευκόλυνση των χρηστών που τις κατασκευάζουν.

Στα σχήματα των οντολογιών βλέπουμε καθαρά την κατανομή των κόμβων στα διάφορα υποδέντρα.

Τα αποτελέσματα του αλγορίθμου ομαδοποίησης φαίνονται αναλυτικά στο παράρτημα Δ. Εκεί φαίνονται καθαρά τα αποτελέσματα των ομαδοποιήσεων για 2,3,…,10 ομάδες.

6.3.3 Authoring attribution problem

Για την λύση του Authoring Attribution Problem όπως αναφέραμε παραπάνω εκτελούμε τον αλγόριθμο ομαδοποίησης με την ελπίδα να μας επιστρέψει σε διαφορετικές ομάδες τα κείμενα των διαφορετικών συγγραφέων.

Η ιδέα αυτής της αντιμετώπισης του authoring attribution problem βασίζεται στην υπόθεση ότι διαφορετικοί συγγραφείς θα πραγματεύονται διαφορετικές έννοιες ή και διαφορετικές ομάδες εννοιών. Σύμφωνα με αυτήν την θεώρηση όμως, θα μπορούσε κανείς να υποθέσει ότι θα ήταν πολύ εύκολο να ξεχωρίσει ο αλγόριθμος τα ελληνικά

ποιήματα από τα γιαπωνέζικα. Αυτό όμως δεν είναι αλήθεια, γιατί όπως εξηγήσαμε παραπάνω υπάρχει μεγάλη συγγένεια και ως προς την μορφή, αλλά και ως προς το περιεχόμενο των κειμένων του Χριστιανόπουλου και των Ιαπώνων, πράγμα που καθιστά τον διαχωρισμό αυτό μη τετριμμένη διαδικασία. Από την άλλη πλευρά, όπως φαίνεται πολύ καθαρά στα σχήματα των οντολογιών, όπου έχουν σημειωθεί με διαφορετικά χρώματα οι κόμβοι που αντιστοιχούν σε έννοιες που ανήκουν σε διαφορετικούς ποιητές, αν και υπάρχει κάποιος διαχωρισμός στους κόμβους αυτούς, εντούτοις παρατηρείται το φαινόμενο μερικοί από τους κόμβους να βρίσκονται διασπαρμένοι σε περιοχές των οντολογιών που ανήκουν σε άλλον συγγραφέα.

6.3.3.1 2 Ομάδες

Ο αλγόριθμος ομαδοποίησης όταν του ζητήθηκε να επιστρέψει δυο ομάδες, επέστρεψε στη μια ομάδα όλα ανεξαιρέτως τα ποιήματα του Χριστιανόπουλου και στην άλλη ομάδα, τα ποιήματα των Ιαπώνων Μπασό και Μπουσόν. Αυτή είναι μια επιθυμητή ομαδοποίηση. Όπως αναφέραμε και παραπάνω αν και τα ποιήματα του Χριστιανόπουλου αναφέρονται κυρίως στον άνθρωπο σε σχέση με το αστικό περιβάλλον, τα ποιήματα των ιαπώνων αναφέρονται, στον άνθρωπο μέσα στο φυσικό περιβάλλον ή απλώς στο φυσικό περιβάλλον, εντούτοις υπάρχουν και ποιήματα του Χριστιανόπουλου που αναφέρονται στο φυσικό περιβάλλον ή σε χρονικές περιόδους όπως «νύχτα» ή «βράδυ γλυκό». Έτσι θα περίμενε κανείς αυτά τα ποιήματα, τα οποία αποτελούν το 25% του συνολικού αριθμού των ποιημάτων του Χριστιανόπουλου να ομαδοποιηθούν μαζί με τα ιαπωνικά ποιήματα. Παρόλα αυτά όμως ο αλγόριθμός μας τα κατέταξε ορθά.

6.3.3.2 3 ομάδες

Όταν ζητήθηκε από αλγόριθμο να επιστρέψει 3,4,5,…,10 ομάδες τότε αυτός επέστρεψε τα ποιήματα ομαδοποιημένα σε νοηματικές ενότητες. Θα σχολιάσουμε παρακάτω τις ομάδες που επέστρεψε ο αλγόριθμος σε κάθε περίπτωση ξεχωριστά.

Θα περιμέναμε ίσως, ζητώντας από τον αλγόριθμο να μας επιστρέψει τρεις ομάδες, να ομαδοποιήσει τα ποιήματα των τριών διαφορετικών ποιητών σε διαφορετικές ομάδες. Ο αλγόριθμός μας όμως δεν πραγματοποιεί αυτόν το διαχωρισμό. Το γεγονός αυτό οφείλεται κατά κύριο λόγο στην συνάφεια και στην πάρα πολύ μεγάλη συγγένεια που υπάρχει μεταξύ των Ιαπώνων ποιητών Μπασό και Μπουσόν. Τα ποιήματα και των δυο αυτών ποιητών έχουν ως εννοιολογικούς άξονες την φύση και το φυσικό περιβάλλον. Θα μπορούσαμε να πούμε ότι τα ποιήματα των Ιαπώνων είναι πολύ πιο κοντά μεταξύ τους εννοιολογικά, από ότι τα ποίηματα των Ιαπώνων από αυτά του Χριστιανόπουλου πράγμα που τα καθιστά πιο δύσκολα διαχωρίσιμα μεταξύ τους. Θα μπορούσε να υποστηρίξει κανείς ότι τα ποιήματα των Ιαπώνων, μπορούν να διαχωριστούν σε ομάδες, μία για κάθε διαφορετικό ποιητή, εάν δημιουργούσαμε μια λίγο διαφορετική οντολογία απ’ όπου θα απουσίαζαν τα ποιήματα του Χριστιανόπουλου και έτσι οι διαφορετικές έννοιες που χρησιμοποιούν οι δυο ποιητές να βρίσκονται κάπως διαχωρισμένες στα δέντρα των οντολογιών.

Η πρώτη ομάδα που μας επέστρεψε ο αλγόριθμος είναι μια μεγάλη ομάδα που περιέχει ποιήματα και των τριών ποιητών. Αυτή η ομάδα αποτελείται από δυο

μικρότερες υποομάδες. Η πρώτη από αυτές τις υποομάδες περιλαμβάνει τα ποιήματα των ιαπώνων που αναφέρονται στην φύση και μάλιστα αυτά που αναφέρονται κυρίως σε έμβιες μορφές της φύσης όπως: λουλούδια, ζώα, έντομα, καθώς και σε κάποια στοιχεία της φύσης. Αυτή η υποομάδα περιλαμβάνει τα ποιήματα στα οποία η φύση πρωταγωνιστεί, ως υποκείμενο αλλά και περιβάλλει τα ποιήματα ως περιβάλλον. Η άλλη υποομάδα περιλαμβάνει τα ποιήματα του Χριστιανόπουλου που αναφέρονται στον άνθρωπο και πιο συγκεκριμένα στο ανθρώπινο σώμα με έννοιες όπως «Κορμί», «Σκέλια» κ.τ.λ. Τα ποιήματα αυτά του Χριστιανόπουλου αναφέρονται κατά κύριο λόγο στο φυσικό περιβάλλον με έννοιες όπως «Χώμα», «Πάρκο» ή «Παλίρροια» ή σε καταστάσεις χρόνου όπως «Νύχτα» ή «Βράδυ Γλυκό». Αυτές είναι έννοιες που χρησιμοποιούνται κυρίως από τους Ιάπωνες ποιητές και εδώ βρίσκουμε την συνάφεια των ποιημάτων του Χριστιανόπουλου με αυτά των Ιαπώνων ποιητών και για αυτόν το λόγο ομαδοποιούνται μαζί. Θα μπορούσε κάποιος να περιγράψει στο σύνολο της, την ομάδα ως τα ποιήματα που έχουν σχέση με την φύση και τον άνθρωπο.

Η δεύτερη ομάδα που μας έδωσε ως αποτέλεσμα ο αλγόριθμος περιλαμβάνει εκείνα τα ποιήματα του Χριστιανόπουλου, όπου πρωταγωνιστής ως υποκείμενο είναι το σώμα. Αλλά αυτά αναφέρονται σε καταστάσεις όπως μοναξιά, ερημιά ή έρωτας και έξαψη. Τα ποιήματα αυτά ομαδοποιήθηκαν ξεχωριστά από τα παραπάνω αφ' ενός γιατί δεν αναφέρονται με κανέναν τρόπο στην φύση και αφετέρου γιατί αποτελούν μια ομάδα ξεχωριστή από μόνα τους.

Η τρίτη ομάδα περιλαμβάνει τα ποιήματα εκείνα που αναφέρονται στα καιρικά φαινόμενα σε σχέση με τον άνθρωπο. Το χαρακτηριστικό «Αντικείμενο» των ποιημάτων αυτής της ομάδας αποτελείται από έννοιες που κατά κύριο λόγο αναφέρονται στον άνθρωπο. Το 70% των περιπτώσεων που ανήκουν σε αυτή την ομάδα αναφέρονται στον άνθρωπο. Από την άλλη πλευρά το 78% των περιπτώσεων διαθέτουν χαρακτηριστικό «Περιβάλλον» που αναφέρεται στα καιρικά φαινόμενα. Είναι αξιοσημείωτο ότι όλα τα ιαπωνικά ποιήματα αυτής της ομάδας αναφέρονται σε καιρικά φαινόμενα. Σε αυτήν την ομάδα περιλαμβάνονται και κάποια ποιήματα του Χριστιανόπουλου που αναφέρονται στον υπόκοσμο με φόντο το αστικό περιβάλλον. Όπως παραδείγματος χάριν «ΓΥΝΑΙΚΑ», «ΔΡΟΜΟΣ». Έτσι και εδώ βλέπουμε ότι η ομάδα αυτή αποτελείται από δυο μικρότερες υποομάδες που μεταφράζονται στα ποιήματα του Χριστιανόπουλου και σε αυτά των Ιαπώνων ποιητών.

Βλέπουμε ότι ο αλγόριθμος μας επέστρεψε τρεις ομάδες οι οποίες δεν είναι καθαρές με την έννοια ότι διαθέτουν μέσα τους και άλλες υποομάδες. Για αυτόν τον λόγο θα συνεχίσουμε να ζητάμε από τον αλγόριθμο να ομαδοποιήσει τα δεδομένα σε περισσότερες ομάδες για να έχουμε μια καλύτερη και λεπτομερέστερη εικόνα των ομάδων που υπάρχουν στα δεδομένα μας.

6.3.3.3 4 Ομάδες

Σε αυτήν την περίπτωση ο αλγόριθμος ανακάλυψε τις δυο υποομάδες που υπήρχαν στην δεύτερη ομάδα της προηγούμενης περίπτωσης δημιουργώντας έτσι δυο καινούργιες ομάδες. Η πρώτη από τις οποίες αναφέρεται στο ανθρώπινο σώμα σε σχέση με καταστάσεις έρωτα όπως για παράδειγμα «Έξαψη», «Έκσταση», «Έρωτας» ενώ η δεύτερη από τις καινούργιες ομάδες αναφέρεται στο ανθρώπινο σώμα αλλά σε σχέση με καταστάσεις μοναξιάς, όπως «Μοναξιά» και «Ερημιά». Μπορούμε να

πούμε ότι ο αλγόριθμός μας ανακάλυψε δυο καθαρές ομάδες, οι οποίες δεν επιδέχονται περαιτέρω ανάλυση κάτι που βλέπουμε να είναι και διαισθητικά σωστό.

Οι άλλες ομάδες ($1^{η}$ και $3^{η}$ ομάδα στην προηγούμενη περίπτωση) μας επιστράφηκαν ως έχουν, χωρίς καμία διαφορά.

6.3.3.4 5 ομάδες

Οι δυο πρώτες ομάδες, που επιστράφηκαν από τον αλγόριθμο αυτήν την φορά ήταν οι ομάδες για τις οποίες μιλήσαμε παραπάνω.

Η Τρίτη ομάδα που επιστράφηκε είναι μια ομάδα όπου το 81% των ποιημάτων,περιέχει το χαρακτηριστικό «Αντικείμενο» αναφέρεται εξ' ολοκλήρου στο φυσικό περιβάλλον χωρίς καμία αναφορά στον άνθρωπο, ενώ το 95% των ποιημάτων αυτής της ομάδας διαθέτει χαρακτηριστικό «περιβάλλον» που είναι εντελώς ανεξάρτητο από τον άνθρωπο, αναφέρεται στο φυσικό περιβάλλον ή σε χρονικές περιόδους. Θα μπορούσαμε να πούμε ότι αυτή η ομάδα είναι που αναφέρεται σε ποιήματα που έχουν να κάνουν με το φυσικό περιβάλλον. Από μια άλλη οπτική γωνία θα μπορούσαμε να πούμε ότι αυτή η ομάδα αποτελείται από δυο υποομάδες που ορίζονται από τα ποιήματα των Ιαπώνων και του Χριστιανόπουλου, που ανήκουν σ' αυτήν, αντίστοιχα. Τα ποιήματα του Χριστιανόπουλου αναφέρονται στο ανθρώπινο κορμί άλλα, σε σχέση με κάποια σκοτεινή ώρα του εικοσιτετραώρου. Αυτό αποτελεί και τον συνδετικό κρίκο αυτών των ποιημάτων με τα ιαπωνικά που ανήκουν σε αυτήν την ομάδα, παρόλου που τα πρώτα αποτελούν μια ξεχωριστή ενότητα.

Το ενοποιητικό στοιχείο των ποιημάτων της τέταρτης ομάδας που μας επέστρεψε ο αλγόριθμος είναι το χαρακτηριστικό «Περιβάλλον» που αναφέρεται εξ' ολοκλήρου στα καιρικά φαινόμενα αφ' ενός και αφ' εταίρου, είναι όλα τα ποιήματα των Ιαπώνων ποιητών. Από την άλλη πλευρά το 61% των ποιημάτων διαθέτουν χαρακτηριστικό «Υποκείμενο» που αναφέρεται έμμεσα ή άμεσα στον ανθρώπινο παράγοντα. Έτσι, μπορούμε να πούμε ότι αυτή η ομάδα αποτελείται από εκείνα τα ποιήματα που αναφέρονται στον άνθρωπο, κυρίως έμμεσα, σε σχέση με τα καιρικά φαινόμενα, κάτι που κάνει την ομάδα αυτήν να διαχωρίζεται εννοιολογικά από την τρίτη ομάδα από την οποία ο ανθρώπινος παράγοντας σχεδόν απουσιάζει.

Η πέμπτη,μ ομάδα αποτελείται εξ' ολοκλήρου από ποιήματα του Χριστιανόπουλου. Εδώ μπορούμε να διακρίνουμε δυο υποομάδες. Αυτά τα ποιήματα που αναφέρονται στο ανθρώπινο σώμα και αυτά που αναφέρονται σε ανθρώπινες οντότητες του υπόκοσμου όπως «Χαφιές» ή «Τσογλάνι». Το ενοποιητικό στοιχείο όλων αυτών των ποιημάτων είναι το γεγονός ότι το χαρακτηριστικό «Περιβάλλον» του συνόλου των ποιημάτων αναφέρεται σε ένα έντονο αστικό περιβάλλον.

6.3.3.5 6 ομάδες

Σε αυτήν την περίπτωση η ομαδοποίηση που μας επέστρεψε ο αλγόριθμος διαχωρίζει την πέμπτη ομάδα της προηγούμενης περίπτωσης σε δυο υποομάδες. Στην πρώτη από αυτές τις ομάδες το 70% των ποιημάτων το χαρακτηριστικό «Υποκείμενο» αναφέρεται στο ανθρώπινο σώμα ενώ μόνο το 30% των περιπτώσεων αναφέρεται σε

ανθρώπινες οντότητες του υποκόσμου. Από την άλλη πλευρά το χαρακτηριστικό «Περιβάλλον» των ποιημάτων αυτής της ομάδας αναφέρονται σε αστικούς χώρους. Το 70% αναφέρεται σε εξωτερικούς αστικούς χώρους όπως «Δρόμος» ή «Πάρκο» ενώ το 30% αναφέρεται σε εσωτερικούς αστικούς χώρους όπως «Πάτωμα» ή «Ντιβάνι». Έτσι μπορούμε να πούμε ότι αυτή η ομάδα περιλαμβάνει ποιήματα που αναφέρονται στον άνθρωπο (ως σώμα ή ως οντότητα) σε σχέση με τους αστικούς χώρους (εσωτερικούς ή εξωτερικούς).

Στην δεύτερη από τις ομάδες που αποτελούσαν την πέμπτη ομάδα της προηγούμενης περίπτωσης το ενοποιητικό στοιχείο των ποιημάτων είναι το χαρακτηριστικό «Περιβάλλον» το οποίο εξ' ολοκλήρου αναφέρεται σε μαγαζιά όπως «Κουρείο», «Σινεμά» ή «Καφενείο». Και σε αυτήν την περίπτωση το χαρακτηριστικό «Υποκείμενο» αναφέρεται στον άνθρωπο είτε ως σώμα είτε ως οντότητα.

6.3.3.6 7 ομάδες

Το καινούργιο στοιχείο που εισάγεται με αυτήν την ομαδοποίηση, είναι ο χωρισμός της ομάδας που περιγράψαμε στην προηγούμενη περίπτωση σε δύο ομάδες. Η πρώτη από αυτές της ομάδες περιλαμβάνει τα ποιήματα όπου το χαρακτηριστικό «Περιβάλλον» αναφέρεται σε εξωτερικούς αστικούς χώρους, όπως «Δρόμος» ή «Χώμα» ή «Πάρκο». Το χαρακτηριστικό «Υποκείμενο» των ποιημάτων αυτής της ομάδας αναφέρεται κατά 57% στο ανθρώπινο κορμί ενώ κατά 43% σε ανθρώπινες οντότητες του υποκόσμου.

Η δεύτερη από αυτές τις ομάδες, περιέχει εκείνα τα ποιήματα που το χαρακτηριστικό τους «Περιβάλλον» αναφέρεται σε εσωτερικούς αστικούς χώρους όπως «Πάτωμα», «Σκάλες» ή «Ντιβάνι». Το χαρακτηριστικό «Υποκείμενο» σε αυτήν την ομάδα αναφέρεται εξ' ολοκλήρου στο ανθρώπινο σώμα. Θα μπορούσε κανείς να πει ότι η ομάδα αυτή περιλαμβάνει τα ποιήματα εκείνα που αναφέρονται σε οικιακούς (αστικούς εσωτερικούς χώρους) σε σχέση με το ανθρώπινο κορμί.

6.3.3.7 8 ομάδες

Σε αυτήν την περίπτωση είχαμε μια κάπως διαφορετική διαμέρηση των δεδομένων. Η πρώτη ομάδα αναφέρεται στο ανθρώπινο σώμα σε σχέση με καταστάσεις όπως η ερημιά, η έξαψη και ο έρωτας. Εδώ βλέπουμε ότι το ποίημα που αναφέρεται στην ερημιά ομαδοποιήθηκε σε αυτήν την ομάδα, ενώ σε όλες τις προηγούμενες φορές είχε ομαδοποιηθεί μαζί με τα ποιήματα που αναφέρονταν στην μοναξιά. Όλα τα ποιήματα αυτής της ομάδας είναι ποιήματα του Χριστιανόπουλου.

Η δεύτερη ομάδα αναφέρεται καθαρά στην κατάσταση της μοναξιάς. Το χαρακτηριστικό «Υποκείμενο» των ποιημάτων αυτής της ομάδας αναφέρεται και πάλι εξ' ολοκλήρου στο ανθρώπινο σώμα. Όλα τα ποιήματα αυτής της ομάδας είναι ποιήματα του Χριστιανόπουλου.

Η τρίτη ομάδα αναφέρεται σε νυχτερινά περιβάλλοντα όπως «Φθινοπωρινό Βράδυ» ή «Νύχτα» ενώ το αντικείμενο αυτών των κειμένων αναφέρεται σε μέρη του σώματος αλλά και σε καταστάσεις όπως η φτώχεια ή η μοναξιά. Το ενοποιητικό στοιχείο των

ποιημάτων αυτής της ομάδας είναι σαφώς οι σκοτεινές ώρες του εικοσιτετραώρου στις οποίες αναφέρονται εξ' ολοκλήρου όλα τα ποιήματα αλλά και το γεγονός ότι το χαρακτηριστικό «Αντικείμενο» των ποιημάτων αναφέρεται στον άνθρωπο είτε άμεσα όπως (ανθρώπινο σώμα ή ανθρώπινη οντότητα) είτε έμμεσα (όπως παραδείγματος χάριν ανθρώπινες καταστάσεις όπως «Φτώχεια» ή «Μοναξιά»). Θα πρέπει να αναφέρουμε ότι το 62% των ποιημάτων, μια σημαντική πλειοψηφία αναφέρεται στο ανθρώπινο σώμα.

Η τέταρτη ομάδα αναφέρεται στα καιρικά φαινόμενα και την έχουμε περιγράψει και παραπάνω. Το υποκείμενο των ποιημάτων αυτής της ομάδας αναφέρεται κατά κύριο λόγο (62% των ποιημάτων) σε φυσικές οντότητες που έχουν άμεση («Ομπρέλα», «Σανδάλι») ή έμμεση (όπως «Βεράντα», «Σκιάχτρο», «Ρυζοχώραφο» ή «Η Γέφυρα του Σέτα») σχέση με τους ανθρώπους. Όλα τα ποιήματα αυτής της ομάδας είναι ιαπωνικά.

Η πέμπτη ομάδα συγγεντρώνει τα ποιήματα που αναφέρονται στο αστικό περιβάλλον (εξωτερικό ή εσωτερικό) σε σχέση όμως με τον άνθρωπο είτε ως οντότητα είτε ως ανθρώπινο σώμα. Όλα τα ποιήματα αυτής της ομάδας είναι ποιήματα του Χριστιανόπουλου.

Η έκτη ομάδα αναφέρεται στον άνθρωπο (είτε ως σώμα, είτε ως οντότητα) αλλά σε σχέση αυτήν την φορά με μαγαζιά της πόλης όπως «Καφενείο». Όλα τα ποιήματα αυτής της ομάδας είναι ποιήματα του Χριστιανόπουλου.

Η έβδομη περιλαμβάνει μόνο το ποίημα «Καθένας», «Παλίρροια» το οποίο ομολογουμένως αποτελεί από μόνο του μια ομάδα, αφού νοηματικά δεν ταιριάζει με κανένα από τα άλλα ποιήματα και με καμία από τις άλλες ομάδες διότι αναφέρεται στον άνθρωπο ως οντότητα, «Καθένας», σε σχέση με το φυσικό περιβάλλον, «Παλίρροια».

Τέλος η όγδοη ομάδα αναφέρεται στην φύση τόσο ως υποκείμενο όσο και ως περιβάλλον. Ο συνδετικός κρίκος των ποιημάτων αυτής της ομάδας είναι αφενός ότι και στα δυο χαρακτηριστικά αυτών των ποιημάτων, δεν γίνεται καμία, ούτε άμεση αλλά ούτε και έμμεση αναφορά στον άνθρωπο και αφ' εταίρου ότι όλα τα ποιήματα αυτά είναι ιαπωνικά.

6.3.3.8 9 ομάδες

Η αλλαγή σε αυτήν την διαμέριση, είναι η αποκοπή του ποιήματος που παριστάνεται από τα χαρακτηριστικά «Όποιος δεν μπορεί να γράψει», «Ανοιξιάτικη Βροχή» από το κυρίως σώμα των ποιημάτων, που αποτελούν τα ποιήματα που αναφέρονται στα καιρικά φαινόμενα. Το ποίημα αυτό δικαιολογημένα αποτελεί μια ξεχωριστή ομάδα ,εφόσον αναφέρεται σε κάποια ανθρώπινη οντότητα «Όποιος δεν μπορεί να γράψει», ενώ όλα τα άλλα ποιήματα αυτής της ομάδας, αναφέρονται καθαρά σε οντότητες της φύσης, χωρίς καμία άμεση ή έμμεση αναφορά στον άνθρωπο. Το ποίημα αυτό αναφέρεται σε μια ανθρώπινη οντότητα, σε σχέση με τα καιρικά φαινόμενα.

6.3.3.9 10 ομάδες

Το καινούργιο στοιχείο που εισάγει αυτή η ομαδοποίηση, είναι ο χωρισμός της ομάδας που αναφέρεται στα καιρικά φαινόμενα σε δυο ομάδες.

Η πρώτη ομάδα, αναφέρεται σε καιρικά φαινόμενα της άνοιξης ή του καλοκαιριού όπως παραδείγματος χάριν «Βροχή του Μάη» ή «Ξαστεριά». Η δεύτερη ομάδα αναφέρεται κυρίως σε καιρικά φαινόμενα του φθινοπώρου και του χειμώνα όπως «Βροχές» ή «Ομίχλη» ή «Αρχές Φθινοπώρου».

6.3.4 Συμπεράσματα

Θα μπορούσαμε να σχολιάσουμε ένα σημείο σχετικά με την πορεία ομαδοποίησης που περιγράψαμε παραπάνω. Παρατηρήσαμε ότι ενώ σε κάποιες διαδοχικές ως προς τον αριθμό των ομάδων ομαδοποιήσεις, ο αλγόριθμος αυτό που έκανε ήταν να εκλεπτύνει την διαμέριση, διαχωρίζοντας μια ομάδα σε δύο μικρότερες ομάδες εντούτοις υπήρχαν και μερικές φορές που ο αλγόριθμος άλλαζε σε μεγάλο βαθμό την ομαδοποίηση περνώντας από μια μικρότερη σε μια μεγαλύτερη ομαδοποίηση. Αυτό το γεγονός μπορεί να ερμηνευτεί ως εξής: Το σύνολο των δεδομένων μας είναι εξαιρετικά δύστροπο. Αποτελείται κατ' αρχάς από πάρα πολύ λίγα αντικείμενα. Από την άλλη πλευρά θα μπορούσε να πει κανείς ότι τα δεδομένα αυτά επιδέχονται διαφορετικές ομαδοποιήσεις. Έτσι ο αλγόριθμος περνώντας σε όλο και πιο εκλεπτυσμένες ομαδοποιήσεις μερικές φορές ανακαλύπτει διαμερίσεις που να είναι λίγο διαφορετικές από τις προηγούμενες, χωρίς αυτό να σημαίνει ότι αυτό είναι ένα μειονέκτημα της μέθοδου. Αυτό οφείλεται κατά κύριο λόγο στην ιδιαιτερότητα των δεδομένων μας.

Σαν ένα γενικό συμπέρασμα, θα μπορούσαμε να αναφέρουμε, τον σχετικά καλό διαχωρισμό των ποιημάτων, σε νοηματικές ενότητες λαμβάνοντας πάντα υπό όψιν το μικρό μέγεθος των ποιημάτων, που ομαδοποιήθηκαν (77), καθώς και ότι δεν υπήρχε εκ των προτέρων μία ομαδοποίηση των ποιημάτων αυτών. Τα ποιήματα επιλέχθηκαν σχεδόν τυχαία από ένα ευρύτερο σύνολο ποιημάτων των ίδιων ποιητών. Θα πρέπει να λάβουμε επίσης υπ' όψιν την αρκετά μεγάλη διασπορά των εννοιών στις οντολογίες καθώς και το μεγάλο αριθμό των διαφορετικών εννοιών που παρουσιάζονται στα ποιήματα, πράγμα που δυσχεραίνει πολύ την διαδικασία ομαδοποίησης, ακόμα και για έναν άνθρωπο.

Θα μπορούσαμε να πούμε δηλαδή ότι ο τρόπος με τον οποίο ομαδοποιήθηκαν τα δεδομένα μας όχι μόνο δεν είναι τετριμμένος, άλλα μπόρεσε να ανακαλύψει πραγματικές εννοιολογικές ομάδες κειμένων μέσα από ένα αντίξοο και σχετικά μικρό σύνολο δεδομένων. Το γεγονός αυτό δίνει μια ακόμα ένδειξη για την στιβαρότητα και την ακρίβεια της μεθόδου ομαδοποίησης κατηγορικών δεδομένων που προτείναμε και μάλιστα στο πεδίο της εξόρυξης κειμένου.

[1] Aldenderfer S. Mark, Blashfield K. Roger, “Cluster Analysis, Quantitative Applications in the Social Sciences”, SAGE Publications, Inc., 1984

[2] Almuallim H., T.G. Dietrich, Learning with many irrelevant features, In Proceedings of the 9th National Conference on Artificial Intelligence and Information Retrieval, pp. 547-552, 1991.

[3] Alsabti Khaled, Ranka Sanjay, Singh Vineet. An efficient k – means Clustering Algorithm. In IPPS: 11th International Parallel Processing Symposium, 1998.

[4] Anderberg M.R. “Cluster Analysis for Applications”. Academic Press, New York,1973.

[5] Apte D., Damareau F., Weiss S.M., Text Mining With Decision Rules and Decision Trees, working Notes of learning from text and the web, Conf. an automated Learning and Discovery CONLAD-98, 1998.

[6] Apte D., Damareau F., Weiss S.M., Toward Language Independent Automated Learning of Text Categorization Model, Proc. Of the 7th Annual International ACM-SIRGIR Conference on Reasearch and Development in Information Retrieval Dublin, 1994.

[7] Armstrong R., Freitag D., Joachims T., Mitchell T., WebWatcher : A learning Apprentice for the WWW, AAAI 1995 Spring Symposium on Information Gathering from Heterogeous, Distributed Environments, Stanford, March 1995.

[8] Autonomy, http://www.autonomy.com/autonomyv3/, October 2001.

[9] Baayen H., Van Halteren H., Twe4edie F.J., Outside the Cave of Shadows: UsingSyntactic Annotation to Enhance Authoship Attribution, Literary and Lingustic Computring, 2, pp. 61-70, 1987.

[10] Balabanovic M., Shoham Y., Learning Information Retrieval Agents: Experiments with Automated Web Browsing, AAAI 1995 Spring Symposium on Information Gathering from Heterogeous, Distributed Environments, Stanford, March 1995.

[11] Bartell B.T., Cottrell G.W., Belew R.K., Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling, Proceedings of the ACMSIG Information Retrieval, Copenhagen, 1992.

[12] Berkhin Pavel. Survey of Clustering Data Mining Techniques. Technical report, Accrue Software, 2002.

[13] Berners – Lee Tim, Hendler James and Lassila Ora, “The semantic Web”, Scientific American, May 2001.

[14] Berry J.A. Michael, Linoff Gordon. Data Mining Techniques. Technical report, Accrue Sofware, 2002

[15] Berry M.W., Dumais S.T., OBrein G.W., Using Linear Algebra for intelligent information retrieval. Siam Review, Vol. 37, No. 4, pp. 573-595, December 1995.

[16] Bock H.H., E. Diday, Analysis of Symbolic Data. Exploratory Methods for Extractiong Statistical Information form Complex Data. Series: Studies in Classifiacation, Data Analysis and Knowledge Organization, Vol. 15, Springer – Verlag, 2000.

[17] B. Boutsinas, D. Tasoulis, M.N. Vrahatis, "An efficient algorithm for Estimating the number of clusters using a windowing technique", Pattern Recognition and Image Analysis, to appear, 16, 2006.

[18] B. Boutsinas, T. Gnardellis, "On Distributing the clustering process", Pattern Recognition Letters, Elsevier Science Publishers B.V., vol. 23, no. 8, 2002, Impact factor: 409/2002, pp. 999-1008.

[19] B. Boutsinas, G. Prassas, G. Antzoulatos, "A methodology for scaling up classification algorithms", International Journal on Artificial Intelligence Tools, 13(3), World Scientific Publishing Company, pp. 623-639, 2004.

[20] Boutsinas B., T. Papastergiou, On Clustering Categorical Objects, submitted in 2005.

[21] Buckley C. and A.F. Lewit, Optimizations of inverted vector searches, In SIGIR '85, pp. 97-100, 1985.

[22] Burrows J.F., Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style, Literary and Linguistic Comp[utingt, 2, pp. 62-70, 1987

[23] Chan K.P., Stolfo J.S., On the Accuracy of meta – Learning for Scalable Data Mining. Journal of Intelligent Information Systems, 8: p. 5 – 28, 2001.

[24] Chan K.P., Stolfo J.S., Toward Parallel and Distributed Learning by meta – Learning. Working Notes AAAI Work. Knowledge Discovery in Databases, p. 227 – 240, 1993.

[25] Chenga V., C.H. Li, J.T. Kwok, C,K. Li, Dissimilarity learning for Nominal Data, Pattern Recognition, in Press.

[26] ClearForest, http://www.clearforest.com/index.asp, November 2001.

[27] Cohen W.W., Learning English Etxt whith ILP methods, Workshop on Inductive Logic Programming, Leuevn, September 1995.

[28] Cost S., S. Salzberg, A weighted Nearest Algorithm for Learning with Symbolic Features, Machine Learning, 10, pp. 57-78, 1993.

[29] Cover T. and P. Hart, Nearest Neighbor Pattern Classification, IEEE Transactions on Information Theory, 13(1), pp.57-78, 1993.

[30] Cox T. F., M.A.A. Cox, Multidimensional Scaling, Chapman & Hall, 1994.

[31] Creecy R. M., Masand B. M., Smith S. J., Waltz D. L. Trading MIPS and Memory of Knowledge Engineering, Communications of the ACM Vol. 35, No. 8, pp. 48-64, August 1992.

[32] Cutting D.R., D.R. Karger, J.O. Pedersen, J.W. Turkey, Scatter/gather: A cluster-based approach to browsing large document collections. In Proc. Of the 1th Annual International ACM SIGIR Conf. On Research and Development in Information Retrieval, Copenhagen, Denmark, pp. 318-329, June 1992.

[33] Dewdney, A.K. "Computer Recreations", Scientific American, January 1986, p.16-25

[34] Dillon R. William, Goldstein Matthew. "Multivariate Analysis Methods and Applications". John Wiley & Sons, Inc., 1984.

[35] dtSearch http://dtsearch.com/dtsoftware.html#anchor412454 , November 2001.

[36] Dubes C.R. Cluster Analysis and Relates Issues. Handbook of Pattern Recognition. Berlin: Spinger – Verlag, 1996.

[37] Dubes R.C., A. K. Jain, Clustering Methodologies in exploratory data analysis, Adv. Comput., 19, 1980, pp.113-228.

[38] Dubes Richard, Jain K. Anil. "Validity Studies in Clustering Methodologies". Pattern Recognition, 11 p.235-254, March 1979.

[39] Duda O. R., Hart E. P. "Pattern Classification and Scene Analysis". John Wiley & Sons, 1973.

[40] Dunn – Rankin P., Scaling Methods, Lautence Erlbaum Associates, 1983.

[41] Esposito F., D. Malebra, G. Semerano, Classification in Noisy Environments Using a Distance Measure Between Structural Symbolic Descriptions, IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(3), pp. 390-402, 1992.

[42] Everitt B.S. "Cluster Analysis", John Wiley & Sons, Inc: New York 1974

[43] Fayyad M. Usama, Piatetsky-Shapiro Gregory, Smyth Padjraic, Uthurusamy Ramasamy. "Advances in Knowledge Discovery and Data Mining". MIT Press/AAAI Press, 1996.

[44] Feldman R. & Dagan I. Knowledge Discovery in textual databases (KDT). In proceedings of the first International Conference on Knowledge Discovery and Data Mining (KDD - 95), Montreal, Canada, August 20 – 21, AAAI Press, p. 112-117.

[45] Feldman R., Fresco M., Hirsh H, et al., Knowledge Management: A text Mining Approach", In proc. Of the 2nd Int. Conf. on practical Aspects of Knowledge Management (PAKM98).

[46] Fisher D.H., Knowledge Acquisition via Incremental Conceptual Clustering, Machine Learning, 2, pp. 139-172, 1987.

[47] Forsyth R.S., Holmes D.I,, Feature – Findingfor Text classification, Liter\ry and Linguistic Computing, 10, 4, 1996.

[48] Frigui H., O. Nasraoui, Simultaneous clustering and attribute discrimination, In Proceedings of the IEEE Conference on Fuzzy Systems, San Antonio, TX, pp. 158-163, 2000.

[49] Frigui H., O. Nasraoui, Simultaneous Clustering and Dynamic Keyword Weighting In Text Documents, Survey in Text Mining, Michael Berry, Ed. Springer, pp. 45-70, 2004.

[50] Fukunga K., Narendra P. M. "A branch and Bound Algorithm for Computing k-nearest Neighbors ". IEEE Transactions on Computers C 24, p.750-753, 1975.

[51] Fukunga K., Short R. D., "Generalized clustering for Problem Localization.". IEEE Transactions on Computers C 27, p.176-181

[52] Fukunga Keinosuke. Statistical Pattern Recognition. Academic Press Inc., 1990.

[53] Gelfand B., Wulfekuhler M., Punch III W.F., Automated Concept Extraction from Plain Text, Working notes of Learning from Text and the Web, Conference on Automated Learning and Discovery CONLAND-98, 1998.

[54] Gower J. C. A General Coefficient of Similarity and some of its Properties. Biometrics, 27: 857 – 872, 1971.

[55] Greenacre M.J., Theory and Applications of Correspondence Analysis, newblock Academic Press, 1984.

[56] Halkidi Maria, Batistakis Yiannis, Vazirigiannis Michalis. "On Clustering Validation Techniques". Journal of Intelligent Information Systems, 17 (2-3), p. 107-145, 2001.

[57] Hearst Marti A., Untangling Text Data Mining, Proceedings of ACL '99: the 37th Annual Meeting of the Association for computational Linguistics, University of Maryland, June 20 – 26, 1999.

[58] Hehenberger M., Coupet P., Text Mining applied to patent Analysis.

[59] Hernandez A. Mauricio, Stolfo J. Salvatore. Real – World Data is Dirty: Data Cleansing and The Merge / Purge Problem. Data Mining and Knowledge Discovery, 2: 9 – 37, 1998

[60] Holmes D., Forsyth R., The Federalist Revisited: New Directions in Authorship Attribution, Literary and Linguistic Computing, 10(2), pp. 172 – 177, 1995.

[61] IBM, Text Mining Technology, Turning Information into Knowledge. A white paper from IBM.

[62] Jain A.K., Dubes R.C. Algorithms for Clustering Data. Prentice Hall, 1988.

[63] Jain K. Anil, Duin P.W. Robert, Mao Jianhang, Statistical Pattern Recognition: A Review, IEEE Transactions On Pattern Recognition And Machine Intelligence, 22 (1): 4 – 37, January 2000.

[64] Jain K.A., Murthy N.M., Flynn J.P., Data Clustering: A Review. CAN Computing Survey, 31 (3): 264 – 323, 1999.

[65] Joachims T., A probabilistic Analysis of Rocchio Algorithm with TFIDF for Text Categorization, Proc. Of the 14th International Conf. on Machine Learning ICML97, pp. 143-151, 1997.

[66] Joachims T., text categorization with support vector machines: learning with many relevant features, Proc. Of the 10th European conf. on Machine learning ECML98, pp. 137-142, 1998.

[67] Johnson S., Hierarchical clustering scemes, Psychometrika, 1967, pp. 241-254.

[68] Karanikas H., Tjortjis C. and Theodoulidis B., An Approach to Text Mining using Information Extraction, PKDD 2000 Knowledge Management: Theory and Applications.

[69] Karanikas Haralambos & Babis Theodoulidis, Knowledge Discovery in Text and Text Mining Software, Technical Report, UMIST Department of Computation, January 2002.

[70] Kaufman L., P.J. Rousseeuw, Finging Groups in Data: An introduction to Cluster Analysi, Wiley Series in Probability and Mathematical Statistics, 1990.

[71] Kira K., L.A. Rendell, The feature selection Problem: Traditional methods and a new algorithm. In proceedings of the 10th Nationla Conference on Artificial Intelligence, pp. 129-134, 1992.

[72] Kodratoff Y., About Knowledge Discovery in Texts: A Definition and an Example. Unpublished Paper.

[73] Kodratoff Y., Introduction to Machine Learning, Pitman, 1988.

[74] Kohavi R., D.Sommerfield. Feature subset selection usisng the wrapper model: Overfitting and dynamic search space topology, In Proceedings of the 1st Internaltional Conference on Knowledge Discovery and Data Mining, pp. 192-197, 1995.

[75] Korfhage R.R., Information Storage and retrieval, Wiley, New York, 1997.

[76] Kosala R. and Blockeel H., Web Mining Research: A Survey. ACM SIGKDD, July 2002.

[77] Kowalski G., Information Retrieval Systems: Theory and Implementation, Kluwer Academic, Hingaham, MA, 1997.

[78] Kruskal J.B., Nonmetric Multidimensional Scaling: a Numerical Method, Psychometrika, 36, 1964, pp. 115-129.

[79] L.A. Rendell, K. Kira, Apractical approach to feature selection. In Proceedings of the International Conference an Machine Learning, pp. 249-256, 1992.

[80] Laan N. M., Stylometry and Method. The case of Euripides, Literary and Linguistic Computing, 10, pp. 271 – 278, 1995.

[81] Lam W., Ho C.Y., Using A Generalized Instance set for Automatic Text Categorization, Proc. Of the 21th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'96), pp. 307-315.

[82] Lam W., Low K.F., Ho, C.Y., Using Bayesian Network Induction Approach for Text Categorization, 15th international Joint Conference on artificial Intelligence IJCAI97, pp. 745-750, 1997.

[83] Lent B>, Agrawal R. and Srikant R., Discovering Trends in Text Databases, IBM Almaden Reasearch center.

[84] Lewis D. D. Gale W. A., A Sequential Algorithm for Training Text Classifiers, Proc. Of the 7th Annual International ACM-SIGIR Conf. on Research and Development in Information Retrieval, Dubline, 1994.

[85] Lewis D.D., Ringuette M., Comparition of two learning algorithms for text categorization, Porc. Of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 1994.

[86] Liere R., Tadepalli P., Active Learning whith Committes: Preliminary Results in Comparing Winnow and Pereptron in Text categorization, Working Notes of Learning from text and the web, conf. on Automated Learning and Discovery CONLAND-98, 1998.

[87] Lowe D., Matthews R., Shakespeare vs. Fletcher: A Srylometric Analysis by Radial Basis Func tions, computers and the humanities, 29, pp. 449-461, 1995.

[88] Lynch Merril, e – Buisness Analytics, In – depth Report, 20 November 2000.

[89] MacQueen J.B., Somme methods for Classification and Analysis of Multivariate Observations, Proceedings of the 5th Berkley Symposium on Mathematics, Statistics and Probability, 1967, pp. 281-297.

[90] Maes P., Agents that Reduse Work and Information Overload, Communications of the ACM Vol. 37, No 7, pp, 30-40, July 1994.

[91] Malerba D., F. Esposito, V. Gioviale, V. Tamma, Comparing Dissimilarity Measures for Symbolic Data Analysis, Proceedings of the Joint Conferences on "New Techniques and Technologies for Statistics" and "Exchange of Technology and Know-how", 2001, pp.207-238.

[92] Martindale C., McKenzie D., On the Utility of Content Analysis in Author Attribution: The Federalist, Computers and the Humanities , 29, pp.259-270,1995.

[93] Mc Elligott M.., Sorensen H., An emergent approach to information filtering, Abakus. U.C.C. Computer Science Journal, Vol. 1, No. 4, December 1993.

[94] Medin D.L. , M.M. Schaffer, Context Theory of Classification Learning, Psychological Review, vol. 85, No. 3, 1987, pp.207-238.

[95] Mendenhall T.C., The Charecteristic Curves of Composition, Science IX, pp. 237-249, 1887.

[96] Merriam T., Matthews R., NeuralCokmputation in Stylometry II:An application to the works of Shakepeareand Marlowe, Literary and Linguistic Computing, 9, pp. 1-6, 1994.

[97] Michalski R.S., R.E. Stepp, Learning from observation: conceptual clustering, In R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), Machine Learning: An Artificial Intelligence Approach (Vol. 1), Palo Alto, CA: Tioga Publishing, 1983.

[98] Mladenic D., Grobelink M., Word sequences as feature in text – learning. Proceedings of the Seventh Electrotechical and Computer Sc. Conference ERK '98, pp. 145-148, Ljubljana, Slovenia: IEEE section, 1998.

[99] Mladenic D., Grobelnik M., 1998. Feature Selection for classification based on text hierarchy., Working notes of Learning from Text and the Web, Conference on Automated Learning and Discovery CONALD – 98, 1998.

[100] Mladenic D., Text – learning and related intelligent agents, IEEE Expert special issue on Applications of Intelligent Information Retrieval, July – August 1999.

[101] Mladenic D., Text leraning and related agents, IEEE Expert, July 1999.

[102] Mladenic DE., Personal WebWatcher: Implementation and Design, Technical Report IJS-DP-7472, October 1996.

[103] Morton A.Q., Literary Detection, New York: Scribners, 1978

[104] Morton A.Q., The Authorship of Greek Prose, Journal of the rayal statistical society (A), 128, pp,. 169-233, 1965.

[105] Mosteller F., Wallace D., Applied Bayasian and Classical Inference: The Case of the Federalist Papers, Addison – Wesley, Reading, MA., 1984.

[106] Mosteller F., Wallace D.L., Inference and Disputed Authorship: The Federalist, Reading: Addison-Wesley, 1964.

[107] Moulinier I., Ganascia J. G., Applying an Existing Machine Learning Algorithm to text categorization, In Connectionist, statistical and symbolic Approaches to Learning for natural Language processing, (S. Wermeter, E. Riloff, G. Scheler Eds), Springer-Verlag, 1996.

[108] Nahm U. Y., Mooney R. J. Using Information Extraction to Aid the Discovery of Prediction Rules from Text.

[109] Nigam K., McCallum A., Poll-Based Active Learning fron text and the web, Conf. on Automated Learning and Discovery CONLAND-98, 1998.

[110] Oracle Text., Application Developer's Guide, Release 9.0.1., June 2001, Part No. A90122-01.

[111] Pazzani M., Billsus D., Learning and Revising User Profiles: The Indetification of Interesting Web Sites, Machine Learning 27, Kluwer Academic Publishers, pp. 313-331. 1997.

[112] Pazzani M., Muramatsu J., Billsus D., Syskill & Webert. Identifying interesting web sites, AAAI Spring Symposium on Machine Learning in Information Acces, Stanford, March 1996 and Proceedings of the Thirteenth National Conference on Artificial Intelligence AAAI 96, pp 54-61, 1996.

[113] Pelleg Dan, Moore Andrew. X – means: Extending k – means whith Efficient Estimation of the Number of Clusters. In Seventeenth International Conference on Machine Learning, p. 727 – 734, 2000.

[114] Plotas D., Tasoulis D., Boutsinas B., Are the Iliad and Odyssey each work of a single poet?, 2nd International conference on Ancient Greece Olympia 2001.

[115] Procopiuc M.C. Applications of Clustering Problems. 1997.

[116] Raatikainen E.E. Kimmo. Cluster Analysis and Workload Classification. Technical Report, DRAFT, 1999.

[117] Rajman M., Besancon R., Text Mining: Natural Language techniques and Text Mining applications, Artificial Intelligence Laboratory, Computer Science Department, Swiss Federal Institute of Technology, IFIP 1997. Published by Chapman & Hall.

[118] Rijsbergen C.J. van. Information Retrieval second edition. Butterworths, London, 1979.

[119] Ruspini E.H., A new approach to Clustering, Information and Control, 15, 1969, pp. 22-32.

[120] Salton G., Buckley C., Term Weighting Approaches in Automatic Text Retrieval, Technical Report, COR-87-881, Department of Computer Science Cornell University, November 1987.

[121] Seber G. A. F. "Multivariate Observations". John Wiley & Sons, Inc., 1984.

[122] Selim S.Z., M.A. Ismail, k- means – Type Algorithms: A Generalized Convergence Theorem and Caractererization of Local Optimality. IEEE Transaction an Pattern Analysis and Machine Intelligence, 6(1), 1984, pp. 81-87.

[123] SemioMap, http://www.semio.com/, October 2001.

[124] Shavlik J., Eliassi-Rad T., Building intelligent agents for Web-based tasks: a theory-refinement approach, Working notes of Learning form Text and the Web, Conference on Automated Learning and Discovery CONLAD-98, 1998.

[125] Sichel H.S., On a Distribution Law for Word Frequencies, Journal of American Statistical Association, 70, pp. 542-547, 1975.

[126] Slattery S., Craven M., Learning to Exploit Document Relationships and Structure: The case for Relational Learning on the web, Working notes of learning from text and the web, Conf. on Automated Learning and Discovery CONLAND-98, 1998.

[127] Smith, M.W.A., An investigation of Morton's Method to Distinguish Elizabethan Playwrights, Computers and the Humanities, 19, 3-21, 1985.

[128] Sokal R., C.D. Michener, A Statistical method for evaluating systematic relationships, University of Kansas Scientific Bulletin, 38, 1958, pp. 1409-1438.

[129] Sorensen H., McElligot M., PSUN:A profiling System for Usnet News, CIKM'95 Intelligent Information Agents Workshop, Baltimore, December 1995.

[130] Spath H. Cluster Analysis Algorithms. Ellis Horwood, 1980.

[131] Stanfill C., D. Waltz, Towards memory – based reasoning, Communication of the ACM, 29(12), pp. 1213 – 1228, 1986.

[132] Sullivan D. Document Warehousing and Text Mining, Wiley Computer Publishing 2001.

[133] Tan Al – H., Text Mining: The State of art and the challenges, in proceedings of the Pacific Asia Conf. on Knowledge Discovery and Data Mining PAKDD '99 workshop on Knowledge Discovery from Advanced Databases.

[134] Texis, Thunderstone, http:///www.thunderstone.com/texis/site/pages , December 2001.

[135] Text Analyst http://www.megaputer.com/products/ta/index.php3 , December 2001.

[136] Theodoridis Sergios, Koutroumbas Konstantinos. “Pattern Recognition”. Academic Press, 1999.

[137] Vrahatis N. M., Boutsinas B., Alevizoz P., Pavlidis G. The new k – Windows Algorithm for Improving the k – means Clustering Algorithm. Journal of Complexity, 18, 2002, pp. 375-391.

[138] Wettschereck D., D.W. Aha, T. Mohri, A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms, Artificial Intelligence Review, 11, pp. 273 – 314, 1997.

[139] WizDoc, WizDoc for Web Sites, users manual.

[140] Yang Y., An evaluation of statistical approaches to text categorization, Journal of Information Retrieval, 1998.

[141] Yang Y., Expert Network: Effective and Efficient Learning form Human Decisions in Text Categorization and Retrieval and Development in Information Retrieval, Dublin 1994.

[142] Yule G.U., On Sentence Length as a Statistical Caracteristic of Style in Prose with Application to Two Cases of Disputed Authorship, Biometrica, 10, pp. 363-390, 1938.

[143] Zamir O., O. Etzioni, O. Madani, R.M. Karp, Fast and intuitive clustering of web documents. In KDD’97, pp. 287 and intuitive clustering of web documents. In KDD’97, pp. 287-290, 1997.

[144] Zhang B., S.N. Srihari, Properties of Binary Vector Dissimilarity Measures, In proceedings of International Conference on Cumputer Vision, Pattern Recognition and Image Processing (CVPRIP), North Carolina, 2003.

[145] Zhexue Huang. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Dept. of Computer Science, The University of British Columbia, 1997.

[146] Zhexue Huang. Extensions to the k-Means Algorithm for Clustering large Sets with Categorical Values. Data Mining and Knowledge Discovery, 2: 283 – 304, 1998.

[147] Zipf G.K., Selected Studies of the principle oa Relative Frequency in Language, Harvard University Press, 1932.

[148] Αντουλάτος Γεράσιμος. Εφαρμογές Αλγορίθμων και Έλεγχοι Αξιοπιστίας Ομαδοποίησης στην Αναγνώριση Προτύπων και στον Καθαρισμό Δεδομένων. Master Thesis, Πάτρα 2003.

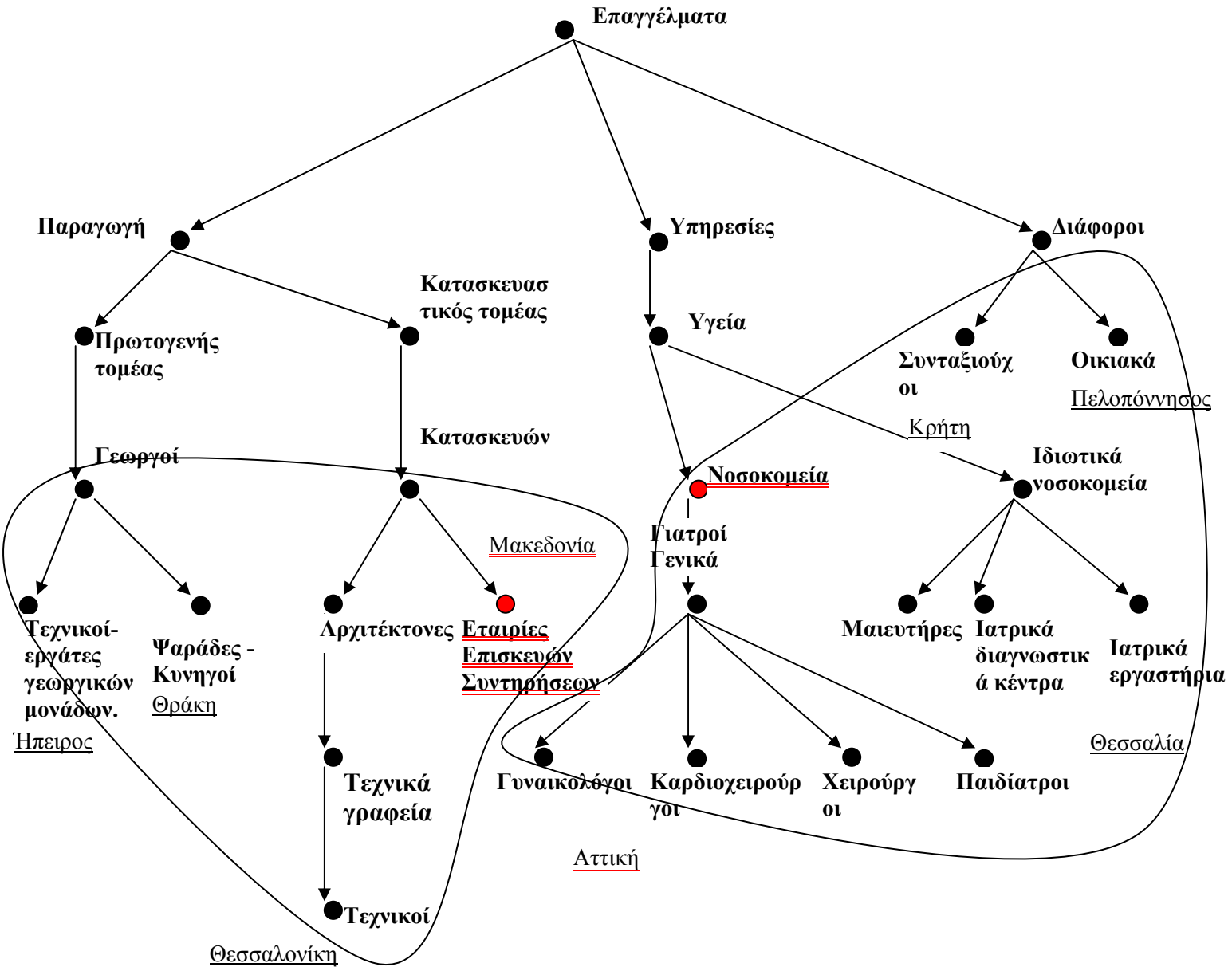
7 Παράρτημα Α

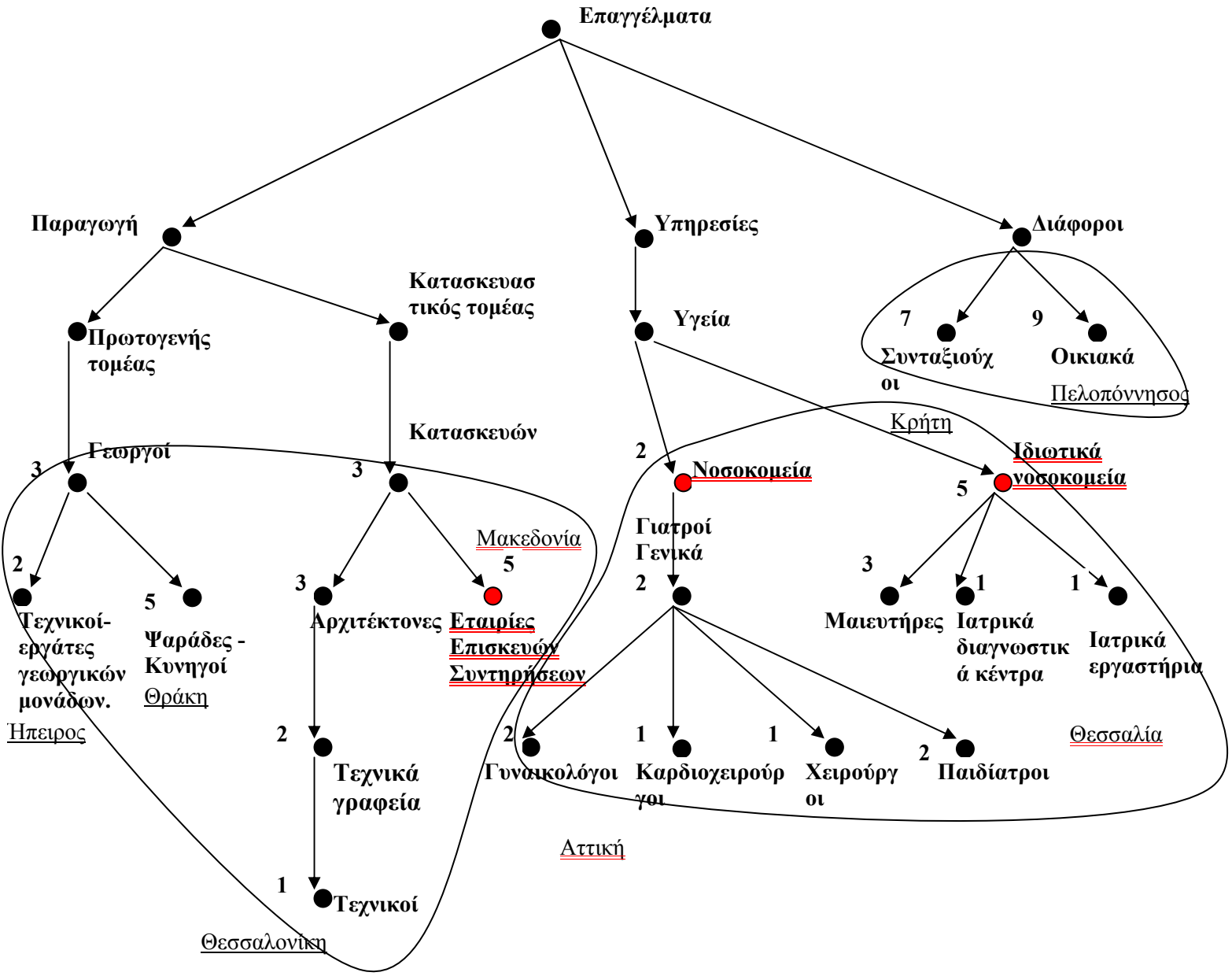
```
"ΗΠΕΙΡΟΣ";"ΓΕΩΡΓΟΣ"
"ΘΕΣΣΑΛΟΝΙΚΗ";"ΚΑΤΑΣΚΕΥΩΝ"
"ΑΤΤΙΚΗ";"ΝΟΣΟΚΟΜΕΙΑ"
"ΘΕΣΣΑΛΙΑ";"ΙΔΙΩΤΙΚΑ ΝΟΣΟΚΟΜΕΙΑ"
"ΠΕΛΟΠΟΝΝΗΣΟΣ";"ΟΙΚΙΑΚΑ"
"ΚΡΗΤΗ";"ΣΥΝΤΑΞΙΟΥΧΟΙ"
"ΑΤΤΙΚΗ";"ΓΙΑΤΡΟΙ ΓΥΝΑΙΚΟΛΟΓΟΙ"
"ΑΤΤΙΚΗ";"ΓΙΑΤΡΟΙ ΠΑΙΔΙΑΤΡΟΙ"
"ΘΕΣΣΑΛΟΝΙΚΗ";"ΤΕΧΝΙΚΟΙ-ΕΡΓΑΤΕΣ"
"ΘΕΣΣΑΛΙΑ";"ΜΑΙΕΥΤΗΡΙΑ"
"ΗΠΕΙΡΟΣ";"ΓΕΩΡΓΟΣ"
"ΗΠΕΙΡΟΣ";"ΓΕΩΡΓΟΣ"
"ΗΠΕΙΡΟΣ";"ΤΕΧΝΙΚΟΙ-ΕΡΓΑΤΕΣ ΓΕΩΡΓ/ΚΤΗΝ.ΜΟΝΑΔΩΝ"
"ΗΠΕΙΡΟΣ";"ΤΕΧΝΙΚΟΙ-ΕΡΓΑΤΕΣ ΓΕΩΡΓ/ΚΤΗΝ.ΜΟΝΑΔΩΝ"
"ΘΡΑΚΗ";"ΚΤΗΝΟΤΡΟΦΟΙ-ΨΑΡΑΔΕΣ-ΚΥΝΗΓΟΙ"
"ΘΡΑΚΗ";"ΚΤΗΝΟΤΡΟΦΟΙ-ΨΑΡΑΔΕΣ-ΚΥΝΗΓΟΙ"
"ΘΡΑΚΗ";"ΚΤΗΝΟΤΡΟΦΟΙ-ΨΑΡΑΔΕΣ-ΚΥΝΗΓΟΙ"
"ΘΡΑΚΗ";"ΚΤΗΝΟΤΡΟΦΟΙ-ΨΑΡΑΔΕΣ-ΚΥΝΗΓΟΙ"
"ΘΡΑΚΗ";"ΚΤΗΝΟΤΡΟΦΟΙ-ΨΑΡΑΔΕΣ-ΚΥΝΗΓΟΙ"
"ΘΕΣΣΑΛΟΝΙΚΗ";"ΚΑΤΑΣΚΕΥΩΝ"
"ΘΕΣΣΑΛΟΝΙΚΗ";"ΚΑΤΑΣΚΕΥΩΝ"
"ΘΕΣΣΑΛΟΝΙΚΗ";"ΑΡΧΙΤΕΚΤΟΝΕΣ-ΜΗΧΑΝΙΚΟΙ"
"ΘΕΣΣΑΛΟΝΙΚΗ";"ΑΡΧΙΤΕΚΤΟΝΕΣ-ΜΗΧΑΝΙΚΟΙ"
"ΘΕΣΣΑΛΟΝΙΚΗ";"ΑΡΧΙΤΕΚΤΟΝΕΣ-ΜΗΧΑΝΙΚΟΙ"
"ΘΕΣΣΑΛΟΝΙΚΗ";"ΤΕΧΝΙΚΟ ΓΡΑΦΕΙΟ"
"ΘΕΣΣΑΛΟΝΙΚΗ";"ΤΕΧΝΙΚΟ ΓΡΑΦΕΙΟ"
"ΜΑΚΕΔΟΝΙΑ";"ΕΤΑΙΡΙΕΣ ΕΠΙΣΚΕΥΩΝ-ΣΥΝΤΗΡΗΣΕΩΝ"
"ΜΑΚΕΔΟΝΙΑ";"ΕΤΑΙΡΙΕΣ ΕΠΙΣΚΕΥΩΝ-ΣΥΝΤΗΡΗΣΕΩΝ"
"ΜΑΚΕΔΟΝΙΑ";"ΕΤΑΙΡΙΕΣ ΕΠΙΣΚΕΥΩΝ-ΣΥΝΤΗΡΗΣΕΩΝ"
"ΜΑΚΕΔΟΝΙΑ";"ΕΤΑΙΡΙΕΣ ΕΠΙΣΚΕΥΩΝ-ΣΥΝΤΗΡΗΣΕΩΝ"
"ΜΑΚΕΔΟΝΙΑ";"ΕΤΑΙΡΙΕΣ ΕΠΙΣΚΕΥΩΝ-ΣΥΝΤΗΡΗΣΕΩΝ"
"ΑΤΤΙΚΗ";"ΝΟΣΟΚΟΜΕΙΑ"
"ΑΤΤΙΚΗ";"ΓΙΑΤΡΟΙ ΓΕΝΙΚΑ"
"ΑΤΤΙΚΗ";"ΓΙΑΤΡΟΙ ΓΕΝΙΚΑ"
"ΑΤΤΙΚΗ";"ΓΙΑΤΡΟΙ ΓΥΝΑΙΚΟΛΟΓΟΙ"
"ΑΤΤΙΚΗ";"ΓΙΑΤΡΟΙ ΚΑΡΔΙΟΛΟΓΟΙ"
"ΑΤΤΙΚΗ";"ΓΙΑΤΡΟΙ ΠΑΙΔΙΑΤΡΟΙ"
"ΑΤΤΙΚΗ";"ΓΙΑΤΡΟΙ ΧΕΙΡΟΥΡΓΟΙ/ΑΝΑΣΘΗΣΙΟΛΟΓΟΙ"
"ΘΕΣΣΑΛΙΑ";"ΙΔΙΩΤΙΚΑ ΝΟΣΟΚΟΜΕΙΑ"
"ΘΕΣΣΑΛΙΑ";"ΙΔΙΩΤΙΚΑ ΝΟΣΟΚΟΜΕΙΑ"
"ΘΕΣΣΑΛΙΑ";"ΙΔΙΩΤΙΚΑ ΝΟΣΟΚΟΜΕΙΑ"
"ΘΕΣΣΑΛΙΑ";"ΙΔΙΩΤΙΚΑ ΝΟΣΟΚΟΜΕΙΑ"
"ΘΕΣΣΑΛΙΑ";"ΜΑΙΕΥΤΗΡΙΑ"
"ΘΕΣΣΑΛΙΑ";"ΜΑΙΕΥΤΗΡΙΑ"
"ΘΕΣΣΑΛΙΑ";"ΙΔΙΩΤΙΚΑ ΙΑΤΡΙΚΑ/ΔΙΑΓΝΩΣΤΚΑ ΚΕΝΤΡΑ"
"ΘΕΣΣΑΛΙΑ";"ΙΑΤΡΙΚΑ ΕΡΓΑΣΤΗΡΙΑ"
"ΚΡΗΤΗ";"ΣΥΝΤΑΞΙΟΥΧΟΙ"
"ΚΡΗΤΗ";"ΣΥΝΤΑΞΙΟΥΧΟΙ"
"ΚΡΗΤΗ";"ΣΥΝΤΑΞΙΟΥΧΟΙ"
"ΚΡΗΤΗ";"ΣΥΝΤΑΞΙΟΥΧΟΙ"
"ΚΡΗΤΗ";"ΣΥΝΤΑΞΙΟΥΧΟΙ"
"ΚΡΗΤΗ";"ΣΥΝΤΑΞΙΟΥΧΟΙ"
"ΠΕΛΟΠΟΝΝΗΣΟΣ";"ΟΙΚΙΑΚΑ"
"ΠΕΛΟΠΟΝΝΗΣΟΣ";"ΟΙΚΙΑΚΑ"
"ΠΕΛΟΠΟΝΝΗΣΟΣ";"ΟΙΚΙΑΚΑ"
"ΠΕΛΟΠΟΝΝΗΣΟΣ";"ΟΙΚΙΑΚΑ"
```

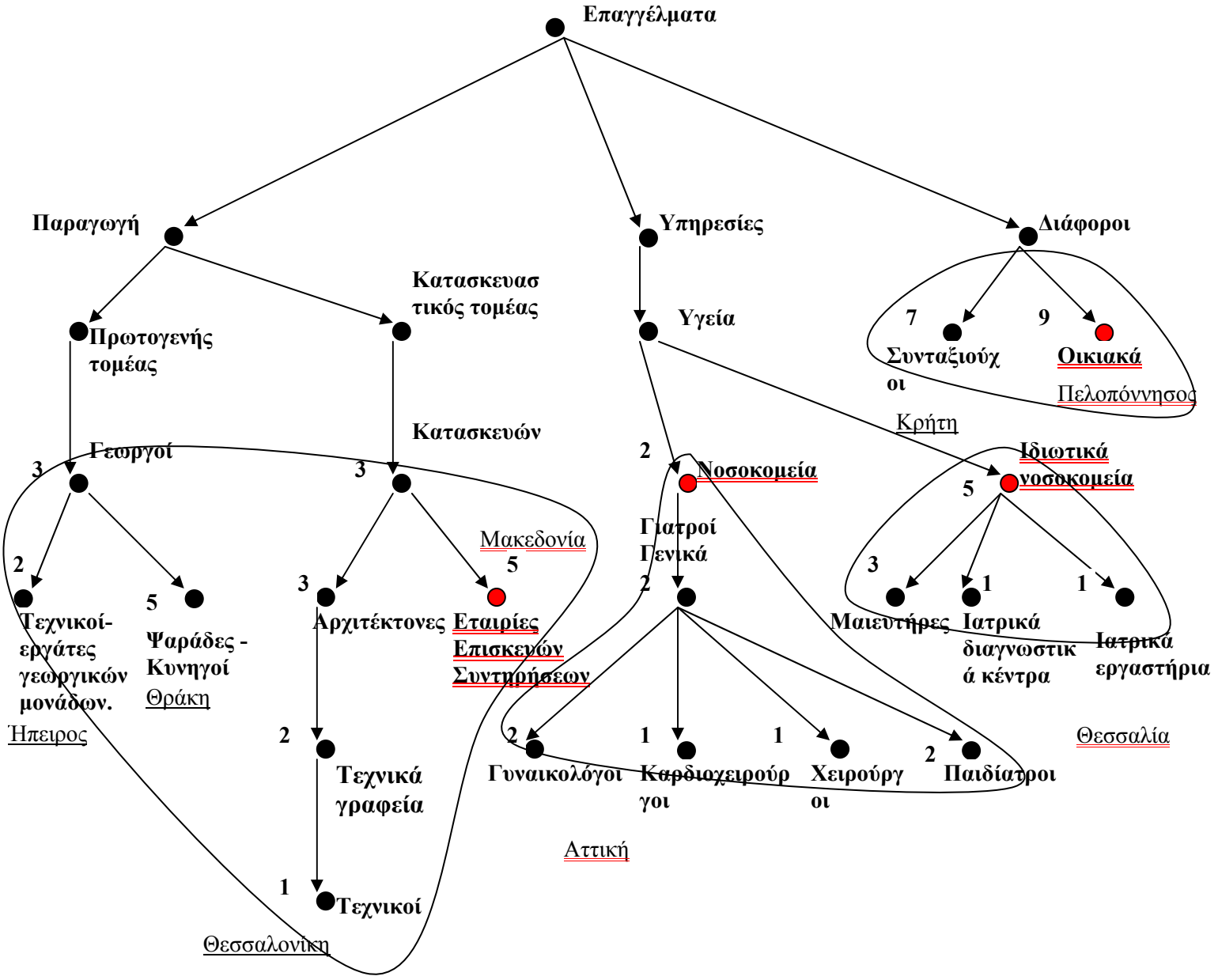
```
"ΠΕΛΟΠΟΝΝΗΣΟΣ";"ΟΙΚΙΑΚΑ"
"ΠΕΛΟΠΟΝΝΗΣΟΣ";"ΟΙΚΙΑΚΑ"
"ΠΕΛΟΠΟΝΝΗΣΟΣ";"ΟΙΚΙΑΚΑ"
"ΠΕΛΟΠΟΝΝΗΣΟΣ";"ΟΙΚΙΑΚΑ"
```

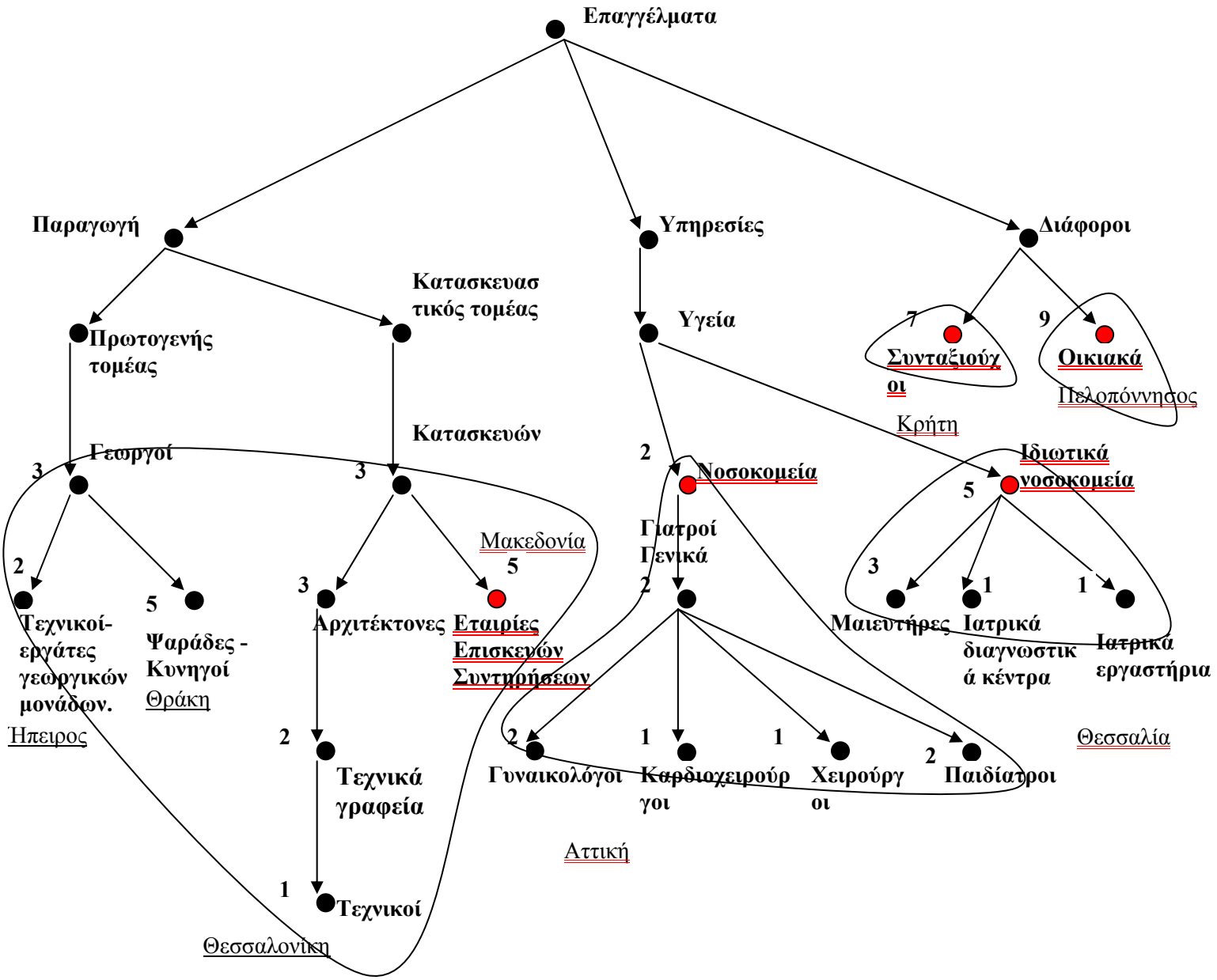
8 Παράρτημα Β

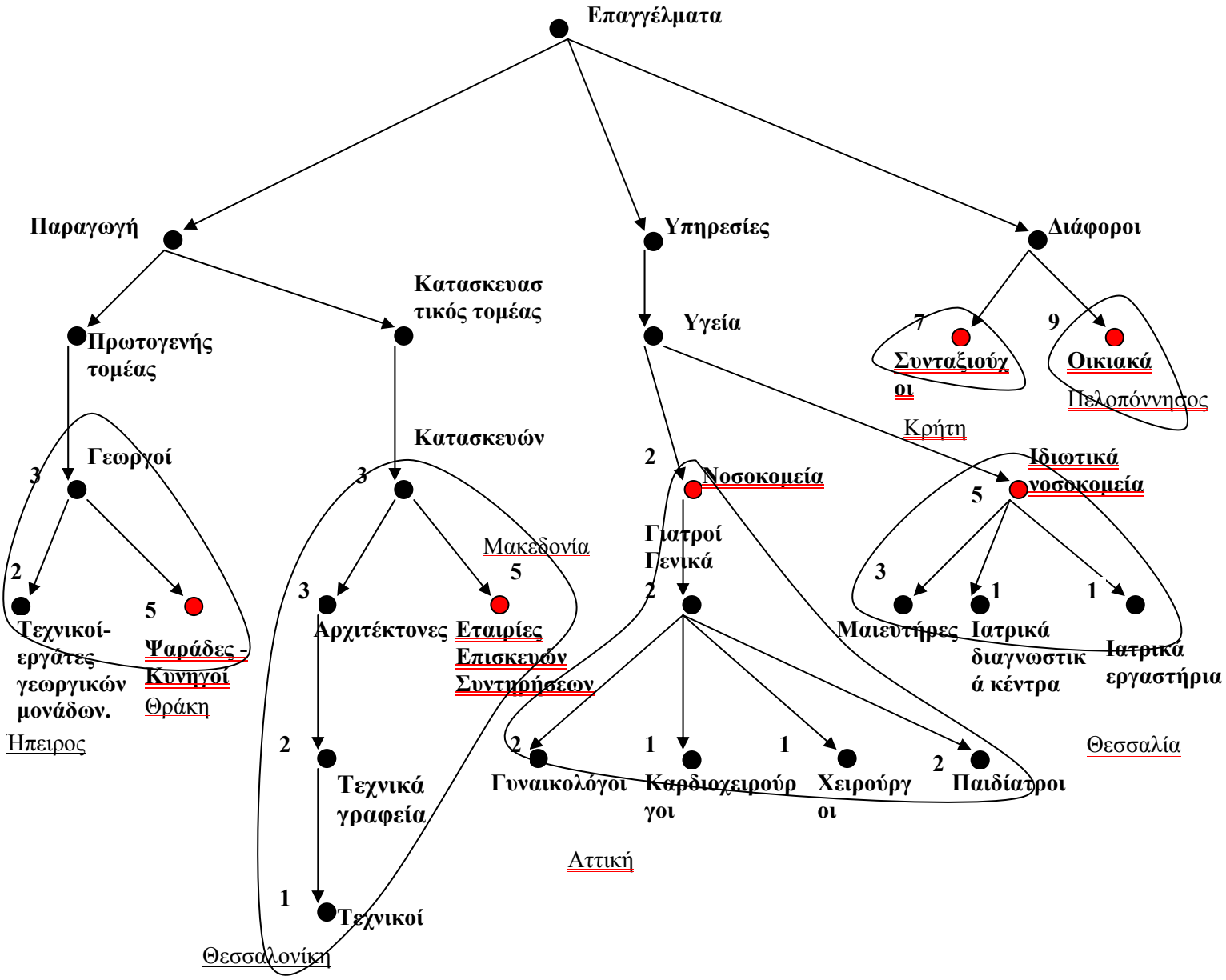
8.1 Προτεινόμενος αλγόριθμος

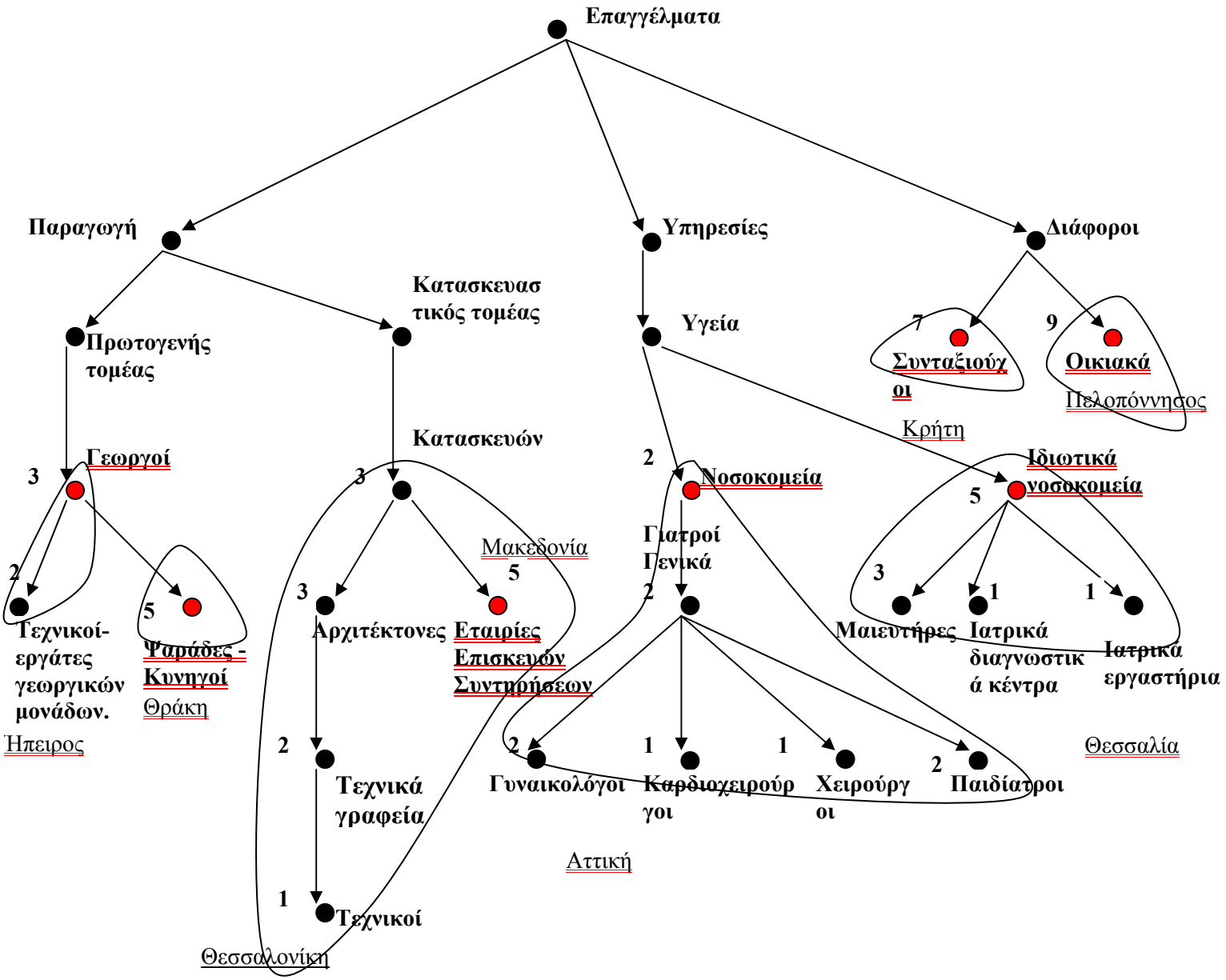



Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Κατασκευαστικός τομέας
Πρωτογενής τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
Κατασκευών
3
Γεωργοί
3
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
3
1
1
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Ηπειρος
2
Τεχνικά γραφεία
2
1
1
2
Θεσσαλία
Γυναικολόγοι
Καρδιοχειρούργοι
Χειρούργοι
Παιδίατροι
Αττική
1
Τεχνικοί
Θεσσαλονίκη

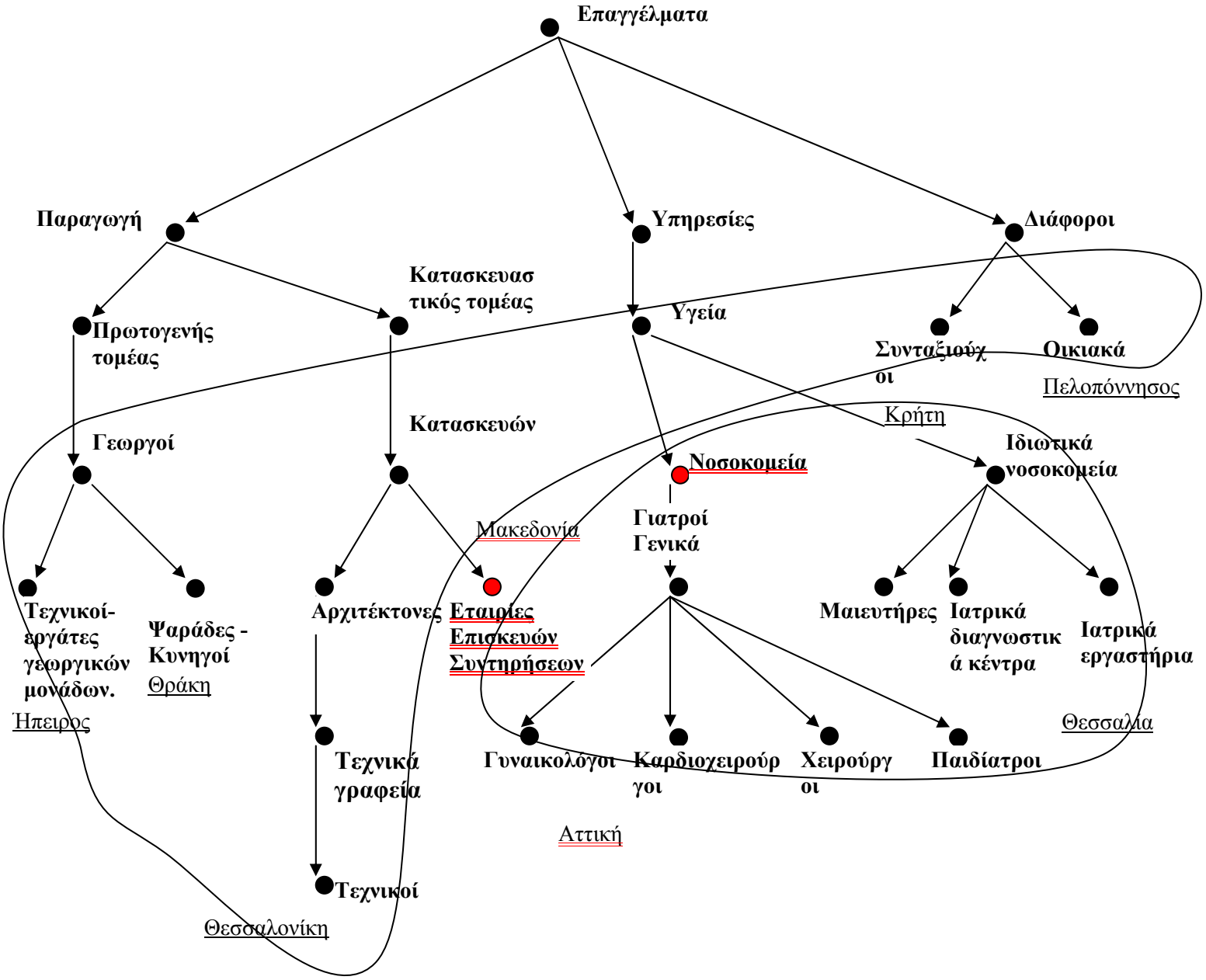

Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
Κατασκευών
3
Γεωργοί
3
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
2
3
1
1
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Ήπειρος
2
Τεχνικά γραφεία
2
1
1
2
Γυναικολόγοι
Καρδιοχειρούργοι
Χειρούργοι
Παιδίατροι
Θεσσαλία
Αττική
1
Τεχνικοί
Θεσσαλονίκη

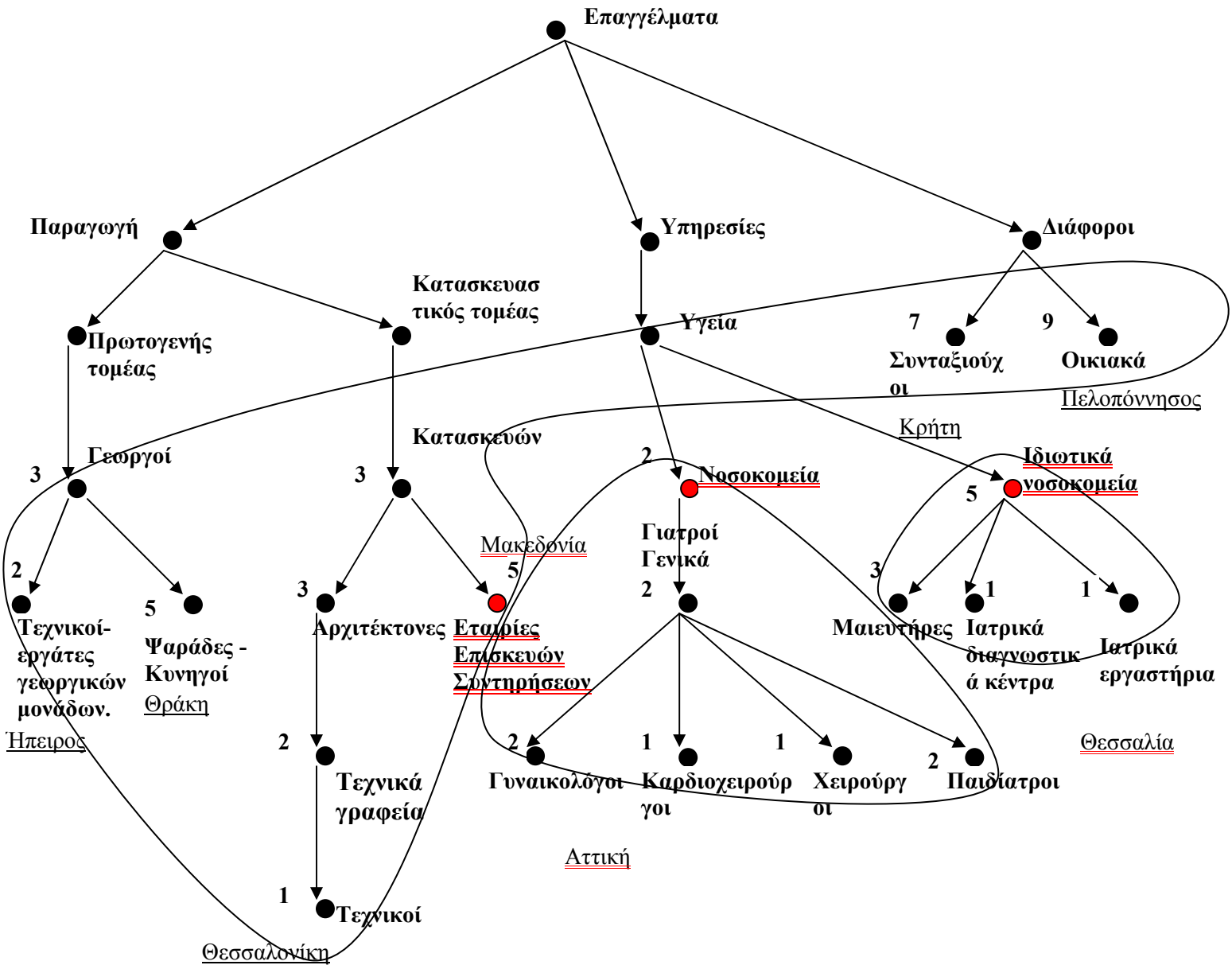

Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Κατασκευαστικός τομέας
Πρωτογενής τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
Κατασκευών
3
Γεωργοί
3
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
3
1
1
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ήπειρος
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
2
Τεχνικά γραφεία
2
1
1
2
Γυναικολόγοι
Καρδιοχειρούργοι
Χειρούργοι
Παιδίατροι
Θεσσαλία
Αττική
1
Τεχνικοί
Θεσσαλονίκη

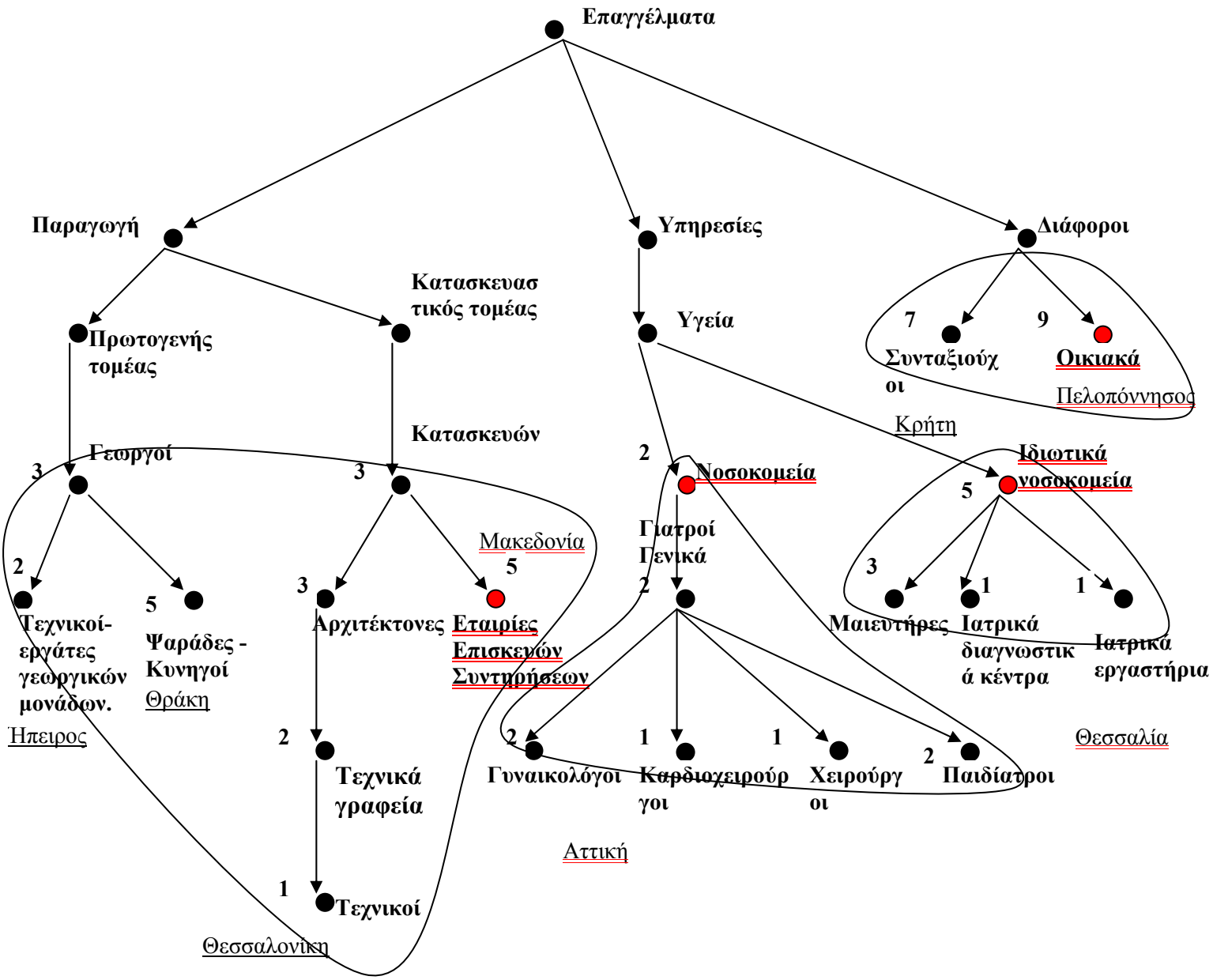

Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
3
1
1
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
5
Ψαράδες - Κυνηγοί
Θράκη
Ήπειρος
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Θεσσαλία
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Αττική
1
Τεχνικοί
Θεσσαλονίκη

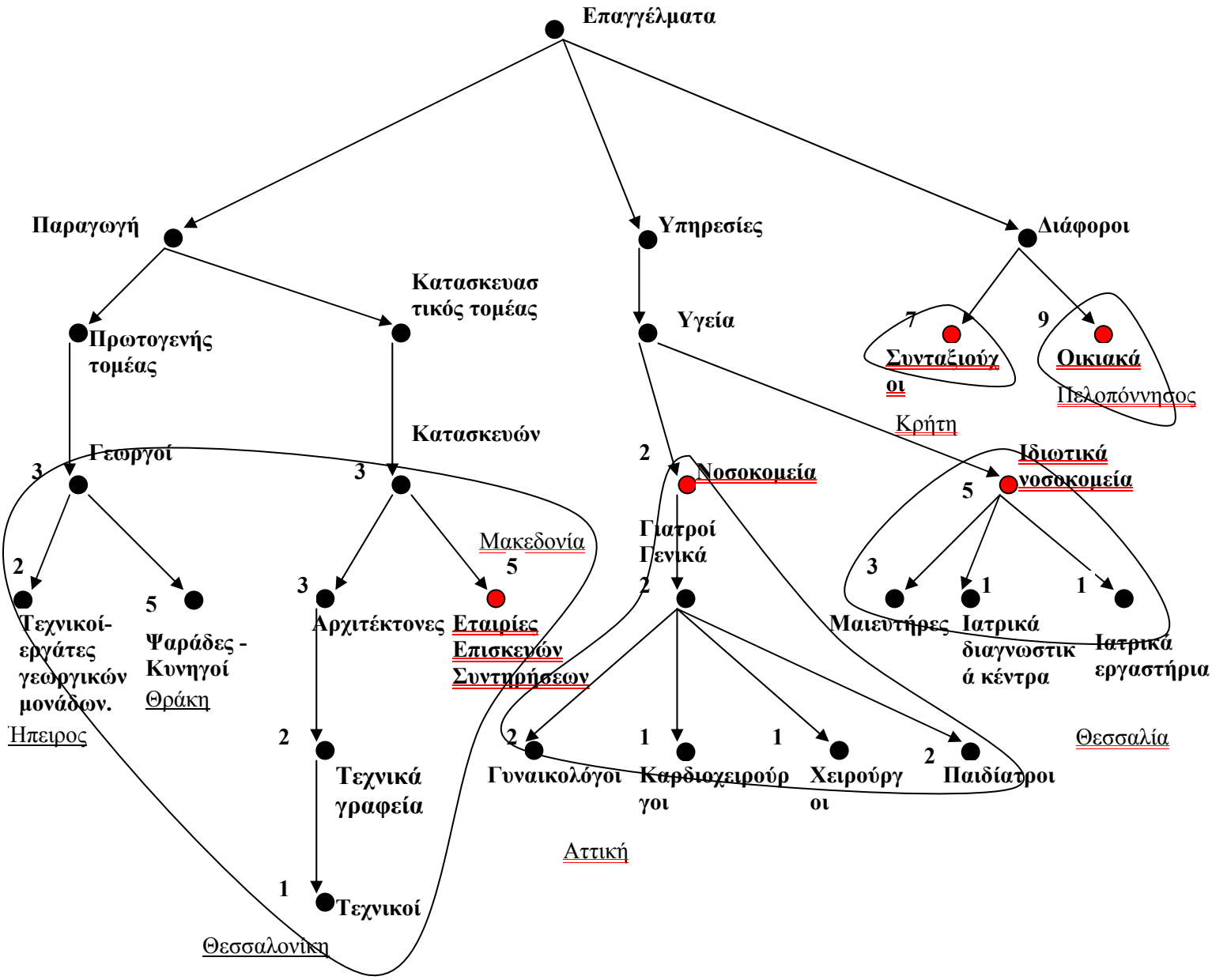

Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
2
3
1
1
2
5
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ηπειρος
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Θεσσαλία
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Αττική
1
Τεχνικοί
Θεσσαλονίκη

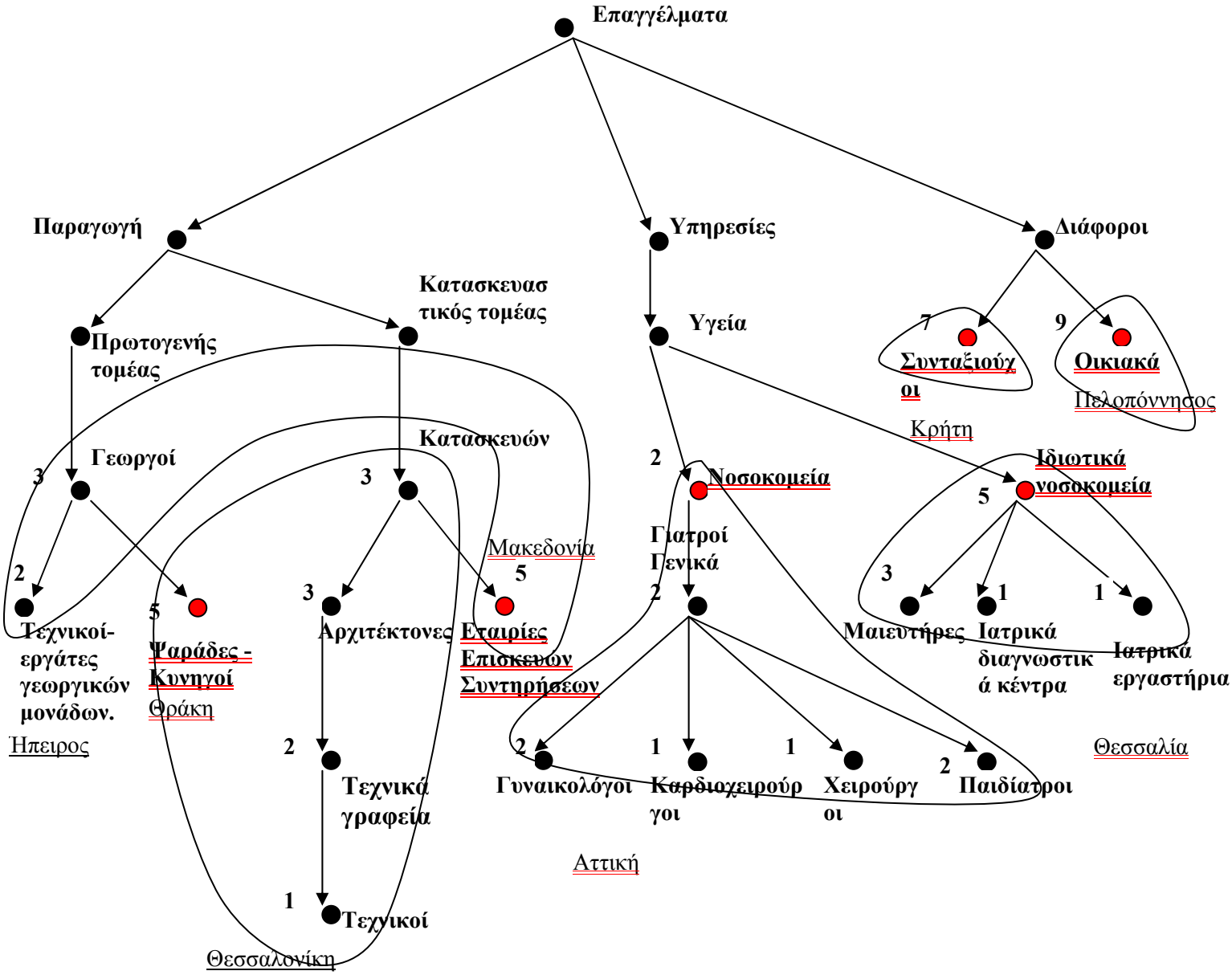
8.2 Kodratoff

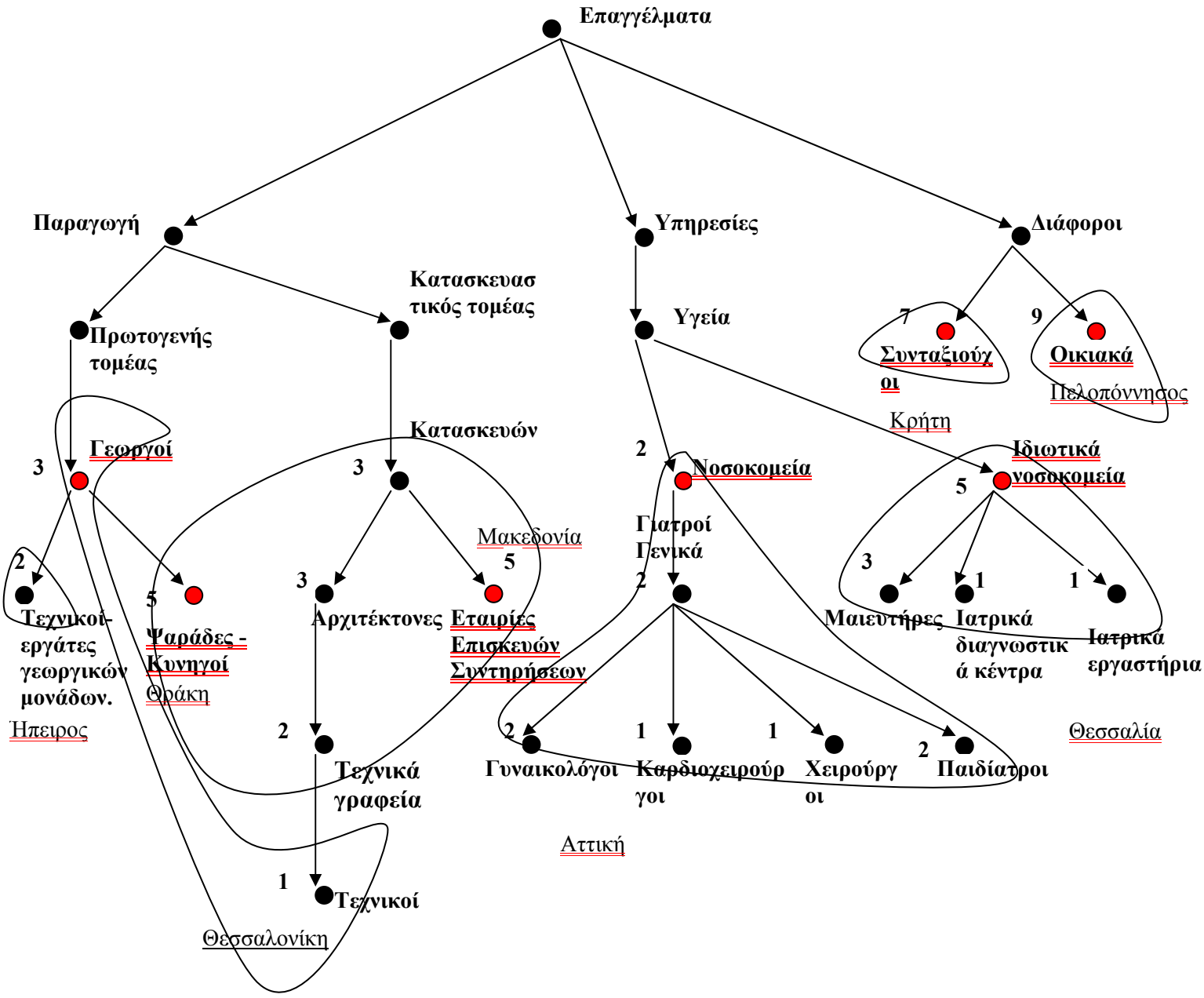



Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
5
Ιδιωτικά νοσοκομεία
Μακεδονία
5
Γιατροί Γενικά
2
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ήπειρος
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
3
Μαιευτήρες
1
Ιατρικά διαγνωστικά κέντρα
1
Ιατρικά εργαστήρια
Θεσσαλία
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Αττική
1
Τεχνικοί
Θεσσαλονίκη

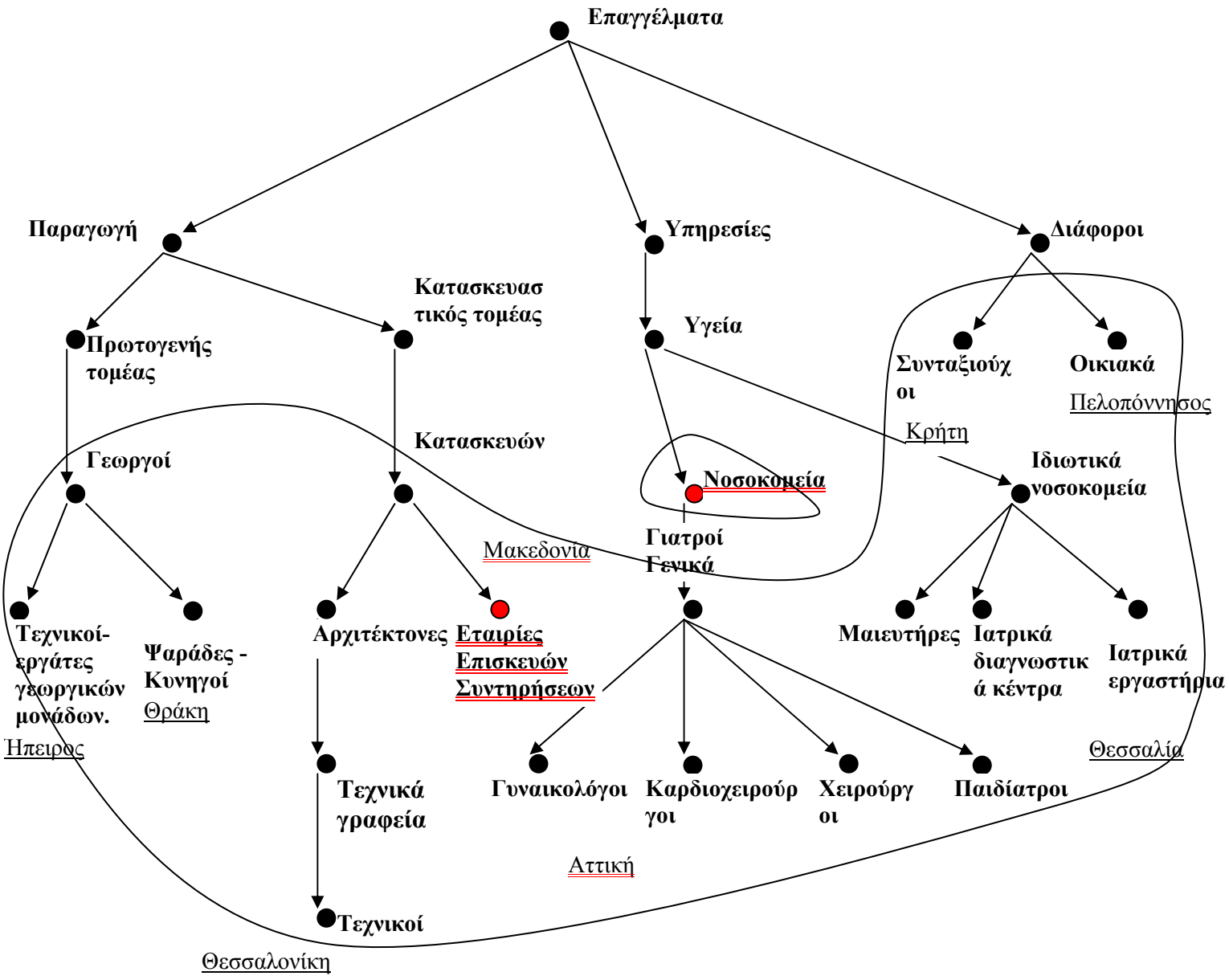

Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
Κατασκευών
3
Γεωργοί
3
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
2
3
1
1
2
5
3
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ψαράδες - Κυνηγοί
Θράκη
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Ήπειρος
2
Τεχνικά γραφεία
2
1
1
2
Γυναικολόγοι
Καρδιοχειρούργοι
Χειρούργοι
Παιδίατροι
Θεσσαλία
Αττική
1
Τεχνικοί
Θεσσαλονίκη

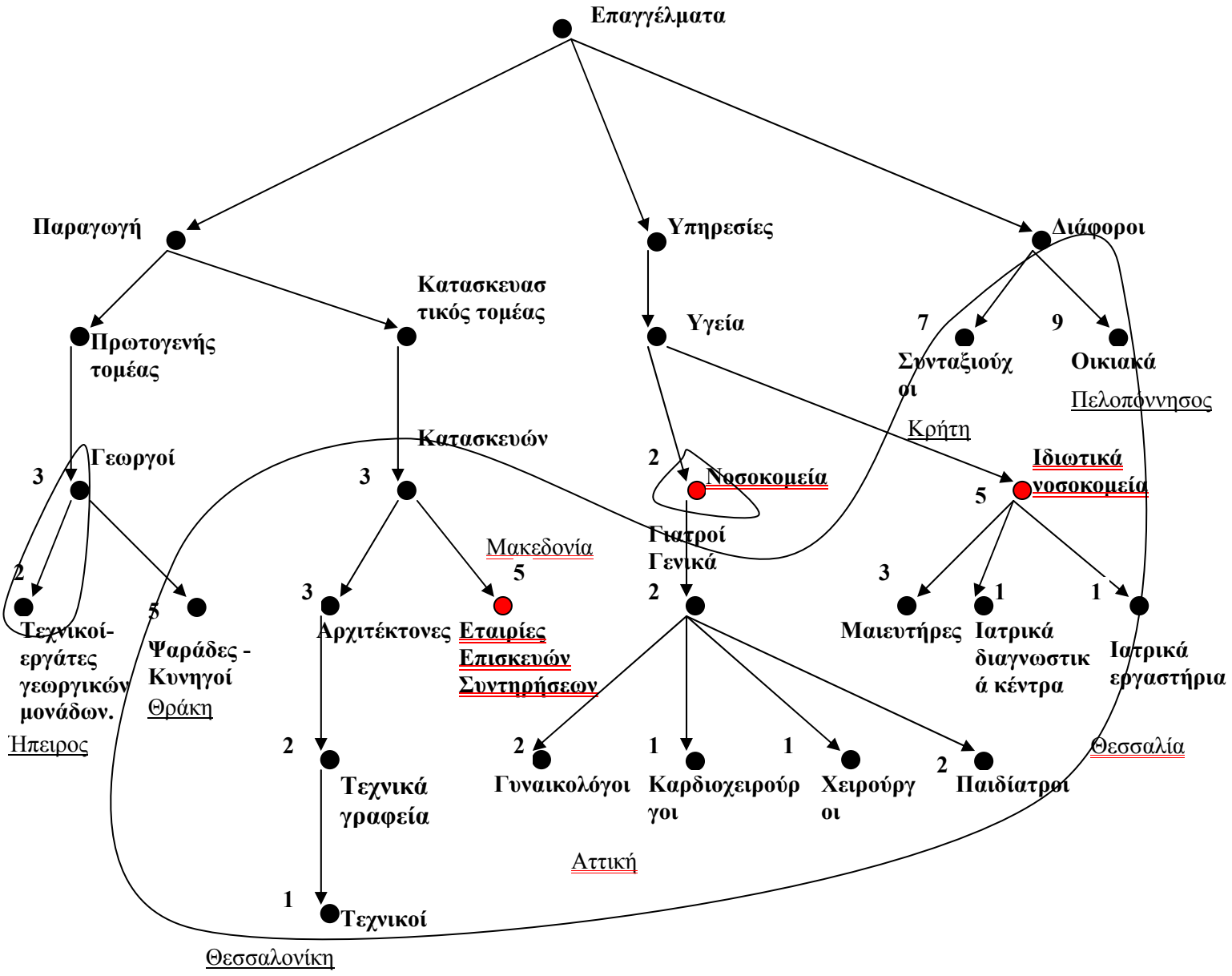

Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
3
2
1
1
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ήπειρος
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Θεσσαλία
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Αττική
1
Τεχνικοί
Θεσσαλονίκη

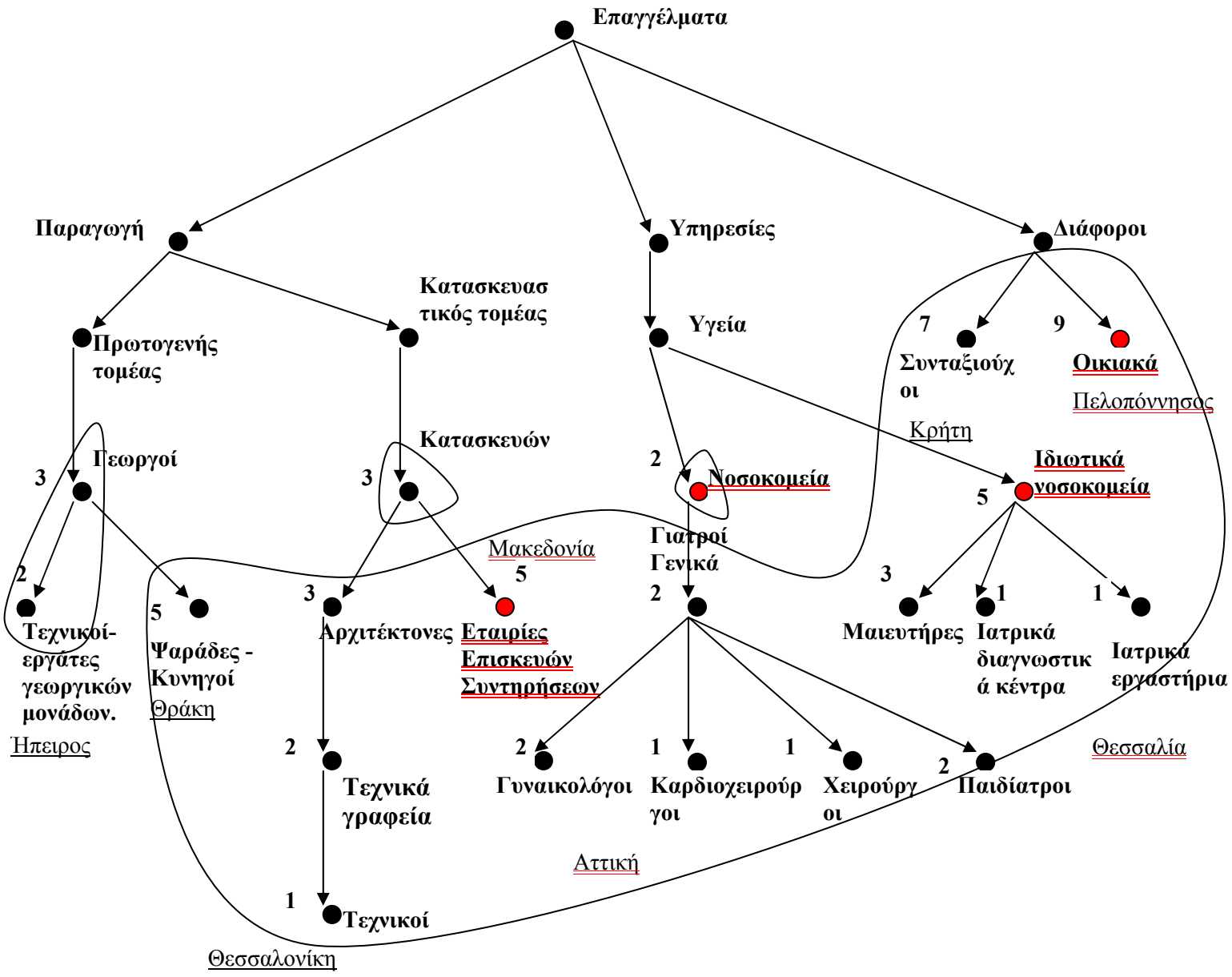

Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
2
3
1
1
2
5
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ήπειρος
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Θεσσαλία
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Αττική
1
Τεχνικοί
Θεσσαλονίκη

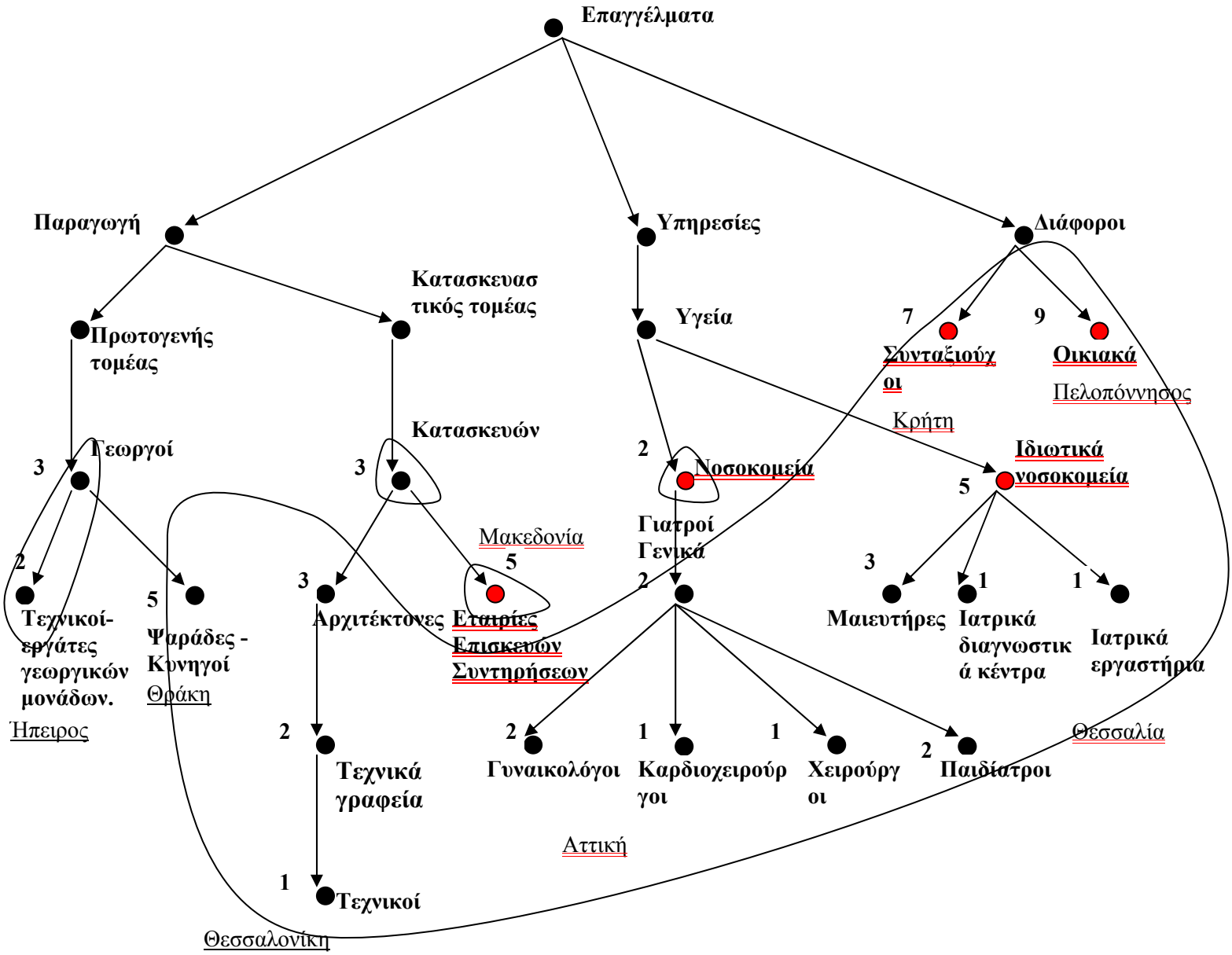

Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
2
3
Μαιευτήρες
1
Ιατρικά διαγνωστικά κέντρα
1
Ιατρικά εργαστήρια
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ήπειρος
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Θεσσαλία
2
Τεχνικά γραφεία
Αττική
1
Τεχνικοί
Θεσσαλονίκη

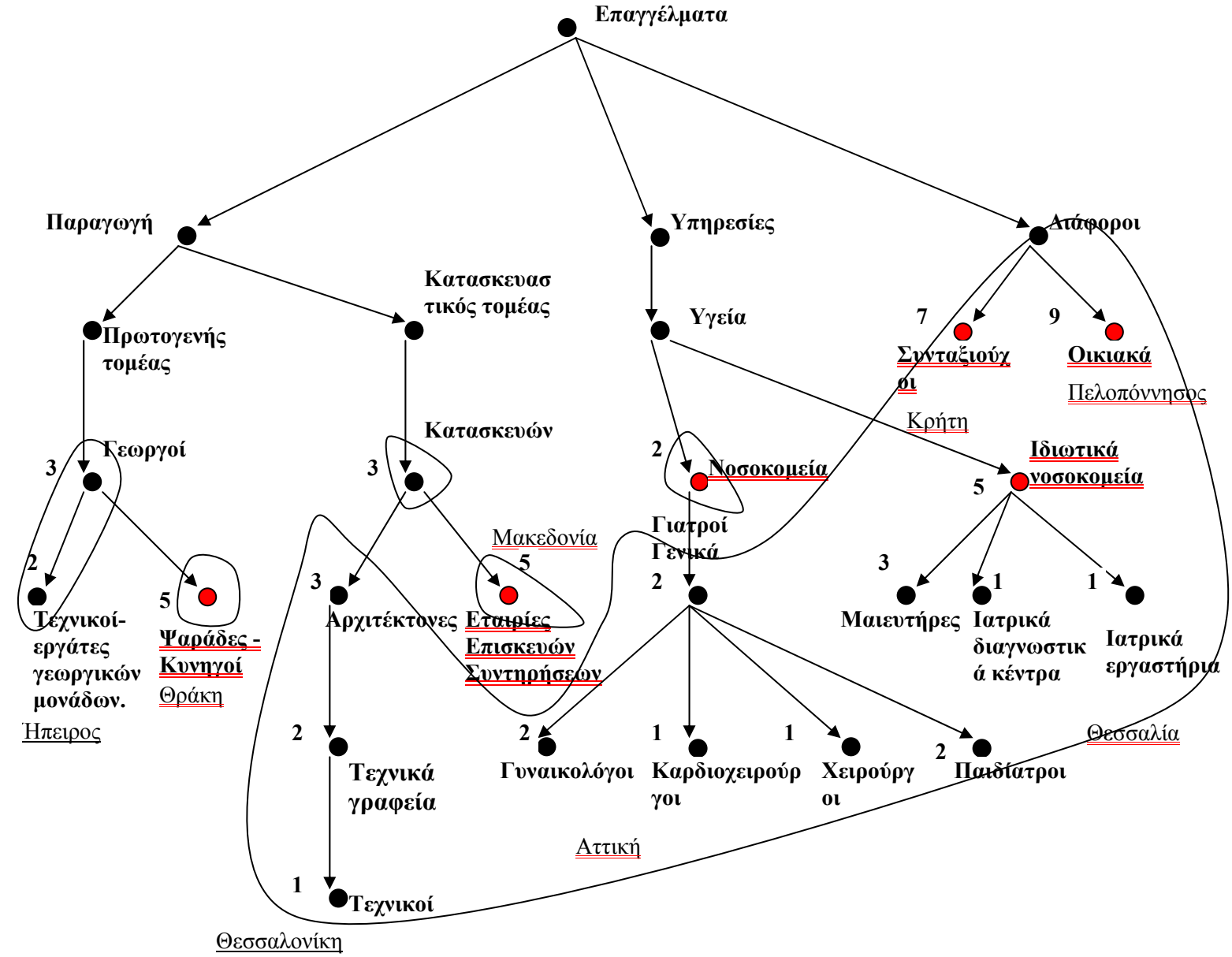
8.3 k – modes

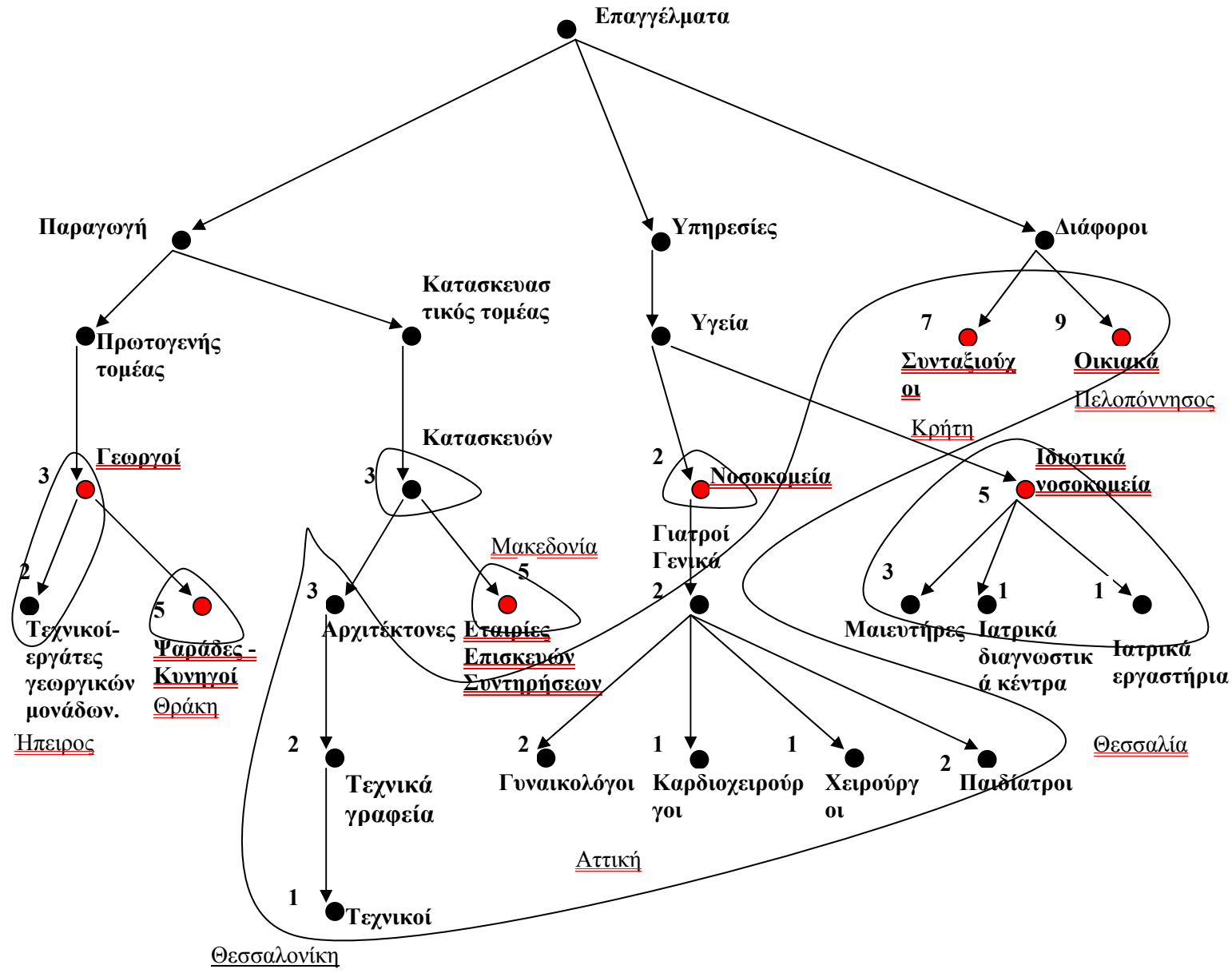

Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
Συνταξιούχοι
Οικιακά
Πελοπόννησος
Κρήτη
Γεωργοί
Κατασκευών
Νοσοκομεία
Ιδιωτικά νοσοκομεία
Μακεδονία
Γιατροί Γενικά
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ψαράδες - Κυνηγοί
Θράκη
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Ήπειρος
Θεσσαλία
Τεχνικά γραφεία
Γυναικολόγοι
Καρδιοχειρούργοι
Χειρούργοι
Παιδίατροι
Αττική
Τεχνικοί
Θεσσαλονίκη


Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
Κατασκευών
3
Γεωργοί
3
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
2
3
1
1
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ήπειρος
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Θεσσαλία
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Αττική
1
Τεχνικοί
Θεσσαλονίκη


Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
Γιατροί Γενικά
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ήπειρος
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
5
Εταιρίες Επισκευών Συντηρήσεων
3
Μαιευτήρες
1
Ιατρικά διαγνωστικά κέντρα
1
Ιατρικά εργαστήρια
Θεσσαλία
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Αττική
1
Τεχνικοί
Θεσσαλονίκη


Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
2
3
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ήπειρος
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
3
Μαιευτήρες
1
Ιατρικά διαγνωστικά κέντρα
1
Ιατρικά εργαστήρια
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Θεσσαλία
Αττική
1
Τεχνικοί
Θεσσαλονίκη


Επαγγέλματα
Παραγωγή
Υπηρεσίες
Άτάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
2
3
1
1
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
Ήπειρος
5
Ψαράδες - Κυνηγοί
Θράκη
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Θεσσαλία
Αττική
1
Τεχνικοί
Θεσσαλονίκη


Επαγγέλματα
Παραγωγή
Υπηρεσίες
Διάφοροι
Πρωτογενής τομέας
Κατασκευαστικός τομέας
Υγεία
7
Συνταξιούχοι
9
Οικιακά
Πελοπόννησος
Κρήτη
3
Γεωργοί
3
Κατασκευών
2
Νοσοκομεία
Ιδιωτικά νοσοκομεία
5
Μακεδονία
5
Γιατροί Γενικά
2
3
1
1
2
Τεχνικοί-εργάτες γεωργικών μονάδων.
5
Ψαράδες - Κυνηγοί
Θράκη
Ήπειρος
3
Αρχιτέκτονες
Εταιρίες Επισκευών Συντηρήσεων
Μαιευτήρες
Ιατρικά διαγνωστικά κέντρα
Ιατρικά εργαστήρια
Θεσσαλία
2
Τεχνικά γραφεία
2
Γυναικολόγοι
1
Καρδιοχειρούργοι
1
Χειρούργοι
2
Παιδίατροι
Αττική
1
Τεχνικοί
Θεσσαλονίκη

9 Παράρτημα Γ

//----------Μπασό----------//
"ΡΟΔΟ ΟΡΕΙΝΟ";"ΚΑΤΑΡΡΑΧΤΗΣ"
"ΦΕΓΓΑΡΙ";"ΒΡΟΧΗ"
"ΒΟΥΝΟ ΦΟΥΤΖΙ";"ΟΜΙΧΛΗ"
"ΓΕΡΑΚΙ";"ΚΑΒΟΣ ΙΡΑΓΚΟ"
"ΜΕΛΙΣΣΑ";"ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ"
"ΠΕΤΑΛΟΥΔΑ";"ΙΤΙΑ"
"ΚΟΡΥΔΑΛΛΟΣ";"ΜΕΡΑ"
"ΣΠΟΥΡΓΙΤΙ";"ΑΓΡΟΣ"
"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ";"ΒΡΟΧΗ ΤΟΥ ΜΑΙ"
"ΛΙΜΠΕΛΟΥΛΑ";"ΧΛΟΗ"
"ΡΥΑΚΙ";"ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ"
"ΛΟΥΛΟΥΔΙ";"ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ"
"ΚΟΡΑΚΙ";"ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ"
"ΚΟΥΚΟΣ";"ΝΗΣΙ"
"ΠΕΤΑΛΟΥΔΑ";"ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ"
"ΒΙΟΛΕΤΑ";"ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ"
"ΡΥΖΟΧΩΡΑΦΟ";"ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ"
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ";"ΛΙΜΝΗ"
"ΜΕΛΙΣΣΑ";"ΠΑΙΩΝΙΑ"
"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ";"ΘΑΛΑΣΣΑ"
"ΚΑΜΕΛΙΑ";"ΒΡΟΧΗ"
"ΒΑΤΡΑΧΟΣ";"ΛΙΜΝΗ"
"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ";"ΒΟΥΝΟ ΑΣΑΜΑ"
//----------Μπουσόν----------//
"ΦΑΣΙΑΝΟΣ";"ΗΣΥΧΗ ΜΕΡΑ"
"ΒΑΤΡΑΧΟΣ";"ΘΑΜΠΟΦΕΓΓΑΡΟ"
"ΑΛΟΓΟ";"ΟΜΙΧΛΗ"
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ";"ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ"
"ΟΜΠΡΕΛΑ";"ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ"
"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ";"ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ"
"ΧΑΡΤΑΕΤΟΣ";"ΟΥΡΑΝΟΣ"
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ";"ΞΑΣΤΕΡΙΑ"
"ΠΕΤΑΛΟΥΔΑ";"ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ"
"ΗΛΙΟΣ";"ΘΑΛΑΣΣΑ"
"ΣΥΝΝΕΦΟ";"ΒΟΥΝΟ ΓΙΟΣΙΝΟ"
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ";"ΝΑΟΣ"
"ΚΑΜΠΙΑ";"ΠΡΩΙΝΟ ΑΕΡΑΚΙ"
"ΒΕΡΑΝΤΑ";"ΒΡΟΧΕΣ"
"ΣΠΟΥΡΓΙΤΙ";"ΞΑΦΝΙΚΗ ΜΠΟΡΑ"
"ΝΕΟ ΦΥΛΛΟ";"ΒΟΥΝΟ ΦΟΥΤΖΙ"
"ΠΑΙΩΝΙΕΣ";"ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ"
"ΦΤΩΧΕΙΑ";"ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ"
"ΜΟΝΑΞΙΑ";"ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ"
"ΒΕΛΟΣ";"ΟΜΙΧΛΗ"
"ΟΙΩΝΟΣ";"ΒΡΟΧΗ"
"ΣΚΙΑΧΤΡΟ";"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ"

"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ";"ΒΟΥΝΟ"
"ΠΕΤΡΑ";"ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ"
"ΣΑΝΔΑΛΙ";"ΧΙΟΝΟΝΕΡΟ"
//--------Χριστιανόπουλος----------//
"ΚΟΡΜΙ";"ΧΩΜΑ"
"ΚΟΡΜΙ";"ΠΑΤΩΜΑ"
"ΚΟΡΜΙ";"ΚΟΥΡΕΙΟ"
"ΚΟΡΜΙ";"ΔΡΟΜΟΣ"
"ΖΩΣΤΗΡΑΣ";"ΒΡΑΔΥ ΓΛΥΚΟ"
"ΚΟΡΜΙ";"ΔΡΟΜΟΣ"
"ΣΚΕΛΙΑ";"ΜΟΝΑΞΙΑ"
"ΚΟΡΜΙ";"ΜΟΝΑΞΙΑ"
"ΟΛΟΙ";"ΕΡΗΜΙΑ"
"ΣΤΗΘΟΣ";"ΜΟΝΑΞΙΑ"
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ";"ΣΚΑΛΕΣ"
"ΜΑΤΙΑ";"ΜΟΝΑΞΙΑ"
"ΠΟΔΙΑ";"ΝΥΧΤΕΣ"
"ΑΦΑΛΟΣ";"ΝΥΧΤΑ"
"ΜΑΤΙΑ";"ΝΥΧΤΑ"
"ΑΓΚΑΛΙΕΣ";"ΝΤΙΒΑΝΙ"
"ΚΑΘΕΝΑΣ";"ΠΑΛΙΡΡΟΙΑ"
"ΣΚΕΛΙΑ";"ΕΞΑΨΗ"
"ΜΑΤΙΑ";"ΕΚΣΤΑΣΗ"
"ΚΑΘΑΡΜΑΤΑ";"ΠΑΡΚΟ"
"ΣΚΕΛΙΑ";"ΠΑΡΚΟ"
"ΚΟΡΜΙ";"ΝΥΧΤΑ"
"ΤΡΙΧΑ";"ΕΡΩΤΑΣ"
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ";"ΣΙΝΕΜΑ"
"ΓΥΝΑΙΚΑ";"ΔΡΟΜΟΣ"
"ΓΥΦΤΟΣ";"ΑΠΟΜΕΣΗΜΕΡΟ"
"ΧΑΦΙΕΣ";"ΚΑΦΕΝΕΙΟ"
"ΤΣΟΓΛΑΝΙ";"ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ"
"ΑΡΒΥΛΑ";"ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ"

10 Παράρτημα Δ

10.1 Αποτελέσματα ομαδοποιήσεων

10.1.1 2 ομάδες:

Cluster No 1

"ΚΟΡΜΙ", "ΧΩΜΑ",
"ΚΟΡΜΙ", "ΠΑΤΩΜΑ",
"ΚΟΡΜΙ", "ΚΟΥΡΕΙΟ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΖΩΣΤΗΡΑΣ", "ΒΡΑΔΥ ΓΛΥΚΟ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΣΚΕΛΙΑ", "ΜΟΝΑΞΙΑ",
"ΚΟΡΜΙ", "ΜΟΝΑΞΙΑ",
"ΟΛΟΙ", "ΕΡΗΜΙΑ",
"ΣΤΗΘΟΣ", "ΜΟΝΑΞΙΑ",
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ", "ΣΚΑΛΕΣ",
"ΜΑΤΙΑ", "ΜΟΝΑΞΙΑ",
"ΠΟΔΙΑ", "ΝΥΧΤΕΣ",
"ΑΦΑΛΟΣ", "ΝΥΧΤΑ",
"ΜΑΤΙΑ", "ΝΥΧΤΑ",
"ΑΓΚΑΛΙΕΣ", "ΝΤΙΒΑΝΙ",
"ΚΑΘΕΝΑΣ", "ΠΑΛΙΡΡΟΙΑ",
"ΣΚΕΛΙΑ", "ΕΞΑΨΗ",
"ΜΑΤΙΑ", "ΕΚΣΤΑΣΗ",
"ΚΑΘΑΡΜΑΤΑ", "ΠΑΡΚΟ",
"ΣΚΕΛΙΑ", "ΠΑΡΚΟ",
"ΚΟΡΜΙ", "ΝΥΧΤΑ",
"ΤΡΙΧΑ", "ΕΡΩΤΑΣ",
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ", "ΣΙΝΕΜΑ",
"ΓΥΝΑΙΚΑ", "ΔΡΟΜΟΣ",
"ΓΥΦΤΟΣ", "ΑΠΟΜΕΣΗΜΕΡΟ",
"ΧΑΦΙΕΣ", "ΚΑΦΕΝΕΙΟ",
"ΤΣΟΓΛΑΝΙ", "ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ",
"ΑΡΒΥΛΑ", "ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ",

Cluster No 2

"ΡΟΔΟ ΟΡΕΙΝΟ", "ΚΑΤΑΡΡΑΧΤΗΣ",
"ΦΕΓΓΑΡΙ", "ΒΡΟΧΗ",
"ΒΟΥΝΟ ΦΟΥΤΖΙ", "ΟΜΙΧΛΗ",
"ΓΕΡΑΚΙ", "ΚΑΒΟΣ ΙΡΑΓΚΟ",
"ΜΕΛΙΣΣΑ", "ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΙΤΙΑ",
"ΚΟΡΥΔΑΛΛΟΣ", "ΜΕΡΑ",
"ΣΠΟΥΡΓΙΤΙ", "ΑΓΡΟΣ",

"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ", "ΒΡΟΧΗ ΤΟΥ ΜΑΙ",
"ΛΙΜΠΕΛΟΥΛΑ", "ΧΛΟΗ",
"ΡΥΑΚΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΛΟΥΛΟΥΔΙ", "ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ",
"ΚΟΡΑΚΙ", "ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ",
"ΚΟΥΚΟΣ", "ΝΗΣΙ",
"ΠΕΤΑΛΟΥΔΑ", "ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ",
"ΒΙΟΛΕΤΑ", "ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ",
"ΡΥΖΟΧΩΡΑΦΟ", "ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΛΙΜΝΗ",
"ΜΕΛΙΣΣΑ", "ΠΑΙΩΝΙΑ",
"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ", "ΘΑΛΑΣΣΑ",
"ΚΑΜΕΛΙΑ", "ΒΡΟΧΗ",
"ΒΑΤΡΑΧΟΣ", "ΛΙΜΝΗ",
"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ", "ΒΟΥΝΟ ΑΣΑΜΑ",
"ΦΑΣΙΑΝΟΣ", "ΗΣΥΧΗ ΜΕΡΑ",
"ΒΑΤΡΑΧΟΣ", "ΘΑΜΠΟΦΕΓΓΑΡΟ",
"ΑΛΟΓΟ", "ΟΜΙΧΛΗ",
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΜΠΡΕΛΑ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΧΑΡΤΑΕΤΟΣ", "ΟΥΡΑΝΟΣ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΞΑΣΤΕΡΙΑ",
"ΠΕΤΑΛΟΥΔΑ", "ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ",
"ΗΛΙΟΣ", "ΘΑΛΑΣΣΑ",
"ΣΥΝΝΕΦΟ", "ΒΟΥΝΟ ΓΙΟΣΙΝΟ",
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ", "ΝΑΟΣ",
"ΚΑΜΠΙΑ", "ΠΡΩΙΝΟ ΑΕΡΑΚΙ",
"ΒΕΡΑΝΤΑ", "ΒΡΟΧΕΣ",
"ΣΠΟΥΡΓΙΤΙ", "ΞΑΦΝΙΚΗ ΜΠΟΡΑ",
"ΝΕΟ ΦΥΛΛΟ", "ΒΟΥΝΟ ΦΟΥΤΖΙ",
"ΠΑΙΩΝΙΕΣ", "ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ",
"ΦΤΩΧΕΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΜΟΝΑΞΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΒΕΛΟΣ", "ΟΜΙΧΛΗ",
"ΟΙΩΝΟΣ", "ΒΡΟΧΗ",
"ΣΚΙΑΧΤΡΟ", "ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ",
"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ", "ΒΟΥΝΟ",
"ΠΕΤΡΑ", "ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ",
"ΣΑΝΔΑΛΙ", "ΧΙΟΝΟΝΕΡΟ",

10.1.2 3 ομάδες:

Cluster No 1

"ΡΟΔΟ ΟΡΕΙΝΟ", "ΚΑΤΑΡΡΑΧΤΗΣ",
"ΓΕΡΑΚΙ", "ΚΑΒΟΣ ΙΡΑΓΚΟ",
"ΜΕΛΙΣΣΑ", "ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΙΤΙΑ",

"ΚΟΡΥΔΑΛΛΟΣ", "ΜΕΡΑ",
"ΣΠΟΥΡΓΙΤΙ", "ΑΓΡΟΣ",
"ΛΙΜΠΕΛΟΥΛΑ", "ΧΛΟΗ",
"ΛΟΥΛΟΥΔΙ", "ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ",
"ΚΟΡΑΚΙ", "ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ",
"ΚΟΥΚΟΣ", "ΝΗΣΙ",
"ΠΕΤΑΛΟΥΔΑ", "ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ",
"ΒΙΟΛΕΤΑ", "ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΛΙΜΝΗ",
"ΜΕΛΙΣΣΑ", "ΠΑΙΩΝΙΑ",
"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ", "ΘΑΛΑΣΣΑ",
"ΒΑΤΡΑΧΟΣ", "ΛΙΜΝΗ",
"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ", "ΒΟΥΝΟ ΑΣΑΜΑ",
"ΦΑΣΙΑΝΟΣ", "ΗΣΥΧΗ ΜΕΡΑ",
"ΒΑΤΡΑΧΟΣ", "ΘΑΜΠΟΦΕΓΓΑΡΟ",
"ΧΑΡΤΑΕΤΟΣ", "ΟΥΡΑΝΟΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ",
"ΗΛΙΟΣ", "ΘΑΛΑΣΣΑ",
"ΣΥΝΝΕΦΟ", "ΒΟΥΝΟ ΓΙΟΣΙΝΟ",
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ", "ΝΑΟΣ",
"ΚΑΜΠΙΑ", "ΠΡΩΙΝΟ ΑΕΡΑΚΙ",
"ΝΕΟ ΦΥΛΛΟ", "ΒΟΥΝΟ ΦΟΥΤΖΙ",
"ΦΤΩΧΕΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΜΟΝΑΞΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ", "ΒΟΥΝΟ",
"ΠΕΤΡΑ", "ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ",
"ΚΟΡΜΙ", "ΧΩΜΑ",
"ΚΟΡΜΙ", "ΠΑΤΩΜΑ",
"ΚΟΡΜΙ", "ΚΟΥΡΕΙΟ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΖΩΣΤΗΡΑΣ", "ΒΡΑΔΥ ΓΛΥΚΟ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ", "ΣΚΑΛΕΣ",
"ΠΟΔΙΑ", "ΝΥΧΤΕΣ",
"ΑΦΑΛΟΣ", "ΝΥΧΤΑ",
"ΜΑΤΙΑ", "ΝΥΧΤΑ",
"ΑΓΚΑΛΙΕΣ", "ΝΤΙΒΑΝΙ",
"ΚΑΘΕΝΑΣ", "ΠΑΛΙΡΡΟΙΑ",
"ΣΚΕΛΙΑ", "ΠΑΡΚΟ",
"ΚΟΡΜΙ", "ΝΥΧΤΑ",
"ΓΥΦΤΟΣ", "ΑΠΟΜΕΣΗΜΕΡΟ",
"ΑΡΒΥΛΑ", "ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ",

Cluster No 2

"ΣΚΕΛΙΑ", "ΜΟΝΑΞΙΑ",
"ΚΟΡΜΙ", "ΜΟΝΑΞΙΑ",
"ΟΛΟΙ", "ΕΡΗΜΙΑ",
"ΣΤΗΘΟΣ", "ΜΟΝΑΞΙΑ",
"ΜΑΤΙΑ", "ΜΟΝΑΞΙΑ",

"ΣΚΕΛΙΑ", "ΕΞΑΨΗ",
"ΜΑΤΙΑ", "ΕΚΣΤΑΣΗ",
"ΤΡΙΧΑ", "ΕΡΩΤΑΣ",

Cluster No 3

"ΦΕΓΓΑΡΙ", "ΒΡΟΧΗ",
"ΒΟΥΝΟ ΦΟΥΤΖΙ", "ΟΜΙΧΛΗ",
"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ", "ΒΡΟΧΗ ΤΟΥ ΜΑΙ",
"ΡΥΑΚΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΡΥΖΟΧΩΡΑΦΟ", "ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ",
"ΚΑΜΕΛΙΑ", "ΒΡΟΧΗ",
"ΑΛΟΓΟ", "ΟΜΙΧΛΗ",
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΜΠΡΕΛΑ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΞΑΣΤΕΡΙΑ",
"ΒΕΡΑΝΤΑ", "ΒΡΟΧΕΣ",
"ΣΠΟΥΡΓΙΤΙ", "ΞΑΦΝΙΚΗ ΜΠΟΡΑ",
"ΠΑΙΩΝΙΕΣ", "ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ",
"ΒΕΛΟΣ", "ΟΜΙΧΛΗ",
"ΟΙΩΝΟΣ", "ΒΡΟΧΗ",
"ΣΚΙΑΧΤΡΟ", "ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ",
"ΣΑΝΔΑΛΙ", "ΧΙΟΝΟΝΕΡΟ",
"ΚΑΘΑΡΜΑΤΑ", "ΠΑΡΚΟ",
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ", "ΣΙΝΕΜΑ",
"ΓΥΝΑΙΚΑ", "ΔΡΟΜΟΣ",
"ΧΑΦΙΕΣ", "ΚΑΦΕΝΕΙΟ",
"ΤΣΟΓΛΑΝΙ", "ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ",

10.1.3 4 ομάδες:

Cluster No 1

"ΣΚΕΛΙΑ", "ΕΞΑΨΗ",
"ΜΑΤΙΑ", "ΕΚΣΤΑΣΗ",
"ΤΡΙΧΑ", "ΕΡΩΤΑΣ",

Cluster No 2

"ΣΚΕΛΙΑ", "ΜΟΝΑΞΙΑ",
"ΚΟΡΜΙ", "ΜΟΝΑΞΙΑ",
"ΟΛΟΙ", "ΕΡΗΜΙΑ",
"ΣΤΗΘΟΣ", "ΜΟΝΑΞΙΑ",
"ΜΑΤΙΑ", "ΜΟΝΑΞΙΑ",

Cluster No 3

"ΡΟΔΟ ΟΡΕΙΝΟ", "ΚΑΤΑΡΡΑΧΤΗΣ",

"ΓΕΡΑΚΙ", "ΚΑΒΟΣ ΙΡΑΓΚΟ",
"ΜΕΛΙΣΣΑ", "ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΙΤΙΑ",
"ΚΟΡΥΔΑΛΛΟΣ", "ΜΕΡΑ",
"ΣΠΟΥΡΓΙΤΙ", "ΑΓΡΟΣ",
"ΛΙΜΠΕΛΟΥΛΑ", "ΧΛΟΗ",
"ΛΟΥΛΟΥΔΙ", "ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ",
"ΚΟΡΑΚΙ", "ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ",
"ΚΟΥΚΟΣ", "ΝΗΣΙ",
"ΠΕΤΑΛΟΥΔΑ", "ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ",
"ΒΙΟΛΕΤΑ", "ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΛΙΜΝΗ",
"ΜΕΛΙΣΣΑ", "ΠΑΙΩΝΙΑ",
"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ", "ΘΑΛΑΣΣΑ",
"ΒΑΤΡΑΧΟΣ", "ΛΙΜΝΗ",
"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ", "ΒΟΥΝΟ ΑΣΑΜΑ",
"ΦΑΣΙΑΝΟΣ", "ΗΣΥΧΗ ΜΕΡΑ",
"ΒΑΤΡΑΧΟΣ", "ΘΑΜΠΟΦΕΓΓΑΡΟ",
"ΧΑΡΤΑΕΤΟΣ", "ΟΥΡΑΝΟΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ",
"ΗΛΙΟΣ", "ΘΑΛΑΣΣΑ",
"ΣΥΝΝΕΦΟ", "ΒΟΥΝΟ ΓΙΟΣΙΝΟ",
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ", "ΝΑΟΣ",
"ΚΑΜΠΙΑ", "ΠΡΩΙΝΟ ΑΕΡΑΚΙ",
"ΝΕΟ ΦΥΛΛΟ", "ΒΟΥΝΟ ΦΟΥΤΖΙ",
"ΦΤΩΧΕΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΜΟΝΑΞΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ", "ΒΟΥΝΟ",
"ΠΕΤΡΑ", "ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ",
"ΚΟΡΜΙ", "ΧΩΜΑ",
"ΚΟΡΜΙ", "ΠΑΤΩΜΑ",
"ΚΟΡΜΙ", "ΚΟΥΡΕΙΟ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΖΩΣΤΗΡΑΣ", "ΒΡΑΔΥ ΓΛΥΚΟ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ", "ΣΚΑΛΕΣ",
"ΠΟΔΙΑ", "ΝΥΧΤΕΣ",
"ΑΦΑΛΟΣ", "ΝΥΧΤΑ",
"ΜΑΤΙΑ", "ΝΥΧΤΑ",
"ΑΓΚΑΛΙΕΣ", "ΝΤΙΒΑΝΙ",
"ΚΑΘΕΝΑΣ", "ΠΑΛΙΡΡΟΙΑ",
"ΣΚΕΛΙΑ", "ΠΑΡΚΟ",
"ΚΟΡΜΙ", "ΝΥΧΤΑ",
"ΓΥΦΤΟΣ", "ΑΠΟΜΕΣΗΜΕΡΟ",
"ΑΡΒΥΛΑ", "ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ",

Cluster No 4

"ΦΕΓΓΑΡΙ", "ΒΡΟΧΗ",
"ΒΟΥΝΟ ΦΟΥΤΖΙ", "ΟΜΙΧΛΗ",

"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ", "ΒΡΟΧΗ ΤΟΥ ΜΑΙ",
"ΡΥΑΚΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΡΥΖΟΧΩΡΑΦΟ", "ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ",
"ΚΑΜΕΛΙΑ", "ΒΡΟΧΗ",
"ΑΛΟΓΟ", "ΟΜΙΧΛΗ",
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΜΠΡΕΛΑ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΞΑΣΤΕΡΙΑ",
"ΒΕΡΑΝΤΑ", "ΒΡΟΧΕΣ",
"ΣΠΟΥΡΓΙΤΙ", "ΞΑΦΝΙΚΗ ΜΠΟΡΑ",
"ΠΑΙΩΝΙΕΣ", "ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ",
"ΒΕΛΟΣ", "ΟΜΙΧΛΗ",
"ΟΙΩΝΟΣ", "ΒΡΟΧΗ",
"ΣΚΙΑΧΤΡΟ", "ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ",
"ΣΑΝΔΑΛΙ", "ΧΙΟΝΟΝΕΡΟ",
"ΚΑΘΑΡΜΑΤΑ", "ΠΑΡΚΟ",
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ", "ΣΙΝΕΜΑ",
"ΓΥΝΑΙΚΑ", "ΔΡΟΜΟΣ",
"ΧΑΦΙΕΣ", "ΚΑΦΕΝΕΙΟ",
"ΤΣΟΓΛΑΝΙ", "ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ",

10.1.4 5 ομάδες:

Cluster No 1

"ΣΚΕΛΙΑ", "ΜΟΝΑΞΙΑ",
"ΚΟΡΜΙ", "ΜΟΝΑΞΙΑ",
"ΟΛΟΙ", "ΕΡΗΜΙΑ",
"ΣΤΗΘΟΣ", "ΜΟΝΑΞΙΑ",
"ΜΑΤΙΑ", "ΜΟΝΑΞΙΑ",

Cluster No 2

"ΣΚΕΛΙΑ", "ΕΞΑΨΗ",
"ΜΑΤΙΑ", "ΕΚΣΤΑΣΗ",
"ΤΡΙΧΑ", "ΕΡΩΤΑΣ",

Cluster No 3

"ΡΟΔΟ ΟΡΕΙΝΟ", "ΚΑΤΑΡΡΑΧΤΗΣ",
"ΓΕΡΑΚΙ", "ΚΑΒΟΣ ΙΡΑΓΚΟ",
"ΜΕΛΙΣΣΑ", "ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΙΤΙΑ",
"ΚΟΡΥΔΑΛΛΟΣ", "ΜΕΡΑ",
"ΣΠΟΥΡΓΙΤΙ", "ΑΓΡΟΣ",
"ΛΙΜΠΕΛΟΥΛΑ", "ΧΛΟΗ",
"ΛΟΥΛΟΥΔΙ", "ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ",
"ΚΟΡΑΚΙ", "ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ",

"ΚΟΥΚΟΣ", "ΝΗΣΙ",
"ΠΕΤΑΛΟΥΔΑ", "ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ",
"ΒΙΟΛΕΤΑ", "ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΛΙΜΝΗ",
"ΜΕΛΙΣΣΑ", "ΠΑΙΩΝΙΑ",
"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ", "ΘΑΛΑΣΣΑ",
"ΒΑΤΡΑΧΟΣ", "ΛΙΜΝΗ",
"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ", "ΒΟΥΝΟ ΑΣΑΜΑ",
"ΦΑΣΙΑΝΟΣ", "ΗΣΥΧΗ ΜΕΡΑ",
"ΒΑΤΡΑΧΟΣ", "ΘΑΜΠΟΦΕΓΓΑΡΟ",
"ΧΑΡΤΑΕΤΟΣ", "ΟΥΡΑΝΟΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ",
"ΗΛΙΟΣ", "ΘΑΛΑΣΣΑ",
"ΣΥΝΝΕΦΟ", "ΒΟΥΝΟ ΓΙΟΣΙΝΟ",
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ", "ΝΑΟΣ",
"ΚΑΜΠΙΑ", "ΠΡΩΙΝΟ ΑΕΡΑΚΙ",
"ΝΕΟ ΦΥΛΛΟ", "ΒΟΥΝΟ ΦΟΥΤΖΙ",
"ΦΤΩΧΕΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΜΟΝΑΞΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ", "ΒΟΥΝΟ",
"ΠΕΤΡΑ", "ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ",
"ΖΩΣΤΗΡΑΣ", "ΒΡΑΔΥ ΓΛΥΚΟ",
"ΠΟΔΙΑ", "ΝΥΧΤΕΣ",
"ΑΦΑΛΟΣ", "ΝΥΧΤΑ",
"ΜΑΤΙΑ", "ΝΥΧΤΑ",
"ΚΑΘΕΝΑΣ", "ΠΑΛΙΡΡΟΙΑ",
"ΚΟΡΜΙ", "ΝΥΧΤΑ",
"ΓΥΦΤΟΣ", "ΑΠΟΜΕΣΗΜΕΡΟ",

Cluster No 4

"ΦΕΓΓΑΡΙ", "ΒΡΟΧΗ",
"ΒΟΥΝΟ ΦΟΥΤΖΙ", "ΟΜΙΧΛΗ",
"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ", "ΒΡΟΧΗ ΤΟΥ ΜΑΙ",
"ΡΥΑΚΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΡΥΖΟΧΩΡΑΦΟ", "ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ",
"ΚΑΜΕΛΙΑ", "ΒΡΟΧΗ",
"ΑΛΟΓΟ", "ΟΜΙΧΛΗ",
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΜΠΡΕΛΑ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΞΑΣΤΕΡΙΑ",
"ΒΕΡΑΝΤΑ", "ΒΡΟΧΕΣ",
"ΣΠΟΥΡΓΙΤΙ", "ΞΑΦΝΙΚΗ ΜΠΟΡΑ",
"ΠΑΙΩΝΙΕΣ", "ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ",
"ΒΕΛΟΣ", "ΟΜΙΧΛΗ",
"ΟΙΩΝΟΣ", "ΒΡΟΧΗ",
"ΣΚΙΑΧΤΡΟ", "ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ",
"ΣΑΝΔΑΛΙ", "ΧΙΟΝΟΝΕΡΟ",

Cluster No 5

"ΚΟΡΜΙ", "ΧΩΜΑ",
"ΚΟΡΜΙ", "ΠΑΤΩΜΑ",
"ΚΟΡΜΙ", "ΚΟΥΡΕΙΟ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ", "ΣΚΑΛΕΣ",
"ΑΓΚΑΛΙΕΣ", "ΝΤΙΒΑΝΙ",
"ΚΑΘΑΡΜΑΤΑ", "ΠΑΡΚΟ",
"ΣΚΕΛΙΑ", "ΠΑΡΚΟ",
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ", "ΣΙΝΕΜΑ",
"ΓΥΝΑΙΚΑ", "ΔΡΟΜΟΣ",
"ΧΑΦΙΕΣ", "ΚΑΦΕΝΕΙΟ",
"ΤΣΟΓΛΑΝΙ", "ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ",
"ΑΡΒΥΛΑ", "ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ",

10.1.5 6 ομάδες:

Cluster No 1

"ΣΚΕΛΙΑ", "ΜΟΝΑΞΙΑ",
"ΚΟΡΜΙ", "ΜΟΝΑΞΙΑ",
"ΟΛΟΙ", "ΕΡΗΜΙΑ",
"ΣΤΗΘΟΣ", "ΜΟΝΑΞΙΑ",
"ΜΑΤΙΑ", "ΜΟΝΑΞΙΑ",

Cluster No 2

"ΣΚΕΛΙΑ", "ΕΞΑΨΗ",
"ΜΑΤΙΑ", "ΕΚΣΤΑΣΗ",
"ΤΡΙΧΑ", "ΕΡΩΤΑΣ",

Cluster No 3

"ΡΟΔΟ ΟΡΕΙΝΟ", "ΚΑΤΑΡΡΑΧΤΗΣ",
"ΓΕΡΑΚΙ", "ΚΑΒΟΣ ΙΡΑΓΚΟ",
"ΜΕΛΙΣΣΑ", "ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΙΤΙΑ",
"ΚΟΡΥΔΑΛΛΟΣ", "ΜΕΡΑ",
"ΣΠΟΥΡΓΙΤΙ", "ΑΓΡΟΣ",
"ΛΙΜΠΕΛΟΥΛΑ", "ΧΛΟΗ",
"ΛΟΥΛΟΥΔΙ", "ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ",
"ΚΟΡΑΚΙ", "ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ",
"ΚΟΥΚΟΣ", "ΝΗΣΙ",
"ΠΕΤΑΛΟΥΔΑ", "ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ",
"ΒΙΟΛΕΤΑ", "ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΛΙΜΝΗ",
"ΜΕΛΙΣΣΑ", "ΠΑΙΩΝΙΑ",

"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ", "ΘΑΛΑΣΣΑ",
"ΒΑΤΡΑΧΟΣ", "ΛΙΜΝΗ",
"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ", "ΒΟΥΝΟ ΑΣΑΜΑ",
"ΦΑΣΙΑΝΟΣ", "ΗΣΥΧΗ ΜΕΡΑ",
"ΒΑΤΡΑΧΟΣ", "ΘΑΜΠΟΦΕΓΓΑΡΟ",
"ΧΑΡΤΑΕΤΟΣ", "ΟΥΡΑΝΟΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ",
"ΗΛΙΟΣ", "ΘΑΛΑΣΣΑ",
"ΣΥΝΝΕΦΟ", "ΒΟΥΝΟ ΓΙΟΣΙΝΟ",
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ", "ΝΑΟΣ",
"ΚΑΜΠΙΑ", "ΠΡΩΙΝΟ ΑΕΡΑΚΙ",
"ΝΕΟ ΦΥΛΛΟ", "ΒΟΥΝΟ ΦΟΥΤΖΙ",
"ΦΤΩΧΕΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΜΟΝΑΞΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ", "ΒΟΥΝΟ",
"ΠΕΤΡΑ", "ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ",
"ΖΩΣΤΗΡΑΣ", "ΒΡΑΔΥ ΓΛΥΚΟ",
"ΠΟΔΙΑ", "ΝΥΧΤΕΣ",
"ΑΦΑΛΟΣ", "ΝΥΧΤΑ",
"ΜΑΤΙΑ", "ΝΥΧΤΑ",
"ΚΑΘΕΝΑΣ", "ΠΑΛΙΡΡΟΙΑ",
"ΚΟΡΜΙ", "ΝΥΧΤΑ",
"ΓΥΦΤΟΣ", "ΑΠΟΜΕΣΗΜΕΡΟ",

Cluster No 4

"ΦΕΓΓΑΡΙ", "ΒΡΟΧΗ",
"ΒΟΥΝΟ ΦΟΥΤΖΙ", "ΟΜΙΧΛΗ",
"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ", "ΒΡΟΧΗ ΤΟΥ ΜΑΙ",
"ΡΥΑΚΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΡΥΖΟΧΩΡΑΦΟ", "ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ",
"ΚΑΜΕΛΙΑ", "ΒΡΟΧΗ",
"ΑΛΟΓΟ", "ΟΜΙΧΛΗ",
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΜΠΡΕΛΑ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΞΑΣΤΕΡΙΑ",
"ΒΕΡΑΝΤΑ", "ΒΡΟΧΕΣ",
"ΣΠΟΥΡΓΙΤΙ", "ΞΑΦΝΙΚΗ ΜΠΟΡΑ",
"ΠΑΙΩΝΙΕΣ", "ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ",
"ΒΕΛΟΣ", "ΟΜΙΧΛΗ",
"ΟΙΩΝΟΣ", "ΒΡΟΧΗ",
"ΣΚΙΑΧΤΡΟ", "ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ",
"ΣΑΝΔΑΛΙ", "ΧΙΟΝΟΝΕΡΟ",

Cluster No 5

"ΚΟΡΜΙ", "ΧΩΜΑ",
"ΚΟΡΜΙ", "ΠΑΤΩΜΑ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",

"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ", "ΣΚΑΛΕΣ",
"ΑΓΚΑΛΙΕΣ", "ΝΤΙΒΑΝΙ",
"ΚΑΘΑΡΜΑΤΑ", "ΠΑΡΚΟ",
"ΣΚΕΛΙΑ", "ΠΑΡΚΟ",
"ΓΥΝΑΙΚΑ", "ΔΡΟΜΟΣ",
"ΤΣΟΓΛΑΝΙ", "ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ",

Cluster No 6

"ΚΟΡΜΙ", "ΚΟΥΡΕΙΟ",
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ", "ΣΙΝΕΜΑ",
"ΧΑΦΙΕΣ", "ΚΑΦΕΝΕΙΟ",
"ΑΡΒΥΛΑ", "ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ",

10.1.6 7 ομάδες:

Cluster No 1

"ΣΚΕΛΙΑ", "ΜΟΝΑΞΙΑ",
"ΚΟΡΜΙ", "ΜΟΝΑΞΙΑ",
"ΟΛΟΙ", "ΕΡΗΜΙΑ",
"ΣΤΗΘΟΣ", "ΜΟΝΑΞΙΑ",
"ΜΑΤΙΑ", "ΜΟΝΑΞΙΑ",

Cluster No 2

"ΣΚΕΛΙΑ", "ΕΞΑΨΗ",
"ΜΑΤΙΑ", "ΕΚΣΤΑΣΗ",
"ΤΡΙΧΑ", "ΕΡΩΤΑΣ",

Cluster No 3

"ΡΟΔΟ ΟΡΕΙΝΟ", "ΚΑΤΑΡΡΑΧΤΗΣ",
"ΓΕΡΑΚΙ", "ΚΑΒΟΣ ΙΡΑΓΚΟ",
"ΜΕΛΙΣΣΑ", "ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΙΤΙΑ",
"ΚΟΡΥΔΑΛΛΟΣ", "ΜΕΡΑ",
"ΣΠΟΥΡΓΙΤΙ", "ΑΓΡΟΣ",
"ΛΙΜΠΕΛΟΥΛΑ", "ΧΛΟΗ",
"ΛΟΥΛΟΥΔΙ", "ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ",
"ΚΟΡΑΚΙ", "ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ",
"ΚΟΥΚΟΣ", "ΝΗΣΙ",
"ΠΕΤΑΛΟΥΔΑ", "ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ",
"ΒΙΟΛΕΤΑ", "ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΛΙΜΝΗ",
"ΜΕΛΙΣΣΑ", "ΠΑΙΩΝΙΑ",
"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ", "ΘΑΛΑΣΣΑ",
"ΒΑΤΡΑΧΟΣ", "ΛΙΜΝΗ",

"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ", "ΒΟΥΝΟ ΑΣΑΜΑ",
"ΦΑΣΙΑΝΟΣ", "ΗΣΥΧΗ ΜΕΡΑ",
"ΒΑΤΡΑΧΟΣ", "ΘΑΜΠΟΦΕΓΓΑΡΟ",
"ΧΑΡΤΑΕΤΟΣ", "ΟΥΡΑΝΟΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ",
"ΗΛΙΟΣ", "ΘΑΛΑΣΣΑ",
"ΣΥΝΝΕΦΟ", "ΒΟΥΝΟ ΓΙΟΣΙΝΟ",
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ", "ΝΑΟΣ",
"ΚΑΜΠΙΑ", "ΠΡΩΙΝΟ ΑΕΡΑΚΙ",
"ΝΕΟ ΦΥΛΛΟ", "ΒΟΥΝΟ ΦΟΥΤΖΙ",
"ΦΤΩΧΕΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΜΟΝΑΞΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ", "ΒΟΥΝΟ",
"ΠΕΤΡΑ", "ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ",
"ΖΩΣΤΗΡΑΣ", "ΒΡΑΔΥ ΓΛΥΚΟ",
"ΠΟΔΙΑ", "ΝΥΧΤΕΣ",
"ΑΦΑΛΟΣ", "ΝΥΧΤΑ",
"ΜΑΤΙΑ", "ΝΥΧΤΑ",
"ΚΑΘΕΝΑΣ", "ΠΑΛΙΡΡΟΙΑ",
"ΚΟΡΜΙ", "ΝΥΧΤΑ",
"ΓΥΦΤΟΣ", "ΑΠΟΜΕΣΗΜΕΡΟ",
Cluster No 4

"ΦΕΓΓΑΡΙ", "ΒΡΟΧΗ",
"ΒΟΥΝΟ ΦΟΥΤΖΙ", "ΟΜΙΧΛΗ",
"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ", "ΒΡΟΧΗ ΤΟΥ ΜΑΙ",
"ΡΥΑΚΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΡΥΖΟΧΩΡΑΦΟ", "ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ",
"ΚΑΜΕΛΙΑ", "ΒΡΟΧΗ",
"ΑΛΟΓΟ", "ΟΜΙΧΛΗ",
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΜΠΡΕΛΑ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΞΑΣΤΕΡΙΑ",
"ΒΕΡΑΝΤΑ", "ΒΡΟΧΕΣ",
"ΣΠΟΥΡΓΙΤΙ", "ΞΑΦΝΙΚΗ ΜΠΟΡΑ",
"ΠΑΙΩΝΙΕΣ", "ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ",
"ΒΕΛΟΣ", "ΟΜΙΧΛΗ",
"ΟΙΩΝΟΣ", "ΒΡΟΧΗ",
"ΣΚΙΑΧΤΡΟ", "ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ",
"ΣΑΝΔΑΛΙ", "ΧΙΟΝΟΝΕΡΟ",

Cluster No 5

"ΚΟΡΜΙ", "ΧΩΜΑ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΚΑΘΑΡΜΑΤΑ", "ΠΑΡΚΟ",
"ΣΚΕΛΙΑ", "ΠΑΡΚΟ",
"ΓΥΝΑΙΚΑ", "ΔΡΟΜΟΣ",

"ΤΣΟΓΛΑΝΙ", "ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ",

Cluster No 6

"ΚΟΡΜΙ", "ΠΑΤΩΜΑ",
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ", "ΣΚΑΛΕΣ",
"ΑΓΚΑΛΙΕΣ", "ΝΤΙΒΑΝΙ",

Cluster No 7

"ΚΟΡΜΙ", "ΚΟΥΡΕΙΟ",
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ", "ΣΙΝΕΜΑ",
"ΧΑΦΙΕΣ", "ΚΑΦΕΝΕΙΟ",
"ΑΡΒΥΛΑ", "ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ",

10.1.7 8 ομάδες:

Cluster No 1

"ΟΛΟΙ", "ΕΡΗΜΙΑ",
"ΣΚΕΛΙΑ", "ΕΞΑΨΗ",
"ΜΑΤΙΑ", "ΕΚΣΤΑΣΗ",
"ΤΡΙΧΑ", "ΕΡΩΤΑΣ",

Cluster No 2

"ΣΚΕΛΙΑ", "ΜΟΝΑΞΙΑ",
"ΚΟΡΜΙ", "ΜΟΝΑΞΙΑ",
"ΣΤΗΘΟΣ", "ΜΟΝΑΞΙΑ",
"ΜΑΤΙΑ", "ΜΟΝΑΞΙΑ",

Cluster No 3

"ΦΤΩΧΕΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΜΟΝΑΞΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΖΩΣΤΗΡΑΣ", "ΒΡΑΔΥ ΓΛΥΚΟ",
"ΠΟΔΙΑ", "ΝΥΧΤΕΣ",
"ΑΦΑΛΟΣ", "ΝΥΧΤΑ",
"ΜΑΤΙΑ", "ΝΥΧΤΑ",
"ΚΟΡΜΙ", "ΝΥΧΤΑ",
"ΓΥΦΤΟΣ", "ΑΠΟΜΕΣΗΜΕΡΟ",

Cluster No 4

"ΦΕΓΓΑΡΙ", "ΒΡΟΧΗ",
"ΒΟΥΝΟ ΦΟΥΤΖΙ", "ΟΜΙΧΛΗ",
"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ", "ΒΡΟΧΗ ΤΟΥ ΜΑΙ",
"ΡΥΑΚΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΡΥΖΟΧΩΡΑΦΟ", "ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ",

"ΚΑΜΕΛΙΑ", "ΒΡΟΧΗ",
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΜΠΡΕΛΑ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΞΑΣΤΕΡΙΑ",
"ΒΕΡΑΝΤΑ", "ΒΡΟΧΕΣ",
"ΠΑΙΩΝΙΕΣ", "ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ",
"ΒΕΛΟΣ", "ΟΜΙΧΛΗ",
"ΟΙΩΝΟΣ", "ΒΡΟΧΗ",
"ΣΚΙΑΧΤΡΟ", "ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ",
"ΣΑΝΔΑΛΙ", "ΧΙΟΝΟΝΕΡΟ",

Cluster No 5

"ΚΟΡΜΙ", "ΧΩΜΑ",
"ΚΟΡΜΙ", "ΠΑΤΩΜΑ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ", "ΣΚΑΛΕΣ",
"ΑΓΚΑΛΙΕΣ", "ΝΤΙΒΑΝΙ",
"ΚΑΘΑΡΜΑΤΑ", "ΠΑΡΚΟ",
"ΣΚΕΛΙΑ", "ΠΑΡΚΟ",
"ΓΥΝΑΙΚΑ", "ΔΡΟΜΟΣ",
"ΤΣΟΓΛΑΝΙ", "ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ",

Cluster No 6

"ΚΟΡΜΙ", "ΚΟΥΡΕΙΟ",
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ", "ΣΙΝΕΜΑ",
"ΧΑΦΙΕΣ", "ΚΑΦΕΝΕΙΟ",
"ΑΡΒΥΛΑ", "ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ",

Cluster No 7

"ΚΑΘΕΝΑΣ", "ΠΑΛΙΡΡΟΙΑ",

Cluster No 8

"ΡΟΔΟ ΟΡΕΙΝΟ", "ΚΑΤΑΡΡΑΧΤΗΣ",
"ΓΕΡΑΚΙ", "ΚΑΒΟΣ ΙΡΑΓΚΟ",
"ΜΕΛΙΣΣΑ", "ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΙΤΙΑ",
"ΚΟΡΥΔΑΛΛΟΣ", "ΜΕΡΑ",
"ΣΠΟΥΡΓΙΤΙ", "ΑΓΡΟΣ",
"ΛΙΜΠΕΛΟΥΛΑ", "ΧΛΟΗ",
"ΛΟΥΛΟΥΔΙ", "ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ",
"ΚΟΡΑΚΙ", "ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ",
"ΚΟΥΚΟΣ", "ΝΗΣΙ",
"ΠΕΤΑΛΟΥΔΑ", "ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ",
"ΒΙΟΛΕΤΑ", "ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ",

"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΛΙΜΝΗ",
"ΜΕΛΙΣΣΑ", "ΠΑΙΩΝΙΑ",
"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ", "ΘΑΛΑΣΣΑ",
"ΒΑΤΡΑΧΟΣ", "ΛΙΜΝΗ",
"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ", "ΒΟΥΝΟ ΑΣΑΜΑ",
"ΦΑΣΙΑΝΟΣ", "ΗΣΥΧΗ ΜΕΡΑ",
"ΒΑΤΡΑΧΟΣ", "ΘΑΜΠΟΦΕΓΓΑΡΟ",
"ΑΛΟΓΟ", "ΟΜΙΧΛΗ",
"ΧΑΡΤΑΕΤΟΣ", "ΟΥΡΑΝΟΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ",
"ΗΛΙΟΣ", "ΘΑΛΑΣΣΑ",
"ΣΥΝΝΕΦΟ", "ΒΟΥΝΟ ΓΙΟΣΙΝΟ",
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ", "ΝΑΟΣ",
"ΚΑΜΠΙΑ", "ΠΡΩΙΝΟ ΑΕΡΑΚΙ",
"ΣΠΟΥΡΓΙΤΙ", "ΞΑΦΝΙΚΗ ΜΠΟΡΑ",
"ΝΕΟ ΦΥΛΛΟ", "ΒΟΥΝΟ ΦΟΥΤΖΙ",
"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ", "ΒΟΥΝΟ",
"ΠΕΤΡΑ", "ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ",

10.1.8 9 ομάδες:

Cluster No 1

"ΟΛΟΙ", "ΕΡΗΜΙΑ",
"ΣΚΕΛΙΑ", "ΕΞΑΨΗ",
"ΜΑΤΙΑ", "ΕΚΣΤΑΣΗ",
"ΤΡΙΧΑ", "ΕΡΩΤΑΣ",

Cluster No 2

"ΣΚΕΛΙΑ", "ΜΟΝΑΞΙΑ",
"ΚΟΡΜΙ", "ΜΟΝΑΞΙΑ",
"ΣΤΗΘΟΣ", "ΜΟΝΑΞΙΑ",
"ΜΑΤΙΑ", "ΜΟΝΑΞΙΑ",

Cluster No 3

"ΦΤΩΧΕΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΜΟΝΑΞΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΖΩΣΤΗΡΑΣ", "ΒΡΑΔΥ ΓΛΥΚΟ",
"ΠΟΔΙΑ", "ΝΥΧΤΕΣ",
"ΑΦΑΛΟΣ", "ΝΥΧΤΑ",
"ΜΑΤΙΑ", "ΝΥΧΤΑ",
"ΚΟΡΜΙ", "ΝΥΧΤΑ",
"ΓΥΦΤΟΣ", "ΑΠΟΜΕΣΗΜΕΡΟ",

Cluster No 4

"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",

Cluster No 5

"ΚΟΡΜΙ", "ΧΩΜΑ",
"ΚΟΡΜΙ", "ΠΑΤΩΜΑ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ", "ΣΚΑΛΕΣ",
"ΑΓΚΑΛΙΕΣ", "ΝΤΙΒΑΝΙ",
"ΚΑΘΑΡΜΑΤΑ", "ΠΑΡΚΟ",
"ΣΚΕΛΙΑ", "ΠΑΡΚΟ",
"ΓΥΝΑΙΚΑ", "ΔΡΟΜΟΣ",
"ΤΣΟΓΛΑΝΙ", "ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ",

Cluster No 6

"ΚΟΡΜΙ", "ΚΟΥΡΕΙΟ",
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ", "ΣΙΝΕΜΑ",
"ΧΑΦΙΕΣ", "ΚΑΦΕΝΕΙΟ",
"ΑΡΒΥΛΑ", "ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ",

Cluster No 7

"ΚΑΘΕΝΑΣ", "ΠΑΛΙΡΡΟΙΑ",

Cluster No 8

"ΡΟΔΟ ΟΡΕΙΝΟ", "ΚΑΤΑΡΡΑΧΤΗΣ",
"ΓΕΡΑΚΙ", "ΚΑΒΟΣ ΙΡΑΓΚΟ",
"ΜΕΛΙΣΣΑ", "ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΙΤΙΑ",
"ΚΟΡΥΔΑΛΛΟΣ", "ΜΕΡΑ",
"ΣΠΟΥΡΓΙΤΙ", "ΑΓΡΟΣ",
"ΛΙΜΠΕΛΟΥΛΑ", "ΧΛΟΗ",
"ΛΟΥΛΟΥΔΙ", "ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ",
"ΚΟΡΑΚΙ", "ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ",
"ΚΟΥΚΟΣ", "ΝΗΣΙ",
"ΠΕΤΑΛΟΥΔΑ", "ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ",
"ΒΙΟΛΕΤΑ", "ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΛΙΜΝΗ",
"ΜΕΛΙΣΣΑ", "ΠΑΙΩΝΙΑ",
"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ", "ΘΑΛΑΣΣΑ",
"ΒΑΤΡΑΧΟΣ", "ΛΙΜΝΗ",
"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ", "ΒΟΥΝΟ ΑΣΑΜΑ",
"ΦΑΣΙΑΝΟΣ", "ΗΣΥΧΗ ΜΕΡΑ",
"ΒΑΤΡΑΧΟΣ", "ΘΑΜΠΟΦΕΓΓΑΡΟ",
"ΧΑΡΤΑΕΤΟΣ", "ΟΥΡΑΝΟΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ",
"ΗΛΙΟΣ", "ΘΑΛΑΣΣΑ",
"ΣΥΝΝΕΦΟ", "ΒΟΥΝΟ ΓΙΟΣΙΝΟ",
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ", "ΝΑΟΣ",

"ΚΑΜΠΙΑ", "ΠΡΩΙΝΟ ΑΕΡΑΚΙ",
"ΝΕΟ ΦΥΛΛΟ", "ΒΟΥΝΟ ΦΟΥΤΖΙ",
"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ", "ΒΟΥΝΟ",
"ΠΕΤΡΑ", "ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ",

Cluster No 9
"ΦΕΓΓΑΡΙ", "ΒΡΟΧΗ",
"ΒΟΥΝΟ ΦΟΥΤΖΙ", "ΟΜΙΧΛΗ",
"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ", "ΒΡΟΧΗ ΤΟΥ ΜΑΙ",
"ΡΥΑΚΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΡΥΖΟΧΩΡΑΦΟ", "ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ",
"ΚΑΜΕΛΙΑ", "ΒΡΟΧΗ",
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΜΠΡΕΛΑ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΞΑΣΤΕΡΙΑ",
"ΒΕΡΑΝΤΑ", "ΒΡΟΧΕΣ",
"ΠΑΙΩΝΙΕΣ", "ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ",
"ΒΕΛΟΣ", "ΟΜΙΧΛΗ",
"ΟΙΩΝΟΣ", "ΒΡΟΧΗ",
"ΣΚΙΑΧΤΡΟ", "ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ",
"ΣΑΝΔΑΛΙ", "ΧΙΟΝΟΝΕΡΟ",

10.1.9 10 ομάδες:

Cluster No 1

"ΟΛΟΙ", "ΕΡΗΜΙΑ",
"ΣΚΕΛΙΑ", "ΕΞΑΨΗ",
"ΜΑΤΙΑ", "ΕΚΣΤΑΣΗ",
"ΤΡΙΧΑ", "ΕΡΩΤΑΣ",

Cluster No 2

"ΣΚΕΛΙΑ", "ΜΟΝΑΞΙΑ",
"ΚΟΡΜΙ", "ΜΟΝΑΞΙΑ",
"ΣΤΗΘΟΣ", "ΜΟΝΑΞΙΑ",
"ΜΑΤΙΑ", "ΜΟΝΑΞΙΑ",

Cluster No 3

"ΦΤΩΧΕΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΜΟΝΑΞΙΑ", "ΦΘΙΝΟΠΩΡΙΝΟ ΒΡΑΔΥ",
"ΖΩΣΤΗΡΑΣ", "ΒΡΑΔΥ ΓΛΥΚΟ",
"ΠΟΔΙΑ", "ΝΥΧΤΕΣ",
"ΑΦΑΛΟΣ", "ΝΥΧΤΑ",
"ΜΑΤΙΑ", "ΝΥΧΤΑ",
"ΚΟΡΜΙ", "ΝΥΧΤΑ",
"ΓΥΦΤΟΣ", "ΑΠΟΜΕΣΗΜΕΡΟ",

Cluster No 4

"ΟΠΟΣ ΔΕΝ ΜΠΟΡΕΙ ΝΑ ΓΡΑΨΕΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",

Cluster No 5

"ΚΟΡΜΙ", "ΧΩΜΑ",
"ΚΟΡΜΙ", "ΠΑΤΩΜΑ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΚΟΡΜΙ", "ΔΡΟΜΟΣ",
"ΦΑΤΣΕΣ ΨΥΧΡΕΣ", "ΣΚΑΛΕΣ",
"ΑΓΚΑΛΙΕΣ", "ΝΤΙΒΑΝΙ",
"ΚΑΘΑΡΜΑΤΑ", "ΠΑΡΚΟ",
"ΣΚΕΛΙΑ", "ΠΑΡΚΟ",
"ΓΥΝΑΙΚΑ", "ΔΡΟΜΟΣ",
"ΤΣΟΓΛΑΝΙ", "ΤΑΝΤΑΛΟΥ ΚΑΙ ΣΑΠΦΟΥΣ ΓΩΝΙΑ",

Cluster No 6

"ΚΟΡΜΙ", "ΚΟΥΡΕΙΟ",
"ΠΑΙΔΙΑ ΠΑΡΑΞΕΝΑ", "ΣΙΝΕΜΑ",
"ΧΑΦΙΕΣ", "ΚΑΦΕΝΕΙΟ",
"ΑΡΒΥΛΑ", "ΠΑΠΟΥΤΣΑΔΙΚΑ ΤΟΥ ΒΑΡΔΑΡΙ",

Cluster No 7

"ΚΑΘΕΝΑΣ", "ΠΑΛΙΡΡΟΙΑ",

Cluster No 8

"ΡΟΔΟ ΟΡΕΙΝΟ", "ΚΑΤΑΡΡΑΧΤΗΣ",
"ΓΕΡΑΚΙ", "ΚΑΒΟΣ ΙΡΑΓΚΟ",
"ΜΕΛΙΣΣΑ", "ΚΑΡΔΙΑ ΤΗΣ ΠΑΙΩΝΙΑΣ",
"ΠΕΤΑΛΟΥΔΑ", "ΙΤΙΑ",
"ΚΟΡΥΔΑΛΛΟΣ", "ΜΕΡΑ",
"ΣΠΟΥΡΓΙΤΙ", "ΑΓΡΟΣ",
"ΛΙΜΠΕΛΟΥΛΑ", "ΧΛΟΗ",
"ΛΟΥΛΟΥΔΙ", "ΑΝΟΙΞΙΑΤΙΚΟ ΦΕΓΓΑΡΙ",
"ΚΟΡΑΚΙ", "ΦΘΙΝΟΠΩΡΙΝΗ ΝΥΧΤΑ",
"ΚΟΥΚΟΣ", "ΝΗΣΙ",
"ΠΕΤΑΛΟΥΔΑ", "ΛΕΥΚΗ ΠΑΠΑΡΟΥΝΑ",
"ΒΙΟΛΕΤΑ", "ΟΡΕΙΝΟ ΜΟΝΟΠΑΤΙ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΛΙΜΝΗ",
"ΜΕΛΙΣΣΑ", "ΠΑΙΩΝΙΑ",
"ΠΟΤΑΜΟΣ ΜΟΓΚΑΜΙ", "ΘΑΛΑΣΣΑ",
"ΒΑΤΡΑΧΟΣ", "ΛΙΜΝΗ",
"ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ", "ΒΟΥΝΟ ΑΣΑΜΑ",
"ΦΑΣΙΑΝΟΣ", "ΗΣΥΧΗ ΜΕΡΑ",
"ΒΑΤΡΑΧΟΣ", "ΘΑΜΠΟΦΕΓΓΑΡΟ",
"ΧΑΡΤΑΕΤΟΣ", "ΟΥΡΑΝΟΣ",

"ΠΕΤΑΛΟΥΔΑ", "ΚΑΜΠΑΝΑ ΤΟΥ ΝΑΟΥ",
"ΗΛΙΟΣ", "ΘΑΛΑΣΣΑ",
"ΣΥΝΝΕΦΟ", "ΒΟΥΝΟ ΓΙΟΣΙΝΟ",
"ΑΝΘΟΣ ΚΕΡΑΣΙΑΣ", "ΝΑΟΣ",
"ΚΑΜΠΙΑ", "ΠΡΩΙΝΟ ΑΕΡΑΚΙ",
"ΝΕΟ ΦΥΛΛΟ", "ΒΟΥΝΟ ΦΟΥΤΖΙ",
"ΘΦΙΝΟΠΩΡΙΝΟ ΦΥΛΛΟ", "ΒΟΥΝΟ",
"ΠΕΤΡΑ", "ΨΥΧΡΟ ΣΕΛΗΝΟΦΩΣ",

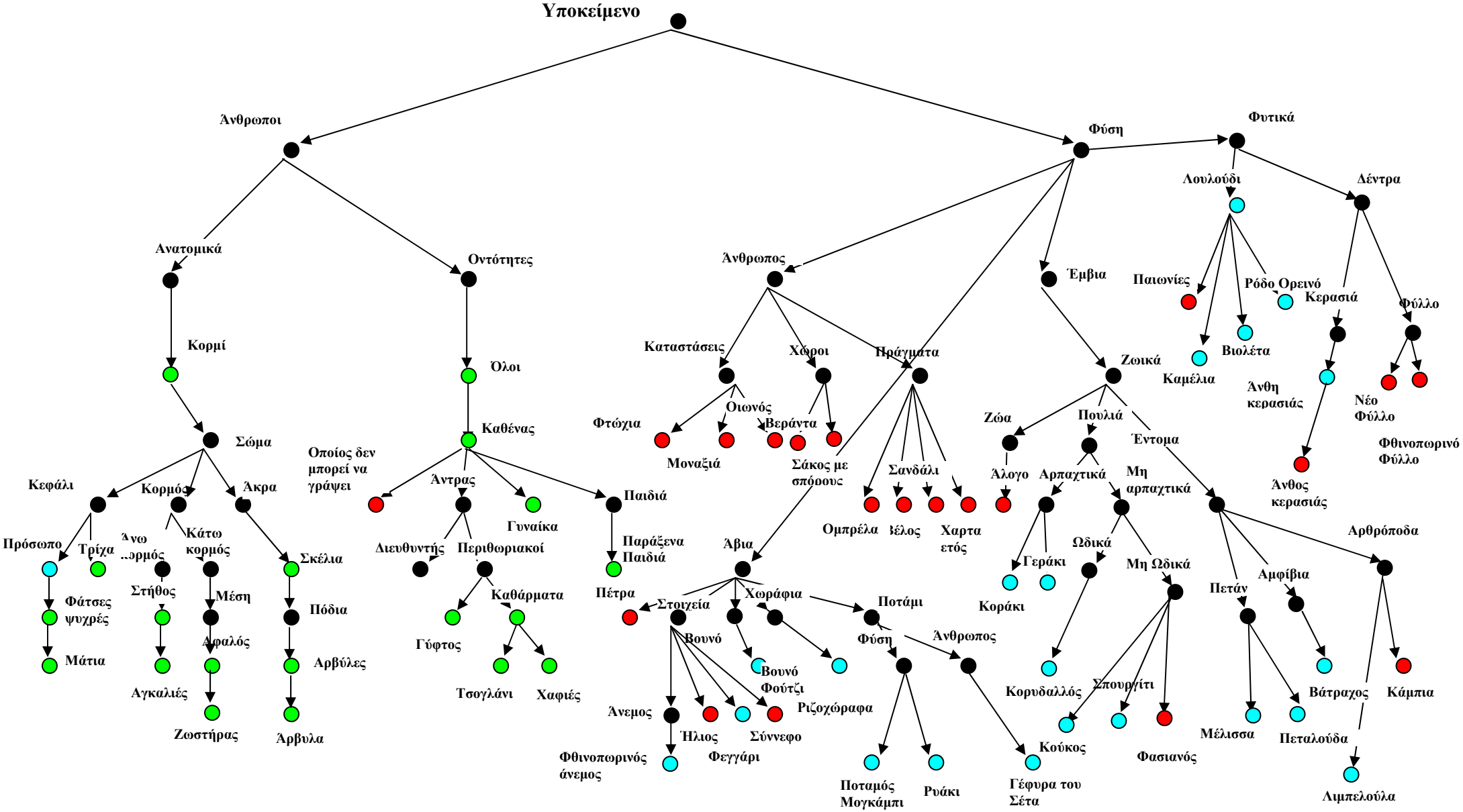
Cluster No 9

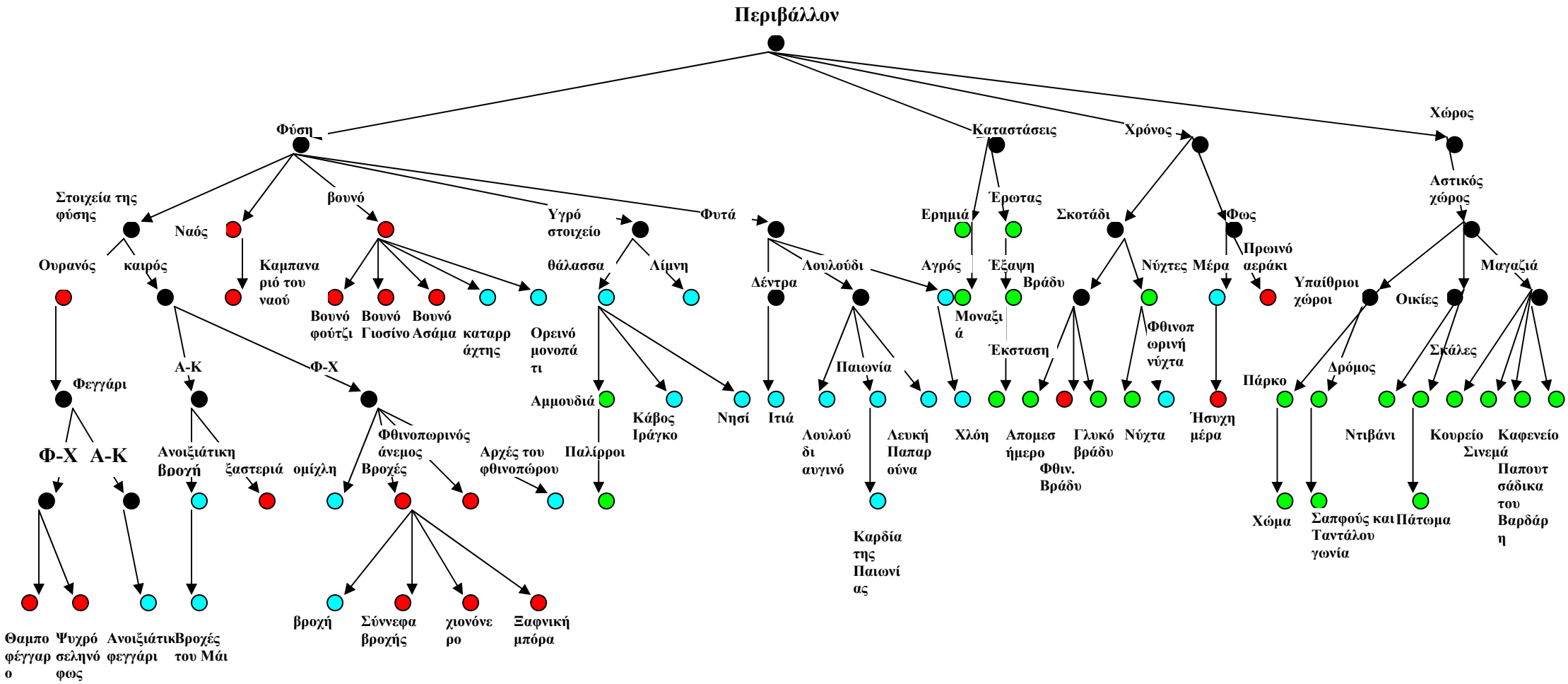
"ΓΕΦΥΡΑ ΤΟΥ ΣΕΤΑ", "ΒΡΟΧΗ ΤΟΥ ΜΑΙ",
"ΡΥΑΚΙ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΣΑΚΟΣ ΜΕ ΣΠΟΡΟΥΣ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΟΜΠΡΕΛΑ", "ΑΝΟΙΞΙΑΤΙΚΗ ΒΡΟΧΗ",
"ΑΝΘΗ ΚΕΡΑΣΙΑΣ", "ΞΑΣΤΕΡΙΑ",

Cluster No 10

"ΦΕΓΓΑΡΙ", "ΒΡΟΧΗ",
"ΒΟΥΝΟ ΦΟΥΤΖΙ", "ΟΜΙΧΛΗ",
"ΡΥΖΟΧΩΡΑΦΟ", "ΑΡΧΕΣ ΦΘΙΝΟΠΩΡΟΥ",
"ΚΑΜΕΛΙΑ", "ΒΡΟΧΗ",
"ΑΛΟΓΟ", "ΟΜΙΧΛΗ",
"ΒΕΡΑΝΤΑ", "ΒΡΟΧΕΣ",
"ΣΠΟΥΡΓΙΤΙ", "ΞΑΦΝΙΚΗ ΜΠΟΡΑ",
"ΠΑΙΩΝΙΕΣ", "ΣΥΝΝΕΦΑ ΒΡΟΧΗΣ",
"ΒΕΛΟΣ", "ΟΜΙΧΛΗ",
"ΟΙΩΝΟΣ", "ΒΡΟΧΗ",
"ΣΚΙΑΧΤΡΟ", "ΦΘΙΝΟΠΩΡΙΝΟΣ ΑΝΕΜΟΣ",
"ΣΑΝΔΑΛΙ", "ΧΙΟΝΟΝΕΡΟ",

11 Παράρτημα Ε


Υποκείμενο
Άνθρωποι
Φύση
Ανατομικά
Οντότητες
Άνθρωπος
Έμβια
Φυτικά
Κορμί
Όλοι
Καταστάσεις
Χώροι
Πράγματα
Ζωικά
Λουλούδι
Δέντρα
Σώμα
Καθένας
Φτώχια
Οιωνός
Βεράντα
Μοναξιά
Σάκος με σπόρους
Ζώα
Πουλιά
Έντομα
Παιωνίες
Ρόδο Ορεινό
Κερασιά
Φύλλο
Καμέλια
Βιολέτα
Άνθη κερασιάς
Νέο Φύλλο
Κεφάλι
Κορμός
Άκρα
Οποίος δεν μπορεί να γράψει
Άντρας
Παιδιά
Σανδάλι
Άλογο
Αρπαχτικά
Μη αρπαχτικά
Φθινοπωρινό Φύλλο
Άνθος κερασιάς
Πρόσωπο
Τρίχα
Άνω κορμός
Κάτω κορμός
Γυναίκα
Διευθυντής
Περιθωριακοί
Παράξενα Παιδιά
Ομπρέλα
Χαρτοπετσέτος
Άβια
Γεράκι
Κοράκι
Ωδικά
Μη Ωδικά
Αρθρόποδα
Φάτσες ψυχρές
Στήθος
Μέση
Σκέλια
Καθάρματα
Πέτρα
Στοιχεία
Χωράφια
Ποτάμι
Πετάν
Αμφίβια
Μάτια
Αφαλός
Πόδια
Γύφτος
Βουνό
Φύση
Άνθρωπος
Κορυδαλλός
Σπουργίτι
Βάτραχος
Κάμπια
Αγκαλιές
Αρβύλες
Τσογλάνι
Χαφιές
Βουνό Φούτζι
Ριζοχώραφα
Άνεμος
Ήλιος
Σύννεφο
Μέλισσα
Πεταλούδα
Ζωστήρας
Άρβυλα
Φθινοπωρινός άνεμος
Φεγγάρι
Ποταμός Μογκάμπι
Ρυάκι
Κούκος
Φασιανός
Γέφυρα του Σέτα
Λιμπελούλα


Περιβάλλον
Φύση
Καταστάσεις
Χρόνος
Χώρος
Στοιχεία της φύσης
Ναός
βουνό
Υγρό στοιχείο
Φυτά
Ερημιά
Έρωτας
Σκοτάδι
Φως
Αστικός χώρος
Ουρανός
καιρός
Καμπαναριό του ναού
Βουνό φούτζι
Βουνό Γιοσίνο
Βουνό Ασάμα
καταρράχτης
Ορεινό μονοπάτι
θάλασσα
Λίμνη
Δέντρα
Λουλούδι
Αγρός
Έξαψη
Βράδυ
Μονάξιά
Έκσταση
Νύχτες
Μέρα
Πρωινό αεράκι
Υπαίθριοι χώροι
Οικίες
Μαγαζιά
Φεγγάρι
Α-Κ
Φ-Χ
Αμμουδιά
Κάβος Ιράγκο
Νησί
Ιτιά
Παιωνία
Χλόη
Απομεσήμερο
Φθιν. Βράδυ
Γλυκό βράδυ
Νύχτα
Φθινοπωρινή νύχτα
Ήσυχη μέρα
Πάρκο
Δρόμος
Σκάλες
Φ-Χ
Α-Κ
Ανοιξιάτικη βροχή
ξαστεριά
ομίχλη
Βροχές
Φθινοπωρινός άνεμος
Αρχές του φθινοπώρου
Παλίρροι
Λουλούδι αυγινό
Λευκή Παπαρούνα
Καρδία της Παιωνίας
Χώμα
Σαπφούς και Ταντάλου γωνία
Πάτωμα
Ντιβάνι
Κουρείο Σινεμά
Καφενείο
Παπουτσάδικα του Βαρδάρη
Θαμπό φέγγαρο
Ψυχρό σεληνόφως
Ανοιξιάτικο φεγγάρι
Βροχές του Μάι
βροχή
Σύννεφα βροχής
χιονόνερο
Ξαφνική μπόρα