ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Αλγόριθμοι ομαδοποίησης και μείωσης διάστασης για δεδομένα του Παγκοσμίου Ιστού

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΕΠΑΜΕΙΝΩΝΔΑ Δ. ΦΡΙΤΖΙΛΑ

Επιβλέπων:
Γ. Παλιούρας
Ερευνητής ΕΚΕΦΕ Δημόκριτος

Αθήνα, Σεπτέμβριος 2003

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Αλγόριθμοι ομαδοποίησης και μείωσης διάστασης για δεδομένα του Παγκοσμίου Ιστού

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΕΠΑΜΕΙΝΩΝΔΑ Δ. ΦΡΙΤΖΙΛΑ

Επιβλέπων:
Γ. Παλιούρας
Ερευνητής ΕΚΕΦΕ Δημόκριτος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την $23^{η}$ Σεπτεμβρίου 2003

Τ. Σελλής	**Α. Σταφυλοπάτης**	**Π. Τσανάκας**
Καθηγητής ΕΜΠ	**Καθηγητής ΕΜΠ**	**Καθηγητής ΕΜΠ**

Αθήνα, Σεπτέμβριος 2003

ΕΠΑΜΕΙΝΩΝΔΑΣ Δ. ΦΡΙΤΖΙΛΑΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών ΕΜΠ

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε από τον Απρίλιο έως τον Σεπτέμβριο του 2003, κατά τη διάρκεια της φοίτησής μου στο $10^ο$ εξάμηνο σπουδών του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ του ΕΜΠ. Το θέμα επιλέχθηκε ύστερα από συζήτηση με τους κ.κ. Γιώργο Παλιούρα, ερευνητή του ΕΚΕΦΕ Δημόκριτος, και Δημήτρη Πιερράκο, υποψήφιο διδάκτορα του Πανεπιστημίου Αθηνών. Σε αυτούς απευθύνω τις ευχαριστίες μου, αφενός γιατί με έφεραν σε επαφή με μια σύγχρονη και ενδιαφέρουσα ερευνητική περιοχή της Πληροφορικής και αφετέρου γιατί μου έδωσαν χρήσιμες υποδείξεις όταν τις χρειάστηκα. Τις ευχαριστίες μου επίσης εκφράζω στον Καθηγητή Σελλή, ο οποίος ήταν πρόθυμος να επισπεύσει τη διαδικασία της προφορικής εξέτασης, προκειμένου να προλάβω κάποια προθεσμία που μου είχε επιβληθεί. Επίσης, ευχαριστώ τα μέλη της οικογένειάς μου για την κατανόηση που έδειξαν, όταν αναγκάστηκα να αφιερώσω σχεδόν όλο τον χρόνο μου στην ολοκλήρωση της εργασίας. Τέλος, ευχαριστώ τις ξαδέρφες μου Ντίνα και Πένυ για τον υπολογιστή που μου δάνεισαν, όταν το δικό μου παλιάς τεχνολογίας PC έφτασε στα όριά του.

Νώντας Φριτζίλας

Περίληψη

Ο κύριος στόχος της παρούσας διπλωματικής εργασίας είναι η υλοποίηση και πειραματική αξιολόγηση δυο συγκεκριμένων αλγορίθμων μείωσης διάστασης, που προορίζονται να χρησιμοποιηθούν πάνω σε δεδομένα του Παγκοσμίου Ιστού. Ο πρώτος αλγόριθμος που εξετάζουμε στηρίζεται στη διαδικασία της ομαδοποίησης, ενώ ο δεύτερος στηρίζεται σε μια τεχνική που ανήκει στην οικογένεια της Πιθανοτικής Ανάλυσης Κρυμμένης Σημασιολογίας. Συνεπώς, η πορεία προς την εκπλήρωση του τελικού στόχου περιλαμβάνει τέσσερα διακριτά στάδια: την υλοποίηση ορισμένων αλγορίθμων ομαδοποίησης, την υλοποίηση του αλγορίθμου Πιθανοτικής Ανάλυσης Κρυμμένης Σημασιολογίας, την αξιοποίηση των παραπάνω στα πλαίσια των δυο αλγορίθμων μείωσης διάστασης και, τέλος, την πειραματική αξιολόγηση των τελευταίων πάνω σε πραγματικά δεδομένα του Ιστού.

Στο πρώτο στάδιο αναπτύσσεται σε Java ένα πλαίσιο εργασίας λογισμικού, που βασίζεται στις κατάλληλες δομές δεδομένων και στον απαραίτητο αντικειμενοστραφή σχεδιασμό, προκειμένου να χρησιμοποιηθεί για την υλοποίηση αλγορίθμων ομαδοποίησης δεδομένων του Ιστού. Επιπλέον, υλοποιούνται τρεις συγκεκριμένοι αλγόριθμοι ομαδοποίησης, η συμπεριφορά των οποίων εξετάζεται πάνω σε ένα συγκεκριμένο σύνολο δεδομένων.

Στο δεύτερο στάδιο υλοποιείται σε Java αλγόριθμος που ανήκει στην ευρύτερη οικογένεια της Πιθανοτικής Ανάλυσης Κρυμμένης Σημασιολογίας. Ο συγκεκριμένος αλγόριθμος υιοθετεί ένα πιθανοτικό μοντέλο παραγωγής των δεδομένων από κάποιες μη ορατές μεταβλητές, το οποίο ονομάζεται Μοντέλο των Όψεων. Στη συνέχεια, χρησιμοποιώντας την τεχνική της Μεγιστοποίηση της Αναμενόμενης Τιμής κατά τη διάρκεια μιας διαδικασίας εκπαίδευσης, καταλήγει σε μια τοπικά βέλτιστη εκτίμηση των παραμέτρων του μοντέλου.

Στο τρίτο στάδιο υλοποιούνται σε Java οι δύο αλγόριθμοι μείωσης διάστασης που αποτελούσαν από την αρχή τον βασικό στόχο της εργασίας. Ο πρώτος αλγόριθμος βασίζεται στην ομαδοποίηση και ονομάζεται Εννοιολογική Δεικτοδότηση, ενώ ο δεύτερος βασίζεται στην εκτίμηση των παραμέτρων του πιθανοτικού Μοντέλου των Όψεων. Ένας αλγόριθμος μείωσης διάστασης επιδιώκει, σε γενικές γραμμές, να απεικονίσει τα πολυδιάστατα διανύσματα των δεδομένων που του παρέχονται σε έναν χώρο μικρότερης διάστασης από τον αρχικό. Τα κίνητρα γι' αυτήν την προσπάθεια είναι αφενός η εξοικονόμηση υπολογιστικών πόρων και αφετέρου η ανακάλυψη συσχετίσεων μεταξύ στοιχείων του συνόλου δεδομένων, που δεν είναι ορατές στην πολυδιάστατη αναπαράσταση.

Στο τελευταίο στάδιο της εργασίας αξιολογείται η αποτελεσματικότητα της μείωσης διάστασης ως διαδικασίας ανακάλυψης κρυμμένων συσχετίσεων. Προφανώς, εξετάζουμε τους δυο συγκεκριμένους αλγορίθμους μείωσης διάστασης που υλοποιούμε στα πλαίσια της εργασίας. Η ποσοτικοποίηση της αξιολόγησης επιτυγχάνεται με τον υπολογισμό ενός δείκτη ποιότητας, που ονομάζεται Βελτίωση Ανάκλησης, πάνω σε ένα σύνολο εγγράφων του Ιστού. Η Βελτίωση Ανάκλησης εκφράζει το κα-

τά πόσο η μείωση διάστασης αυξάνει την ακρίβεια της ανάκλησης εγγράφων βάσει ερωτημάτων κειμένου.

Λέξεις κλειδιά: Ομαδοποίηση, Πιθανοτική Ανάλυση Κρυμμένης Σημασιολογίας, Μοντέλο των Όψεων, Μείωση Διάστασης, Εννοιολογική Δεικτοδότηση, Βελτίωση Ανάκλησης, Παγκόσμιος Ιστός, Java.

Abstract

The main goal of this diploma thesis is the implementation and experimental evaluation of two dimensionality reduction algorithms, which are intended to be used for the processing of Web data. The first algorithm we focus on is based on clustering, while the second one is based on an algorithm belonging to the family of Probabilistic Latent Semantic Analysis (PLSA). Thus, the fulfillment of our goal is accomplished through four discrete stages: the implementation of three clustering algorithms, the implementation of the PLSA algorithm, the integration of the above into the two dimensionality reduction algorithms and finally, the experimental evaluation of the latter on real world data.

At the first stage we implement in Java a software framework that is based on the appropriate data structures and the necessary object oriented design, so as to be used for the further development of clustering algorithms for Web data. Furthermore, we develop three specific clustering algorithms and test their behaviour on a certain dataset.

At the second stage we implement in Java an algorithm that belongs to the family of Probabilistic Latent Semantic Analysis. This algorithm models the data generation process using a probabilistic model with latent variables, called Aspect Model. Then it computes an optimal estimation of the model's parameters, using the optimization technique called Expectation Maximization during a training process on some part of the available data.

At the third stage we implement in Java two dimensionality reduction algorithms, which also was our initial goal. The first algorithm, called Concept Indexing, is based on clustering, while the second one is based on the estimation of the probabilistic parameters that appear in the Aspect Model. In general, a dimensionality reduction algorithm attempts to map the high dimensional vectors of the supplied data to a space of lower dimensionality than the original one. The motivation for this attempt is twofold: firstly, we wish to minimize the computational resources required to process the data and secondly, we hope to discover correlations between the elements of the dataset, which are not visible in the high dimensional representation.

At the final stage of the thesis we evaluate the effectiveness of the two dimensionality reduction algorithms we have implemented, as far as their capability of discovering latent correlations is concerned. The quantification of the evaluation is achieved by computing a quality index, called Retrieval Improvement. This index expresses the extent to which dimensionality reduction improves the precision of document retrieval based on free text queries. For the purpose of evaluation we use a real world document corpus originating from the Web.

Keywords: Clustering, Probabilistic Latent Semantic Analysis, Aspect Model, Dimensionality Reduction, Concept Indexing, World Wide Web, Retrieval Improvement, Java.

Κεφάλαιο 1

Εισαγωγή

1.1 Εξόρυξη δεδομένων και μηχανική μάθηση

Στη σημερινή εποχή, η ποσότητα των δεδομένων στα οποία έχει πρόσβαση ο μέσος πολίτης μιας αναπτυγμένης τεχνολογικά χώρας μοιάζει να μην έχει φραγμούς. Οι φθηνοί προσωπικοί υπολογιστές καθιστούν εξαιρετικά εύκολη την αποθήκευση πληροφοριών, που παλιότερα θα σβήνονταν αμέσως μετά την πρώτη χρήση τους. Σκληροί δίσκοι πολλών Gigabytes λύνουν το δίλημμα άλλων εποχών σχετικά με το τι είναι χρήσιμο να αποθηκευτεί για μελλοντική χρήση και τι όχι: σήμερα απλώς αγοράζουμε έναν επιπλέον δίσκο και δεν πετάμε το παραμικρό κομμάτι αποθηκευμένης πληροφορίας. Από την άλλη μεριά, οι αγοραστικές επιλογές μας και οι κινήσεις μας στο Διαδίκτυο, συχνά καταγράφονται και αποθηκεύονται σε τεράστιες βάσεις δεδομένων, προκειμένου να αποτελέσουν μελλοντικά αντικείμενο στατιστικής ανάλυσης. Πιο σημαντικό απ' όλα, ο Παγκόσμιος Ιστός αποτελεί για τον καθένα μια ανεξάντλητη πηγή πληροφοριών, αναμφισβήτητα πολύτιμη, αλλά και ασύλληπτα πολύπλοκη στη διαχείριση της.

Καθώς ο όγκος των διαθέσιμων δεδομένων αυξάνεται, γίνεται ολοένα και πιο δύσκολο για τον άνθρωπο να εξάγει χρήσιμα συμπεράσματα από αυτά, αφήνοντας έτσι μέσα στην τεράστια ποσότητα των δεδομένων σημαντική ποσότητα κρυμμένης, και συνεπώς χαμένης, πληροφορίας. Σε αυτό το σημείο υπεισέρχεται η εξόρυξη δεδομένων, που ορίζεται σαν η διαδικασία ανακάλυψης προτύπων μέσα στα δεδομένα με έναν τρόπο αυτοματοποιημένο ή συνηθέστερα ημιαυτοματοποιημένο. Τα πρότυπα που αποκαλύπτονται θα πρέπει να είναι χρήσιμα, με την έννοια ότι θα πρέπει να οδηγούν σε κάποιο πλεονέκτημα αυτούς που κατέχουν την προκύπτουσα απ' αυτά γνώση. Ένα συνηθισμένο επεξηγηματικό παράδειγμα για τα παραπάνω είναι η περίπτωση μιας εταιρίας που επιθυμεί να κάνει σχεδιασμό της μακροπρόθεσμης στρατηγικής της στον τομέα των υπηρεσιών και των προϊόντων που παρέχει στους καταναλωτές. Χρησιμοποιώντας την τεράστια ποσότητα των αποθηκευμένων δεδομένων, και με την προϋπόθεση ότι κατέχει την απαιτούμενη τεχνογνωσία, η εν λόγω εταιρία μπορεί να εξάγει πολύ χρήσιμα συμπεράσματα για τις τρέχουσες αγο-

ραστικές συνήθειες των καταναλωτών, την απήχηση των διάφορων προϊόντων στις διάφορες ομάδες αγοραστικού κοινού, τις μελλοντικές τάσεις των καταναλωτών κ.α. Σε αυτήν την περίπτωση, τα δεδομένα μοιάζουν να είναι η πρώτη ύλη που τροφοδοτεί την ανάπτυξη της εν λόγω επιχείρησης, αν βέβαια η γνώση που βρίσκεται κρυμμένη μέσα σε αυτά μπορεί να εξορυχθεί.

Εξαιτίας του μεγάλου όγκου των δεδομένων και της μεγάλης πολυπλοκότητας των κρυμμένων συσχετίσεων, είναι προφανές ότι η μαζική εξόρυξη των δεδομένων μπορεί να γίνει μόνο με αυτοματοποιημένο τρόπο, δηλαδή με τη βοήθεια της Πληροφορικής. Οι περισσότεροι αλγόριθμοι που χρησιμοποιούνται γι' αυτόν τον σκοπό ανήκουν στην οικογένεια των αλγορίθμων μηχανικής μάθησης. Χρησιμοποιούμε τον όρο μηχανική μάθηση, με την έννοια της ανάπτυξης προγραμμάτων υπολογιστών, ικανών να ανακαλύπτουν πρότυπα μέσα στα δεδομένα, να παρουσιάζουν την αποκτημένη γνώση σε μορφή εύληπτη για τον άνθρωπο και τέλος, να κάνουν προβλέψεις για περιπτώσεις δεδομένων που τους παρουσιάζονται για πρώτη φορά. Με τη διαδικασία της μηχανικής μάθησης είναι συνήθως συνυφασμένη μια διαδικασία εκπαίδευσης, δηλαδή τροφοδοσίας του προγράμματος με υπάρχοντα δεδομένα, για τα πρότυπα των οποίων υπάρχει κάποια γνώση η οποία και μεταφέρεται ρητά στο πρόγραμμα. Σε αυτήν την περίπτωση επιδιώκουμε κυρίως να εμφυσήσουμε στο πρόγραμμα μηχανικής μάθησης την ικανότητα να αναγνωρίζει περιπτώσεις δεδομένων που του παρουσιάζονται για πρώτη φορά και που προφανώς δεν ανήκουν στα δεδομένα εκπαίδευσης. Βέβαια, στην οικογένεια της μηχανικής μάθησης εντάσσονται και αλγόριθμοι που δεν απαιτούν καμιά προϋπάρχουσα γνώση σχετικά με τα πρότυπα των δεδομένων, όπως είναι για παράδειγμα οι αλγόριθμοι ομαδοποίησης. Σε αυτήν την περίπτωση ενδιαφερόμαστε κυρίως για την απο μηδενική βάση εξερεύνηση της δομής και των συσχετίσεων που υπάρχουν κρυμμένα στα δεδομένα, τα πορίσματα της οποίας μπορούν να χρησιμοποιηθούν σε κάποιο μετέπειτα στάδιο εκπαίδευσης. Σε κάθε περίπτωση, το πεδίο της μηχανικής μάθησης έχει να επιδείξει τα τελευταία χρόνια πολλά αξιόλογα αποτελέσματα, όπως η ανίχνευση προσπαθειών απάτης με πιστωτικές κάρτες, τα συστήματα που μαθαίνουν τις αναγνωστικές προτιμήσεις των χρηστών τους, τα οχήματα που κινούνται μόνα τους σε αυτοκινητόδρομους. Ταυτόχρονα έχουν υπάρξει σημαντικές πρόοδοι στη θεωρία των αλγορίθμων που στηρίζουν τις πρακτικές εφαρμογές.

1.2 Μείωση διάστασης για την εξόρυξη του Ιστού

Δεν είναι δύσκολο να φανταστεί κανείς μια από τις σημαντικότερες προκλήσεις που καλούνται να αντιμετωπίσουν σήμερα οι επιστήμονες που ασχολούνται με την εξόρυξη δεδομένων. Πρόκειται για την εξόρυξη του Παγκόσμιου Ιστού, της τεράστιας συλλογής ηλεκτρονικών εγγράφων που τείνει να καλύπτει ολοένα και μεγαλύτερο ποσοστό της δημοσιευμένης γνώσης. Το πλήθος των σελίδων που αποτελούν τον Ιστό είναι σήμερα της τάξης του δισεκατομμυρίου, ενώ η πληροφορία που περιέχεται σε αυτές μπορεί να διαχωριστεί σε πληροφορία κειμένου και σε πληροφορία δομής.

Η πληροφορία κειμένου σχετίζεται με τη σημασιολογία των όσων γράφονται στις ιστοσελίδες, ενώ η πληροφορία δομής σχετίζεται με τον τρόπο που οι ιστοσελίδες αλληλοσυνδέονται με τη χρήση των υπερσυνδέσεων. Είναι περιττό να αναφέρουμε ότι οι ιστοσελίδες υπόκεινται σε πολύ συχνές αλλαγές, καθιστώντας έτσι τον Ιστό μια πλήρως δυναμική δομή. Από την άλλη μεριά, ο τρόπος που οι χρήστες του Ιστού αλληλεπιδρούν με τις ιστοσελίδες μπορούμε να θεωρήσουμε ότι εμπλουτίζει το περιεχόμενο των τελευταίων, διότι τους προσδίδει μια αξία όσον αφορά την προτίμηση που τους δείχνει το κοινό. Συμμετρικά ισχύει ότι και οι χρήστες χαρακτηρίζονται από το είδος των ιστοσελίδων που προτιμούν και από τη συχνότητα με την οποία τις επισκέπτονται.

Σε γενικές γραμμές, η εξόρυξη του Παγκοσμίου Ιστού έχει τρεις διακεκριμένες κατευθύνσεις, ανάλογα με το είδος της γνώσης που προσπαθεί να εξαγάγει και τον σκοπό που φιλοδοξεί να εκπληρώσει. Η εξόρυξη περιεχομένου χαρακτηρίζει τις ιστοσελίδες βάσει του κειμένου που περιέχεται σε αυτές και ο βασικός σκοπός της είναι η ανάπτυξη τεχνικών που να επιτρέπουν την αναζήτηση εγγράφων του Ιστού βάσει ερωτημάτων κειμένου. Η εξόρυξη δομής στοχεύει στον ποιοτικό χαρακτηρισμό των ιστοσελίδων βάσει των υπερσυνδέσεων που τις αλληλοσυνδέουν. Ένας τέτοιος χαρακτηρισμός επιτρέπει στη συνέχεια την απόδοση συντελεστών σημαντικότητας στις διάφορες ιστοσελίδες, που με τη σειρά τους μπορούν να χρησιμοποιηθούν για την εκλέπτυνση της διαδικασίας ανάκλησης. Τέλος, η εξόρυξη προτύπων χρήσης εστιάζει στους χρήστες του Ιστού και συγκεκριμένα στην τάση τους να επισκέπτονται συγκεκριμένες σελίδες και ίσως με συγκεκριμένη σειρά. Τα πρότυπα χρήσης μπορούν στη συνέχεια να χρησιμοποιηθούν για την ανάπτυξη συστημάτων πρόβλεψης των προτιμήσεων του κοινού ή για την προσαρμογή των ηλεκτρονικών υπηρεσιών στις ανάγκες του κάθε χρήστη βάσει κάποιου προφίλ που έχει κατασκευαστεί ειδικά γι' αυτόν.

Για να ανταποκριθούν στις παραπάνω προκλήσεις, οι χρησιμοποιούμενοι αλγόριθμοι μηχανικής μάθησης καλούνται να λάβουν υπόψη τους τις ιδιαιτερότητες των δεδομένων που προέρχονται από τον Ιστό. Ένα σημαντικό χαρακτηριστικό αυτών των δεδομένων είναι η υψηλή διάσταση, δηλαδή το μεγάλο πλήθος των χαρακτηριστικών που απαιτούνται για τον χαρακτηρισμό του καθενός στοιχείου του συνόλου των δεδομένων. Για να γίνουμε πιο σαφείς υπενθυμίζουμε ότι οι περισσότεροι αλγόριθμοι μηχανικής μάθησης θεωρούν μια διανυσματική αναπαράσταση του συνόλου των δεδομένων, όπου το κάθε στοιχείο αναπαρίσταται σαν ένα διάνυσμα d συνιστωσών: $\vec{x} = (x_1, x_2, \ldots, x_d)$. Οι συνιστώσες x_i του διανύσματος $\vec{x}$ ονομάζονται χαρακτηριστικά και αντιστοιχούν στη μέτρηση διαφόρων ιδιοτήτων που έχουμε την πεποίθηση ότι χαρακτηρίζουν τα στοιχεία του συνόλου δεδομένων. Η διανυσματική αναπαράσταση των δεδομένων είναι ένας πολύ χρήσιμος φορμαλισμός, διότι επιτρέπει την εισαγωγή πολλών μεθόδων από τα Μαθηματικά στο πεδίο της μηχανικής μάθησης.

Σαν μείωση διάστασης ορίζεται η διαδικασία κατά την οποία τα διανύσματα των δεδομένων απεικονίζονται μέσω κάποιου μετασχηματισμού σε έναν χώρο χα-

μηλότερης διάστασης από τον αρχικό, πράγμα που πρακτικά σημαίνει ότι για τον χαρακτηρισμό των δεδομένων είναι πλέον απαραίτητο ένα μικρότερο πλήθος χαρακτηριστικών. Είναι προφανές, ότι ο μετασχηματισμός μείωσης διάστασης θα πρέπει να γίνεται με κάποιον μεθοδικό τρόπο, ώστε τα χαρακτηριστικά των δεδομένων στον μειωμένης διάστασης χώρο να εκφράζουν όσο το δυνατόν ουσιαστικότερες ιδιότητες και συσχετίσεις των στοιχείων του συνόλου δεδομένων. Τα κίνητρα για τη μείωση διάστασης είναι δυο: πρώτον, εξοικονόμηση υπολογιστικών πόρων κατά την επεξεργασία των δεδομένων από αλγορίθμους μηχανικής μάθησης και δεύτερον, εξεύρεση των ουσιαστικών χαρακτηριστικών που είναι σε θέση να αποκαλύψουν τις συσχετίσεις μεταξύ των στοιχείων του συνόλου δεδομένων. Αυτή η δεύτερη επιδίωξη επηρεάζει προφανώς και την ποιότητα των αποτελεσμάτων της μηχανικής μάθησης. Για παράδειγμα, ας θεωρήσουμε την περίπτωση που το σύνολο των δεδομένων αποτελείται από έγγραφα που αναπαρίστανται σαν διανύσματα λέξεων. Μπορούμε να φανταστούμε μια στρατηγική μείωσης διάστασης που θα προσπαθούσε να απεικονίσει τα διανύσματα των εγγράφων σε έναν χαμηλότερης διάστασης χώρο, όπου οι συνιστώσες θα αντιστοιχούσαν σε θεματικές ενότητες αντί σε μεμονωμένες λέξεις. Αν αυτό ήταν εφικτό, θα μπορούσαμε να ελπίζουμε σε ταχύτερα και ποιοτικότερα αποτελέσματα από την πλευρά των αλγορίθμων μηχανικής μάθησης.

Μετά τα όσα αναφέρθηκαν παραπάνω γίνεται πλέον σαφές το πώς η μείωση διάστασης μπορεί να ενταχθεί στο πλαίσιο της εξόρυξης του Παγκοσμίου Ιστού. Εξαιτίας του πλήθους των ιστοσελίδων, της πολυπλοκότητας της δομής και του πλήθους των χρηστών του Ιστού, είναι λογικό τα ακατέργαστα δεδομένα που χρησιμοποιούνται για την εξόρυξη του Ιστού να έχουν διανυσματικές αναπαραστάσεις πολύ υψηλής διάστασης. Η μείωση διάστασης εμφανίζεται λοιπόν σαν ένα στάδιο προεπεξεργασίας, προκειμένου οι αλγόριθμοι μηχανικής μάθησης των επόμενων σταδίων να μπορέσουν αφενός να χειριστούν τα δεδομένα από άποψη υπολογιστικών πόρων και αφετέρου να στηριχτούν σε έναν όσο το δυνατόν ουσιαστικότερο χαρακτηρισμό των δεδομένων. Αναλυτικότερη περιγραφή της μείωσης διάστασης και των γενικών μεθοδολογιών που ακολουθούνται υπάρχει στην ενότητα 2.4.

1.3 Στόχος της εργασίας

Ο κύριος στόχος της παρούσας διπλωματικής εργασίας είναι η υλοποίηση και αξιολόγηση δυο συγκεκριμένων τεχνικών μείωσης διάστασης για την επεξεργασία δεδομένων του Παγκοσμίου Ιστού. Τα δεδομένα που μας ενδιαφέρουν εμπίπτουν στην κατηγορία των δεδομένων δυαδικών συσχετίσεων, κάτι το οποίο σημαίνει ότι αναφέρονται σε κοινές εμφανίσεις αντικειμένων που προέρχονται από δυο διαφορετικές κλάσεις, όπως είναι τα ζευγάρια εγγράφων-λέξεων, εγγράφων-αναγνωστών ή καταναλωτών-προϊόντων. Η πιο άμεση αναπαράσταση των δεδομένων αυτού του τύπου πραγματοποιείται με τη χρήση διανυσμάτων, τα οποία όμως, εξαιτίας της φύσης των δεδομένων, είναι πολυδιάστατα και αραιά. Σε αυτό το σημείο υπεισέρχεται η μείωση της διάστασης, που καλείται να απεικονίσει τα διανύσματα των δεδομένων

σε έναν χώρο πολύ μικρότερης διάστασης από τον αρχικό, χωρίς να χάσει χρήσιμη πληροφορία. Στην πραγματικότητα επιδιώκεται μάλιστα το αντίθετο: μέσω της διαδικασίας μείωσης διάστασης να αποκαλυφθούν συσχετίσεις μεταξύ των δεδομένων, οι οποίες δεν είναι φανερές στην πολυδιάστατη αναπαράστασή τους. Αν τελικά επιτευχθεί αυτό, η μείωση διάστασης λειτουργεί σαν μια διαδικασία εξόρυξης χρήσιμης γνώσης από τα δεδομένα.

Η πρώτη από τις τεχνικές μείωσης διάστασης που εξετάζουμε, χρησιμοποιεί τη διαδικασία της ομαδοποίησης με τη λογική του 'μαύρου κουτιού'. Συνεπώς προκύπτει η ανάγκη υλοποίησης ορισμένων αλγορίθμων ομαδοποίησης, που αποδίδουν ικανοποιητικά αποτελέσματα για δεδομένα δυαδικών συσχετίσεων που προέρχονται από πηγές του Παγκοσμίου Ιστού. Η σχεδίαση και υλοποίηση του απαιτούμενου λογισμικού γίνεται έτσι, ώστε να ορίζει ένα ευρύτερο πλαίσιο εργασίας. Αυτή η γενίκευση παρέχει μια υποδομή που μπορεί να χρησιμοποιηθεί για τη μελλοντική υλοποίηση και άλλων αλγορίθμων ομαδοποίησης, καθώς και για την υποστήριξη δεδομένων προς ομαδοποίηση, που αναπαρίστανται με διαφορετικό τρόπο από αυτόν που χρησιμοποιούμε εμείς.

Για την υλοποίηση της δεύτερης τεχνικής μείωσης διάστασης που εξετάζουμε, απαιτείται ένας αλγόριθμος που ανήκει στην οικογένεια της Πιθανοτικής Ανάλυσης Κρυμμένης Σημασιολογίας και πραγματοποιεί μια βέλτιστη εκτίμηση των παραμέτρων ενός πιθανοτικού μοντέλου βάσει κάποιου δεδομένου συνόλου παρατηρήσεων [11]. Για την υλοποίηση αυτού του αλγορίθμου χρησιμοποιούμε ως βάση το πακέτο λογισμικού PennAspect, που διανέμεται υπό τους όρους της δημόσιας άδειας GNU από την ομάδα εξόρυξης δεδομένων του πανεπιστημίου της Πενσυλβάνια [8].

Τελικά εκπληρώνουμε τον αρχικό σκοπό της εργασίας, αξιοποιώντας τους αλγορίθμους ομαδοποίησης και ΠΑΚΣ που έχουμε αναπτύξει, στο ευρύτερο πλαίσιο δυο διαδικασιών μείωσης διάστασης. Σε κάθε περίπτωση, αξιολογούμε τα αποτελέσματα της μείωσης διάστασης πάνω σε ένα σύνολο εγγράφων του Παγκοσμίου Ιστού, που είναι χωρισμένο σε γνωστές εκ των προτέρων κατηγορίες. Η αξιολόγηση γίνεται με τον υπολογισμό ενός δείκτη που ονομάζεται Βελτίωση Ανάκλησης και συγκρίνει την ακρίβεια της ανάκλησης σχετικών εγγράφων στην περίπτωση της πολυδιάστατης και της μειωμένης διάστασης διανυσματικής αναπαράστασης των εγγράφων [16].

1.4 Οργάνωση του τόμου

Στα κεφάλαια που ακολουθούν αναπτύσσονται όσες ιδέες είναι απαραίτητες για την κατανόηση των θεμάτων, με τα οποία ασχοληθήκαμε στα πλαίσια αυτής της εργασίας. Συγκεκριμένα, η οργάνωση των επόμενων κεφαλαίων έχει ως εξής:

- *Κεφάλαιο 2:* Ορίζεται ο τύπος των δεδομένων που ονομάζουμε δεδομένα δυαδικών συσχετίσεων. Επίσης, παρουσιάζεται η ιδέα της αναπαράστασης τέτοιων δεδομένων σαν διανύσματα και εξηγούνται ορισμένα εγγενή χαρακτηριστικά αυτής της διανυσματικής αναπαράστασης, όπως είναι η χαμηλή πυκνότητα και

η υψηλή διάσταση. Στη συνέχεια παραθέτουμε ορισμένα μέτρα ομοιότητας που έχουν αποδειχθεί κατάλληλα για τον υπολογισμό αποστάσεων μεταξύ τέτοιων πολυδιάστατων διανυσμάτων. Ο ορισμός των κατάλληλων μέτρων ομοιότητας είναι απαραίτητος, προκειμένου να λειτουργήσουν αποτελεσματικά διαδικασίες επεξεργασίας των δεδομένων που στηρίζονται στην έννοια της απόστασης, όπως είναι η διαδικασία της ομαδοποίησης. Τέλος, γίνεται μια γενική παρουσίαση της μείωσης διάστασης από δυο διαφορετικές οπτικές γωνίες: πρώτον, ως σταδίου προεπεξεργασίας των δεδομένων και δεύτερον, ως μεθόδου για την εξόρυξη γνώσης από αυτά.

- *Κεφάλαιο 3:* Αρχικά δίνονται ορισμένες γενικές πληροφορίες για τη διαδικασία της ομαδοποίησης και τις δυσκολίες που αυτή καλείται να αντιμετωπίσει. Στη συνέχεια, ακολουθεί μια ακριβής περιγραφή των τριών συγκεκριμένων αλγορίθμων ομαδοποίησης που υλοποιούνται στην παρούσα εργασία: του αλγορίθμου Συμπαγών Ομάδων, του αλγορίθμου Jarvis-Patrick και του αλγορίθμου Πυκνότητας Κοινών Κοντινών Γειτόνων. Τέλος, παρουσιάζεται μια συγκεκριμένη τεχνική με την οποία κάποιος αλγόριθμος ομαδοποίησης μπορεί να χρησιμοποιηθεί στα πλαίσια μιας ευρύτερης διαδικασίας μείωσης διάστασης. Αυτή η τεχνική ονομάζεται Εννοιολογική Δεικτοδότηση και είναι η μια από τις δυο μεθόδους μείωσης διάστασης που εξετάζουμε στην παρούσα εργασία.

- *Κεφάλαιο 4:* Αρχικά περιγράφεται το Μοντέλο των Όψεων, το οποίο παρέχει μια πιθανοτική περιγραφή της διαδικασίας παραγωγής των δεδομένων δυαδικών συσχετίσεων από κάποιες μη ορατές μεταβλητές. Στη συνέχεια περιγράφεται η τεχνική βελτιστοποίησης που είναι γνωστή σαν Μεγιστοποίηση της Αναμενόμενης Τιμής, καθώς και το πώς αυτή μπορεί να εφαρμοστεί στην περίπτωση του Μοντέλου των Όψεων. Ο στόχος σε αυτήν την περίπτωση είναι να επιτευχθεί μια όσο το δυνατόν καλύτερη προσαρμογή των παραμέτρων του πιθανοτικού μοντέλου δεδομένων των παρατηρήσεων που έχουν πραγματοποιηθεί. Τέλος, παρουσιάζεται η δυνατότητα του Μοντέλου των Όψεων να χρησιμοποιηθεί στα πλαίσια μιας διαδικασίας μείωσης διάστασης.

- *Κεφάλαιο 5:* Περιγράφονται τα διάφορα στάδια ανάπτυξης του λογισμικού που χρησιμοποιήθηκε στα πλαίσια της εργασίας. Συγκεκριμένα, περιγράφονται τα στάδια της ανάλυσης, της σχεδίασης, της υλοποίησης και του ελέγχου. Επίσης, γίνονται ορισμένες αναφορές σε δυνατές μελλοντικές επεκτάσεις του λογισμικού, δεδομένης της υποδομής που έχει ήδη αναπτυχθεί.

- *Κεφάλαιο 6:* Παρατίθενται τα αποτελέσματα που προέκυψαν από την εκτέλεση των αλγορίθμων ομαδοποίησης και ΠΑΚΣ πάνω σε ένα σύνολο δεδομένων που αποτελείται από ζευγάρια ιστοσελίδων - λέξεων. Τα αποτελέσματα συνοδεύονται από σύντομο σχολιασμό, όσον αφορά τη συμπεριφορά των αλγορίθμων για τις διάφορες τιμές των ρυθμιστικών παραμέτρων. Στη συνέχεια αξιολογούνται οι διαδικασίες μείωσης διάστασης παρουσιάζοντας τις μετρήσεις που

πραγματοποιήθηκαν για τη Βελτίωση Ανάκλησης πάνω σε ένα σύνολο δεδομένων που αποτελείται από ζευγάρια κειμένων ηλεκτρονικού ταχυδρομείου και λέξεων. Τα αποτελέσματα αποδεικνύουν, ότι οι διαδικασίες μείωσης διάστασης που εξετάσαμε έχουν τα αναμενόμενα αποτελέσματα.

- *Κεφάλαιο 7:* Σε αυτό το τελευταίο κεφάλαιο γίνεται μια σύντομη σύνοψη των όσων παρουσιάστηκαν στην εργασία. Επιπλέον, δίνεται ένας κατάλογος δυνατών μελλοντικών επεκτάσεων, όσον αφορά τόσο ορισμένα θεωρητικά τμήματα, όσο και ορισμένα πρακτικά θέματα υλοποίησης λογισμικού.

Κεφάλαιο 2

Βασικές Έννοιες

2.1 Δεδομένα δυαδικών συσχετίσεων

Στα πλαίσια της εργασίας ασχολούμαστε με την επεξεργασία μιας συγκεκριμένης κατηγορίας δεδομένων που ονομάζονται δεδομένα δυαδικών συσχετίσεων. Τα δεδομένα δυαδικών συσχετίσεων αναφέρονται σε ένα χώρο δυο αφηρημένων συνόλων $\mathcal{X}$ και $\mathcal{Y}$, στον οποίο οι παρατηρήσεις πραγματοποιούνται για ζεύγη αντικειμένων $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Στην απλούστερη περίπτωση, μια παρατήρηση αποτελείται από την ίδια την ύπαρξη του ζεύγους (x, y), δηλαδή το γεγονός ότι τα δυο συγκεκριμένα αντικείμενα εμφανίστηκαν ταυτόχρονα, ενώ σε μια πιο πολύπλοκη περίπτωση, το ζευγάρι (x, y) μπορεί να αντιπροσωπεύεται από μια βαθμωτή τιμή $w(x, y) \in \mathbb{R}$, δηλαδή ένα είδος αριθμητικής αξιολόγησης του γεγονότος ότι τα δυο αντικείμενα εμφανίστηκαν σαν ζευγάρι. Μερικές ενδεικτικές επιστημονικές περιοχές στις οποίες εμφανίζονται δεδομένα δυαδικών συσχετίσεων είναι:

- Η μηχανική όραση και ιδιαίτερα ο τομέας της κατάτμησης εικόνας. Εκεί το $\mathcal{X}$ αντιστοιχεί σε περιοχές της εικόνας, το $\mathcal{Y}$ αντιστοιχεί σε χαρακτηριστικά της εικόνας που λαμβάνουν συνεχείς ή διακριτές τιμές και μια παρατήρηση $w(x, y)$ δηλώνει την ύπαρξη ενός συγκεκριμένου χαρακτηριστικού, με κάποια δεδομένη τιμή, σε μια συγκεκριμένη περιοχή της εικόνας.

- Η βασισμένη στο κείμενο ανάκληση πληροφοριών, όπου το $\mathcal{X}$ αντιπροσωπεύει μια συλλογή εγγράφων, το $\mathcal{Y}$ ένα σύνολο λέξεων, και το ζευγάρι (x, y) την εμφάνιση μιας λέξης μέσα στο σώμα κάποιου κειμένου.

- Η ανάλυση των προτιμήσεων και των καταναλωτικών τάσεων ενός κοινού. Το $\mathcal{X}$ αντιστοιχεί σε ένα σύνολο ατόμων, το $\mathcal{Y}$ σε ένα σύνολο προϊόντων, και το ζεύγος (x, y) στην προτίμηση ενός συγκεκριμένου προϊόντος από ένα συγκεκριμένο άτομο. Αυτή η περίπτωση δεδομένων είναι η πρώτη ύλη για μια τεχνική πρόβλεψης των προτιμήσεων, που είναι γνωστή σαν συνεργατική διήθηση [2].

Η ενοποίηση όλων των παραπάνω περιπτώσεων δεδομένων κάτω από την κοινή οπτική των δεδομένων δυαδικών συσχετίσεων μας επιτρέπει να συνεχίσουμε τη συζήτηση μας σε ένα αφηρημένο και ανεξάρτητο από την εκάστοτε εφαρμογή επίπεδο.

Η απλούστερη δομή που μπορεί να αναπαραστήσει ένα σύνολο δεδομένων δυαδικών συσχετίσεων είναι ένας δισδιάστατος πίνακας, οι γραμμές του οποίου αντιστοιχούν στα αντικείμενα του ενός διακριτού συνόλου και οι στήλες στα αντικείμενα του άλλου συνόλου. Οι εγγραφές του πίνακα αντιπροσωπεύουν το ζευγάρι αντικειμένων που προκύπτει από τον αντίστοιχο συνδυασμό γραμμής και στήλης. Ο τύπος των εγγραφών εξαρτάται από τον τρόπο που έχουμε αποφασίσει να μετρήσουμε και να κωδικοποιήσουμε τη συσχέτιση αυτών των αντικειμένων. Μπορούμε να φανταστούμε λοιπόν τις ακόλουθες περιπτώσεις:

- Οι εγγραφές του πίνακα είναι δίτιμες λογικές μεταβλητές, δηλαδή παίρνουν μόνο τις τιμές 'αληθής' και 'ψευδής'. Σε αυτήν την περίπτωση, το μόνο που καταφέρνουμε να μετρήσουμε ή το μόνο που μας ενδιαφέρει να αναπαραστήσουμε, είναι το γεγονός, ότι τα αντικείμενα x και y εμφανίζονται ταυτόχρονα.

- Οι εγγραφές του πίνακα είναι ακέραιοι αριθμοί. Σε αυτήν την περίπτωση αναφερόμαστε πλέον σε πλήθος ταυτόχρονων εμφανίσεων των αντικειμένων x και y, και όχι μόνο στο γεγονός της κοινής εμφάνισης. Δηλαδή, η κάθε εγγραφή του πίνακα αποδίδει πλέον και κάποιο βάρος στο αντίστοιχο ζεύγος αντικειμένων, δεδομένου ότι το πλήθος των παρατηρήσεων είναι πεπερασμένο.

- Οι εγγραφές του πίνακα είναι πραγματικοί αριθμοί. Σε αυτήν την περίπτωση το πλήθος ταυτόχρονων εμφανίσεων έχει αντικατασταθεί από μια κανονικοποιημένη συχνότητα, μέσα από μια διαδικασία που αποσκοπεί στο να τροποποιήσει τα βάρη των ζευγαριών, με τρόπο που τα τελικά βάρη να λαμβάνουν καλύτερα υπόψη τους το σύνολο των παρατηρήσεων. Για παράδειγμα, ένας συνηθισμένος μετασχηματισμός που εφαρμόζεται σε πίνακες εγγράφων-λέξεων λαμβάνει υπόψη του για τον υπολογισμό των τελικών βαρών τόσο το πλήθος εμφανίσεων της αντίστοιχης λέξης στο αντίστοιχο έγγραφο, όσο και το συνολικό πλήθος των εγγράφων που περιέχουν τη συγκεκριμένη λέξη. Ο λόγος που συμβαίνει αυτό είναι ότι όσο πιο κοινή είναι μια λέξη, δηλαδή σε όσο περισσότερα έγγραφα χρησιμοποιείται, με τόσο λιγότερη βαρύτητα συνεισφέρει στον χαρακτηρισμό των εγγράφων που την περιέχουν. Αντίθετα, μια λέξη που είναι σπάνια αποκτάει μεγάλη βαρύτητα, διότι επιβάλλει έναν σαφή διαχωρισμό των λίγων εγγράφων που την περιέχουν από όλα τα υπόλοιπα που δεν την περιέχουν.

2.2 Μοντέλο του διανυσματικού χώρου

Είναι γνωστό από πολλές εφαρμογές των Μαθηματικών, ότι ένας δισδιάστατος πίνακας $M \times N$ μπορεί να θεωρηθεί είτε σαν ένα σύνολο διανυσμάτων-γραμμών διάστασης N, είτε σαν ένα σύνολο διανυσμάτων-στηλών διάστασης M. Από την άλλη

μεριά είναι επίσης γνωστό, ότι μεταξύ διανυσμάτων της ίδιας διάστασης μπορούμε να ορίσουμε διάφορες αλγεβρικές πράξεις, καθώς και συναρτήσεις απόστασης, που διαισθητικά εκφράζουν το πόσο κοντά βρίσκονται τα αντίστοιχα αντικείμενα στον πολυδιάστατο χώρο. Συνεπώς ένας δισδιάστατος πίνακας περιέχει σε συμπυκνωμένη μορφή την πληροφορία που χρειάζεται, ώστε να εξάγουμε συμπεράσματα τόσο για τον χώρο των διανυσμάτων-γραμμών, όσο και για τον χώρο των διανυσμάτων-στηλών. Από αυτή την οπτική γωνία, οι πίνακες δυαδικών συσχετίσεων που περιγράψαμε στην προηγούμενη ενότητα εισάγουν έναν χρήσιμο φορμαλισμό, διότι ορίζουν μια διανυσματική αναπαράσταση των οντοτήτων που εμπλέκουν. Η διανυσματική αναπαράσταση των δεδομένων δυαδικών συσχετίσεων ονομάζεται Μοντέλο του Διανυσματικού Χώρου και δίνει τη δυνατότητα επεξεργασίας αυτών των δεδομένων με τη χρήση μεθόδων από τα Μαθηματικά. Εξάλλου, είναι γνωστό ότι οι περισσότεροι αλγόριθμοι μηχανικής μάθησης βασίζονται σε μια διανυσματική αναπαράσταση των δεδομένων, προκειμένου να μπορέσουν να θεμελιωθούν θεωρητικά πάνω σε ένα μαθηματικό υπόβαθρο.

Εξετάζοντας λεπτομερέστερα την περίπτωση ενός πίνακα εγγράφων-λέξεων μπορούμε να πούμε ότι μέσω των περιεχομένων του πίνακα, το κάθε έγγραφο χαρακτηρίζεται από το σύνολο των λέξεων που περιέχει και το βάρος με το οποίο η κάθε λέξη υπεισέρχεται σε αυτό, ενώ η κάθε λέξη, κατ' αναλογία, χαρακτηρίζεται από το σύνολο των εγγράφων στα οποία εμφανίζεται και φυσικά από τα αντίστοιχα βάρη αυτών των εγγράφων. Από την άλλη μεριά, η διαίσθησή μας υποδεικνύει, ότι δυο έγγραφα που ανήκουν στην ίδια θεματική ενότητα έχουν μάλλον αυξημένη πιθανότητα να περιέχουν τις ίδιες λέξεις και αντίστοιχα, ότι δυο λέξεις με παρόμοια σημασία έχουν μάλλον αυξημένη πιθανότητα να εμφανίζονται στα ίδια έγγραφα. Συνεπώς θα μπορούσαμε να χρησιμοποιήσουμε μια μαθηματική συνάρτηση απόστασης μεταξύ διανυσμάτων, προκειμένου να αποφανθούμε για το κατά πόσο δύο έγγραφα ή λέξεις είναι σχετικά.

Πάνω σε αυτήν ακριβώς την ιδέα βασίζονται πολλές τεχνικές αυτόματης ανάκλησης κειμένων βάσει λέξεων κλειδιών ή βάσει ερωτημάτων κειμένου. Η στρατηγική αυτών των τεχνικών είναι ότι για ένα δεδομένο ερώτημα κειμένου τα έγγραφα που πρέπει να ανακληθούν ως σχετικά είναι αυτά που βρίσκονται κοντά στο εν λόγω ερώτημα, όπως προκύπτει από την εφαρμογή κάποιας συνάρτησης απόστασης στις διανυσματικές αναπαραστάσεις των εγγράφων. Σε αυτό το σημείο βέβαια, πρέπει να τονίσουμε ότι αυτή η ιδέα, αν και κομψή, παρουσιάζει δυο πολύ σοβαρές αδυναμίες, που είναι κατ' επέκταση και περιοριστικοί παράγοντες στην επίδοση των αντίστοιχων τεχνικών ανάκλησης. Η πρώτη είναι η αδυναμία να χειριστεί σωστά το φαινόμενο της συνωνυμίας, δηλαδή το γεγονός ότι διαφορετικές λέξεις αναφέρονται στο ίδιο θέμα, όπως κάνουν, για παράδειγμα, οι λέξεις 'καθηγητής' και 'εκπαιδευτικός'. Εξαιτίας αυτής της αδυναμίας, δυο έγγραφα που το ένα μιλάει για 'καθηγητές' και το άλλο για 'εκπαιδευτικούς' δεν θα χαρακτηριστούν συναφή, μιας και ο πίνακας δυαδικών συσχετίσεων περιέχει πληροφορία για καθεμία λέξη ξεχωριστά και καμιά πληροφορία για τις πιθανές συνωνυμίες. Η δεύτερη αδυναμία της προσέγγισης

έγκειται στο φαινόμενο της πολυσημίας, δηλαδή την ιδιότητα ορισμένων λέξεων να έχουν διαφορετικές σημασίες μέσα σε διαφορετικά συμφραζόμενα, όπως γίνεται με τη λέξη 'δίσκος'. Εξαιτίας της πολυσημίας, δυο έγγραφα που περιέχουν εμφανίσεις της ίδιας λέξης αλλά με διαφορετική σημασία το καθένα, μπορεί να χαρακτηριστούν συναφή, ενώ στην πραγματικότητα να μην είναι.

Για την επίλυση των παραπάνω προβλημάτων έχουν προταθεί διάφορες λύσεις που εμπίπτουν στην οικογένεια των αλγορίθμων μείωσης διάστασης [6, 14, 3]. Η βασική ιδέα αυτών των αλγορίθμων είναι η διανυσματική αναπαράσταση των εγγράφων σε ένα νέο 'σύστημα αξόνων', με τρόπο που να λαμβάνονται υπόψη τα συμφραζόμενα των λέξεων μέσα στα διάφορα έγγραφα. Αυτό είναι σε συμφωνία με τα όσα αναφέρθηκαν στην ενότητα 1.2 σχετικά με την προσπάθεια που κάνει κάποια διαδικασία μείωσης διάστασης προκειμένου να βρει ένα μικρότερο και ποιοτικότερο σύνολο χαρακτηριστικών για τη διανυσματική αναπαράσταση των δεδομένων.

2.3 Μέτρα ομοιότητας

Από την προηγούμενη συζήτηση προκύπτει, ότι ο ορισμός μιας συνάρτησης ομοιότητας μεταξύ των διανυσμάτων-στηλών ή των διανυσμάτων-γραμμών του πίνακα δυαδικών συσχετίσεων είναι καθοριστικής σημασίας για διαδικασίες στις οποίες υπεισέρχεται η έννοια της απόστασης μεταξύ των αντικειμένων που αντιπροσωπεύουν τα διανύσματα του πίνακα, όπως είναι για παράδειγμα η διαδικασία της ομαδοποίησης. Τα μέτρα ομοιότητας, ή ισοδύναμα οι συναρτήσεις απόστασης, που μπορεί να ορίσει κανείς μεταξύ διανυσμάτων είναι πάρα πολλά και ανάλογα με τη μορφή τους κατέχουν διαφορετικές μαθηματικές ιδιότητες, περισσότερο ή λιγότερο επιθυμητές [18]. Στα πλαίσια της παρούσας ανάλυσης θα περιοριστούμε στη σύντομη παρουσίαση μόνο μερικών από αυτά.

Το μέτρο ομοιότητας με το οποίο είμαστε περισσότερο εξοικειωμένοι από τη φυσική μας εμπειρία στον τρισδιάστατο χώρο είναι η ευκλείδεια απόσταση. Αυτή ορίζεται στη γενική περίπτωση για N-διάστατα διανύσματα με αριθμητικές συνιστώσες από τον τύπο: $d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N}(x_i - yi)^2}$, όπου x_i, y_i είναι η i-οστή συνιστώσα των διανυσμάτων x και y αντίστοιχα. Αν και η ευκλείδεια απόσταση έχει πολλές χρήσιμες μαθηματικές ιδιότητες, όπως το γεγονός ότι καθιστά ισχύουσα την τριγωνική ανισότητα, στην περίπτωση αραιών διανυσμάτων παρουσιάζει ένα σοβαρό μειονέκτημα. Αυτό παρουσιάζεται με ένα χαρακτηριστικό παράδειγμα, στο οποίο φαίνεται ότι η ευκλείδεια μετρική χάνει την χρησιμότητα της όταν τα εμπλεκόμενα διανύσματα είναι αραιά και πολυδιάστατα. Έστω το ζεύγος 10-διάστατων σημείων P_1 και P_2, που φαίνονται στον πίνακα 2.1. Η ευκλείδεια απόσταση αυτών των δυο σημείων είναι ίση με 5. Τώρα ας θεωρήσουμε ένα άλλο ζευγάρι σημείων P_3 και P_4, που επίσης φαίνονται στον πίνακα 2.1 και των οποίων η ευκλείδεια απόσταση είναι επίσης 5. Ωστόσο παρατηρούμε το εξής παράδοξο: αν τα σημεία αντιπροσωπεύουν για παράδειγμα έγγραφα και η κάθε συνιστώσα είναι το πλήθος των εμφανίσεων της αντίστοιχης λέξης, τότε, διαισθητικά, τα σημεία P_3 και P_4, που έχουν 7 λέξεις κοινές

μοιάζουν να είναι πιο όμοια απ' ό,τι είναι τα σημεία P_1 και P_2, που δεν έχουν καμία κοινή λέξη. Γενικεύοντας αυτήν την παρατήρηση, συμπεραίνουμε ότι για αραιά και πολυδιάστατα διανύσματα η ομοιότητα μοιάζει να ορίζεται με πιο χρήσιμο τρόπο, αν αγνοηθούν οι συνιστώσες που είναι μηδενικές και για τα δυο διανύσματα. Αυτό εξηγείται από το γεγονός, ότι το πληροφοριακό περιεχόμενο ενός αραιού και πολυδιάστατου διανύσματος συγκεντρώνεται στις λίγες μη μηδενικές συνιστώσες του και συνεπώς η σύγκριση δύο διανυσμάτων θα πρέπει να λαμβάνει υπόψη της αυτές τις συνιστώσες με μεγαλύτερη βαρύτητα απ' ό,τι τις μηδενικές. Ωστόσο, όπως φαίνεται και από το παράδειγμα που προηγήθηκε, η ευκλείδεια απόσταση δεν λαμβάνει υπόψη της αυτό το γεγονός, αφού βασίζεται στο ποσοστό σύμπτωσης των αντίστοιχων συνιστωσών των δύο διανυσμάτων, ανεξάρτητα αν αυτές είναι μηδενικές ή όχι.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
P_1	3	0	0	0	0	0	0	0	0	0
P_2	0	0	0	0	0	0	0	0	0	4
P_3	3	2	4	0	1	2	3	1	2	0
P_4	0	2	4	0	1	2	3	1	2	4

Πίνακας 2.1: 4 σημεία με 10 ακέραιες συνιστώσες.

Συνεπώς θα επιθυμούσαμε να έχουμε ένα μέτρο ομοιότητας, που να λαμβάνει υπόψη του τις κοινές μη μηδενικές συνιστώσες των διανυσμάτων με μεγαλύτερη βαρύτητα απ' ότι τις μηδενικές, και μάλιστα με έναν τρόπο που να πραγματοποιεί κάποιου είδους κανονικοποίηση, έτσι ώστε το συνολικό πλήθος των μη μηδενικών συνιστωσών των διανυσμάτων να λαμβάνεται επίσης υπόψη. Ένα τέτοιο μέτρο ομοιότητας είναι το Μέτρο του Συνημιτόνου που για N-διάστατα αριθμητικά διανύσματα ορίζεται ως:

$$d(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|x\| \, \|y\|} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2} \cdot \sqrt{\sum_{i=1}^{N} y_i^2}}$$

Η γεωμετρική ερμηνεία του για την τρισδιάστατη περίπτωση είναι ότι αντιστοιχεί στο συνημίτονο της γωνίας που σχηματίζουν τα δυο διανύσματα. Παρατηρούμε ότι το μέτρο του συνημιτόνου λαμβάνει υπόψη μόνο τις συνιστώσες που είναι μη μηδενικές και για τα δυο διανύσματα και επίσης ότι κανονικοποιεί την τιμή της απόστασης, έτσι ώστε να βρίσκεται μέσα στο διάστημα $[0, 1]$. Στην περίπτωση του πίνακα 2.1 και με χρήση του Μέτρου Συνημιτόνου, η ομοιότητα των σημείων P_1 και P_2 προκύπτει ίση με 0, ενώ η ομοιότητα των P_3 και P_4 είναι ίση με 0.759, δηλαδή οι τιμές του μέτρου αντανακλούν τη διαισθητική μας αντίληψη περί ομοιότητας.

Ωστόσο και το Μέτρο Συνημιτόνου παρουσιάζει μια σημαντική αδυναμία που κάνει την εμφάνισή της σε ορισμένες περιπτώσεις πολυδιάστατων διανυσμάτων. Συγκεκριμένα, από τον μαθηματικό ορισμό του μέτρου δεν προκύπτει ότι καθιστά ισχύουσα την τριγωνική ανισότητα, όπως για παράδειγμα συμβαίνει με την ευκλείδια απόσταση. Αυτό σημαίνει ότι το γεγονός ότι δύο σημεία βρίσκονται κοντά

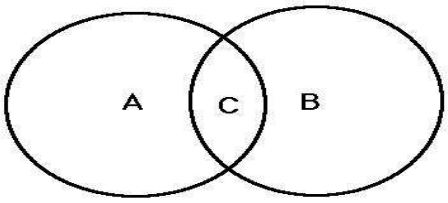
σε κάποιο τρίτο δεν συνεπάγεται αυτόματα ότι βρίσκονται κοντά και μεταξύ τους, πράγμα που έρχεται σε αντίθεση με τη διαισθητική μας αντίληψη περί απόστασης. Το φαινόμενο που περιγράφουμε εμφανίζεται στα δεδομένα του πίνακα 2.2, στον οποίο παρουσιάζονται τρία 10-διάστατα σημεία με ακέραιες συνιστώσες. Σύμφωνα με το Μέτρο Συνημιτόνου τα σημεία P_1 και P_2 βρίσκονται κοντά μεταξύ τους, όπως και τα σημεία P_2 και P_3. Παρόλαυτά, η ομοιότητα των σημείων P_1 και P_3 είναι ίση με 0. Αυτό οφείλεται στο γεγονός ότι η ομοιότητα του ζευγαριού P_1 και P_2 και του ζευγαριού P_2 και P_3 προέρχονται από διαφορετικά υποσύνολα συνιστωσών.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
P_1	1	2	2	3	1	0	0	0	0	0
P_2	0	1	2	1	3	2	3	1	2	0
P_3	0	0	0	0	0	2	3	1	2	1

Πίνακας 2.2: 3 σημεία με 10 ακέραιες συνιστώσες.

Επαναλαμβάνουμε ότι το Μέτρο Συνημιτόνου που περιγράφηκε παραπάνω ορίζεται για την περίπτωση διανυσμάτων με αριθμητικές συνιστώσες. Όπως έχει αναφερθεί όμως και στην προηγούμενη ενότητα, υπάρχει η περίπτωση οι εγγραφές του πίνακα δυαδικών συσχετίσεων να μην είναι αριθμητικές, αλλά λογικές τιμές, οπότε προκύπτει η ανάγκη για τον ορισμό μέτρων ομοιότητας ικανών να χειριστούν και αυτή την περίπτωση δεδομένων. Μια απλή σκέψη είναι να αντιστοιχήσουμε το λογικό 'ψευδές' στον αριθμό 0 και το λογικό 'αληθές' στον αριθμό 1 και να χρησιμοποιήσουμε καποιο μέτρο ομοιότητας προορισμένο για αριθμητικά διανύσματα. Ωστόσο, αυτή η αντιστοίχηση θα γινόταν μάλλον καταχρηστικά, διότι η αναπαράσταση των διανυσμάτων με λογικές συνιστώσες είναι εξαρχής προορισμένη για μια συνολοθεωρητική περιγραφή των αντίστοιχων αντικειμένων, που δεν είναι αυτονόητο ότι έχει κάποια ισοδύναμη αριθμητική περιγραφή. Συνεπώς, τα μέτρα ομοιότητας για διανύσματα λογικών συνιστωσών θα πρέπει να είναι σε θέση να εκφράσουν με κάποιον τρόπο την ομοιότητα μεταξύ των αντίστοιχων συνόλων. Για παράδειγμα, θα μπορούσαν να χρησιμοποιούν το βαθμό επικάλυψης των δύο εμπλεκόμενων συνόλων πραγματοποιώντας επιπλέον κάποιου είδους κανονικοποίηση ως προς την πληθικότητά τους.

Έστω ότι δίνονται τα διανύσματα $\vec{x}$ και $\vec{y}$ με τις συνιστώσες τους να είναι δίτιμες λογικές μεταβλητές. Έστω, επίσης, ότι το σύνολο A είναι το σύνολο των 'αληθών' συνιστωσών του $\vec{x}$, το B είναι το σύνολο των 'αληθών' συνιστωσών του $\vec{y}$, και το C είναι το σύνολο των συνιστωσών που είναι 'αληθείς' και στα δύο λογικά διανύσματα. Εποπτικά τα παραπάνω φαίνονται στο σχήμα 2.1. Τα περισσότερα από τα γνωστά μέτρα ομοιότητας για διανύσματα λογικών συνιστωσών βασίζονται στην πληθικότητα των παραπάνω συνόλων, διαφέροντας μόνο ως προς τον συντελεστή που χρησιμοποιούν, προκειμένου να πραγματοποιήσουν την κανονικοποίηση του μέτρου, ώστε η τιμή του να πέσει εντός του διαστήματος $[0, 1]$. Ενδεικτικά παραθέτουμε



Σχήμα 2.1: Πράξεις μεταξύ συνόλων.

τέσσερα συχνά χρησιμοποιούμενα μέτρα ομοιότητας:

$$\text{Συντελεστής Jaccard:} \quad \frac{|A \cap B|}{|A \cup B|} = \frac{|C|}{|A| + |B| - |C|}$$

$$\text{Συντελεστής Dice:} \quad \frac{2 \times |A \cap B|}{|A| + |B|} = \frac{2 \times |C|}{|A| + |B|}$$

$$\text{Συντελεστής Επικάλυψης:} \quad \frac{|A \cap B|}{min(|A|, |B|)} = \frac{|C|}{min(|A|, |B|)}$$

$$\text{Μέτρο Συνημιτόνου:} \quad \frac{|A \cap B|}{\sqrt{|A| \times |B|}} = \frac{|C|}{\sqrt{|A| \times |B|}}$$

Παρατηρούμε ότι το Μέτρο Συνημιτόνου εμφανίζεται και στην περίπτωση διανυσμάτων με λογικές συνιστώσες. Ο τύπος του προκύπτει από τον τύπο που ισχύει για αριθμητικά διανύσματα με την αντικατάσταση του 'ψευδούς' από τον αριθμό 0 και του 'αληθούς' από τον αριθμό 1. Βέβαια, σε αυτήν την περίπτωση δεν επιδέχεται πλέον τη γεωμετρική ερμηνεία του συνημιτόνου κάποιας γωνίας, αλλά ορίζει ένα μέτρο ομοιότητας που εκφράζει την κανονικοποιημένη επικάλυψη των δύο συνόλων.

2.4 Μείωση διάστασης

Όπως έχει ήδη αναφερθεί, ένα εγγενές χαρακτηριστικό της αναπαράστασης δεδομένων δυαδικών συσχετίσεων με το Μοντέλο του Διανυσματικού Χώρου είναι το γεγονός, ότι τα διανύσματα που προκύπτουν είναι πολυδιάστατα και αραιά. Υπενθυμίζουμε ότι η πυκνότητα ενός διανύσματος ορίζεται σαν το ποσοστό των συνιστωσών του που έχουν μη μηδενικές, ή γενικότερα μη προκαθορισμένες τιμές. Η χαμηλή πυκνότητα των διανυσμάτων συνεπάγεται την ανάγκη ειδικών χειρισμών, όπως είναι για παράδειγμα η χρήση εξειδικευμένων συναρτήσεων απόστασης που είναι σε θέση να λαμβάνουν υπόψη τους με αυξημένη βαρύτητα τις μη μηδενικές συνιστώσες των διανυσμάτων. Από την άλλη μεριά, η υψηλή διάσταση των διανυσμάτων δημιουργεί επιπρόσθετες δυσκολίες στην αποδοτική επεξεργασία των δεδομένων, διότι από τη μια αυξάνει κατακόρυφα τους υπολογιστικούς πόρους που απαιτούνται και

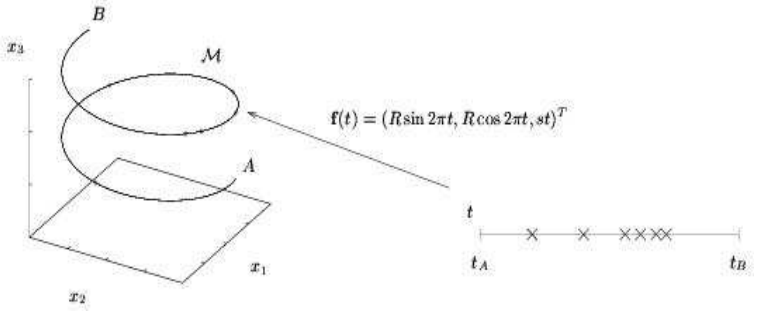
από την άλλη αποκρύπτει εκείνα τα συγκεκριμένα χαρακτηριστικά που είναι σε θέση να αποκαλύψουν τις ουσιαστικές συσχετίσεις που υπάρχουν μέσα στα δεδομένα. Η αντιμετώπιση των δυο τελευταίων προκλήσεων είναι το κίνητρο μιας διαδικασίας που είναι γνωστή σαν μείωση διάστασης, όπως αναλυτικότερα εξηγείται στα ακόλουθα.

Έστω μια εφαρμογή, στην οποία ένα σύστημα επεξεργάζεται δεδομένα που βρίσκονται στη μορφή ενός συνόλου διανυσμάτων, όπως είναι το σήμα φωνής, οι εικόνες, ή γενικότερα ένα σύνολο μετρήσεων κάποιου πολυμεταβλητού μεγέθους. Ας υποθέσουμε ότι το σύστημα μπορεί να ανταποκριθεί από άποψη υπολογιστικών πόρων, μόνο αν η διάσταση του κάθε διανύσματος δεν είναι πολύ υψηλή, όπου το κατώφλι αυτό βέβαια εξαρτάται από τη συγκεκριμένη εφαρμογή. Όταν όμως τα δεδομένα προς επεξεργασία είναι υψηλότερης διάστασης απ' ό,τι είναι ανεκτό, προκύπτει η ανάγκη να μειώσουμε τη διάσταση τους σε ένα ανεκτό μέγεθος, κρατώντας όσο περισσότερη από την αρχική πληροφορία μπορούμε, και στη συνέχεια να τροφοδοτήσουμε το σύστημα επεξεργασίας με τα χαμηλής διάστασης δεδομένα. Κάτω από αυτό το πρίσμα, η διαδικασία της μείωσης διάστασης εμφανίζεται σαν ένα στάδιο προεπεξεργασίας που είναι απαραίτητο για τη λειτουργία του συνολικού συστήματος.

Συχνά, ένα φαινόμενο που με μια πρώτη ματιά φαίνεται ότι απαιτεί για την περιγραφή του έναν μεγάλο αριθμό μεταβλητών, μπορεί να διέπεται στην πραγματικότητα από έναν μικρότερο αριθμό μεταβλητών, οι οποίες ωστόσο δεν είναι δυνατό να μετρηθούν άμεσα, είναι δηλαδή κατά κάποιον τρόπο κρυμμένες μέσα στα δεδομένα. Ένα χαρακτηριστικό παράδειγμα είναι η φυσική ομιλία, μια αναμφίβολα πολύπλοκη διαδικασία, για την οποία ωστόσο έχει διατυπωθεί η εικασία ότι η μοντελοποίησή της μπορεί να επιτευχθεί με τη χρησιμοποίηση πέντε μόνο μεταβλητών. Σε τέτοιες περιπτώσεις, η μείωση διάστασης εμφανίζεται σαν ένα ισχυρό εργαλείο για τη μοντελοποίηση των αντίστοιχων φαινομένων, δεδομένου ότι ανακαλύπτει τις κρυμμένες μεταβλητές που τα διέπουν και συνεισφέρει με αυτόν τον τρόπο στην καλύτερη κατανόησή τους. Ακόμα όμως και στην περίπτωση που οι σημαντικές για την περιγραφή του φαινομένου μεταβλητές δεν είναι κρυμμένες, αλλά απλώς 'θαμμένες' μέσα στο πλήθος όλων των υπόλοιπων, άσχετων με το συγκεκριμένο φαινόμενο μεταβλητών, η μείωση διάστασης στην ιδεατή περίπτωση αναλαμβάνει να τις ξεχωρίσει και να στρέψει τη προσοχή μας σε αυτές.

Προκειμένου να γίνει περισσότερη κατανοητή η συζήτηση, στο σχήμα 2.2 παραθέτουμε ένα παράδειγμα μείωση διάστασης από τη Γεωμετρία, που κατά κάποιον τρόπο οπτικοποιεί την προηγούμενη γενική περιγραφή. Στο συγκεκριμένο παράδειγμα, ένα σύνολο σημείων του χώρου $\mathbb{R}^3$, που στη γενική περίπτωση απαιτούν τρεις συνιστώσες για την διανυσματική αναπαράστασή τους, απεικονίζεται σε ένα σύνολο σημείων του χώρου $\mathbb{R}$, που απαιτούν μόνο μια συνιστώσα. Το κρυμμένο αίτιο που επιτρέπει αυτή την απεικόνιση είναι το γεγονός ότι τα σημεία ανήκουν όλα σε μια καμπύλη σπείρας, δηλαδή είναι συσχετισμένα μεταξύ τους με έναν τρόπο που δεν είναι άμεσα φανερός.

Κατά καιρούς έχουν προταθεί διάφορες τεχνικές για τη μείωση διάστασης, που μπορούν να χωριστούν σε δυο κατηγορίες, όσον αφορά τη γενική μεθοδολογία προ-

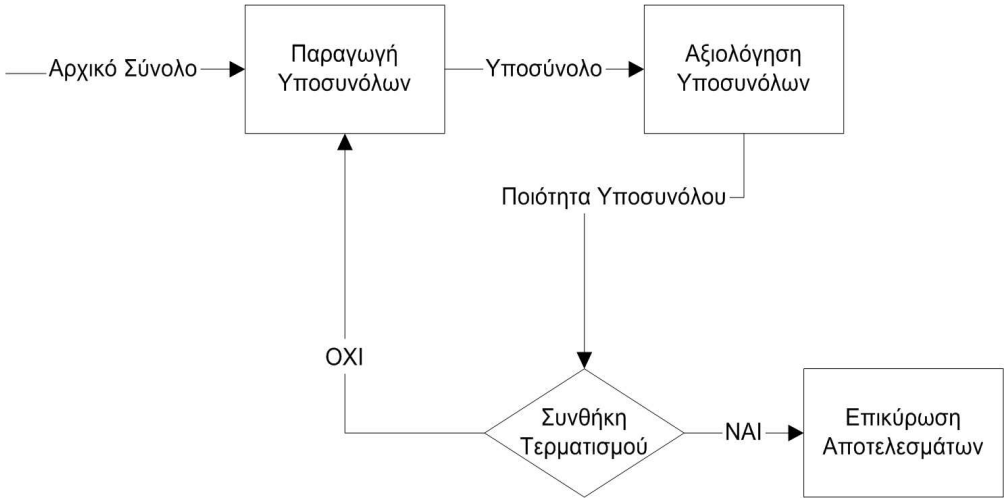


Σχήμα 2.2: Γεωμετρικό παράδειγμα μείωσης διάστασης [3].

σέγγισης του προβλήματος. Η πρώτη μεθοδολογία ονομάζεται επιλογή χαρακτηριστικών και βασίζεται στην επιλογή ενός υποσυνόλου χαρακτηριστικών από αυτά που ήδη εμφανίζονται και στην αρχική διανυσματική αναπαράσταση των δεδομένων. Η δεύτερη μεθοδολογία ονομάζεται εξαγωγή χαρακτηριστικών και επιδιώκει να βρει ένα νέο σύνολο χαρακτηριστικών για την αναπαράσταση των δεδομένων, ξένο προς αυτό που εμφανίζεται στην αρχική αναπαράσταση. Πίσω από αυτήν την προσέγγιση κρύβεται η πεποίθηση ότι οι νέες συντεταγμένες που εισάγονται ενσωματώνουν σε μια συμπυκνωμένη αναπαράσταση τις συσχετίσεις που υπάρχουν στα αρχικά δεδομένα. Σε αυτό το σημείο απλώς αναφέρουμε, ότι οι αλγόριθμοι μείωσης διάστασης που υλοποιούμε στην παρούσα εργασία ανήκουν στη δεύτερη κατηγορία, δηλαδή πραγματοποιούν εξαγωγή χαρακτηριστικών.

Το γενικό πλαίσιο ενός αλγορίθμου επιλογής χαρακτηριστικών φαίνεται στο σχήμα 2.3. Η διαδικασία αποτελείται από τέσσερα διακριτά στάδια: παραγωγή των υποσυνόλων των χαρακτηριστικών, αξιολόγηση των υποσυνόλων, έλεγχο κάποιου κριτηρίου τερματισμού και επικύρωση της ποιότητας των αποτελεσμάτων. Τα τρία πρώτα στάδια ορίζουν μια διαδικασία αναζήτησης του βέλτιστου υποσυνόλου στον χώρο όλων των δυνατών υποσυνόλων που μπορούν να προκύψουν από το αρχικό σύνολο των χαρακτηριστικών. Η αναζήτηση αυτή είναι συνήθως ευρετική, διότι μια εξαντλητική αναζήτηση απαιτεί την εξέταση όλων των δυνατών υποσυνόλων, των οποίων το πλήθος αυξάνεται εκθετικά συναρτήσει του αρχικού πλήθους των χαρακτηριστικών. Οι αλγόριθμοι επιλογής χαρακτηριστικών κατηγοριοποιούνται ανάλογα με την συγκεκριμένη μέθοδο που εφαρμόζεται σε καθένα από τα τρία στάδια, δηλαδή ανάλογα με τον τρόπο που μεταβαίνουν από τη μία κατάσταση του χώρου αναζήτησης στην επόμενη, ανάλογα με το ποσοτικό μέτρο που χρησιμοποιούν προκειμένου να αξιολογήσουν τις καταστάσεις και ανάλογα με το κριτήριο που χρησιμοποιούν προκειμένου να διακόψουν την αναζήτηση. Για μια εποπτική περιγραφή της γενικής μεθοδολογίας επιλογής χαρακτηριστικών και πολλών συγκεκριμένων αλγορίθμων παραπέμπουμε στο [17].

Η εξαγωγή χαρακτηριστικών από την άλλη μεριά αποσκοπεί στην εκ νέου απεικόνιση των διανυσμάτων των δεδομένων σε ένα νέο σύστημα αξόνων. Γι' αυτόν τον



Σχήμα 2.3: Γενική μεθοδολογία επιλογής χαρακτηριστικών.

σκοπό συχνά χρησιμοποιούνται τεχνικές από τη Γραμμική 'Αλγεβρα, που στοχεύουν στον μετασχηματισμό του αρχικού δισδιάστατου πίνακα των δεδομένων σε έναν πίνακα μικρότερης διάστασης, ο οποίος πληρεί κάποιο στατιστικό κριτήριο βελτιστότητας. Παραδείγματα τέτοιων μεθόδων είναι η Ανάλυση Πρωτευουσών Συνιστωσών και η Αποσύνθεση Ιδιαζουσών Τιμών. 'Αλλες μέθοδοι εξαγωγής χαρακτηριστικών στοχεύουν σε μια απεικόνιση των αρχικών διανυσμάτων σε έναν χώρο χαμηλότερης διάστασης, με τρόπο που οι σχετικές αποστάσεις μεταξύ των σημείων του συνόλου δεδομένων να διατηρούνται. Για μια εποπτική παρουσίαση ορισμένων αλγορίθμων εξαγωγής χαρακτηριστικών παραπέμπουμε στο [3].

Εξειδικεύοντας τη συζήτησή μας για την περίπτωση δεδομένων κειμένου, δηλαδή εγγράφων και λέξεων, τονίζουμε ότι η μείωση διάστασης με τη μεθοδολογία της εξαγωγής χαρακτηριστικών αποτελεί μια δυνατότητα αντιμετώπισης των προβλημάτων που δημιουργούνται εξαιτίας της πολυσημίας και της συνωνυμίας κατά την προσπάθεια ανάκλησης σχετικών εγγράφων. Συγκεκριμένα, μέσω της μείωσης διάστασης είναι θεωρητικά δυνατό να ανακαλυφθούν κάποιες νέες μεταβλητές που αντιστοιχούν στις θεματικές ενότητες που θίγονται στο σύνολο των εγγράφων. Η διανυσματική αναπαράσταση των εγγράφων στον χώρο που ορίζουν αυτές οι νέες μεταβλητές παρέχει τη δυνατότητα σύγκρισης της ομοιότητας των εγγράφων με πιο αξιόπιστο τρόπο απ' αυτόν που παρέχει η αναπαράστασή τους στον χώρο που ορίζουν οι μεμονωμένες λέξεις.

Κεφάλαιο 3

Ομαδοποίηση

3.1 Εισαγωγή

Η ομαδοποίηση είναι μια διαδικασία επεξεργασίας δεδομένων, η οποια έχει σαν στόχο τον διαμερισμό του συνόλου των δεδομένων σε υποσύνολα. Η βασική ιδιότητα που πρέπει να πληρεί ο διαμερισμός αυτός, απαιτεί οποιοδήποτε στοιχείο που έχει ταξινομηθεί σε κάποιο υποσύνολο, να μοιάζει περισσότερο με τα υπόλοιπα στοιχεία του ίδιου υποσυνόλου απ' ό,τι με τα στοιχεία των άλλων υποσυνόλων. Σε αυτό το σημείο βέβαια, ήδη υπάρχει μια ασάφεια ως προς τον τρόπο με τον οποίο ορίζεται η ομοιότητα μεταξύ των στοιχείων του συνόλου των δεδομένων, αλλά αυτό δεν θα μας απασχολήσει στα πλαίσια του προηγούμενου άτυπου ορισμού. Με πιο απλά λόγια, η ομαδοποίηση είναι μια διαδικασία που χρησιμοποιεί κανείς προκειμένου να ανακαλύψει ομοιογενείς ομάδες μέσα σε δεδομένα, για την δομή των οποίων έχει αρχικά λίγη διαθέσιμη πληροφορία. Η διαδικασία της ομαδοποίησης είναι βασικό στοιχείο πολλών συστημάτων ανάκλησης πληροφορίας, αναγνώρισης προτύπων ή επεξεργασίας βιολογικών δεδομένων. Ειδικά τα τελευταία χρόνια, που η υπολογιστική ισχύς αυξήθηκε κατακόρυφα και η εξόρυξη μεγάλου όγκου δεδομένων έγινε εφικτή, η ομαδοποίηση αποτέλεσε ισχυρό εργαλείο στα χέρια των ερευνητών, ενώ η εύρεση αποδοτικών αλγορίθμων αποτέλεσε και αποτελεί σημαντικό ερευνητικό πεδίο.

Δεδομένου ότι ένας αλγόριθμος ομαδοποίησης ξεκινάει την εξερεύνηση ενός συνόλου δεδομένων έχοντας αρχικά ελάχιστη ή καθόλου πληροφορία, δεν μπορεί παρά να στηριχτεί στο μεγαλύτερο μέρος του στον υπολογισμό κάποιου μέτρου ομοιότητας, ή ισοδύναμα κάποιας συνάρτησης απόστασης, για διάφορα ζευγάρια στοιχείων του συνόλου των δεδομένων. Κατά συνέπεια, ο ορισμός ενός επαρκούς για την εκάστοτε περίπτωση μέτρου ομοιότητας είναι πολύ σημαντικός, προκειμένου ο αλγόριθμος της ομαδοποίησης να σχηματίσει όσο το δυνατόν πιο ομοιογενείς ομάδες. Ειδικότερα, ένα πετυχημένο μέτρο ομοιότητας θα πρέπει να είναι σε θέση να αντιμετωπίσει αποτελεσματικά τα φαινόμενα της υψηλής διάστασης και της χαμηλής πυκνότητας των διανυσμάτων των δεδομένων. Για μια συνοπτική παρουσίαση και σύγκριση ορισμένων μέτρων ομοιότητας, τα οποία έχουν αποδειχθεί αποτελεσματικά

για αραιά δεδομένα δυαδικών συσχετίσεων, παραπέμπουμε στην ενότητα 2.3. Ωστόσο, ακόμα και στην περίπτωση που έχει οριστεί ένα αποτελεσματικό μέτρο ομοιότητας, ένας αλγόριθμος ομαδοποίησης καλείται να αντιμετωπίσει ποικίλα προβλήματα, που προκύπτουν από την άγνοια της δομής του χώρου που ορίζουν τα στοιχεία του συνόλου δεδομένων με την συγκεκριμένη συνάρτηση απόστασης. Συγκεκριμένα, τα προβλήματα της διαφορετικής πυκνότητας, του διαφορετικού σχήματος και του διαφορετικού μεγέθους των ομάδων, ήταν και είναι καθοριστικοί περιοριστικοί παράγοντες για τη διαδικασία της ομαδοποίησης. Συνεπώς, ένας ιδεατός αλγόριθμος καλείται να σχηματίσει όσο το δυνατόν πιο ομοιογενείς ομάδες, ακόμα και αν αυτές διαφέρουν μεταξύ τους σε μέγεθος, σχήμα ή πυκνότητα.

Κλείνοντας την παρούσα εισαγωγική ενότητα, θα παρουσιάσουμε μια συνοπτική ταξινόμηση των αλγορίθμων ομαδοποίησης, προκειμένου οι αλγόριθμοι που θα παρουσιαστούν στις επόμενες ενότητες να ενταχθούν σε ένα γενικότερο πλαίσιο. Όσον αφορά την υφή των υποσυνόλων στα οποία διαχωρίζουν τα αρχικά δεδομένα, οι αλγόριθμοι ομαδοποίησης χαρακτηρίζονται σαν σκληροί ή ασαφείς. Οι πρώτοι διαμερίζουν το αρχικό σύνολο δεδομένων σε ξένα μεταξύ τους υποσύνολα, δηλαδή κάθε στοιχείο ανήκει μετά τον τερματισμό του αλγορίθμου σε μια και μόνο ομάδα, ενώ οι δεύτεροι καταλήγουν σε μια ποσοστιαία τοποθέτηση του κάθε στοιχείου μέσα στις σχηματιζόμενες ομάδες. Οι σκληροί αλγόριθμοι διακρίνονται περαιτέρω σε διαχωριστικούς και ιεραρχικούς, ανάλογα με το αν οι ομάδες που σχηματίζουν είναι ανεξάρτητες μεταξύ τους ή σχηματίζουν μια δενδρική δομή συνόλων, στην οποία κάθε ομάδα είναι υπερσύνολο όλων των παιδιών της. Δηλαδή, οι ιεραρχικοί αλγόριθμοι καταλήγουν σε μια ταξινόμηση που βαίνει από το γενικότερο προς το ειδικότερο. Τέλος, όσον αφορά τον τρόπο με τον οποίο οι ιεραρχικοί αλγόριθμοι οδεύουν προς τον σχηματισμό των τελικών ομάδων, διακρίνονται σε διαιρετικούς και συνενωτικούς, ανάλογα με το αν αρχίζουν από μεγάλες ομάδες που σταδιακά διαιρούν σε μικρότερες βάσει κάποιων κριτηρίων, ή αν αρχίζουν από μικρές ομάδες που σταδιακά συνενώνουν σε μεγαλύτερες.

Είναι βέβαια περιττό να αναφέρουμε, ότι οι αλγόριθμοι ομαδοποίησης που έχουν κατά καιρούς προταθεί είναι πάρα πολλοί, με διαφορετικά χαρακτηριστικά όσον αφορά την υπολογιστική πολυπλοκότητα, τις ρυθμιστικές παραμέτρους που δέχονται και τις ιδιότητες των ομάδων που τείνουν να σχηματίζουν. Στις επόμενες ενότητες θα παρουσιάσουμε τρεις σκληρούς, διαχωριστικούς αλγορίθμους ομαδοποίησης, που υλοποιήθηκαν στα πλαίσια της παρούσας διπλωματικής εργασίας και χρησιμοποιήθηκαν για την επεξεργασία δεδομένων του Παγκοσμίου Ιστού. Κοινό χαρακτηριστικό και των τριών αλγορίθμων είναι η ακριβής εύρεση των k κοντινότερων γειτόνων για κάθε στοιχείο του αρχικού συνόλου δεδομένων και γι' αυτόν τον λόγο ανήκουν στην οικογένεια των αλγορίθμων ομαδοποίησης, που λέμε ότι βασίζονται στην αρχή των k Κοντινότερων Γειτόνων. Είναι προφανές, ότι αυτή η αρχή απαιτεί τον υπολογισμό του μέτρου ομοιότητας για κάθε δυνατό ζευγάρι στοιχείων του συνόλου δεδομένων και άρα προκύπτει άμεσα ότι η πολυπλοκότητα των αντίστοιχων αλγορίθμων είναι $O(n^2)$, όπου n είναι η πληθικότητα του συνόλου των δεδομένων.

Επειδή για περιπτώσεις μεγάλων συνόλων δεδομένων η τετραγωνική πολυπλοκότητα είναι απαγορευτική, απλώς αναφέρουμε ότι έχουν προταθεί πολλοί άλλοι ευριστικοί αλγόριθμοι ομαδοποίησης με χαμηλότερη πολυπλοκότητα.

3.2 Αλγόριθμος Συμπαγών Ομάδων

Αν θεωρήσουμε ένα σύνολο σημείων του επιπέδου που σχηματίζουν διακριτές ομάδες, είναι λογικό να φανταστούμε μια διαδικασία ομαδοποίησης, η οποία θα προσπαθήσει πρώτα να εντοπίσει τα μεγαλύτερα και πιο συμπαγή νέφη σημείων. Σε αυτή την ιδέα βασίζεται ο αλγόριθμος Συμπαγών Ομάδων που περιγράφεται στην παρούσα ενότητα. Τα βήματα του αλγορίθμου είναι τα εξής:

1. Για κάθε σημείο βρες τους k κοντινότερους γείτονές του βάσει κάποιου κατά βούληση επιλεγμένου μέτρου ομοιότητας. Το κάθε σημείο γίνεται εν δυνάμει κέντρο κάποιας ομάδας. Το k είναι ρυθμιστική παράμετρος του αλγορίθμου.

2. Σχημάτισε την πιο συμπαγή ομάδα, δηλαδή αυτή που το κέντρο της απέχει τη μικρότερη απόσταση από το πιο απομακρυσμένο σημείο της.

3. Σχημάτισε την πολυπληθέστερη ομάδα που είναι το λιγότερο τόσο συμπαγής όσο η ομάδα που σχηματίστηκε τελευταία.

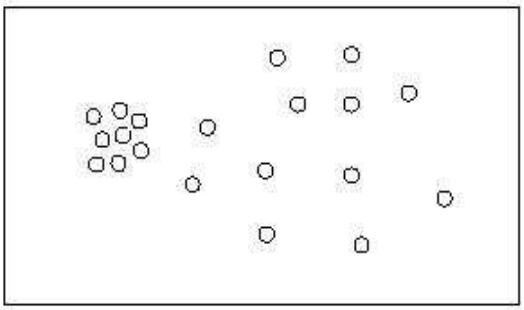
4. Επανάλαβε το βήμα 3 έως ότου να μην μπορούν να σχηματιστούν άλλες ομάδες, παρά μόνο με μεμονωμένα σημεία.

Το σημείο κλειδί του αλγορίθμου εντοπίζεται στο βήμα 3 και συγκεκριμένα στην απαίτηση κάθε νέα ομάδα που σχηματίζεται να είναι πιο συμπαγής από την αμέσως προηγούμενη ομάδα που είχε σχηματιστεί. Δηλαδή, το βασικό κριτήριο που χρησιμοποιεί ο αλγόριθμος για τον σχηματισμό των ομάδων είναι το πόσο συμπαγείς είναι αυτές. Είναι προφανές βέβαια, ότι σε αντιστάθμισμα για την ολοένα και μεγαλύτερη συμπάγεια των σχηματιζόμενων ομάδων η πληθικότητά τους γίνεται ολοένα και μικρότερη. Συμπερασματικά, ο αλγόριθμος σχηματίζει μια ακολουθία ομάδων με ολοένα και μεγαλύτερη συμπάγεια και ολοένα και μικρότερο μέγεθος.

Όσον αφορά τη ρυθμιστική παράμετρο k, αυτή καθορίζει το μέγεθος της γειτονιάς που βρίσκουμε αρχικά για το κάθε σημείο και κατ' επέκταση την αρχική τιμή του κατωφλίου που χρησιμοποιείται στη συνέχεια για το σχηματισμό ολοένα και πιο συμπαγών ομάδων. Συνεπώς, αναμένουμε ότι αύξηση της τιμής του k οδηγεί στον σχηματισμό περισσότερων ομάδων, διότι η ελάχιστη τιμή συμπάγειας που υπολογίζεται στην πρώτη επανάληψη του αλγορίθμου αυξάνει και αυτή με την αύξηση του k. Η πιο χαλαρή αρχική τιμή κατωφλίου δίνει με τη σειρά της την ευκαιρία σε περισσότερες ομάδες να είναι υποψήφιες προς επιλογή σύμφωνα με τα κριτήρια του βήματος 3. Συμπερασματικά μπορούμε να πούμε, ότι μεσω του k ρυθμίζεται το πλήθος και η συμπάγεια των σχηματιζόμενων ομάδων. Σε αυτό το σημείο αξίζει να παρατηρήσουμε, ότι ο αλγόριθμος Συμπαγών Ομάδων δεν οδηγεί στη γενική

περίπτωση σε 100% κάλυψη του αρχικού συνόλου σημείων, δηλαδή δεν επιβάλλει την απαίτηση όλα τα σημεία να καταλήξουν τοποθετημένα σε κάποια ομάδα. Τα σημεία που καταλήγουν να μείνουν μεμονωμένα μπορούν, ανάλογα με την γνώση που υπάρχει για το αρχικό σύνολο δεδομένων, είτε να θεωρηθούν θόρυβος και να αγνοηθούν, είτε να θεωρηθούν σημαντικά για την περιγραφή των δεδομένων και να αποτελέσουν ομάδες του ενός στοιχείου.

Η απλή λογική του αλγορίθμου τού επιβάλλει ωστόσο έναν πολύ σημαντικό περιορισμό. Συγκεκριμένα, ας φανταστούμε την περίπτωση μιας κατανομής σημείων, στην οποία υπάρχει μια περιοχή υψηλής πυκνότητας, δηλαδή μια περιοχή με πολλά σημεία τα οποία βρίσκονται πολύ κοντά μεταξύ τους, όπως φαίνεται στο σχήμα 3.1. Κατά την εκτέλεση του αλγορίθμου, και ήδη από την πρώτη επανάληψη, η ανώτατη επιτρεπόμενη τιμή της συμπάγειας θα καθοριστεί από την πυκνή ομάδα των σημείων και συνεπώς θα πάρει μια πολύ μικρή τιμή. Αυτό έχει σαν αποτέλεσμα, οι ομάδες σημείων που θα σχηματιστούν στα επόμενα βήματα να πρέπει να είναι τουλάχιστον τόσο πυκνές. Το τελικό αποτέλεσμα θα είναι ο σχηματισμός λίγων ομάδων υψηλής πυκνότητας και η αγνόηση άλλων ομάδων, που πιθανώς παρουσιάζουν ομοιογένεια, αλλά έχουν χαμηλότερη πυκνότητα. Δηλαδή, ο αλγόριθμος Συμπαγών Ομάδων αδυνατεί να ανιχνεύσει ομάδες με μεγάλες διαφορές στην πυκνότητα τους, αφού καταλήγει να αγνοεί τις λιγότερο πυκνές από αυτές.



Σχήμα 3.1: Ομάδες διαφορετικής πυκνότητας.

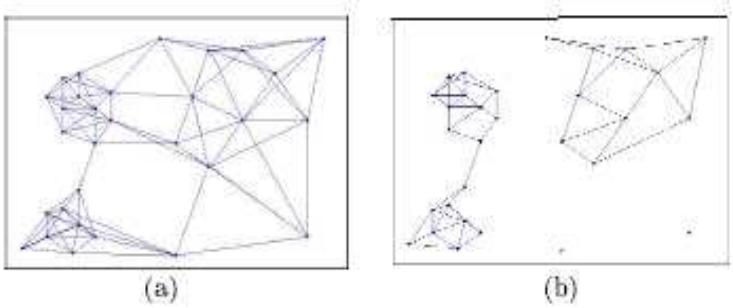
3.3 Αλγόριθμος Jarvis-Patrick

Στην ενότητα 2.3 παρουσιάστηκαν ορισμένα μέτρα ομοιότητας που μπορούν να φανούν αποτελεσματικά κατά την επεξεργασία αραιών και πολυδιάστατων δεδομένων, και επίσης εξηγήθηκε με ένα παράδειγμα γιατί η κλασική ευκλείδεια συνάρτηση απόστασης δεν αποτελεί σε αυτήν την περίπτωση μια αξιόπιστη λύση. Ωστόσο, και μέτρα ομοιότητας όπως ο συντελεστής Jaccard ή το Μετρο Συνημιτόνου, αν και πιο αξιόπιστα από την ευκλείδεια απόσταση, δεν παύουν να έχουν αδυναμίες. Για παράδειγμα, ένα πρόβλημα που προκύπτει από την υψηλή διάσταση των διανυσμάτων

είναι το γεγονός ότι οι αποστάσεις μεταξύ των διαφορετικών σημείων τείνουν να είναι όλες εν γένει μεγάλες, χωρίς μεγάλες διαφορές μεταξύ τους. Αυτή η εξομοίωση των τιμών των αποστάσεων μεταξύ των ζευγαριών σημείων μειώνει την διακριτική ικανότητα του μέτρου ομοιότητας και καθιστά συνεπώς δυσκολότερο το έργο των αλγορίθμων ομαδοποίησης, που βασίζονται στον υπολογισμό αυτών των αποστάσεων. Επιπλέον, ένα άλλο πρόβλημα πολλών μέτρων ομοιότητας είναι το γεγονός ότι δεν καθιστούν ισχύουσα την τριγωνική ανισότητα και συνεπώς το γεγονός ότι δυο σημεία βρίσκονται κοντά σε ένα τρίτο δεν συνεπάγεται αυτόματα ότι βρίσκονται κοντά και μεταξύ τους (βλ. ενότητα 2.3). Η τελευταία διαπίστωση έρχεται σε αντίθεση με την διαισθητική αντίληψή μας, σύμφωνα με την οποία οποιοδήποτε ζευγάρι σημείων που ανήκουν στην ίδια ομάδα θα πρέπει να βρίσκονται κοντά μεταξύ τους.

Ένας εναλλακτικός τρόπος για να ορίσουμε την ομοιότητα μεταξύ δυο σημείων είναι να στηριχθούμε στο πλήθος των κοινών κοντινών γειτόνων των εν λόγω σημείων, όπως αυτοί προκύπτουν με τη χρήση ενός παραδοσιακού μέτρου ομοιότητας. Συγκεκριμένα, ορίζουμε σαν ομοιότητα k Κοινών Κοντινών Γειτόνων (k-ΚΚΓ) δυο σημείων το πλήθος των κοινών γειτόνων που έχουν αυτά τα σημεία ανάμεσα στους k κοντινότερούς τους. Η ομοιότητα k-ΚΚΓ μεταξύ δύο σημείων ορίζεται μόνο στην περίπτωση που το καθένα απο τα δύο σημεία βρίσκεται στη λίστα των k κοντινότερων γειτόνων του άλλου. Η ιδέα πίσω από αυτόν τον ορισμό είναι ότι αν τα σημεία P_1 και P_2 είναι κοντινά μεταξύ τους βάσει κάποιου μέτρου ομοιότητας, και αν επιπλέον και τα δυο είναι κοντινά σε όλα τα σημεία ενός συνόλου S, τότε μπορούμε να είμαστε περισσότερο βέβαιοι για την ομοιότητα των P_1 και P_2 απ' ό,τι αν χρησιμοποιούσαμε απλώς την τιμή του μέτρου ομοιότητας.

Το σχήμα 3.2 απεικονίζει δυο σημαντικές ιδιότητες του μέτρου ομοιότητας k-ΚΚΓ στην περίπτωση ενός συνόλου δισδιάστατων σημείων. Στο σχήμα 3.2a, για κάθε σημείο έχουν σχεδιαστεί ακμές που το ενώνουν με τους πέντε κοντινότερους γείτονές του. Από την άλλη μεριά, στο σχήμα 3.2b υπάρχει ακμή μεταξύ δυο σημείων, μόνο αν αυτά βρίσκονται το ένα μέσα στους πέντε κοντινότερους γείτονες του άλλου, ενώ για απλότητα αγνοείται το ακριβές πλήθος των ΚΚΓ που κανονικά θα έμπαινε σαν βάρος της αντίστοιχης ακμής του γραφήματος. Μπορούμε να κάνουμε



Σχήμα 3.2: Μέτρο ομοιότητας k Κοινών Κοντινών Γειτόνων [9].

δυο παρατηρήσεις σχετικά με αυτό το σχηματικό παράδειγμα. Πρώτον, τα σημεία που βρίσκονται μακριά από οποιαδήποτε ομάδα, και που πιθανώς αντιστοιχούν σε θόρυβο, καταλήγουν να μην συνδέονται με κάποιο σημείο διότι δεν βρίσκονται στη λίστα των ίδιων τους των γειτόνων. Συνεπώς, η χρήση του μέτρου ομοιότητας των k-ΚΚΓ μοιάζει να αντιμετωπίζει αποτελεσματικά τον θόρυβο. Δεύτερον, οι συνδέσεις που αντιστοιχούν σε περιοχές ομοιόμορφης πυκνότητας διατηρούνται, ανεξαρτήτως αν η πυκνότητα είναι μεγάλη ή μικρή, ενώ αντίθετα οι συνδέσεις που αντιστοιχούν σε περιοχές μετάβασης μεταξύ περιοχών διαφορετικών πυκνοτήτων απαλείφονται. Αυτή η ιδιότητα είναι σημαντική, διότι η ανίχνευση ομάδων με μεγάλες διαφορές στην πυκνότητα είναι μια από τις πιο σημαντικές προκλήσεις για τους αλγορίθμους ομαδοποίησης.

Η ιδέα του ορισμού ενός μέτρου ομοιότητας βασισμένου στο πλήθος των κοινών κοντινών γειτόνων χρησιμοποιήθηκε για πρώτη φορά από τους Jarvis και Patrick [15], και το όνομα αυτών των ερευνητών πήρε επίσης ο αλγόριθμος ομαδοποίησης που παρουσιάζουμε παρακάτω:

1. Για κάθε σημείο βρες του k κοντινότερους γείτονές του βάσει κάποιου κατά βούληση επιλεγμένου μέτρου ομοιότητας. Το k είναι ρυθμιστική παράμετρος του αλγορίθμου.

2. Για κάθε δυνατό ζευγάρι σημείων δοκίμασε αν το ένα βρίσκεται στην λίστα των γειτόνων του άλλου. Αν συμβαίνει αυτό, τότε πήγαινε στο βήμα 3, αλλιώς συνέχισε με το επόμενο ζευγάρι σημείων.

3. Αν τα σημεία του ζευγαριού έχουν στη λίστα των γειτόνων τους τουλάχιστον m κοινά σημεία, τότε τοποθέτησε τα στην ίδια ομάδα. Σε διαφορετική περίπτωση, γύρισε στο βήμα 2 και συνέχισε με το επόμενο ζευγάρι σημείων. Το m είναι ρυθμιστική παράμετρος του αλγορίθμου.

Το χαρακτηριστικό του αλγορίθμου Jarvis-Patrick είναι η χρήση του μέτρου ομοιότητας k-ΚΚΓ με όλα τα πλεονεκτήματα που αυτό συνεπάγεται. Επιπλέον, οι δύο ρυθμιστικές παράμετροι k και m παρέχουν τη δυνατότητα προσαρμογής του αλγορίθμου σε διάφορες περιπτώσεις δεδομένων, χωρίς αυτό να σημαίνει βέβαια ότι ο καθορισμός των βέλτιστων τιμών γι' αυτές τις παραμέτρους είναι κάτι εύκολο. Σε γενικές γραμμές, για σταθερή τιμή του k αύξηση του m αναμένουμε να οδηγήσει σε μείωση της κάλυψης και ταυτόχρονη αύξηση του πλήθους των σχηματιζόμενων ομάδων. Υπενθυμίζουμε ότι το m αντιπροσωπεύει μια τιμή κατωφλίου, η οποία καθορίζει την ελάχιστη τιμή του μέτρου ομοιότητας k-ΚΚΓ που πρέπει να έχουν δύο σημεία, προκειμένου να τοποθετηθούν στην ίδια ομάδα. Συνεπώς, η αύξηση του m κάνει τον αλγόριθμο πιο επιλεκτικό, με την έννοια ότι απαιτείται μεγαλύτερη τιμή ομοιότητας μεταξύ δύο σημείων προκειμένου αυτά να τοποθετηθούν στην ίδια ομάδα. Αυτό έχει σαν αποτέλεσμα να συγχωνεύονται λιγότερες ομάδες, άρα το πλήθος των ομάδων να αυξάνεται, ενώ την ίδια στιγμή λιγότερα σημεία να πληρούν την προϋπόθεση προς ένταξη σε κάποια ομάδα. Από την άλλη μεριά και για τους

ίδιους λόγους, για δεδομένη τιμή του m αύξηση του k αναμένουμε να προκαλέσει αύξηση της κάλυψης και μείωση του πλήθους των ομάδων. Αυτό συμβαίνει, διότι η λίστα των κοντινών γειτόνων για κάθε σημείο είναι μεγαλύτερη και συνεπώς για κάποιο ζευγάρι σημείων είναι πιο πιθανό να εντοπιστούν περισσότεροι κοινοί γείτονες, άρα τα σημεία να έχουν μεγαλύτερη ομοιότητα βάσει του μέτρου k-ΚΚΓ.

Πάντως σε γενικές γραμμές έχει παρατηρηθεί ότι ο αλγόριθμος Jarvis - Patrick έχει μια ροπή προς τον σχηματισμό 'επιμηκών' ομάδων, δηλαδή ομάδων όπου το πιο απομακρυσμένο ζευγάρι σημείων απέχουν πολύ μεταξύ τους. Η αιτία αυτού του φαινομένου είναι ότι κατά τη διαδικασία κατανομής των σημείων στις ομάδες ο αλγόριθμος δεν διστάζει να τοποθετήσει κάποιο σημείο P σε κάποια ομάδα C, όταν ανακαλύψει έστω κι ένα σημείο της C που είναι κοντινό στο P, δηλαδή που έχει τιμή μέτρου ομοιότητας k-ΚΚΓ μεγαλύτερη από m. Θα μπορούσε κανείς να φανταστεί παραλλαγές του αλγορίθμου, όπου το κριτήριο τοποθέτησης κάποιου σημείου σε μια ομάδα θα λάμβανε υπόψη του την ομοιότητα του εν λόγω σημείου με περισσότερα του ενός σημεία της ομάδας.

3.4 Αλγόριθμος Πυκνότητας Κοινών Κοντινών Γειτόνων

Ο αλγόριθμος ομαδοποίησης που περιγράφεται σ' αυτήν την ενότητα παρουσιάζεται στο [9] και αποτελεί επέκταση του αλγορίθμου Jarvis-Patrick, καθώς και αυτός υιοθετεί το μέτρο ομοιότητας των k-ΚΚΓ. Από την άλλη μεριά, η λογική σχηματισμού των ομάδων είναι πιο πολύπλοκη από αυτή του αλγορίθμου Jarvis-Patrick, διότι χρησιμοποιεί επιπλέον τη νέα έννοια της πυκνότητας των σημείων. Συγκεκριμένα, ο αλγόριθμος Πυκνότητας Κοινών Κοντινών Γειτόνων (ΠΚΚΓ) στηρίζεται σε ένα κλασικό μέτρο ομοιότητας και επαναϋπολογίζει την ομοιότητα των σημείων βάσει του μέτρου k-ΚΚΓ, ακριβώς όπως κάνει και ο αλγόριθμος Jarvis-Patrick. Στη συνέχεια, χρησιμοποιώντας τις νέες τιμές των αποστάσεων βρίσκει τα πυρηνικά σημεία, δηλαδή τα σημεία που συγκεντρώνουν γύρω τους έναν αρκετά μεγάλο αριθμό άλλων σημείων. Σε αυτό ακριβώς το σημείο υπεισέρχεται η έννοια της πυκνότητας ενός σημείου P, η οποία ορίζεται σαν το πλήθος των σημείων που βάσει του μέτρου ομοιότητας k-ΚΚΓ έχουν τιμή ομοιότητας με το P μεγαλύτερη ή ίση από μια καθορισμένη τιμή r. Οι ομάδες σχηματίζονται γύρω από τα πυρηνικά σημεία με την τοποθέτηση των υπόλοιπων σημείων στην ομάδα του κοντινότερού τους πυρηνικού σημείου. Τελικά οι ομάδες κοντινών πυρηνικών σημείων συγχωνεύονται. Τα ακριβή βήματα του αλγορίθμου ΠΚΚΓ δίνονται παρακάτω:

1. Για κάθε σημείο βρες τους k κοντινότερους γείτονές του βάσει κάποιου κατά βούληση επιλεγμένου μέτρου ομοιότητας. Το k είναι ρυθμιστική παράμετρος του αλγορίθμου.

2. Για κάθε σημείο P, βρες τα σημεία που έχουν μαζί του τουλάχιστον m ΚΚΓ. Αυτό ισοδυναμεί με την εύρεση των σημείων, που βάσει του μέτρου ομοιότητας

k-ΚΚΓ απέχουν από το σημείο P λιγότερο από m. Το m είναι ρυθμιστική παράμετρος του αλγορίθμου. Το πλήθος των σημείων που πληρούν την παραπάνω ιδιότητα ονομάζεται πυκνότητα του σημείου P.

3. Βρες τα σημεία με πυκνότητα μεγαλύτερη από r. Αυτά τα σημεία ονομάζονται πυρηνικά. Το r είναι ρυθμιστική παράμετρος του αλγορίθμου.

4. Αν δυο πυρηνικά σημεία έχουν τουλάχιστον m ΚΚΓ, τότε τοποθετούνται στην ίδια ομάδα.

5. Για κάθε μη-πυρηνικό σημείο εξέτασε αν έχει τουλάχιστον m ΚΚΓ με κάποιο πυρηνικό σημείο. Αν συμβαίνει αυτό, τότε τοποθέτησε το μη-πυρηνικό σημείο στην ίδια ομάδα που ανήκει και το αντίστοιχο πυρηνικό σημείο. Σε διαφορετική περίπτωση, το μη-πυρηνικό σημείο χαρακτηρίζεται σαν θόρυβος και δεν τοποθετείται σε καμία ομάδα.

Συμπερασματικά, επαναλαμβάνουμε ότι ο παραπάνω αλγόριθμος συνδυάζει το αποτελεσματικό μέτρο ομοιότητας των k-ΚΚΓ με την ιδέα τον πυρηνικών σημείων υψηλής πυκνότητας. Με το πρώτο στοιχείο αντιμετωπίζει τα προβλήματα της διαφορετικής πυκνότητας των ομάδων και της αναξιοπιστίας των κλασικών μέτρων ομοιότητας για διανύσματα υψηλής διάστασης, ενώ με το δεύτερο στοιχείο αντιμετωπίζει τα προβλήματα του διαφορετικού σχήματος των ομάδων.

3.5 Η ομαδοποίηση ως τεχνική μείωσης διάστασης

Όπως αναφέρθηκε και στην ενότητα 2.4, μια από τις βασικές επιδιώξεις της μείωσης διάστασης είναι η εύρεση των συσχετίσεων που υπάρχουν μέσα στα δεδομένα, αλλά δεν αναφέρονται ρητά, και η χρησιμοποίησή τους προς την κατεύθυνση μιας πιο συμπαγούς αναπαράστασης των δεδομένων. Από την άλλη μεριά, η διαδικασία της ομαδοποίησης αποσκοπεί στην εύρεση ομοιογενών ομάδων στα δεδομένα, οι οποίες υπό μια έννοια αντιστοιχούν σε όχι ρητά διατυπωμένες συσχετίσεις. Συνεπώς, ήδη μπορούμε να φανταστούμε ότι η ομαδοποίηση θα μπορούσε να χρησιμοποιηθεί σαν μια τεχνική μείωσης διάστασης, αρκεί βέβαια η αντίστοιχη διαδικασία να βασιστεί σε έναν λογικά θεμελιωμένο φορμαλισμό. Στην παρούσα ενότητα παρουσιάζουμε έναν τέτοιο, βασισμένο στην ομαδοποίηση αλγόριθμο μείωσης διάστασης, ο οποίος παρουσιάζεται στο [16] και ονομάζεται Εννοιολογική Δεικτοδότηση. Για λόγους απλοποίησης της ορολογίας, η περιγραφή που ακολουθεί διατυπώνεται για την περίπτωση που τα δεδομένα προς επεξεργασία είναι δυαδικά δεδομένα συσχετίσεων της φυσικής γλώσσας, δηλαδή έγγραφα και λέξεις. Ωστόσο, η ίδια τεχνική θα μπορούσε να χρησιμοποιηθεί αυτούσια και στην περίπτωση δεδομένων δυαδικών συσχετίσεων άλλου τύπου, διότι ο μαθηματικός φορμαλισμός δεν θα άλλαζε σε κανένα σημείο.

Έστω ένας πίνακας δυαδικών συσχετίσεων για ζευγάρια εγγράφων-λέξεων, όπως περιγράφεται στην ενότητα 2.1. Σύμφωνα με το Μοντέλο του Διανυσματικού Χώρου

το κάθε έγγραφο αναπαρίσταται σαν ένα πολυδιάστατο και αραιό διάνυσμα λέξεων, με συνιστώσες που έχουν είτε λογικές είτε αριθμητικές τιμές. Στη συνέχεια θα θεωρήσουμε ότι η αναπαράσταση των διανυσμάτων εμπίπτει στην δεύτερη περίπτωση, διότι ακόμα και τα διανύσματα με δίτιμες συνιστώσες μπορούν να μετατραπούν σε διανύσματα με αριθμητικές συνιστώσες αντικαθιστώντας το λογικό 'αληθές' με τον αριθμό 1 και το λογικό 'ψευδές' με τον αριθμό 0. Βέβαια αυτή η αντιστοίχηση γίνεται μάλλον καταχρηστικά, αλλά δεδομένου ότι ο μαθηματικός φορμαλισμός που ακολουθεί απαιτεί διανύσματα αριθμητικών συνιστωσών, είναι ό,τι καλύτερο μπορούμε να κάνουμε με τη δεδομένη πληροφορία που διαθέτουμε.

Δεδομένου ενός συνόλου εγγράφων S και των αντίστοιχων διανυσματικών τους αναπαραστάσεων, ορίζουμε σαν κεντροειδές του συνόλου S το διάνυσμα $\vec{C}$ που δίνεται από τον τύπο: $\vec{C} = \frac{1}{|S|} \sum_{d \in S} \vec{d}$. Το διάνυσμα $\vec{C}$ είναι ο διανυσματικός μέσος όρος όλων των εγγράφων του συνόλου S, ό,τι είναι δηλαδή και το κέντρο βάρους για ένα σύστημα σημειακών μαζών. Διαισθητικά, το κεντροειδές $\vec{C}$ παρέχει έναν μηχανισμό περίληψης των περιεχομένων των εγγράφων του αντίστοιχου συνόλου, διότι οι πιο σημαντικές συνιστώσες του αντιστοιχούν στις λέξεις που εμφανίζονται με τα μεγαλύτερα βάρη στα έγγραφα του συνόλου S.

Έχοντας παρουσιάσει τον ορισμό του κεντροειδούς διανύσματος ενός συνόλου εγγράφων, μπορούμε πλέον να προχωρήσουμε στην ακριβέστερη περιγραφή της μεθοδολογίας μείωσης διάστασης. Έστω ότι έχουμε έναν πίνακα εγγράφων-λέξεων διαστάσεων $N \times M$, δηλαδή υπάρχουν N έγγραφα που το καθένα αναπαρίσταται σαν ένα διάνυσμα M λέξεων. Τα βήματα που ακολουθεί ο αλγόριθμος της Εννοιολογικής Δεικτοδότησης είναι τα εξής:

1. Πάνω στο σύνολο των εγγράφων εκτελείται κάποιος αλγόριθμος ομαδοποίησης, που τελικά σχηματίζει k ομάδες. Η τιμή του k μπορεί, ανάλογα με τον χρησιμοποιούμενο αλγόριθμο, είτε να επιβάλλεται άμεσα, είτε να καθορίζεται έμμεσα από τις τιμές άλλων ρυθμιστικών παραμέτρων, όπως συμβαίνει με τους αλγόριθμους που παρουσιάσαμε στα προηγούμενα.

2. Για κάθε ομάδα εγγράφων υπολογίζεται το αντίστοιχο κεντροειδές διάνυσμα και κανονικοποιείται ώστε να έχει μοναδιαίο μήκος. Τελικά καταλήγουμε στο σύνολο των κανονικοποιημένων κεντροειδών διανυσμάτων $\{\vec{C_1}, \vec{C_2}, \ldots, \vec{C_k}\}$.

3. Δεδομένης της πολυδιάστατης αναπαράστασης $\vec{d_H}$ ενός εγγράφου η k-διαστατή αναπαράσταση του είναι η εξής: $\vec{d_L} = (\vec{d_H} \cdot \vec{C_1}, \vec{d_H} \cdot \vec{C_2}, \ldots, \vec{d_H} \cdot \vec{C_k})$.

Με ορολογία Γεωμετρίας, η παραπάνω διαδικασία αντιστοιχεί στην εύρεση ενός k-διάστατου συστήματος μοναδιαίων αξόνων και την προβολή των πολυδιάστατων διανυσμάτων των εγγράφων πάνω σε αυτούς τους άξονες. Το χαμηλής διάστασης σύστημα συντεταγμένων υπολογίζεται από τα αποτελέσματα της διαδικασίας ομαδοποίησης και ο κάθε άξονας του αντιπροσωπεύει τη θεματική ενότητα που καλύπτεται από την αντίστοιχη ομάδα εγγράφων. Στη συνέχεια, η προβολή του διανύσματος

$\vec{d_H}$ πάνω στον άξονα $\vec{C_j}$, καθορίζει την j-οστή συνιστώσα της διανυσματικής αναπαράστασης του εγγράφου d στον μειωμένης διάστασης χώρο και εκφράζει το κατά πόσο το συγκεκριμένο έγγραφο εμπίπτει στην θεματική ενότητα που αντιπροσωπεύει ο άξονας $\vec{C_j}$. Τελικά, στην χαμηλής διάστασης αναπαράσταση τα έγγραφα δεν χαρακτηρίζονται βάσει των λέξεων που περιέχουν, αλλά βάσει της ομοιότητάς τους με τους θεματικούς πυρήνες που έχουν ανιχνευθεί στο σύνολο των εγγράφων.

Κεφάλαιο 4

Πιθανοτική Ανάλυση Κρυμμένης Σημασιολογίας

4.1 Μοντέλο των Όψεων

Η Πιθανοτική Ανάλυση Κρυμμένης Σημασιολογίας (ΠΑΚΣ) είναι μια γενική μεθοδολογία πιθανοτικής μοντελοποίησης της διαδικασίας παραγωγής ενός συνόλου δυαδικών δεδομένων. Προκειμένου να επιτύχει μια όσο δυνατόν καλύτερη προσαρμογή του πιθανοτικού μοντέλου για κάποιο δεδομένο σύνολο παρατηρήσεων, η ΠΑΚΣ προβλέπει μια διαδικασία εκπαίδευσης, κατά την οποία πραγματοποιείται μια βέλτιστη εκτίμηση των παραμέτρων του μοντέλου. Τα πιθανοτικά μοντέλα που μπορεί να υιοθετήσει κανείς μέσα στο γενικότερο πλαίσιο της ΠΑΚΣ είναι τρία [11], αλλά εμείς για τους σκοπούς της εργασίας χρησιμοποιούμε μόνο ένα από αυτά, που ονομάζεται Μοντέλο των Όψεων. Κατά την περιγραφή που ακολουθεί θεωρούμε ότι τα δεδομένα που επεξεργαζόμαστε είναι ζευγάρια εγγράφων-λέξεων, προκειμένου να απλοποιηθεί η ορολογία που χρησιμοποιούμε. Ωστόσο, όσα αναφέρονται μπορούν να εφαρμοστούν αυτούσια για οποιαδήποτε περίπτωση δυαδικών δεδομένων, αρκεί βέβαια να είναι λογική η παραδοχή ότι οι κοινές εμφανίσεις των εμπλεκόμενων οντοτήτων μπορούν να ερμηνευθούν από το Μοντέλο των Όψεων.

Έστωσαν ένα σύνολο εγγράφων $D = \{d_1, d_2, \ldots, d_N\}$, ένα σύνολο λέξεων $W = \{w_1, w_2, \ldots, w_M\}$, καθώς και ο αντίστοιχος πίνακας δυαδικών συσχετίσεων των εγγράφων και των λέξεων. Το Μοντέλο των Όψεων βασίζεται στην παραδοχή ότι κάθε παρατήρηση, δηλαδή κάθε εμφάνιση κάποιας συγκεκριμένης λέξης σε κάποιο συγκεκριμένο έγγραφο, σχετίζεται με την ύπαρξη κάποιας κρυμμένης μεταβλητής z_k που ανήκει στο σύνολο $Z = \{z_1, z_2, \ldots, z_A\}$, με $A << N, M$. Ορίζουμε τις ακόλουθες πιθανότητες:

- $P(d_i)$: η πιθανότητα να εμφανιστεί το έγγραφο d_i στον πίνακα δυαδικών συσχετίσεων, δηλαδή η πιθανότητα το συγκεκριμένο έγγραφο να περιέχει μια οποιαδήποτε λέξη.
- $P(z_k|d_i)$: η δεσμευμένη πιθανότητα που έχει η κρυμμένη μεταβλητή z_k να

σχετίζεται με ένα ζευγάρι, στο οποίο είναι ήδη σίγουρο ότι υπάρχει το έγγραφο d_i.

- $P(w_j|z_k)$: η δεσμευμένη πιθανότητα που έχει η λέξη w_j να αποτελεί μέρος ενός ζευγαριού, το οποίο είναι ήδη σίγουρο ότι σχετίζεται με την κρυμμένη μεταβλητή z_k.

Χρησιμοποιώντας τους παραπάνω ορισμούς, μπορεί κανείς να περιγράψει ένα πιθανοτικό μοντέλο παραγωγής των δεδομένων ζευγαριών εγγράφων - λέξεων ως εξής:

1. Διάλεξε ένα έγγραφο d_i με πιθανότητα $P(d_i)$.
2. Διάλεξε μια κρυμμένη μεταβλητή z_k με πιθανότητα $P(z_k|d_i)$.
3. Διάλεξε μια λέξη w_j με πιθανότητα $P(w_j|z_k)$.

Σαν αποτέλεσμα της παραπάνω διαδικασίας, μπορούμε να υπολογίσουμε την πιθανότητα να παρατηρηθεί ένα συγκεκριμένο ζευγάρι εγγράφου - λέξης (d_i, w_j), αθροίζοντας τις πιθανότητες για όλες τις δυνατές επιλογές των κρυμμένων μεταβλητών, που μπορούν δεδομένου του εγγράφου d_i να δημιουργήσουν την λέξη w_j.

$$P(d_i, w_j) = P(d_i)P(w_j|d_i) = P(d_i)\sum_{k=1}^{A} P(w_j|z_k)P(z_k|d_i)$$

Σε αυτό το σημείο αξίζει να σημειωθεί ότι μια ισοδύναμη διατύπωση της παραπάνω σχέσης μπορεί να προκύψει με εφαρμογή του τύπου του Bayes ως εξής:

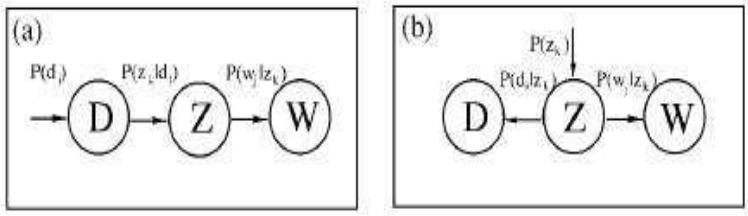
$$P(d_i, w_j) = P(d_i)\sum_{k=1}^{A} P(w_j|z_k)P(z_k|d_i) =$$
$$= P(d_i)\sum_{k=1}^{A} P(w_j|z_k)\frac{P(z_k)P(d_i|z_k)}{P(d_i)} = \sum_{k=1}^{A} P(z_k)P(d_i|z_k)P(w_j|z_k)$$

Παρατηρούμε ότι σε αυτή τη δεύτερη σχέση υπολογισμού των πιθανοτήτων $P(d_i, w_j)$, οι δυο οντότητες, έγγραφα και λέξεις, εξαρτώνται με τον ίδιο συμμετρικό τρόπο από την αντίστοιχη κρυμμένη μεταβλητή. Σε αυτήν την περίπτωση, η διαδικασία παραγωγής των παρατηρούμενων ζευγαριών μπορεί να διατυπωθεί ως εξής:

1. Διάλεξε μια κρυμμένη μεταβλητή z_k με πιθανότητα $P(z_k)$.
2. Διάλεξε ένα έγγραφο d_i με πιθανότητα $P(d_i|z_k)$.
3. Διάλεξε μια λέξη w_j με πιθανότητα $P(w_j|z_k)$.

Οι δυο ισοδύναμες διατυπώσεις του Μοντέλου των Όψεων και οι αντίστοιχοι μηχανισμοί παραγωγής των παρατηρήσεων απεικονίζονται εποπτικά στο σχήμα 4.1. Η δεύτερη συμμετρική παραμέτριση προσφέρεται επιπλέον και για μια διαισθητική ερμηνεία της λογικής του συγκεκριμένου πιθανοτικού μοντέλου: κάθε ζευγάρι εγγράφου-λέξης που παρατηρείται θεωρείται ότι έχει σαν γενεσιουργό αίτιο κάποιον θεματικό

πυρήνα που θίγει, και ο οποίος αντιπροσωπεύεται από την αντίστοιχη κρυμμένη μεταβλητή z_k. Είναι λογικό, δεδομένου ενός θέματος, οι πιθανότητες εμφάνισης των διάφορων εγγράφων και των διάφορων λέξεων να ποικίλλουν και είναι επίσης λογικό σε ένα σύνολο εγγράφων τα διάφορα θέματα να εμφανίζονται με διαφορετική βαρύτητα. Δηλαδή, το Μοντέλο των Όψεων φαίνεται ότι μπορεί να παράσχει έναν τρόπο αντιμετώπισης των προβλημάτων της συνωνυμίας και της πολυσημίας, διότι συσχετίζει το κάθε ζευγάρι εγγράφου - λέξης με κάποια κρυμμένη μεταβλητή που αντιπροσωπεύει το νόημα της συγκεκριμένης λέξης στο συγκεκριμένο έγγραφο, με άλλα λόγια τα συμφραζόμενα της εν λόγω λέξης.



Σχήμα 4.1: Γραφική απεικόνιση του Μοντέλου των Όψεων [12].

Βέβαια, η θεωρητική περιγραφή του Μοντέλου των Όψεων που προηγήθηκε δεν το καθιστά αυτόματα χρήσιμο στην πράξη, διότι καμία από τις πιθανότητες που εισήχθησαν δεν είναι με κάποιον τρόπο γνωστή εκ των προτέρων. Οι εν λόγω πιθανότητες είναι άγνωστες παράμετροι του μοντέλου, που θα πρέπει να εκτιμηθούν βάσει κάποιου κριτηρίου βελτιστοποίησης. Η λύση στο πρόβλημα ευρεσης μιας όσο το δυνατόν καλύτερης εκτίμησης των παραμέτρων του μοντέλου, δεδομένων των παρατηρήσεων που έχουν πραγματοποιηθεί, αναπτύσσεται στις επόμενες δυο ενότητες.

4.2 Μεγιστοποίηση της Αναμενόμενης Τιμής

Έστω το σύνολο X, επί του οποίου ορίζεται η συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) $p(x, \vec{\Theta}), \forall x \in X$. Θεωρούμε ότι η σ.π.π. p εξαρτάται από κάποιες παραμέτρους, που για λόγους απλοποίησης του συμβολισμού συνοψίζουμε στο διάνυσμα $\vec{\Theta}$. Έστω επίσης ένα σύνολο ανεξάρτητων μεταξύ τους παρατηρήσεων $\mathcal{X}$, του οποίου τα στοιχεία ανήκουν στο X και ακολουθούν όλα την σ.π.π. $p(x, \vec{\Theta})$. Η πιθανοφάνεια $\mathcal{L}(\mathcal{X}, \vec{\Theta})$ του συνόλου των παρατηρήσεων ορίζεται σαν η πιθανότητα του γεγονότος να πραγματοποιηθούν όλες οι παρατηρήσεις. Δεδομένης της ανεξαρτησίας των ενδεχομένων αυτή η πιθανότητα ισούται με το γινόμενο των αντίστοιχων πιθανοτήτων, δηλαδή $\mathcal{L}(\mathcal{X}, \vec{\Theta}) = \prod_{i=1}^{N} p(x_i, \vec{\Theta})$, όπου $p(x_i, \vec{\Theta})$ είναι η πιθανότητα να πραγματοποιηθεί η i-οστή παρατήρηση και N είναι το συνολικό πλήθος των παρατηρήσεων. Η Εκτίμηση Μέγιστης Πιθανοφάνειας είναι το πρόβλημα εύρεσης των κατάλληλων τιμών των παραμέτρων $\vec{\Theta}$ της σ.π.π. $p(x, \vec{\Theta})$ έτσι, ώστε η πιθανοφάνεια

των παρατηρήσεων να μεγιστοποιείται. Διαισθητικά, αυτό σημαίνει ρύθμιση των άγνωστων παραμέτρων $\vec{\Theta}$, ώστε η υιοθετούμενη συνάρτηση πυκνότητας πιθανότητας να δικαιολογεί όσο το δυνατόν καλύτερα τις δεδομένες παρατηρήσεις. Πολλές φορές για λόγους απλοποίησης των υπολογισμών, αντί της ποσότητας $\mathcal{L}(\mathcal{X}, \vec{\Theta})$ μεγιστοποιούμε την ποσότητα $log\ \mathcal{L}(\mathcal{X}, \vec{\Theta})$, που ονομάζεται λογαριθμική πιθανοφάνεια των δεδομένων. Επειδή η λογαριθμική συνάρτηση είναι γνησίως αύξουσα, είναι προφανές ότι αν σε κάποιο σημείο η λογαριθμική πιθανοφάνεια παρουσιάζει ακρότατο, τότε και η απλή πιθανοφάνεια παρουσιάζει ακρότατο του ίδιου τύπου.

Σε ορισμένες εφαρμογές ωστόσο, η μοντελοποίηση της διαδικασίας παραγωγής των δεδομένων επιβάλλει την ύπαρξη μεταβλητών, τις οποίες κατά τη διαδικασία καταγραφής των παρατηρήσεων αδυνατούμε να μετρήσουμε. Δηλαδή, σε κάθε στοιχείο των παρατηρούμενων δεδομένων αντιστοιχεί και ένα συμπληρωματικό στοιχείο με τις τιμές των κρυμμένων μεταβλητών, το οποίο όμως είναι άγνωστο. Σε αυτήν την περίπτωση θεωρούμε ότι υπάρχουν δυο σύνολα X και Y, επί του καρτεσιανού γινομένου των οποίων ορίζεται η συνάρτηση πυκνότητας πιθανότητας $p(x, y, \vec{\Theta}) \forall (x, y) \in X \times Y$. Επίσης υπάρχει το σύνολο των παρατηρήσεων $\mathcal{X}$ με στοιχεία που ανήκουν όλα στο X, καθώς και το σύνολο των κρυμμένων δεδομένων $\mathcal{Y}$ με στοιχεία που ανήκουν όλα στο Y. Το πλήρες σύνολο δεδομένων $\mathcal{Z}$ μπορούμε να θεωρήσουμε ότι είναι η παράθεση των παρατηρήσεων $\mathcal{X}$ και των κρυμμένων δεδομένων $\mathcal{Y}$, δηλαδή $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ και η πιθανοφάνεια του πλήρους συνόλου δεδομένων $\mathcal{L}(\mathcal{X}, \mathcal{Y}, \vec{\Theta})$ ορίζεται σε αυτήν την περίπτωση σαν η πιθανότητα να παρατηρηθεί το πλήρες σύνολο δεδομένων $(\mathcal{X}, \mathcal{Y})$. Επειδή όμως το διάνυσμα των κρυμμένων δεδομένων είναι άγνωστο, η ποσότητα $\mathcal{L}(\mathcal{X}, \mathcal{Y}, \vec{\Theta})$ δεν είναι πια συνάρτηση μόνο των παραμέτρων $\vec{\Theta}$, αλλά και του διανύσματος $\mathcal{Y}$ που μπορεί να θεωρηθεί σαν μια τυχαία μεταβλητή. Κατά συνέπεια, σε αυτήν τη περίπτωση δεν μπορούμε να επιδιώξουμε μεγιστοποίηση της πιθανοφάνειας του πλήρους συνόλου δεδομένων ως προς τις παραμέτρους $\vec{\Theta}$, όπως κάναμε στην προηγούμενη απλή περίπτωση. Το καλύτερο που μπορούμε να κάνουμε είναι να επιδιώξουμε την μεγιστοποίηση της αναμενόμενης τιμής της πιθανοφάνειας (ή της λογαριθμικής πιθανοφάνειας) του πλήρους συνόλου δεδομένων πάνω στην κατανομή της τυχαίας μεταβλητής $\mathcal{Y}$, δηλαδή τη μεγιστοποίηση της ποσότητας $E[log\ \mathcal{L}(\mathcal{X}, \mathcal{Y}, \vec{\Theta})]$.

Ωστόσο, η αναλυτική λύση του προηγούμενου προβλήματος είναι αδύνατη εξαιτίας του γεγονότος ότι η κατανομή που ακολουθεί η τυχαία μεταβλητή $\mathcal{Y}$ εξαρτάται από τις παραμέτρους $\vec{\Theta}$, οι οποίες όμως είναι επίσης άγνωστες. Σε αυτό το σημείο υπεισέρχεται ο αλγόριθμος Μεγιστοποίησης της Αναμενόμενης Τιμής (MAT) [7], που στην πραγματικότητα είναι μάλλον μια γενική μεθοδολογία προσέγγισης του παραπάνω προβλήματος βελτιστοποίησης παρά ένας αλγόριθμος με την έννοια της αυστηρά καθορισμένης αλληλουχίας βημάτων. Η μεθοδολογία MAT υπολογίζει μια ακολουθία τιμών για το διάνυσμα των παραμέτρων $\vec{\Theta}$, η οποία αποδεικνύεται ότι συγκλίνει σε ένα τοπικό τουλάχιστον μέγιστο της αναμενόμενης πιθανοφάνειας των πλήρων δεδομένων. Ο τρόπος με τον οποίο γίνεται ο υπολογισμός της συγκλίνουσας ακολουθίας $\{\vec{\Theta}_0, \vec{\Theta}_1, \ldots, \vec{\Theta}_t, \ldots\}$ είναι ο εξής:

1. Στις αρχικές παραμέτρους $\vec{\Theta}_0$ δίνονται κάποιες τυχαίες ή κατ' εκτίμηση καλές τιμές.

2. Για δεδομένα $\mathcal{X}$ και $\vec{\Theta}_t$ υπολογίζεται η κατανομή της τυχαίας μεταβλητής $\mathcal{Y}$ και με χρήση αυτής της κατανομής υπολογίζεται στη συνέχεια η αναμενόμενη τιμή της πιθανοφάνειας των πλήρων δεδομένων:

$$E_{\mathcal{X},\vec{\Theta}_t}[log\ \mathcal{L}(\mathcal{X},\mathcal{Y},\vec{\Theta})] = Q(\vec{\Theta})$$

3. Μεγιστοποιείται η ποσότητα $Q(\vec{\Theta})$ και τίθεται $\vec{\Theta_{t+1}} \leftarrow argmax\ Q(\vec{\Theta})$.

Τα βήματα 2 και 3 επαναλαμβάνονται για κάποιον προκαθορισμένο αριθμό επαναλήψεων ή έως ότου δυο διαδοχικές τιμές των παραμέτρων $\vec{\Theta}$ ικανοποιήσουν κάποιο κριτήριο σύγκλισης. Δηλαδή, ο αλγόριθμος MAT υπολογίζει μια ολοένα και καλύτερη εκτίμηση των παραμέτρων χρησιμοποιώντας στη διαδικασία της μεγιστοποίησης της αναμενόμενης τιμής της πιθανοφάνειας των δεδομένων την προηγούμενη τιμή των παραμέτρων. Η απόδειξη της σύγκλισης του αλγορίθμου, καθώς και οι μαθηματικές ιδιότητες αυτής ξεφεύγουν από τους σκοπούς της παρουσίασης και μπορούν να αναζητηθούν στα [7, 4].

Σαν τελευταίο σχόλιο, επαναλαμβάνουμε ότι ο αλγόριθμος MAT αποτελεί μια γενική μεθοδολογία προσέγγιση του προβλήματος της μεγιστοποίησης της αναμενόμενης πιθανοφάνειας των πλήρων δεδομένων. Η πρακτική εφαρμογή του αλγορίθμου MAT, δηλαδή ο ακριβής προσδιορισμός των βημάτων 2 και 3 εξαρτάται από το εκάστοτε πρόβλημα που καλούμαστε να επιλύσουμε.

4.3 Εφαρμογή της MAT στο Μοντέλο των Όψεων

Ύστερα από όσα αναφέρθηκαν στην προηγούμενη ενότητα, φαίνεται ότι σαν κριτήριο βελτιστοποίησης για την εκτίμηση των πιθανοτήτων που εμφανίζονται στο Μοντέλο των Όψεων μπορεί να χρησιμοποιηθεί η μεγιστοποίηση της αναμενόμενης πιθανοφάνειας των πλήρων δεδομένων. Ειδικότερα, στην περίπτωση του Μοντέλου των Όψεων τα παρατηρούμενα δεδομένα $\mathcal{X}$ είναι κοινές εμφανίσεις εγγράφων και λέξεων της μορφής (d^n, w^n) με $1 \leq n \leq N$, όπου ο εκθέτης n υποδηλώνει τον αύξοντα αριθμό της παρατήρησης. Η κρυμμένη μεταβλητή z^n, που σύμφωνα με την υπόθεση του μοντέλου είναι το γενεσιουργό αίτιο της αντίστοιχης παρατήρησης, είναι το τμήμα των πλήρων δεδομένων που διαφεύγει της αντίληψης μας και άρα το σύνολο των κρυμμένων δεδομένων $\mathcal{Y}$ είναι το $\{z^1, z^2, \ldots, z^N\}$. Δηλαδή, τα πλήρη δεδομένα $\mathcal{Z}$ αποτελούνται από τριάδες της μορφής (d^n, w^n, z^n). Τέλος, οι παράμετροι $\vec{\Theta}$ που εμφανίζονται στη γενική περιγραφή του αλγορίθμου MAT, στην περίπτωση του Μοντέλου των Όψεων είναι οι πιθανότητες $P(z_k), P(d_i|z_k), P(w_j|z_k)$, για όλες τις δυνατές τιμές i, j, k.

Στα [11, 12] σκιαγραφείται η εφαρμογή της μεθοδολογίας MAT στην περίπτωση του Μοντέλου των Όψεων, η οποία ωστόσο παρουσιάζει αρκετές τεχνικές δυσκολίες. Γι' αυτό το λόγο, στην παρούσα ανάπτυξη θα περιοριστούμε στην παράθεση του τελικού αποτελέσματος, δηλαδή των τύπων που δίνουν τη νέα εκτίμηση των παραμέτρων $\vec{\Theta_{t+1}}$ συναρτήσει της τρέχουσας εκτίμησης $\vec{\Theta_t}$.

$$P_{t+1}(z_k) = \frac{1}{N}\sum_{n=1}^{N}\frac{P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)}{\sum_{r=1}^{A}P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)}$$

$$P_{t+1}(d_i|z_k) = \frac{\sum_{n:d_n=d_i}\frac{P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)}{\sum_{r=1}^{A}P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)}}{\sum_{n=1}^{N}\frac{P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)}{\sum_{r=1}^{A}P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)}}$$

$$P_{t+1}(w_j|z_k) = \frac{\sum_{n:w_n=w_j}\frac{P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)}{\sum_{r=1}^{A}P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)}}{\sum_{n=1}^{N}\frac{P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)}{\sum_{r=1}^{A}P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)}}$$

Όπως ήδη αναφέρθηκε στην προηγούμενη ενότητα, οι διαδοχικές εκτιμήσεις των παραμέτρων που πραγματοποιεί ο αλγόριθμος MAT μπορεί να συνεχίζονται έως ότου εκπληρωθεί κάποιο κριτήριο σύγκλισης. Ωστόσο, μπορεί να χρησιμοποιηθεί και μια άλλη τεχνική γνωστή σαν πρόωρος τερματισμός, η οποία αποσκοπεί στον περιορισμό του φαινομένου της υπερπροσαρμογής. Το φαινόμενο αυτό είναι ένα από τα βασικά προβλήματα οποιουδήποτε αλγορίθμου μηχανικής μάθησης και εμφανίζεται όταν η επίδοση κάποιου μοντέλου σε καινούργια δεδομένα είναι πολύ χειρότερη από την επίδοση του στα δεδομένα με τα οποία έχει εκπαιδευτεί. Η τεχνική του πρόωρου τερματισμού απαιτεί τον τυχαίο διαχωρισμό των διαθέσιμων δεδομένων σε δυο ανεξάρτητα τμήματα: το μεγαλύτερο τμήμα χρησιμοποιείται για την εκπαίδευση του πιθανοτικού μοντέλου, δηλαδή την εκτίμηση των εμπλεκόμενων πιθανοτήτων, και το δεύτερο τμήμα χρησιμοποιείται για τον έλεγχο της απόδοσης του εκπαιδευμένου μοντέλου σε καινούργια δεδομένα. Όταν εφαρμόζεται η τεχνική του πρόωρου τερματισμού στην εκπαίδευση του Μοντέλου των Όψεων, οι επαναλήψεις του αλγορίθμου MAT σταματούν, όταν για την τρέχουσα εκτίμηση των παραμέτρων η πιθανοφάνεια των δοκιμαστικών δεδομένων είναι μικρότερη απ' ό,τι ήταν για την αμέσως προηγούμενη εκτίμηση, δηλαδή όταν η επίδοση του μοντέλου στα δοκιμαστικά δεδομένα αρχίζει να χειροτερεύει.

Στα [11, 12, 13, 14] προτείνεται επίσης μια διαδικασία εκτίμησης των παραμέτρων του Μοντέλου των Όψεων, ελαφρώς τροποποιημένη από την εκδοχή που δίνει η εφαρμογή της κλασικής μεθοδολογίας MAT που έχει παρουσιαστεί στα προηγούμενα. Συγκεκριμένα, προτείνεται μια μεθοδολογία που κατά κάποιον τρόπο συνδυάζει τις μαθηματικές ιδιότητες της λύσης που δίνει ο κλασικός αλγόριθμος MAT με τις ιδιότητες που παρουσιάζει το φυσικό φαινόμενο της ανόπτησης των μετάλλων και ονομάζεται Θερμική Μεγιστοποίηση της Αναμενόμενης Τιμής (ΘMAT). Στις εξισώσεις εκτίμησης των παραμέτρων που δίνει ο αλγόριθμος ΘMAT υπεισέρχεται μια επιπλέον σταθερά β, που σε αναλογία με τα φυσικά συστήματα ονομάζεται

αντίστροφη υπολογιστική θερμοκρασία. Οι εξισώσεις επαναληπτικής εκτίμησης των πιθανοτήτων στην περίπτωση της εφαρμογής του ΘMAT στο Μοντέλο των Όψεων γίνονται:

$$P_{t+1}(z_k) = \frac{1}{N}\sum_{n=1}^{N}\frac{[P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)]^\beta}{\sum_{r=1}^{A}[P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)]^\beta}$$

$$P_{t+1}(d_i|z_k) = \frac{\sum_{n:d_n=d_i}\frac{[P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)]^\beta}{\sum_{r=1}^{A}[P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)]^\beta}}{\sum_{n=1}^{N}\frac{[P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)]^\beta}{\sum_{r=1}^{A}[P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)]^\beta}}$$

$$P_{t+1}(w_j|z_k) = \frac{\sum_{n:w_n=w_j}\frac{[P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)]^\beta}{\sum_{r=1}^{A}[P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)]^\beta}}{\sum_{n=1}^{N}\frac{[P_t(z_k)P_t(d^n|z_k)P_t(w^n|z_k)]^\beta}{\sum_{r=1}^{A}[P_t(z_r)P_t(d^n|z_r)P_t(w^n|z_r)]^\beta}}$$

Επιπλέον τα βήματα του αλγορίθμου ΘMAT είναι ελαφρώς τροποποιημένα από αυτά του κλασικού αλγορίθμου MAT και έχουν ως εξής:

1. Στο β δίνεται αρχική τιμή 1 και αρχίζει η διαδοχή των επαναληπτικών εκτιμήσεων των παραμέτρων. Παρατηρούμε ότι για $\beta = 1$ οι εξισώσεις του ΘMAT συμπίπτουν με τις εξισώσεις του κλασικού MAT.

2. Όταν η επίδοση του μοντέλου στα δοκιμαστικά δεδομένα χειροτερεύσει, η θερμοκρασία β μειώνεται κατά έναν συντελεστή $\eta < 1$ και εκτελείται μια επανάληψη της εκτίμησης των παραμέτρων.

3. Αν η επίδοση στα δοκιμαστικά δεδομένα βάσει των νέων τιμών των παραμέτρων βελτιώνεται, τότε οι επαναλήψεις συνεχίζονται με αυτήν την τιμή του β. Σε διαφορετική περίπτωση επιστρέφουμε στο βήμα 2.

4. Ο αλγόριθμος σταματάει, όταν ο συντελεστής β μειωθεί πέρα από κάποιο όριο ή όταν συμπληρωθεί κάποιος προκαθορισμένος αριθμός επαναλήψεων.

4.4 Η ΠΑΚΣ ως τεχνική μείωσης διάστασης

Στην προηγούμενη ενότητα παρουσιάστηκε ο τρόπος με τον οποίο μπορεί να γίνει μια βέλτιστη εκτίμηση των παραμέτρων του Μοντέλου των Όψεων, δεδομένων των παρατηρήσεων που έχουν πραγματοποιηθεί. Όταν τελικά γίνει αυτό, και επειδή το πλήθος των κρυμμένων μεταβλητών θεωρούμε ότι είναι είναι πολύ μικρότερο τόσο από το πλήθος των εγγράφων όσο και από το πλήθος των λέξεων, τότε με χρήση των παραμέτρων του Μοντέλου των Όψεων μπορούμε να ορίσουμε έναν μετασχηματισμό μείωσης διάστασης. Συγκεκριμένα, ένα έγγραφο d παύει να αναπαρίσταται με το αντίστοιχο αραιό και πολυδιάστατο διάνυσμα των λέξεων $\vec{d_H}$ και απεικονίζεται στο πυκνό και χαμηλής διάστασης διάνυσμα

$$\vec{d_L} = (P(d|z_1), P(d|z_2), \ldots, P(d|z_A))$$

Ο παραπάνω μετασχηματισμός βασίζεται στην αντίληψη ότι οι κρυμμένες μεταβλητές z_k αντιστοιχούν στους θεματικούς πυρήνες που υπάρχουν στο σύνολο των εγγράφων και άρα μπορούν να αποτελέσουν τους άξονες ενός χαμηλής διάστασης 'συστήματος συντεταγμένων'. Από την άλλη μεριά, η συνολική πιθανότητα εμφάνισης κάποιου εγγράφου d δίνεται από τον τύπο:

$$P(d) = P(d|z_1)P(z_1) + P(d|z_2)P(z_2) + , \ldots , + P(d|z_A)P(z_A)$$

Δηλαδή, η συνολική πιθανότητα $P(d)$ είναι γραμμικός συνδυασμός των πιθανοτήτων των διάφορων θεματικών πυρήνων $P(z_k)$. Ο συντελεστής $P(d|z_k)$ του γραμμικού συνδυασμού εκφράζει το ποσοστό που ο θεματικός πυρήνας z_k καθορίζει τη συνολική πιθανότητα εμφάνισης του εγγράφου d. Με αυτήν την έννοια, οι συντελεστές $P(d|z_k)$ χαρακτηρίζουν το εν λόγω έγγραφο στον μειωμένης διάστασης χώρο.

Κεφάλαιο 5

Ανάπτυξη λογισμικού

5.1 Ανάλυση

Στην παρούσα ενότητα παρουσιάζεται η ανάλυση του λογισμικού που αναπτύχθηκε για τους σκοπούς της εργασίας. Συγκεκριμένα, για καθένα από τα τμήματα του λογισμικού παρουσιάζονται με συνοπτικό τρόπο οι λειτουργικές απαιτήσεις που αυτό καλείται να εκπληρώσει, χωρίς ωστόσο να γίνεται καμία αναφορά σε λύσεις που μπορούν να δοθούν ή σε λεπτομέρειες υλοποίησης τέτοιων λύσεων. Επίσης, παρουσιάζονται οι περιορισμοί που επιβάλλουν ο όγκος και ο τύπος των δεδομένων και οι οποίοι πρέπει να ληφθούν υπόψη κατά τη σχεδίαση, που είναι το επόμενο βήμα στη διαδικασία ανάπτυξης του λογισμικού.

5.1.1 Ομαδοποίηση

Η υλοποίηση αλγορίθμων ομαδοποίησης υπενθυμίζουμε ότι είναι βασικό τμήμα της εργασίας, διότι η διαδικασία της ομαδοποίησης χρησιμοποιείται από την Εννοιολογική Δεικτοδότηση με σκοπό την μείωση διάστασης. Ακολουθεί ένας κατάλογος με τις λειτουργικές απαιτήσεις που έχουμε από το υπεύθυνο για την ομαδοποίηση τμήμα του λογισμικού.

1. Εισαγωγή των δεδομένων πάνω στα οποία πρόκειται να εκτελεστεί κάποιος αλγόριθμος ομαδοποίησης από αρχείο τυποποιημένης μορφής. Στην περίπτωση μας, τα δεδομένα που μας ενδιαφέρουν είναι περιεχόμενα πινάκων δυαδικών συσχετίσεων για έγγραφα και λέξεις, όπου οι εγγραφές των πινάκων είναι λογικές μεταβλητές. Τα αρχεία εισόδου είναι απλά αρχεία κειμένου, στα οποία αποθηκεύονται ανά γραμμή τα διανύσματα των εγγράφων με παράθεση των δεικτών των αληθών συνιστωσών τους.

2. Δυνατότητα υπολογισμού διαφόρων μέτρων ομοιότητας τόσο μεταξύ εγγράφων όσο και μεταξύ λέξεων.

3. Εκτέλεση τριών διαφορετικών αλγορίθμων ομαδοποίησης, που θα πρέπει να δουλεύουν ανεξάρτητα από το είδος των οντοτήτων που χειρίζονται (έγγραφα ή λέξεις), καθώς και από το μέτρο ομοιότητας που χρησιμοποιείται για τον καθορισμό των αποστάσεων. Οι ρυθμιστικές παράμετροι των αλγορίθμων πρέπει να καθορίζονται από τον χρήστη-προγραμματιστή. Μετά το πέρας κάποιου αλγορίθμου ομαδοποίησης, τα αποτελέσματα της διαδικασίας θα πρέπει να είναι διαθέσιμα για περαιτέρω χρήση από τον χρήστη-προγραμματιστή.

Οι περιορισμοί που επιβάλλονται στο τμήμα του λογισμικού που αναλαμβάνει την ομαδοποίηση προκύπτουν από τη δομή και τον όγκο των δεδομένων που επεξεργαζόμαστε. Συγκεκριμένα, οι πίνακες εγγράφων-λέξεων που καλούμαστε να χειριστούμε είναι πολύ μεγάλοι ως προς τις διαστάσεις τους, αλλά και πολύ αραιοί, δηλαδή μόνο ένα πολύ μικρό ποσοστό των εγγραφών τους έχει τιμή 'αληθής'. Επιπλέον, και οι τρεις αλγόριθμοι ομαδοποίησης που πρέπει να υλοποιήσουμε έχουν πολυπλοκότητα $O(n^2)$ ως προς τον υπολογισμό του μέτρου ομοιότητας μεταξύ των στοιχείων του συνόλου δεδομένων. Δεδομένου του μεγάλου αριθμού των εγγράφων και των λέξεων, αναμένουμε ότι, εξαιτίας της τετραγωνικής πολυπλοκότητας, ο χρόνος εκτέλεσης των αλγορίθμων θα είναι απαγορευτικά μεγάλος.

5.1.2 Εννοιολογική Δεικτοδότηση

Η Εννοιολογική Δεικτοδότηση υπενθυμίζουμε ότι είναι μια τεχνική μείωσης διάστασης που χρησιμοποιεί κάποιον αλγόριθμο ομαδοποίησης με τη λογική του 'μαύρου κουτιού'. Οι λειτουργίες που πρέπει να φέρνει σε πέρας το αντίστοιχο τμήμα του λογισμικού είναι:

1. Εισαγωγή των αρχικών πολυδιάστατων δεδομένων από αρχείο τυποποιημένης μορφής.

2. Καθορισμός από τον χρήστη-προγραμματιστή κάποιου αλγορίθμου ομαδοποίησης που πρόκειται να χρησιμοποιηθεί από την Εννοιολογική Δεικτοδότηση. Η υπόλοιπη διαδικασία της μείωσης διάστασης θα πρέπει να είναι ανεξάρτητη από τον χρησιμοποιούμενο αλγόριθμο ομαδοποίησης.

3. Εκτέλεση της μείωσης διάστασης πάνω στα αρχικά δεδομένα και διάθεση των χαμηλής διάστασης διανυσμάτων που προκύπτουν στον χρήστη-προγραμματιστή για περαιτέρω χρήση.

5.1.3 Εκπαίδευση Μοντέλου των Όψεων

Υπενθυμίζουμε ότι το συγκεκριμένο τμήμα λογισμικού υλοποιεί την επαναληπτική διαδικασία εκτίμησης των πιθανοτήτων, που προκύπτει από την εφαρμογή της μεθοδολογίας MAT στο Μοντέλο των Όψεων. Οι λειτουργικές απαιτήσεις από το τμήμα λογισμικού είναι:

1. Εισαγωγή των δεδομένων από αρχείο τυποποιημένης μορφής.

2. Διαχωρισμός του συνόλου των δεδομένων σε δύο υποσύνολα, από τα οποία το ένα θα είναι προορισμένο για την εκπαίδευση του μοντέλου και το άλλο για τη δοκιμή των επιδόσεων του σε νέα δεδομένα. Ο διαχωρισμός θα πρέπει να γίνεται με τυχαία επιλογή.

3. Εκτέλεση του επαναληπτικού αλγορίθμου που υπολογίζει τη συγκλίνουσα ακολουθία για τις πιθανότητες που εμφανίζονται στο μοντέλο. Ορισμένες ρυθμιστικές παράμετροι του αλγορίθμου θα πρέπει να εισάγονται από τον χρήστη.

Οι περιορισμοί που επιβάλλονται σε αυτό το τμήμα του λογισμικού οφείλονται αφενός στον όγκο των δεδομένων που καλούμαστε να επεξεργαστούμε και αφετέρου στην αυξημένη πολυπλοκότητα του αλγορίθμου που πραγματοποιεί την εκπαίδευση του Μοντέλου των Όψεων. Συγκεκριμένα, η πολυπλοκότητα κάθε επανάληψης του αλγορίθμου είναι $O(N \cdot M \cdot K)$, όπου N είναι το πλήθος των εγγράφων, M το πλήθος των λέξεων και K το πλήθος των όψεων. Επιπλέον, κατά τον υπολογισμών των διαδοχικών τιμών χρησιμοποιείται η πράξη της ύψωσης πραγματικού αριθμού σε πραγματικό εκθέτη που, ως γνωστόν, είναι υπολογιστικά χρονοβόρα. Δεδομένου ότι οι πίνακες εγγράφων-λέξεων που χειριζόμαστε είναι πολύ υψηλών διαστάσεων, αναμένουμε ότι ακόμα και για μικρές τιμές του K ο χρόνος εκτέλεσης του αλγορίθμου θα είναι υπερβολικά μεγάλος.

5.1.4 Αξιολόγηση της μείωσης διάστασης

Η αξιολόγηση γίνεται με τον δείκτη της Βελτίωσης Ανάκλησης, ο υπολογισμός του οποίου απαιτεί την εύρεση των N κοντινότερων γειτόνων για κάθε έγγραφο στην περίπτωση της πολυδιάστατης και της ολιγοδιάστατης διανυσματικής αναπαράστασης. Οι λειτουργικές απαιτήσεις από το αντίστοιχο τμήμα λογισμικού είναι:

1. Εισαγωγή των αρχικών δεδομένων από αρχείο τυποποιημένης μορφής. Στο αρχείο εισόδου, εκτός από τα ζευγάρια εγγράφων-λέξεων, θα πρέπει να δηλώνεται και η κατηγορία στην οποία ανήκει το κάθε έγγραφο. Οι κατηγορίες των εγγράφων είναι γνωστές εκ των προτέρων.

2. Εφαρμογή πάνω στα αρχικά δεδομένα κάποιας μεθόδου μείωσης διάστασης που καθορίζει ο χρήστης-προγραμματιστής.

3. Εύρεση των N κοντινότερων γειτόνων κάθε εγγράφου στην περίπτωση της πολυδιάστατης και της ολιγοδιάστατης διανυσματικής αναπαράστασης. Τελικά, υπολογισμός της Βελτίωσης Ανάκλησης.

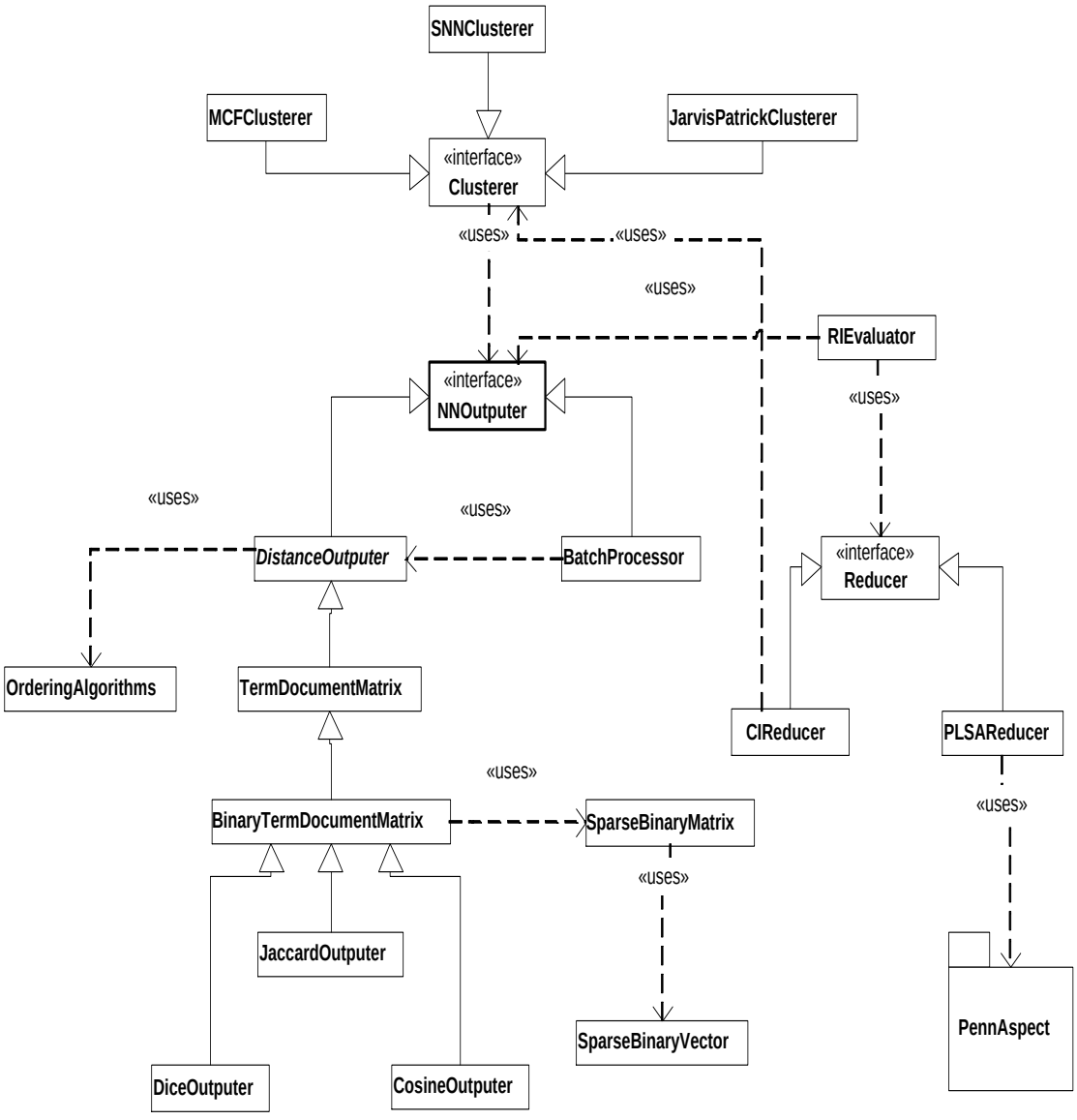
5.2 Σχεδίαση

Το δεύτερο βήμα στην ανάπτυξη του λογισμικού, μετά την ανάλυση των απαιτήσεων, είναι η σχεδίαση. Στην παρούσα ενότητα παρουσιάζουμε τις σημαντικότερες από τις σχεδιαστικές αποφάσεις που λάβαμε, χωρίς όμως να προχωρούμε σε λεπτομέρειες της υλοποίησης. Συγκεκριμένα, η σχεδίαση του λογισμικού γίνεται σύμφωνα με τις αρχές του αντικειμενοστραφούς παραδείγματος προγραμματισμού και σ' αυτήν την ενότητα παρουσιάζεται σε ένα αφηρημένο επίπεδο, ανεξάρτητο από τη γλώσσα υλοποίησης. Η σχεδίαση γίνεται με δυο στόχους: πρώτον, την εκπλήρωση των απαιτήσεων που προέκυψαν από το στάδιο της ανάλυσης και δεύτερον, την όσο το δυνατόν μεγαλύτερη επεκτασιμότητα του συστήματος. Η εκπλήρωση του δεύτερου στόχου επιτρέπει την αξιοποίηση της υπάρχουσας υποδομής για τον μελλοντικό εμπλουτισμό του λογισμικού με επιπλέον λειτουργίες.

Στο σχήμα 5.1 φαίνονται οι πιο σημαντικές κλάσεις που εμπλέκονται στο λογισμικό σύστημα και οι συσχετίσεις τους. Χάριν απλοποιήσεως του σχήματος, σκόπιμα παραλείπονται ορισμένες βοηθητικές κλάσεις που υλοποιούν κάποια πολύ συγκεκριμένη λειτουργικότητα, όπως είναι για παράδειγμα η μετατροπή ενός αρχείου εισόδου από μια μορφή σε κάποια άλλη. Επίσης, το τμήμα του λογισμικού που υλοποιεί την εκπαίδευση του Μοντέλου των Όψεων αναπαρίσταται σαν ένα ενιαίο πακέτο, χωρίς καμία αναφορά στις εσωτερικές του λεπτομέρειες. Αυτό συμβαίνει, διότι αυτό το τμήμα δεν αναπτύχθηκε στα πλαίσια της εργασίας, αλλά ελήφθη έτοιμο προς χρήση από την ιστοσελίδα της ομάδας εξόρυξης δεδομένων του πανεπιστημίου της Πενσυλβάνια [8]. Βέβαια, χρειάστηκαν να γίνουν ορισμένες επεμβάσεις στην αρχική υλοποίηση, προκειμένου το πακέτο λογισμικού να ταιριάξει με τις δικές μας δομές δεδομένων και να βελτιστοποιηθεί ως προς την ταχύτητα εκτέλεσης για τον τύπο δεδομένων της δικής μας περίπτωσης. Ωστόσο, οι λεπτομέρειες αυτών των επεμβάσεων κρίνεται ότι δεν έχουν θέση μέσα σε μια γενική περιγραφή της σχεδίασης.

Στη συνέχεια ακολουθεί μια συνοπτική περιγραφή των κλάσεων που εμφανίζονται στο σχήμα 5.1. Η ακριβής περιγραφή των μεθόδων της κάθε κλάσης μπορεί να αναζητηθεί στην τεκμηρίωση του πηγαίου κώδικα.

- *NNOutputer:* Παρέχει τον ενδιάμεσο μηχανισμό αφαίρεσης που είναι απαραίτητος, προκειμένου να ανεξαρτητοποιηθούν οι αλγόριθμοι ομαδοποίησης από το είδος των αντικειμένων που ομαδοποιούνται, καθώς και από τα μέτρα ομοιότητας που χρησιμοποιούνται. Η μόνη υποχρέωση που επιβάλλει η διαπροσωπεία NNOutputer σε κάποια κλάση που την υλοποιεί, είναι να παρέχει η τελευταία τη δυνατότητα εύρεσης των k κοντινότερων γειτόνων για κάθε σημείο του συνόλου δεδομένων. Σε αυτό το σημείο, ξεκαθαρίζουμε ότι τα σημεία που πρόκειται να ομαδοποιηθούν χαρακτηρίζονται το καθένα από έναν μοναδικό ακέραιο. Η αναφορά σε κάποιο σημείο του συνόλου δεδομένων, όποτε αυτό είναι απαραίτητο, γίνεται λοιπόν με χρήση του αντίστοιχου αναγνωριστικού ακεραίου.



Σχήμα 5.1: UML διάγραμμα των βασικών κλάσεων.

- *Clusterer:* Ορίζει ένα σύνολο μεθόδων που πρέπει να δημοσιεύει κάθε κλάση που υλοποιεί κάποιον αλγόριθμο ομαδοποίησης. Οι κλάσεις MCFClusterer, JarvisPatrickClusterer και SNNClusterer είναι υλοποιήσεις αυτής της διαπροσωπείας για τους αλγορίθμους Συμπαγών Ομάδων, Jarvis-Patrick και Πυκνότητας Κοινών Κοντινών Γειτόνων αντίστοιχα. Η αφαίρεση που επιτυγχάνεται μέσω της διαπροσωπείας Clusterer είναι χρήσιμη για την υλοποίηση της Εννοιολογικής Δεικτοδοτησης, που με αυτόν τον τρόπο ανεξαρτητοποιείται από τον χρησιμοποιούμενο αλγόριθμο ομαδοποίησης. Από την άλλη μεριά, η ανεξαρτησία των αλγορίθμων ομαδοποίησης από το ακριβές είδος των αντικειμένων που ομαδοποιούνται, καθώς και από τα χρησιμοποιούμενα μέτρα ομοιότητας οφείλεται στο γεγονός ότι τα αντικείμενα τύπου Clusterer συνδιαλέγονται με αντικείμενα του γενικού τύπου NNOutputer.

- *DistanceOutputer:* Πρόκειται για μια αφηρημένη κλάση, η οποία ορίζει ένα σύνολο μεθόδων που πρέπει να δημοσιεύει οποιαδήποτε κλάση πρόκειται να χρησιμοποιηθεί για τον υπολογισμό αποστάσεων μεταξύ ζευγαριών σημείων. Δεδομένου ότι η δυνατότητα υπολογισμού της απόστασης για οποιοδήποτε ζευγάρι σημείων δίνει τη δυνατότητα για εύρεση των k κοντινότερων γειτόνων για κάθε σημείο, προκύπτει ότι αυτή η αφηρημένη κλάση μπορεί να υλοποιήσει την διαπροσωπεία NNOutputer. Η εύρεση των k κοντινότερων γειτόνων γίνεται με τον αλγόριθμο Εύρεσης των k Πρώτων [5], που υλοποιείται σαν στατική μέθοδος της κλάσης OrderingAlgorithms.

- *TermDocumentMatrix:* Πρόκειται για μια αφηρημένη κλάση που αντιπροσωπεύει κάποιον πίνακα εγγράφων-λέξεων. Ορίζει μεθόδους που δίνουν πληροφορίες για τις διαστάσεις του πίνακα και μεθόδους που αναλαμβάνουν να γεμίσουν τον πίνακα από αρχεία εισόδου τυποποιημένης μορφής. Τέλος, περιέχει μια μέθοδο που επιτρέπει τον καθορισμό του τρόπου χρήσης του πίνακα για τον υπολογισμό της ομοιότητας είτε μεταξύ εγγράφων είτε μεταξύ λέξεων.

- *BinaryTermDocumentMatrix:* Αυτή η αφηρημένη κλάση εξειδικεύει την προηγούμενη κλάση και αντιστοιχεί σε έναν πίνακα εγγράφων-λέξεων με εγγραφές λογικού τύπου. Τα παιδιά αυτής της κλάσης, JaccardOutputer, DiceOutputer και CosineOutputer, παρέχουν μια πλήρως λειτουργική υλοποίηση ενός πίνακα εγγράφων-λέξεων, ορίζοντας επιπλέον πάνω σε αυτόν ένα συγκεκριμένο μέτρο ομοιότητας μεταξύ των εγγράφων ή των λέξεων. Σε μια μελλοντική επέκταση του συστήματος, θα μπορούσαμε να φανταστούμε μια 'αδελφή' κλάση NumericTermDocumentMatrix, που θα αντιστοιχούσε σε έναν πίνακα εγγράφων-λέξεων με αριθμητικές εγγραφές και θα είχε τα δικά της παιδιά για τις εξειδικευμένες συναρτήσεις απόστασης, που μπορεί να ορίσει κανείς για διανύσματα αριθμητικών συνιστωσών.

- *SparseBinaryMatrix:* Είναι η υλοποίηση ενός αραιού πίνακα με λογικές εγγραφές, με τις απαραίτητες μεθόδους για τον χειρισμό των στοιχείων του

πίνακα. Σαν δομικό στοιχείο χρησιμοποιεί την κλάση SparseBinaryVector, η οποία αντιπροσωπεύει ένα αραιό διάνυσμα με λογικές συνιστώσες. Με τη σειρά της, η κλάση SparseBinaryMatrix χρησιμοποιείται σαν δομικό στοιχείο από την κλάση BinaryTermDocumentMatrix.

- *BatchProcessor:* Με αυτήν την κλάση επιλύουμε το πρόβλημα του μεγάλου χρόνου εκτέλεσης που δημιουργεί η τετραγωνική πολυπλοκότητα των αλγορίθμων ομαδοποίησης. Συγκεκριμένα, αυτή η κλάση αναλαμβάνει να αντλήσει από ένα αντικείμενο τύπου DistanceOutputer τους k κοντινότερους γείτονες για κάθε σημείο καθώς και τις αποστάσεις αυτών από το εν λόγω σημείο. Στη συνέχεια, αποθηκεύει αυτήν την πληροφορία σε ένα δυαδικό αρχείο για μελλοντική χρήση. Με αυτόν τον τρόπο επιτυγχάνουμε την ανεξαρτητοποίηση του χρονοβόρου υπολογισμού των k κοντινών γειτόνων για κάθε σημείο από τη διαδικασία της ομαδοποίησης, που απλώς χρησιμοποιεί αυτήν την πληροφορία. Η ανάγνωση των αρχείων των κοντινών γειτόνων πραγματοποιείται και αυτή από αντικείμενα της κλάσης BatchProcessor, με τρόπο μάλιστα που υλοποιεί τη διαπροσωπεία NNOutputer. Έτσι, κάποιος Clusterer μπορεί να χρησιμοποιήσει έναν BatchProcessor προκειμένου να αντλήσει από ένα έτοιμο αρχείο τις πληροφορίες που χρειάζεται για τους k κοντινότερους γείτονες των σημείων που εξετάζει. Αυτό πραγματοποιείται σε χρόνο απείρως μικρότερο από αυτόν που θα χρειαζόταν αν ο υπολογισμός των αποστάσεων και η εύρεση των κοντινών γειτόνων γινόταν σε πραγματικό χρόνο. Συμπερασματικά, ένα τυπικό σενάριο προβλέπει πρώτα την αρκετά χρονοβόρα δημιουργία ενός αρχείου κοντινών γειτόνων και αργότερα τη χρήση αυτού του αρχείου με διάφορους αλγορίθμους ομαδοποίησης και διάφορους συνδυασμούς τιμών για τις ρυθμιστικές παραμέτρους των αλγορίθμων.

- *Reducer:* Είναι μια διαπροσωπεία η οποία ορίζει ένα σύνολο μεθόδων που πρέπει να δημοσιεύει κάθε κλάση που υλοποιεί κάποιον αλγόριθμο μείωσης διάστασης. Στην περίπτωση μας, αυτή η διαπροσωπεία υλοποιείται από δυο κλάσεις, μια για την κάθε μέθοδο μείωσης διάστασης που εξετάζουμε. Η κλάση CIReducer αντιστοιχεί στην μέθοδο της Εννοιολογικής Δεικτοδότησης και κλάση PLSAReducer αντιστοιχεί στην μέθοδο που κάνει χρήση του Μοντέλου των Όψεων. Παρατηρούμε ότι αυτή η δεύτερη κλάση χρησιμοποιεί το έτοιμο πακέτο λογισμικού PennAspect.

- *RIEvaluator:* Υπολογίζει το δείκτη της Βελτίωσης Ανάκλησης για κάποια μέθοδο μείωσης διάστασης. Εκτός από ένα αντικείμενο τύπου Reducer, αυτή η κλάση χρησιμοποιεί και ένα αντικείμενο τύπου NNOutputer, προκειμένου να βρίσκει τους κοντινούς γείτονες και στην περίπτωση της αρχικής πολυδιάστατης αναπαράστασης των δεδομένων.

5.3 Υλοποίηση

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε για την υλοποίηση του λογισμικού ήταν η Java. Συγκεκριμένα, χρησιμοποιήσαμε την έκδοση 1.4.2 του πακέτου ανάπτυξης JSDK (Java Software Development Kit) για περιβάλλον Windows 98, η οποία διανέμεται δωρεάν από την εταιρία Sun. Το πακέτο JSDK περιλαμβάνει διάφορα εργαλεία ανάπτυξης για την Java, καθώς και τη μηχανή εκτέλεσης που διερμηνεύει τον ενδιάμεσο κώδικα που παράγει ο μεταγλωττιστής. Ο μεταγλωττιστής που χρησιμοποιήσαμε δεν ήταν ωστόσο ο javac, που διανέμεται από την Sun μαζί με το JSDK, αλλά ο πολύ πιο γρήγορος jikes, που διανέμεται δωρεάν από την εταιρία IBM. Οι βιβλιοθήκες που συμπεριλαμβάνονται στο JSDK αποδείχτηκαν ιδιαίτερα χρήσιμες για την υλοποίηση, διότι παρέχουν μεταξύ άλλων ορισμένες εξειδικευμένες δομές δεδομένων, καθώς και πολλές ευκολίες για ταχύτατη μεταφορά δεδομένων από και προς αρχεία του σκληρού δίσκου.

Ένα βασικό πρόβλημα που χρειάστηκε να επιλύσουμε ήδη από τα πρώτα στάδια της ανάπτυξης του λογισμικού ήταν ο τρόπος με τον οποίο θα υλοποιούσαμε μια δομή δεδομένων ικανή να αποθηκεύσει τα περιεχόμενα κάποιου πολυδιάστατου και αραιού πίνακα με λογικές εγγραφές. Είναι προφανές, ότι η απλούστατη υλοποίηση με χρήση ενός δισδιάστατου πίνακα με εγγραφές τύπου boolean δεν είναι εφικτή, διότι θα απαιτούσε διαθέσιμη μνήμη RAM της τάξης μερικών GByte. Δεδομένου ότι ένας δισδιάστατος πίνακας δεν είναι άλλο από ένας μονοδιάστατος πίνακας μονοδιάστατων πινάκων, προκύπτει ότι το παραπάνω πρόβλημα ανάγεται στην υλοποίηση μια δομής ικανής να αποθηκεύσει οικονομικά τα περιεχόμενα ενός αραιού μονοδιάστατου πίνακα με λογικές εγγραφές.

Μια δυνατή λύση θα ήταν να χρησιμοποιήσουμε την κλάση BitSet, που συμπεριλαμβάνεται στο JSDK και υλοποιεί ακριβώς αυτό που ζητάμε, αναπαριστώντας την κάθε εγγραφή του πίνακα με ένα και μοναδικό bit, αντί για ένα byte που απαιτεί ο ενσωματωμένος τύπος boolean της Java. Ωστόσο, για τον γρήγορο υπολογισμό των διάφορων μέτρων ομοιότητας μεταξύ των αραιών διανυσμάτων που χειριζόμαστε, είναι απαραίτητο να υπάρχει η δυνατότητα για εύρεση των αληθών συνιστωσών κάποιου διανύσματος χωρίς να χρειάζεται να γίνεται σειριακή αναζήτηση πάνω σε όλο το μήκος του διανύσματος.

Μιας και με τη δομή BitSet δεν μπορεί κανείς να αποφύγει τη σειριακή αναζήτηση, τελικά καταλήγουμε σε μια άλλη λύση, η οποία βασίζεται στους πίνακες κατακερματισμού. Συγκεκριμένα, αποθηκεύουμε ένα αραιό λογικό διάνυσμα κρατώντας μόνο τους δείκτες των αληθών συνιστωσών του με τη μορφή ακεραίων. Προκειμένου μάλιστα να είναι δυνατή η γρήγορη τυχαία προσπέλαση σε οποιαδήποτε θέση του διανύσματος, χρησιμοποιούμε μια συνάρτηση κατακερματισμού ορισμένη πάνω στις πιθανές τιμές των δεικτών. Ειδικότερα, χρησιμοποιούμε την κλάση HashSet του JSDK, η οποία υλοποιεί την αφηρημένη δομή ενός συνόλου με χρήση ενός πίνακα κατακερματισμού. Με αυτόν τον τρόπο, ο έλεγχος για την ύπαρξη ή όχι κάποιου στοιχείου σε ένα σύνολο, στην περίπτωση μας ενός δείκτη που αντιστοιχεί σε αληθή συνιστώσα, έχει σχεδόν σταθερή πολυπλοκότητα. Επιπλέον, η ιδέα αυτής της

υλοποίησης θα μπορούσε εύκολα να γενικευθεί και για την περίπτωση διανυσμάτων με αριθμητικές συνιστώσες. Σε αυτήν την περίπτωση, θα χρησιμοποιούσαμε την κλάση HashMap του JSDK, που θα αντιστοίχιζε τους δείκτες των μη μηδενικών συνιστωσών του διανύσματος με το αριθμητικό περιεχόμενο της αντίστοιχης θέσης.

Μια άλλη απόφαση που ελήφθη κατά τη διάρκεια της υλοποίησης αφορούσε τη μορφή των αρχείων που επρόκειτο να δημιουργούν τα αντικείμενα του τύπου BatchProcesssor. Υπενθυμίζουμε ότι αυτά τα αρχεία χρησιμοποιούνται για την αποθήκευση των k κοντινότερων γειτόνων για κάθε σημείο του συνόλου δεδομένων, καθώς και των αποστάσεων των γειτόνων από το εν λόγω σημείο. Δεδομένου ότι τα αρχεία προορίζονται για εσωτερική χρήση από την εφαρμογή μας και δεν απαιτείται να έχουν κάποια τυποποιημένη μορφή κειμένου, επιλέξαμε να δημιουργούμε δυαδικά αρχεία, που είναι πολύ πιο οικονομικά από τα αρχεία κειμένου ως προς το μέγεθός τους. Ειδικότερα, αποθηκεύουμε για κάθε σημείο του συνόλου δεδομένων k εγγραφές, που η καθεμία αποτελείται από τα 4 byte του ακεραίου που χαρακτηρίζει τον αντίστοιχο κοντινό γείτονα και από τα 4 byte του πραγματικού που αντιστοιχεί στην απόσταση του εν λόγω σημείου από αυτόν τον γείτονα. Για την ανάγνωση και την εγγραφή των δυαδικών αρχείων χρησιμοποιούμε ορισμένες κλάσεις και μεθόδους του πακέτου java.nio, που εντάχθηκε για πρώτη φορά στο JSDK με την έκδοση 1.4. Αυτές οι μέθοδοι είναι υλοποιημένες έτσι, ώστε να αξιοποιούν με αποδοτικό τρόπο το σύστημα διαχείρισης αρχείων του εκάστοτε λειτουργικού συστήματος, με σκοπό την ελαχιστοποίηση του πλήθους των προσβάσεων στον σκληρό δίσκο.

Κεφάλαιο 6

Πειραματικά αποτελέσματα

6.1 Αποτελέσματα ομαδοποίησης

Σε αυτήν την ενότητα παρουσιάζουμε ορισμένα αποτελέσματα που προέκυψαν από την εφαρμογή των αλγορίθμων ομαδοποίησης που υλοποιήσαμε πάνω σε πραγματικά δεδομένα του Παγκοσμίου Ιστού. Η δοκιμή των αλγορίθμων πάνω σε πραγματικά δεδομένα ήταν απαραίτητη για εμάς, προκειμένου να αποκτήσουμε εμπιστοσύνη στην ορθότητα της υλοποίησής μας. Ο λόγος που παρουσιάζουμε ορισμένα από αυτά τα πειραματικά αποτελέσματα και στον αναγνώστη είναι για να σκιαγραφήσουμε μια εικόνα των δυνατοτήτων και των περιορισμών που χαρακτηρίζουν τους συγκεκριμένους αλγορίθμους ομαδοποίησης. Επί τη ευκαιρία, γίνεται μια σύντομη συζήτηση σχετικά με την επίδραση των ρυθμιστικών παραμέτρων των αλγορίθμων πάνω στα αποτελέσματα της ομαδοποίησης, προκειμένου να συμπληρωθεί η θεωρητική περιγραφή που έγινε στο κεφάλαιο 3. Πάντως, δεν πρόκειται σε καμία περίπτωση για μια διεξοδική πειραματική αξιολόγηση των αλγορίθμων ομαδοποίησης , διότι κάτι τέτοιο αφενός δεν εμπίπτει στους στόχους της εργασίας και αφετέρου είναι από μόνο του αρκετά πολύπλοκο θέμα.

Το σύνολο δεδομένων του Παγκοσμίου Ιστού που χρησιμοποιήθηκε για την δοκιμή των αλγορίθμων ομαδοποίησης ήταν το σώμα κειμένων CYTA. Πρόκειται για ένα σύνολο δεδομένων αποτελούμενο από ζευγάρια ιστοσελίδων-λέξεων, όπου το κάθε ζευγάρι αντιστοιχεί στην εμφάνιση κάποιας λέξης μέσα σε κάποια ιστοσελίδα. Το CYTA προέκυψε από μια αυτοματοποιημένη διαδικασία ανάγνωσης κάποιων ιστοσελίδων, που βρέθηκαν να εμφανίζονται μέσα στα αρχεία καταγραφής των εξυπηρετητών ενός Κυπριακού παροχέα υπηρεσιών Internet. Σε αυτό το σημείο, απλώς αναφέρουμε ότι το σύνολο δεδομένων που χρησιμοποιήσαμε εμείς είναι αποτέλεσμα μιας προεπεξεργασίας των ιστοσελίδων, η οποία μεταξύ άλλων περιελάμβανε εξαγωγή των λέξεων, απαλοιφή των πολύ κοινών λέξεων, καθώς και των εντολών HTML. Όσον αφορά το μέγεθός του, το CYTA περιλαμβάνει 12355 ιστοσελίδες και 5086 λέξεις, δηλαδή βάσει του Μοντέλου του Διανυσματικού Χώρου κάθε ιστοσελίδα αναπαρίσταται σαν διάνυσμα 5086 λέξεων και κάθε λέξη αναπαρίσταται σαν διάνυσμα

12355 ιστοσελίδων. Στους πίνακες 6.1 και 6.2 φαίνονται ορισμένες χαρακτηριστικές ομάδες λέξεων και ιστοσελίδων αντίστοιχα, οι οποίες παρατίθενται απλώς για να δώσουν μια εικόνα των αποτελεσμάτων της ομαδοποίησης. Είναι βέβαια αυτονόητο, ότι προέκυψαν και άλλες ομάδες με όχι τόσο ξεκαθαρισμένο θεματικό προσανατολισμό.

Σε αυτό το σημείο είναι χρήσιμο να αναφέρουμε, ότι η μεθοδολογία που ακολουθήσαμε κατά την εκτέλεση των πειραμάτων ομαδοποίησης περιελάμβανε δυο στάδια: πρώτον, τη δημιουργία ενός αρχείου με την πληροφορία για τους k κοντινότερους γείτονες για κάθε στοιχείο του συνόλου δεδομένων και δεύτερον, την χρησιμοποίηση αυτού του αρχείου για την εκτέλεση των πειραμάτων με διαφορετικούς αλγορίθμους και διαφορετικούς συνδυασμούς παραμέτρων. Η δημιουργία του αρχείου των κοντινότερων γειτόνων χρειαζόταν περίπου μιάμιση ώρα για να ολοκληρωθεί, σε PC με επεξεργαστή συχνότητας ρολογιού 300 MHz, μνήμη RAM 380 MB και μηχανή εκτέλεσης για την Java, αυτή που συμπεριλαμβάνεται στην έκδοση 1.4.2 του JSDK. Η εκτέλεση των αλγορίθμων ομαδοποίησης, δεδομένου ότι το αρχείο των κοντινότερων γειτόνων ήταν ήδη έτοιμο, χρειαζόταν από μερικά δευτερόλεπτα έως 15 λεπτά, ανάλογα με το αν εφαρμοζόταν πάνω στις λέξεις ή στις ιστοσελίδες, που είναι πολύ περισσότερες, και ανάλογα με τον χρησιμοποιούμενο αλγόριθμο ομαδοποίησης και τις παραμέτρους του.

abnormal	associate
clinical	disease
disorder	findings
formation	identified
indicated	laboratory
occurs	presenting
subsequent	survival
tissue	typically
adriana	aguilera
alicia	anderson
aniston	ashley
barrymore	britney
butterfly	cameron
carmen	catherine
christina	christy
claudia	mariah

academic	admissions
alumni	courses
employers	faculty
altogether	arguments
augustine	belief
blessed	centuries
christ	christianity
christians	concrete
confessions	conscious
creator	creatures
darkness	destroy
diversity	divine
endless	equally
essence	eternal
existence	expressions
fallen	fathers

aquarius	astrology
capricorn	gemini
pisces	sagittarius
scorpio	taurus
zodiac	libra
thanksgiving	divorce
anniversary	easter
encouragement	father‘s
remembrance	halloween
invitations	mother‘s
occasions	saving
graduation	scenery
sweetest	sympathy
valentine‘s	voyage
football	basketball
racing	season

Πίνακας 6.1: Μερικές χαρακτηριστικές ομάδες λέξεων.

Στους επόμενους πίνακες παραθέτουμε τα αποτελέσματα ορισμένων ενδεικτικών εκτελέσεων των τριών αλγορίθμων που υλοποιήσαμε, όσον αφορά τα στατιστικά χαρακτηριστικά της ομαδοποίησης που επιτυγχάνουν. Με τους πίνακες προσπαθούμε να δώσουμε μια εικόνα του κάθε αλγορίθμου, όσον αφορά τα εγγενή χαρακτηριστικά

artists.mp3s.com/artists/101/christina_aguilera.html
artists.mp3s.com/artists/14/standing_in_the_sun.html
artists.mp3s.com/artists/27/nee.html
artists.mp3s.com/artists/60/metallica2.html
artists.mp3s.com/artists/60/psdie.html
genres.mp3.com/music
genres.mp3.com/music/metal
genres.mp3.com/music/metal/death_metal
genres.mp3.com/music/metal/gothic_metal
genres.mp3.com/music/metal/heavy_metal
genres.mp3.com/newsongs/metal/black_metal
www.mp3.com/newartist
active.macromedia.com/flash2/cabs/swflash.cab
download.macromedia.com/pub/shockwave/cabs/director/swdir8d196.cab
download.macromedia.com/pub/shockwave/cabs/flash/swflash4r28.cab
www.macromedia.com/shockwave/download/triggerpages/default.html
www.macromedia.com/shockwave/download/triggerpages/flash.html
de.news.yahoo.com/notfound.html
future.quarta.ru/icars/subaru/subaru.html
www.challenger.com.au
www.angelfire.com/la/angleo/index.html
www.clarion.co.jp
www.challenger.com.au
www.fiat.com
www.fiat.com/eng/main.htm
www.fiat.com/ita/main.htm
www.gmx.at
www.home.aone.net.au/melbournewankers/chat.html
www.kartshop.ch
www.ucy.ac.cy/faculty/cp/buttonbar2.html

Πίνακας 6.2: Μερικές χαρακτηριστικές ομάδες ιστοσελίδων.

του και την επίδραση των ρυθμιστικών παραμέτρων του. Η κάθε γραμμή ενός πίνακα αντιστοιχεί σε μια διαφορετική εκτέλεση του αντίστοιχου αλγορίθμου ομαδοποίησης, με τιμές των ρυθμιστικών παραμέτρων που δίνονται στην αριστερή στήλη. Για κάθε εκτέλεση δίνεται το πλήθος των ομάδων που τελικά σχηματίζονται, καθώς και η κάλυψη που επιτυγχάνεται. Υπενθυμίζουμε ότι η κάλυψη ορίζεται σαν το ποσοστό των στοιχείων του αρχικού συνόλου δεδομένων που μετά το πέρας του αλγορίθμου βρίσκονται τοποθετημένα σε κάποια ομάδα. Κανένας από τους τρεις αλγορίθμους που υλοποιήσαμε δεν εγγυάται 100% κάλυψη, κάτι που σε ορισμένες περιπτώσεις είναι επιθυμητό. Ο λόγος είναι ότι η ύπαρξη θορύβου σε κάποιο σύνολο δεδομένων είναι ένα πολύ πιθανό ενδεχόμενο και συνεπώς η μη ένταξη των αντίστοιχων σημείων σε κάποια ομάδα είναι κάτι το επιθυμητό, αν φυσικά ο χρησιμοποιούμενος αλγόριθμος είναι σε θέση να απομονώνει αποτελεσματικά τον θόρυβο. Με αυτήν την έννοια, η κάλυψη από μόνη της δεν είναι ένα μέγεθος που μπορεί να χαρακτηρίσει την ποιότητα της ομαδοποίησης, από τη στιγμή μάλιστα που θα μπορούσαμε να χρησιμοποιήσουμε αλγορίθμους που επιβάλλουν πλήρη κάλυψη σε κάθε περίπτωση δεδομένων. Διαισθητικά πάντως, η επίτευξη μεγάλης κάλυψης από αλγορίθμους που δεν την επιβάλλουν, συνεπάγεται την ύπαρξη μικρής ποσότητας θορύβου και συνεπώς την ύπαρξη ομοιογενών ομάδων.

Οι πίνακες 6.3 και 6.4 περιέχουν τα αποτελέσματα της ομαδοποίησης λέξεων και ιστοσελίδων, αντίστοιχα, για μερικές εκτελέσεις του αλγορίθμου Συμπαγών Ομάδων. Όπως αναφέρθηκε και στην ενότητα 3.2, το χαρακτηριστικό του αλγορίθμου έγκειται στην αδυναμία του να ανακαλύπτει αποτελεσματικά ομάδες διαφορετικών πυκνοτήτων. Στην περίπτωση της ομαδοποίησης των λέξεων, αυτό φαίνεται ξεκάθαρα από το γεγονός ότι το πλήθος των ομάδων που σχηματίζονται είναι πολύ μικρό. Συγκεκριμένα, η ύπαρξη κάποιας πυκνής ομάδας λέξεων επιβάλλει από την αρχή ένα κάτω φράγμα στην συμπάγεια των ομάδων που πρόκειται να σχηματιστούν στη συνέχεια. Το αποτέλεσμα είναι να σχηματίζονται τελικά λίγες ομάδες, που είναι τουλάχιστον τόσο συμπαγής όσο αυτή που σχηματίστηκε αρχικά. Επιπλέον, από το γεγονός ότι κάλυψη είναι πρακτικά ανεξάρτητη από την ρυθμιστική παράμετρο k, συμπεραίνουμε ότι αυτή η πυκνή ομάδα είναι και πολυπληθής, διότι καταφέρνει να επιβάλλει τον περιορισμό της ακόμα και για μεγάλες τιμές του k. Στην περίπτωση της ομαδοποίησης των ιστοσελίδων, παρατηρούμε ότι επιτυγχάνεται πολύ μεγαλύτερη κάλυψη απ' ό,τι στην περίπτωση των λέξεων. Από αυτό συμπεραίνουμε, ότι στον διανυσματικό χώρο των ιστοσελίδων υπάρχουν αρκετές μεγάλες και εξίσου συμπαγείς ομάδες. Έτσι, ο αλγόριθμος Συμπαγών Ομάδων ανακαλύπτει χωρίς πρόβλημα τις πιο συμπαγείς από αυτές, αγνοώντας και πάλι βέβαια τις λιγότερο συμπαγείς ομάδες. Το τελευταίο αποδεικνύεται από το γεγονός ότι και σε αυτήν την περίπτωση η κάλυψη που επιτυγχάνεται είναι πρακτικά ανεξάρτητη από την παράμετρο k.

Οι πίνακες 6.5 και 6.6 περιέχουν τα αποτελέσματα της ομαδοποίησης λέξεων και ιστοσελίδων, αντίστοιχα, για μερικές εκτελέσεις του αλγορίθμου Jarvis-Patrick. Παρατηρούμε ότι στην περίπτωση της ομαδοποίησης των λέξεων, όπου ο αλγόριθ-

Παράμετροι		
k	Πλήθος ομάδων	Κάλυψη (%)
10	8	0.708
30	9	1.652
40	9	1.868
50	9	2.143
60	9	2.340
80	9	2.812
100	9	3.421

Πίνακας 6.3: Ομαδοποίηση λέξεων με τον αλγόριθμο Συμπαγών Ομάδων.

Παράμετροι		
k	Πλήθος ομάδων	Κάλυψη (%)
10	1139	35.686
30	1016	37.871
40	995	38.130
50	983	38.235
60	977	38.179
80	1030	39.587
100	1016	39.619

Πίνακας 6.4: Ομαδοποίηση ιστοσελίδων με τον αλγόριθμο Συμπαγών Ομάδων.

μος Συμπαγών Ομάδων αντιμετωπίζει πρόβλημα, ο αλγόριθμος Jarvis-Patrick επιτυγχάνει μεγαλύτερη κάλυψη και σχηματίζει περισσότερες ομάδες. Αυτό σημαίνει ότι ξεπερνά αποτελεσματικά το πρόβλημα των διαφορετικών πυκνοτήτων, που είναι περιοριστικός παράγοντας για τον αλγόριθμο Συμπαγών Ομάδων. Από την άλλη μεριά, στην ομαδοποίηση των ιστοσελίδων παρατηρούμε, ότι για περιπτώσεις εκτέλεσης, όπου το πλήθος των ομάδων δεν διαφέρει πολύ από αυτό στο οποίο καταλήγει και ο αλγόριθμος Συμπαγών Ομάδων, ο αλγόριθμος Jarvis-Patrick εμφανίζεται να πετυχαίνει μεγαλύτερη κάλυψη. Διαισθητικά αυτό σημαίνει, ότι οι ομάδες που σχηματίζονται είναι πιο ανομοιογενείς. Όσον αφορά τις ρυθμιστικές παραμέτρους του αλγορίθμου παρατηρούμε ότι για σταθερό λόγο των παραμέτρων k και m, η αύξηση του k οδηγεί στον σχηματισμό πιο ανομοιογενών ομάδων. Αυτό το συμπεραίνουμε από το ότι το πλήθος των ομάδων μειώνεται, ενώ την ίδια στιγμή η κάλυψη αυξάνεται. Δηλαδή, όταν το k ξεπεράσει κάποια τιμή, ακόμα και το μέτρο ομοιότητας των k Κοντινών Γειτόνων ενδεχομένως χάνει την αξιοπιστία του, διότι σε όλες τις λίστες k κοντινότερων γειτόνων έχουν παρεισφρήσει πάρα πολλά άσχετα σημεία.

Παράμετροι			
k	m	Πλήθος ομάδων	Κάλυψη (%)
10	6	184	16.614
15	5	282	44.534
15	10	143	16.477
20	10	194	32.088
50	25	139	39.717
70	35	119	41.231
100	50	100	43.728

Πίνακας 6.5: Ομαδοποίηση λέξεων με τον αλγόριθμο Jarvis-Patrick.

Παράμετροι			
k	m	Πλήθος ομάδων	Κάλυψη (%)
10	6	1164	38.575
15	5	1149	62.558
15	10	1166	41.109
20	10	1131	56.058
50	25	957	64.435
70	35	923	66.394
100	50	901	69.057

Πίνακας 6.6: Ομαδοποίηση ιστοσελίδων με τον αλγόριθμο Jarvis-Patrick.

Οι πίνακες 6.7 και 6.8 περιέχουν τα αποτελέσματα της ομαδοποίησης λέξεων

και ιστοσελίδων, αντίστοιχα, για μερικές εκτελέσεις του αλγορίθμου Πυκνότητας Κοινών Κοντινών Γειτόνων. Στην περίπτωση της ομαδοποίησης των λέξεων, παρατηρούμε ότι ο αλγόριθμος ΠΚΚΓ καταλήγει στον σχηματισμό περισσότερων ομάδων από τον αλγόριθμο Jarvis-Patrick για τις περιπτώσεις, όπου η κάλυψη είναι περίπου ίση και για τους δυο. Αυτό σημαίνει, ότι στην περίπτωση του ΠΚΚΓ υπάρχουν ξεχωριστές ομάδες που στην περίπτωση του Jarvis-Patrick έχουν συγχωνευθεί. Από την άλλη μεριά, στην περίπτωση της ομαδοποίησης των ιστοσελίδων μοιάζει να συμβαίνει το αντίθετο, δηλαδή για ίδια κάλυψη ο αλγόριθμος ΠΚΚΓ σχηματίζει λιγότερες ομάδες από τον Jarvis-Patrick. Η αντίθεση αυτή οφείλεται μάλλον στο διαφορετικό σχήμα των ομάδων που υπάρχουν μέσα στους διανυσματικούς χώρους των λέξεων και των ιστοσελίδων.

Παράμετροι				
k	m	r	Πλήθος ομάδων	Κάλυψη (%)
20	10	17	206	17.853
30	15	15	222	20.527
50	20	20	284	32.874
50	25	20	221	20.802
50	25	30	181	17.361
70	25	30	323	42.312
100	40	30	307	37.200

Πίνακας 6.7: Ομαδοποίηση λέξεων με τον αλγόριθμο ΠΚΚΓ.

Παράμετροι				
k	m	r	Πλήθος ομάδων	Κάλυψη (%)
20	10	17	646	36.754
30	15	15	647	39.838
50	20	20	575	50.433
50	25	20	593	38.535
50	25	30	594	36.026
70	25	30	510	54.844
100	40	30	575	51.906

Πίνακας 6.8: Ομαδοποίηση ιστοσελίδων με τον αλγόριθμο ΠΚΚΓ.

6.2 Αποτελέσματα ΠΑΚΣ

Στην παρούσα ενότητα παρουσιάζουμε ορισμένα ενδεικτικά αποτελέσματα που προέκυψαν από την εκπαίδευση του Μοντέλου των Όψεων πάνω στο σύνολο ιστοσελίδων-

λέξεων CYTA. Συγκεκριμένα, πραγματοποιήσαμε την εκπαίδευση του πιθανοτικού μοντέλου ορίζοντας το πλήθος των κρυμμένων μεταβλητών z_k να είναι ίσο με 10. Σε αυτό το σημείο, υπενθυμίζουμε ότι οι κρυμμένες μεταβλητές αντιστοιχούν διαισθητικά σε θεματικούς πυρήνες που εμφανίζονται στο σύνολο των δεδομένων. Όσον αφορά τις λεπτομέρειες της εκπαίδευσης, αναφέρουμε ότι χρησιμοποιήσαμε το 90% του συνολικού πλήθους των ζευγαριών ιστοσελίδων-λέξεων για την εκπαίδευση του μοντέλου και το υπόλοιπο 10% για τον έλεγχο της επίδοσης του μοντέλου πάνω σε νέα δεδομένα, όπως απαιτεί ο αλγόριθμος. Επίσης, ορίσαμε ένα μέγιστο πλήθος επαναλήψεων ίσο με 100, διότι έχει πειραματικά διαπιστωθεί ότι γύρω στις 50 επαναλήψεις είναι συνήθως αρκετές για να επιτευχθεί η σύγκλιση της αναμενόμενης πιθανοφάνειας σε κάποιο τοπικό μέγιστο [14]. Τέλος, αναφέρουμε ότι η διαδικασία εκπαίδευσης ολοκληρώθηκε μέσα σε έξι ώρες, σε PC με επεξεργαστή των 300MHz.

Στους πίνακες που ακολουθούν παραθέτουμε τις δέκα πιο πιθανές λέξεις και τις δέκα πιο πιθανές ιστοσελίδες για τρεις από τις δέκα συνολικά κρυμμένες μεταβλητές, θέλοντας με αυτόν τον τρόπο να σκιαγραφήσουμε τον θεματικό πυρήνα που αντιπροσωπεύει η καθεμία από αυτές. Δηλαδή, για την κρυμμένη μεταβλητή z δίνουμε τις δέκα ιστοσελίδες με τις μεγαλύτερες πιθανότητες $P(d|z)$, καθώς και τις δέκα λέξεις με τις μεγαλύτερες πιθανότητες $P(w|z)$. Τα περιεχόμενα των πινάκων πραγματικά φανερώνουν την ύπαρξη μιας συνοχής γύρω από ένα κοινό για την κάθε περίπτωση θέμα.

Λέξεις	Ιστοσελίδες
research	www.watchman.org/cat95.htm
international	www.fas.org/sgp/othergov/naracia.html
national	www.cyna.org.cy/monthly.htm
public	www.csis.org/pubs/pubsecur.html
through	www.ibb.be/docs2/codetax.html
history	www.hri.org/news/greek/ana/2000/00-04-04.ana.html
including	www.mp3.com/news/daily.html
university	www.csis.org/polmil
european	www.scrc.umanitoba.ca/SCRC/profiles.html
resources	home.aafp.org/afp/991101ap/1969.html

Πίνακας 6.9: Όψη με πιθανότητα $P(z) = 0.1894$.

Λέξεις	Ιστοσελίδες
software	www.memepool.com/Subject/Computing
products	www.mp3.com/news/daily.html
windows	www.webattack.com/shareware/security/swfirewall.shtml
download	www.win98central.com
services	www.fileflash.com
system	www.memepool.com/Subject/Web
request	www.reactorcritical.com
computer	www.aberdeeninc.com
product	xchat.org/changelog.txt
microsoft	www.hri.org/fonts/w95

Πίνακας 6.10: Όψη με πιθανότητα $P(z) = 0.1189$.

Λέξεις	Ιστοσελίδες
travel	www.windowoncyprus.com/friends-of-the-cyprus-donkey.htm
business	www.windowoncyprus.com/ayianapa1.htm
sports	www.windowoncyprus.com/junior-school-in-nicosia-cyprus.htm
people	www.windowoncyprus.com/bunjee.htm
health	www.windowoncyprus.com/links.htm
shopping	www.windowoncyprus.com/safari-tours-excursions-cyprus-trips.htm
directory	www.sidereal-horoscopes.com/scorpio.htm
entertainment	www.windowoncyprus.com/weekly-predictions.htm
policy	www.windowoncyprus.com/entertai.htm
education	www.windowoncyprus.com/ayia-.htm

Πίνακας 6.11: Όψη με πιθανότητα $P(z) = 0.0799$.

6.3 Αξιολόγηση της μείωσης διάστασης

Η παρούσα ενότητα αποτελεί κατά κάποιον τρόπο το τελευταίο στάδιο πριν την εκπλήρωση του αρχικού στόχου της εργασίας. Έχοντας υλοποιήσει τις δυο μεθόδους μείωσης διάστασης που περιγράφηκαν αναλυτικά στα προηγούμενα κεφάλαια, σε αυτήν την ενότητα πραγματοποιούμε μια αξιολόγησή τους πάνω σε πραγματικά δεδομένα του Παγκόσμιου Ιστού. Όπως αναφέρθηκε και στην ενότητα 2.4, τα κίνητρα πίσω από τη διαδικασία μείωσης διάστασης είναι δυο: πρώτον, η εξοικονόμηση υπολογιστικών πόρων και δεύτερον, η ανακάλυψη συσχετίσεων που δεν είναι ορατές στην πολυδιάστατη αναπαράσταση των δεδομένων. Για μεγαλύτερη σαφήνεια ας δούμε την αξία του καθενός από αυτά τα κίνητρα σε μια συγκεκριμένη εφαρμογή, όπως είναι η ανάκληση των N πιο σχετικών εγγράφων μέσα από ένα σύνολο εγγράφων, δοθέντος κάποιου ερωτήματος κειμένου. Σε αυτήν την περίπτωση, η μείωση διάστασης της διανυσματική αναπαράσταση των εγγράφων θα προσκόμιζε δυο οφέλη. Το πρώτο θα ήταν η μεγαλύτερη ταχύτητα με την οποία θα γινόταν η εύρεση των N πιο σχετικών εγγράφων και το δεύτερο θα ήταν η καλύτερη ποιότητα των αποτελεσμάτων, δηλαδή μεγαλύτερο ποσοστό πραγματικά σχετικών εγγράφων επί του συνόλου των ανακληθέντων εγγράφων.

Στην αξιολόγηση που πραγματοποιούμε σ' αυτήν την ενότητα στοχεύουμε κυρίως στην ποσοτική μέτρηση των επιδόσεων της μείωσης διάστασης σαν διαδικασίας ανακάλυψης κρυμμένων συσχετίσεων. Το σκέλος της μείωσης διάστασης που αφορά την εξοικονόμηση υπολογιστικών πόρων δεν το αξιολογούμε ξεχωριστά, διότι η αξία του είναι μάλλον αυτονόητη. Για τον σκοπό της αξιολόγησης χρησιμοποιούμε έναν δείκτη ποιότητας που ονομάζεται Βελτίωση Ανάκλησης και προτείνεται στο [16]. Προκειμένου να χρησιμοποιήσει κανείς αυτόν τον δείκτη, χρειάζεται ένα σύνολο εγγράφων, το οποίο είναι εκ των προτέρων χωρισμένο από ειδικούς σε γνωστές κατηγορίες. Αυτή η πληροφορία λειτουργεί σαν ένα πλαίσιο σταθερής αλήθειας που στη συνέχεια χρησιμοποιείται, προκειμένου να μετρηθεί η ακρίβεια της ανάκλησης σχετικών εγγράφων. Συγκεκριμένα, η Βελτίωση Ανάκλησης υπολογίζεται ως εξής:

1. Για κάθε έγγραφο d_i βρες τα N πιο σχετικά έγγραφα, όπως αυτά προκύπτουν στην περίπτωση της πολυδιάστατης διανυσματικής αναπαράστασης, με χρήση κάποιου κατά βούληση επιλεγμένου μέτρου ομοιότητας. Έστω n_i το πλήθος των ανακληθέντων εγγράφων που ανήκουν στην ίδια κατηγορία με το d_i.

2. Άθροισε τα επιμέρους αποτελέσματα του προηγούμενου βήματος και υπολόγισε την ποσότητα $R_{orig} = \sum_{i=1}^{N} n_i$, όπου N είναι το πλήθος των εγγράφων. Η ποσότητα R_{orig} είναι ένα μέτρο για την ακρίβεια της ανάκλησης, όταν χρησιμοποιείται η πολυδιάστατη διανυσματική αναπαράσταση των εγγράφων.

3. Επανάλαβε τα βήματα 1 και 2 για την περίπτωση της ολιγοδιάστατης αναπαράστασης των εγγράφων, όπως αυτή έχει προκύψει με την εφαρμογή κάποιας διαδικασίας μείωσης διάστασης. Υπολόγισε την ποσότητα R_{red}.

4. Η Βελτίωση Ανάκλησης ορίζεται ώς: $BA = \frac{R_{red} - R_{orig}}{R_{orig}}$

Η Βελτίωση Ανάκλησης διαισθητικά εκφράζει το κατά πόσο η μείωση διάστασης φέρνει πιο κοντά τα έγγραφα που είναι στην πραγματικότητα σχετικά μεταξύ τους. Ειδικότερα, όπως μαρτυρά και ο ίδιος ο όρος, η Βελτίωση Ανάκλησης εκφράζει τη βελτίωση που επιτυγχάνουμε με τη μείωση διάστασης στην ακρίβεια ανάκλησης σχετικών εγγράφων, όταν σαν ερωτήματα χρησιμοποιούνται τα ίδια τα έγγραφα. Η ακρίβεια της ανάκλησης εκφράζεται από το ποσοστό των ανακληθέντων εγγράφων που ανήκουν στην ίδια κατηγορία με το έγγραφο-ερώτημα και συνεπώς είναι στην πραγματικότητα σχετικά με αυτό.

Δεδομένου ότι για την μέτρηση της Βελτίωσης Ανάκλησης απαιτείται ένα σύνολο εκ των προτέρων κατηγοριοποιημένων εγγράφων, δεν μπορούσαμε να χρησιμοποιήσουμε το CYTA, που παρουσιάστηκε στις προηγούμενες ενότητες. Έτσι, χρησιμοποιήσαμε ένα άλλο σώμα κειμένων με την ονομασία LINGSPAM. Αυτό αποτελείται συνολικά από 2893 κείμενα και 27088 λέξεις. Από το σύνολο των κειμένων τα 481 είναι spam μηνύματα ηλεκτρονικού ταχυδρομείου και τα υπόλοιπα 2412 είναι μηνύματα που εμφανίστηκαν σε μια ομάδα νέων σχετική με τη γλωσσολογία, δηλαδή τα κείμενα είναι εκ των προτέρων διαχωρισμένα σε δυο κατηγορίες. Όπως και στην περίπτωση του CYTA, τα έγγραφα αναπαρίστανται σαν λογικά διανύσματα λέξεων. Το LINGSPAM ήταν το σύνολο δεδομένων πάνω στο οποίο εφαρμόσαμε τις δυο μεθόδους μείωσης διάστασης που αναπτύξαμε, προκειμένου τελικά να είμαστε σε θέση να αξιολογήσουμε τις επιδόσεις τους με τον υπολογισμό της Βελτίωσης Ανάκλησης. Τα πειραματικά αποτελέσματα παρουσιάζονται στους πίνακες που ακολουθούν.

Ο πίνακας 6.12 προορίζεται για την αξιολόγηση της μεθόδου μείωσης διάστασης που βασίζεται στην εκπαίδευση του Μοντέλου των Όψεων. Ειδικότερα, περιέχει τις τιμές της Βελτίωσης Ανάκλησης, εκφρασμένες σαν επί τοις εκατό ποσοστό, για διάφορες τιμές του πλήθους των όψεων του μοντέλου, καθώς και για διάφορες τιμές του πλήθους των σχετικών εγγράφων που ανακαλούνται. Υπενθυμίζουμε ότι το πλήθος των ανακαλούμενων εγγράφων υπεισέρχεται στον υπολογισμό των συντελεστών R_{orig} και R_{red} και κατ' επέκταση στον υπολογισμό της Βελτίωσης Ανάκλησης. Κατ' αρχήν παρατηρούμε ότι το Μοντέλο των Όψεων με 6 κρυμμένες μεταβλητές επιτυγχάνει τις καλύτερες επιδόσεις. Δεδομένου ότι οι κρυμμένες μεταβλητές αντιστοιχούν σε θεματικούς πυρήνες, το αποτέλεσμα αυτό θα μπορούσε να σημαίνει ότι στο σύνολο των κειμένων εμφανίζονται κατά προσέγγιση έξι διαφορετικά θέματα. Λιγότερες από έξι κρυμμένες μεταβλητές αδυνατούν να κατατάξουν αποτελεσματικά τα κείμενα σε θεματικούς πυρήνες, ενώ περισσότερες από έξι κρυμμένες μεταβλητές επιβάλλουν μια υπερβολικά εξειδικευμένη διάκριση των κειμένων.

Όσον αφορά την επίδραση της παραμέτρου N, παρατηρούμε ότι η αύξηση της τιμής της οδηγεί σε αύξηση της Βελτίωσης Ανάκλησης, ακόμα και στις περιπτώσεις που η συνολική επίδοση του μοντέλου δεν είναι καλή. Αυτό το αποτέλεσμα οφείλεται στο γεγονός, ότι για τις μεγάλες τιμές του N, η πολυδιάστατη αναπαράσταση των εγγράφων στερεί από το χρησιμοποιούμενο μέτρο ομοιότητας την αξιοπιστία του, όσον αφορά την εύρεση σχετικών εγγράφων. Κατ' επέκταση, το ποσοστό των

πραγματικά σχετικών εγγράφων στο σύνολο των N ανακληθέντων μειώνεται, προκαλώντας τη μείωση της ποσότητας R_{orig}.

Βέβαια, για $N = 10$ η πολυδιάστατη αναπαράσταση των εγγράφων αποδίδει καλύτερα, όπως φαίνεται από την πρώτη γραμμή του πίνακα. Αυτό οφείλεται στο γεγονός, ότι ο καθορισμός της ομοιότητας των εγγράφων βάσει των κοινών τους λέξεων είναι επαρκής για την περίπτωση που το ζητούμενο είναι η εύρεση των πρώτων πολύ κοντινών γειτόνων. Ωστόσο, όταν το απαιτούμενο πλήθος σχετικών εγγράφων αυξάνεται η μειωμένης διάστασης αναπαράσταση, με τις κρυμμένες συσχετίσεις που ενσωματώνει, αποδεικνύεται πιο αποτελεσματική. Όταν μάλιστα η βελτίωση που επιφέρει η μείωση διάστασης εμφανίζεται και για σχετικά μικρές τιμές του N, όπως συμβαίνει στην περίπτωση μας για το μοντέλο των έξι όψεων, η αξία της μείωσης διάστασης σαν διαδικασίας εξόρυξης γνώσης γίνεται μεγαλύτερη.

	Διάσταση του Μοντέλου Όψεων					
N	4	6	8	10	16	32
10	-1.077	-0.957	-0.972	-1.530	-1.889	-1.903
20	-0.282	0.023	0.062	-0.996	-1.798	-1.758
30	0.330	0.749	0.583	-0.696	-2.088	-1.742
40	0.874	1.295	1.125	-0.399	-2.166	-1.676
50	1.372	1.806	1.639	-0.160	-2.205	-1.606
60	1.818	2.180	2.049	0.098	-2.138	-1.514
100	2.924	3.355	3.065	0.661	-2.258	-0.571
150	3.525	4.207	3.903	1.041	-2.119	1.018
200	4.334	5.184	5.003	1.976	-0.872	2.299
250	5.316	6.345	6.193	3.485	0.980	3.668
300	6.019	7.501	7.371	4.534	2.266	4.599

Πίνακας 6.12: Βελτίωση Ανάκλησης (%) για μοντέλα διαφόρων διαστάσεων.

Στους πίνακες 6.13, 6.14 και 6.15 παρουσιάζονται οι μετρήσεις της Βελτίωσης Ανάκλησης για τη μέθοδο της Εννοιολογικής Δεικτοδότησης, όταν αυτή χρησιμοποιεί τους αλγορίθμους Συμπαγών Ομάδων, Jarvis-Patrick και Πυκνότητας Κοινών Κοντινών Γειτόνων, αντίστοιχα. Ειδικότερα, παρουσιάζονται οι τιμές της Βελτίωσης Ανάκλησης, εκφρασμένες σαν επί τοις εκατό ποσοστό, για διάφορες τιμές των ρυθμιστικών παραμέτρων των αλγορίθμων, καθώς και για δυο τιμές του πλήθους N των ανακαλούμενων εγγράφων, μια σχετικά μικρή και μια σχετικά μεγάλη. Σε αυτό το σημείο τονίζουμε, ότι δεν έγινε καμία προσπάθεια για την εύρεση ενός βέλτιστου συνδυασμού των παραμέτρων, διότι κάτι τέτοιο θα αποτελούσε από μόνο του ένα ξεχωριστό θέμα μελέτης. Πάντως, η γενική τάση των αποτελεσμάτων φαίνεται καθαρά και για τους συνδυασμούς ρυθμιστικών παραμέτρων που επιλέχθηκαν από εμάς και εμφανίζονται στους πίνακες. Ένα σχόλιο που ισχύει και για τις τρεις περιπτώσεις είναι, ότι η επιλογή μεγάλης τιμής για το N καθιστά τη μειωμένης διά-

στασης αναπαράσταση των εγγράφων αποτελεσματικότερη από την πολυδιάστατη, όσον αφορά την ακρίβεια της ανάκλησης σχετικών εγγράφων. Αυτό ισχύει ακόμα και στις περιπτώσεις που η πολυδιάστατη αναπαράσταση είναι αποτελεσματικότερη για μικρές τιμές του N και ο λόγος που συμβαίνει αυτό εξηγήθηκε παραπάνω.

Ειδικότερα, στην περίπτωση της χρήσης του αλγορίθμου Συμπαγών Ομάδων από την Εννοιολογική Δεικτοδότηση παρατηρούμε χαμηλές επιδόσεις για $N = 50$ στην πλειοψηφία των επιλογών της ρυθμιστικής παραμέτρου k. Αυτό οφείλεται στην κακή επίδοση του ίδιου του αλγορίθμου ομαδοποίησης, τα αίτια της οποίας αναλύονται λεπτομερέστερα στις ενότητες 3.2 και 6.1. Αντίθετα, οι άλλοι δυο αλγόριθμοι ομαδοποίησης αποδίδουν ικανοποιητικά αποτελέσματα σχεδόν σε όλες τις περιπτώσεις επιλογής των ρυθμιστικών παραμέτρων τους. Ο αλγόριθμος Jarvis-Patrick αποδίδει άσχημα σε μια μόνο περίπτωση, όπου η μειωμένη διάσταση, που είναι ίση με το πλήθος των σχηματισμένων ομάδων, είναι μικρή και η αντίστοιχη διανυσματική αναπαράσταση των εγγράφων μάλλον ανεπαρκής όσον αφορά τη διακριτική ακρίβεια που παρέχει στα μέτρα ομοιότητας. Διαισθητικά πάντως, αναμένουμε να υπάρχει ένα βέλτιστο πλήθος συνιστωσών, αρκετά μεγάλο για να επιτρέπει την αποτελεσματική διάκριση μεταξύ των στοιχείων του συνόλου δεδομένων, αλλά και αρκετά μικρό, ώστε να παρέχει το πλεονέκτημα της συμπαγούς αναπαράστασης, που είναι εξάλλου και ο βασικός σκοπός της μείωσης διάστασης. Σε μια μοναδική επίσης περίπτωση εκτέλεσης αποδίδει άσχημα και ο αλγόριθμος ΠΚΚΓ, όταν η τιμή της παραμέτρου r είναι μικρή. Σε αυτήν την περίπτωση μάλλον πρόκειται για κακή ποιότητα των σχηματισμένων ομάδων, όσον αφορά την ομοιογένειά τους.

Παράμετροι			
k	Μειωμένη διάσταση	N=50	N=250
20	42	-0.565	0.966
40	43	-0.914	1.016
80	41	-1.134	0.193
100	40	-1.206	-0.064
120	36	-1.133	0.203
180	33	-1.751	-0.974
200	35	-1.113	0.137
260	37	0.314	1.938
300	36	-0.125	2.045
380	40	1.234	2.392
400	34	0.889	2.272
450	34	0.913	2.291

Πίνακας 6.13: Βελτίωση Ανάκλησης (%) για ομαδοποίηση με τον αλγόριθμο Συμπαγών Ομάδων.

Παράμετροι				
k	m	Μειωμένη διάσταση	N=50	N=250
20	10	221	1.277	2.257
25	10	204	1.508	2.284
40	15	159	1.623	1.780
40	20	192	1.161	1.428
40	30	153	1.015	2.496
50	20	147	0.157	0.527
50	25	183	0.936	1.560
50	30	193	1.447	1.860
70	30	137	0.833	2.059
70	40	176	1.409	1.950
100	40	190	0.870	1.496
100	50	141	0.565	1.714
100	60	174	1.254	1.898
150	50	46	-4.740	-2.930
150	75	114	0.471	1.256
200	100	106	1.283	2.713
250	125	92	0.343	1.406
300	150	87	-0.535	0.509

Πίνακας 6.14: Βελτίωση Ανάκλησης (%) για ομαδοποίηση με τον αλγόριθμο Jarvis-Patrick.

Παράμετροι					
k	m	r	Μειωμένη διάσταση	N=50	N=250
30	10	50	150	1.759	2.958
30	15	20	87	0.953	2.455
50	20	10	224	1.313	2.170
50	20	30	174	1.081	2.248
50	20	50	145	1.124	1.550
100	40	10	208	-9.933	-3.598
100	40	30	224	0.243	0.754
100	40	50	173	0.618	1.456
200	90	10	176	0.887	2.110
200	90	20	218	0.350	0.862
200	90	30	206	1.823	3.309
200	90	50	199	1.350	2.298
200	90	70	190	1.030	2.171

Πίνακας 6.15: Βελτίωση Ανάκλησης (%) για ομαδοποίηση με τον αλγόριθμο ΠΚΚΓ.

Κεφάλαιο 7

Επίλογος

7.1 Σύνοψη

Επιχειρώντας να κάνουμε έναν τελικό απολογισμό, μπορούμε να πούμε ότι ο αρχικός στόχος της εργασίας εκπληρώθηκε με επιτυχία. Σε πρώτη φάση καταφέραμε να υλοποιήσουμε τις δυο τεχνικές μείωσης διάστασης που είχαμε αρχικά επιλέξει: την Εννοιολογική Δεικτοδοτηση, που στηρίζεται στη διαδικασία της ομαδοποίησης, και την Πιθανοτική Μείωση Διάστασης, που στηρίζεται στην Πιθανοτική Ανάλυση Κρυμμένης Σημασιολογίας. Στη συνέχεια, αξιολογήσαμε τις επιδόσεις των δυο αυτών τεχνικών πάνω σε πραγματικά δεδομένα του Παγκοσμίου Ιστού, λαμβάνοντας μάλιστα αποτελέσματα που συμφωνούν με ό,τι αναμέναμε και θεωρητικά.

Στην πορεία προς την επίτευξη του βασικού μας στόχου ασχοληθήκαμε με τη διαδικασία της ομαδοποίησης, την οποία και χρησιμοποιήσαμε σαν ανεξάρτητη διαδικασία στο ευρύτερο πλαίσιο της Εννοιολογικής Δεικτοδότησης. Ειδικότερα όσον αφορά την ομαδοποίηση, υλοποιήσαμε ένα πλαίσιο εργασίας λογισμικού, βασισμένο στον κατάλληλο σχεδιασμό και τις κατάλληλες δομές δεδομένων, ώστε να μπορεί να χρησιμοποιηθεί για την ανάπτυξη μιας ολόκληρης σειράς αλγορίθμων ομαδοποίησης. Επιπλέον, στηριζόμενοι σε αυτήν την υποδομή αναπτύξαμε τρεις συγκεκριμένους αλγορίθμους, των οποίων τη συμπεριφορά δοκιμάσαμε πάνω σε πραγματικά δεδομένα του Παγκοσμίου Ιστού.

Επίσης, στην πορεία προς τον τελικό στόχο χρειάστηκε να χρησιμοποιήσουμε τη μεθοδολογία της Πιθανοτικής Ανάλυσης Κρυμμένης Σημασιολογίας και συγκεκριμένα τον αλγόριθμο βέλτιστης εκτίμησης των παραμέτρων του Μοντέλου των Όψεων. Αν και δεν ήταν απαραίτητο να υλοποιήσουμε τον εν λόγω αλγόριθμο από το μηδέν, εντούτοις χρειάστηκε να τροποποιήσουμε το πακέτο λογισμικού PennAspect, προκειμένου να το προσαρμόσουμε στις δικές μας δομές δεδομένων και να το βελτιστοποιήσουμε ως προς την ταχύτητα για τον δικό μας τύπο δεδομένων. Προκειμένου να κατανοήσουμε καλύτερα τη συμπεριφορά του αλγορίθμου, τον δοκιμάσαμε πάνω σε πραγματικά δεδομένα του Παγκοσμίου Ιστού, λαμβάνοντας λογικά αποτελέσματα.

Όσον αφορά την αξιολόγηση των δυο συγκεκριμένων τεχνικών μείωσης διάστασης, που ήταν και ο βασικός στόχος της εργασίας, αυτή έγινε βάσει του δείκτη που ονομάζεται Βελτίωση Ανάκλησης. Θεωρώντας ότι ο δείκτης αυτός είναι ένα αξιόπιστο μέτρο ποιότητας, προκύπτει από τα αποτελέσματα ότι τόσο η τεχνική της Εννοιολογικής Δεικτοδότησης, όσο και η τεχνική της Πιθανοτικής Μείωσης Διάστασης είναι αποτελεσματικές μέθοδοι μείωσης διάστασης. Αυτό σημαίνει ότι εκπληρώνουν δυο σκοπούς: πρώτον, παρέχουν μια πιο οικονομική αναπαράσταση των δεδομένων από άποψη υπολογιστικών πόρων και δεύτερον, ανακαλύπτουν συσχετίσεις μεταξύ των δεδομένων που δεν ήταν φανερές στην πολυδιάστατη αναπαράστασή τους.

7.2 Μελλοντικές επεκτάσεις

Ως γνωστόν, ο χρόνος μιας διπλωματικής εργασίας είναι περιορισμένος και συνεπώς όχι αρκετός για να υλοποιηθούν και να δοκιμαστούν στην πράξη όλα όσα φαίνονται ενδιαφέροντα. Ορισμένες κατευθύνσεις προς τις οποίες θα μπορούσε να επεκταθεί το αντικείμενο της παρούσας εργασίας δίνονται παρακάτω:

- Υποστήριξη πινάκων δυαδικών συσχετίσεων με ακέραιες ή πραγματικές εγγραφές. Με αυτήν την προσθήκη, οι αλγόριθμοι ομαδοποίησης, ο αλγόριθμος ΠΑΚΣ και οι αλγόριθμοι μείωσης διάστασης θα μπορούσαν να εφαρμοστούν αυτούσιοι και στην περίπτωση δεδομένων δυαδικών συσχετίσεων, που αναπαρίστανται σαν διανύσματα αριθμητικών συνιστωσών (βλ. ενότητα 2.1). Δεδομένου ότι αυτή η αναπαράσταση είναι πιο περιεκτική σε πληροφορία, οι επιδόσεις των αλγορίθμων θα ήταν μάλλον καλύτερες απ' αυτές που παρατηρήσαμε εμείς για τα δικά μας δεδομένα.

- Διεξοδική μελέτη της συμπεριφοράς των αλγορίθμων ομαδοποίησης και εύρεση ενός συστηματικού τρόπου καθορισμού των ρυθμιστικών παραμέτρων τους. Επειδή η απόδοση της ομαδοποίησης επηρεάζει έμμεσα και την απόδοση της Εννοιολογικής Δεικτοδότησης, προκύπτει ότι η εύρεση ενός βέλτιστου συνδυασμού για τις ρυθμιστικές παραμέτρους των αλγορίθμων ομαδοποίησης θα οδηγούσε στην βελτίωση των επιδόσεων της διαδικασίας μείωσης διάστασης. Ένας συστηματικός τρόπος ρύθμισης των παραμέτρων θα μπορούσε να ξεκινήσει από την στατιστική ανάλυση των αποστάσεων μεταξύ των στοιχείων του συνόλου δεδομένων.

- Αξιολόγηση των αλγορίθμων μείωσης διάστασης με χρήση και άλλων δεικτών, εκτός από τη Βελτίωση Ανάκλησης. Ένα τέτοιος δείκτης θα μπορούσε να είναι το ποσοστό επιτυχούς ανάκλησης εγγράφων βάσει ερωτημάτων κειμένου. Η ανάκληση θα βασιζόταν στην εύρεση των N κοντινότερων στο εκάστοτε ερώτημα κειμένου εγγράφων. Η αξιολόγηση της μείωσης διάστασης θα μπορούσε να γίνει συγκρίνοντας την ακρίβεια και την πληρότητα της ανάκλησης στην

πολυδιάστατη και στην μειωμένης διάστασης περίπτωση διανυσματικής αναπαράστασης των εγγράφων. Ένας άλλος τρόπος αξιολόγησης της μείωσης διάστασης θα μπορούσε να υλοποιηθεί χρησιμοποιώντας κάποιον αλγόριθμο ταξινόμησης πάνω σε ένα σύνολο εγγράφων χωρισμένων εκ των προτέρων σε γνωστές κατηγορίες. Συγκρίνοντας την ακρίβεια της ταξινόμησης για την πολυδιάστατη και την μειωμένης διάστασης διανυσματική αναπαράσταση των εγγράφων θα μπορούσαμε να εκτιμήσουμε το κατά πόσο η μείωση διάστασης φέρνει πιο κοντά τα έγγραφα που είναι στην πραγματικότητα σχετικά μεταξύ τους.

- Αξιολόγηση των αλγορίθμων μείωσης διάστασης πάνω σε περισσότερα σύνολα δεδομένων, κατά προτίμηση χωρισμένων σε περισσότερες από δυό κλάσεις. Η αποτελεσματικότητα των συγκεκριμένων αλγορίθμων μπορεί να θεμελιωθεί μόνο μετά από μια σειρά πειραμάτων σε δεδομένα με διαφορετικά χαρακτηριστικά.

- Η μελέτη κάποιας μεθόδου μείωσης διάστασης βασισμένης στην ομαδοποίηση των λέξεων αντί στην ομαδοποίηση των εγγράφων. Μια τέτοια μέθοδος θα μπορούσε να επιδιώκει τη συγχώνευση μιας ολόκληρης ομάδας λέξεων σε μια μοναδική νέα συνιστώσα, μειώνοντας έτσι και τη διάσταση της διανυσματικής αναπαράστασης των εγγράφων.

- Υλοποίηση περισσότερων αλγορίθμων ομαδοποίησης προκειμένου να δοκιμαστεί η επίδοση τους στα πλαίσια της Εννοιολογικής Δεικτοδότησης. Επειδή τα αποτελέσματα της ομαδοποίησης εξαρτώνται σε μεγάλο βαθμό από το κατά πόσο ο χρησιμοποιούμενος αλγόριθμος μπορεί να αντεπεξέλθει στις ιδιομορφίες του εκάστοτε συνόλου δεδομένων, θα ήταν χρήσιμο να υπάρχουν αρκετές εναλλακτικές επιλογές αλγορίθμων ομαδοποίησης.

Παράρτημα Α

Απόδοση ξενόγλωσσων όρων

Ανάλυση Πρωτευουσών Συνιστωσών	Primary Component Analysis
ανόπτηση	annealing
Αποσύνθεση Ιδιαζουσών Τιμών	Singular Value Decomposition
Βελτίωση Ανάκλησης	Retrieval Improvement
δεδομένα δυαδικών συσχετίσεων	dyadic data
εννοιολογική δεικτοδότηση	concept indexing
εξόρυξη δεδομένων	data mining
Μεγιστοποίηση της Αναμενόμενης Τιμής	Expectation Maximization
μηχανική μάθηση	machine learning
μοντέλο διανυσματικού χώρου	vector space model
ομαδοποίηση	clustering
Πιθανοτική Ανάλυση Κρυμμένης Σημασιολογίας (ΠΑΚΣ)	Probabilistic Latent Semantic Analysis (PLSA)
πιθανοφάνεια	likelihood
συνεργατική διήθηση	collaborative filtering
υπερπροσαρμογή	overfitting

Βιβλιογραφία

[1] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models. Technical Report, University of Berkeley, ICSI-TR-97-021, 1997.

[2] J. Breese, D. Heckerman, C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pages 43-52, July 1998.

[3] M. A. Carreira-Perpiñan. A review of dimension reduction techniques. Technical Report, Dept. of Computer Science, University of Sheffield, January 1997.

[4] M. Collins. The EM algorithm. www.cis.upenn.edu/ mcollins/papers/wpeII.4.ps.

[5] T. H. Cormen, C. E. Leiserson, R. L. Rivest. Introduction to algorithms. ASIN 0262031418, MIT Press, 1990.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

[7] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39, 1-38, 1977.

[8] A. I. Schein, A. Popescul, L. H. Ungar. PennAspect: two-way aspect model implementation. Technical Report MS-CIS-01-25, Department of Computer and Information Science, The University of Pennsylvania.

[9] L. Ertöz, M. Steinbach, V.Kumar. Finding clusters of different sizes, shapes and densities in noisy, high dimensional data. Technical Report, 2002.

[10] S. Guha, R. Rastogi, K. Shim. ROCK: a robust clustering algorithm for categorical attributes. In Proceedings of the 15th International Conference on Data Engineering.

[11] T. Hofmann, J. Puzicha. Unsupervised learning from dyadic data. Technical Report TR-98-042, International Computer Science Insitute, Berkeley, CA.

[12] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning Journal, 42(1):177-196, 2001.

[13] T. Hofmann. Probabilistic latent semantic analysis. In Proceedings of the 15th Conference on Uncertainty in AI, 1999.

[14] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 15th Conference on Uncertainty in AI, 1999.

[15] R. A. Jarvis, E. A. Patrick. Clustering using a similarity measure based on shared nearest neighbours. IEEE Trans. Comput. C-22, 1025-1034 (1973).

[16] G. Karypis, E. Han. Concept Indexing, a fast dimensionality reduction algorithm with applications to document retrieval & categorization. Technical Report TR-00-0016, University of Minnesota, 2000.

[17] H. Liu, L. Yu. Feature selection for data mining.

[18] C. Michel. Cardinal, nominal or ordinal similarity measures in comparative evaluation of information retrieval process. Proceedings of the second international conference on language resources and evaluation, Athens, June 2000.

[19] T. M. Mitchell. Machine learning. ISBN 0-07-042807-7, McGraw-Hill, 1997.

[20] I. H. Witten, E. Frank. Data Mining, practical machine learning tools and techniques with Java implementations. ISBN 1-55860-552-5, Morgan Kaufmann, 1999.