ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ ΚΕΙΜΕΝΩΝ

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβλήθηκε στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Πληροφορικής
Εξεταστική Επιτροπή

από τον

ΑΡΓΥΡΗ ΚΑΛΟΓΕΡΑΤΟ

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Οκτώβριος 2007

ΠΡΟΛΟΓΟΣ

Η παρούσα διατριβή εκπονήθηκε στα πλαίσια του Μεταπτυχιακού Προγράμματος Σπουδών του Τμήματος Πληροφορικής, του Πανεπιστημίου Ιωαννίνων. Πραγματοποιήθηκε υπό την επίβλεψη του Αν. Καθ. κ. Α. Λύκα, τον οποίο ευχαριστώ για τις συμβουλές του και την επιστημονική ανησυχία την οποία μου μετέδωσε κατά τη συνεργασία μας.

Το πρόβλημα που εξετάζεται είναι η ομαδοποίηση κειμένων, η οποία αφορά τον αυτόματο διαχωρισμό των κειμένων μίας συλλογής σε ομάδες, ώστε αυτές να περιέχουν κείμενα με συναφές περιεχόμενο. Με την ανάπτυξη και ευρεία χρήση των ηλεκτρονικών υπολογιστών ανέκυψαν ζητήματα αυτόματης οργάνωσης και διαχείρισης δεδομένων. Η ομαδοποίηση είναι μια διαδικασία οργάνωσης δεδομένων η οποία μπορεί να εφαρμοστεί στα κείμενα χωρίς επίβλεψη και να βελτιώσει την αποδοτικότητα και συντηρισιμότητα συστημάτων που διαχειρίζονται μεγάλους όγκους ηλεκτρονικών εγγράφων. Υπάρχουν πολλές ακόμα εφαρμογές της ομαδοποίησης κειμένων όπως η ομαδοποίηση των αποτελεσμάτων ενός ερωτήματος χρήστη σε μία μηχανή αναζήτησης, ώστε να είναι περισσότερο ερμηνεύσιμα και ευπαρουσίαστα.

Η εργασία αυτή πέρα από παραδοσιακές και προσεγγίσεις προτείνει διάφορες τεχνικές που μπορούν να βελτιώσουν το αποτέλεσμα της ομαδοποίησης ή να προετοιμάσουν κατάλληλα τα δεδομένα εισόδου πριν την εφαρμογή οποιασδήποτε μεθόδου ομαδοποίησης.

Ιωάννινα, Οκτώβριος 2007

Αργύρης Καλογεράτος

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Αργύρης Καλογεράτος του Οδυσσέα και της Σβετλάνας.
ΜΕ, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Νοέμβριος, 2007.
Τίτλος: Μέθοδοι Ομαδοποίησης Κειμένων.
Επιβλέπων: Αριστείδης Λύκας.

Στην εργασία αυτή εξετάζεται το πρόβλημα της ομαδοποίησης κειμένων. Ανάμεσα στα ζητήματα που μας απασχολούν είναι οι ενδιαφέρουσες στατιστικές ιδιότητες των κειμένων, η τεχνική αλλά και θεωρητική πλευρά της προεπεξεργασίας και φιλτραρίσματος των δεδομένων, τα μοντέλα που είναι ικανά να αναπαραστήσουν τη σύνθετη νοηματική πληροφορία τους χρήσει αποδοτικών δομών δεδομένων, τα μέτρα ομοιότητας και φυσικά οι αλγόριθμοι και τεχνικές ομαδοποίησης που μπορεί να εφαρμόσει ένα σύστημα για να λάβει έναν διαμερισμό των κειμένων σε ομάδες.

Μελετώνται λοιπόν οι βασικές συνιστώσες του προβλήματος και διάφορες *state-of-the-art* προσεγγίσεις που το αντιμετωπίζουν. Παράλληλα γίνεται μία προσπάθεια να καλυφθούν κάποια κενά της βιβλιογραφίας εντοπίζοντας μειονεκτήματα στις παραδοσιακές προσεγγίσεις και αναφέροντας τις κατευθύνσεις για την περαιτέρω διερεύνηση του προβλήματος.

Οι αλγοριθμικές μέθοδοι που εξετάζονται θεωρούνται παραδοσιακές για το πρόβλημα. Συγκεκριμένα εφαρμόζονται δύο κατηγορίες μεθόδων: α) η ιεραρχική συσσωρευτική ομαδοποίηση (*HAC*) και β) η οικογένεια μεθόδων K-Μέσων.

Ιδιαίτερη αναφορά αξίζει να κάνουμε στη μελέτη της τοπικής πληροφορίας της *K-NN* γειτονιάς των δεδομένων την οποία ενσωματώνουμε σε δύο τεχνικές. Η πληροφορία αυτή δεν είναι προφανές πως μπορεί να χρησιμοποιηθεί σε ένα πρόβλημα μάθησης χωρίς επίβλεψη και από όσο μπορούμε να γνωρίζουμε δεν υπάρχει εκτενής σχετική αναφορά στη βιβλιογραφία, για μεθόδους ομαδοποίσης πέραν αυτών που βασίζονται στην πυκνότητα των αντικειμένων στο χώρο (*density-*

based methods). Η πρώτη τεχνική που αναπτύχθηκε προετοιμάζει τα δεδομένα για την κύρια διαδικασία εκπαίδευσης απαλείφοντας όρους από τα μοντέλα των κειμένων, εξετάζοντας τη γειτονιά κάθε κειμένου ώστε να εκτιμηθεί η σημαντικότητά των όρων που περιέχει. Τα χαρακτηριστικά μιας γειτονιάς μπορούν να ενθαρρύνουν ή να περιορίσουν το φιλτράρισμα όρων, αντίθετα με άλλες μεθόδους που δεν διαθέτουν κάποιο παρόμοιο κριτήριο. Η διαδικασία αυτή έμμεσα οδηγεί και στη μείωη της διάστασης του προβλήματος.

Η δεύτερη αφορά την ομαδοποίηση των κειμένων σε έναν υποχώρο των χαρακτηριστικών τον οποίο «μαθαίνει» ο προτεινόμενος αλγόριθμος δυναμικά κατά την διαδικασία εκπαίδευσης. Η πρότυπη μέθοδος αυτή, εισάγει τον ορισμό Συνθετικών Κέντρων – Αντιπροσώπων των ομάδων και τροποποιεί κατάλληλα το γενικό αλγόριθμο Κ-Μέσων. Οι αντιπρόσωποι μπορούν να κατασκευάζονται χρησιμοποιώντας ένα κέντρο αναφοράς, όπως ο αριθμητικός μέσος ο ενδιάμεσος ή ο ενδιάμεσος μαζί με την *K-NN* γειτονιά από κείμενα της ίδιας ομάδας, καθώς και ένα σχήμα φιλτραρίσματος πάνω στο κέντρο αναφοράς. Η προτεινόμενη μέθοδος καλείται Κ-Συνθετικών Κέντρων.

Πειραματικά η μέθοδος αυτή παρουσιάζει πολύ καλά αποτελέσματα, τα οποία είναι ενθαρρυντικά και για την περαιτέρω μελέτη της προσέγγισης αυτής. Η προσέγγιση των Συνθετικών Κέντρων χρησιμοποιείται και με τον αλγόριθμο Γενικευμένου Κ-Μέσων ο οποίος παρουσιάζει τα καλύτερα πειραματικά αποτελέσματα σε όλες τις συλλογές κειμένων που χρησιμοποιήθηκαν. Εξετάστηκε και η τεχνική των Κ-Ενδιαμέσων και Γενικευμένου Κ-Ενδιαμέσων η οποία, αν και χρήσιμη, στις περισσότερες περιπτώσεις είναι ξεκάθαρα υποδεέστερη από τις αντίστοιχες των Συνθετικών Κέντρων.

Συνοψίζοντας, ένα σημαντικό αποτέλεσμα της διατριβής αυτής έχει να κάνει λοιπόν, με την ενσωμάτωση της τοπικής πληροφορίας στο πρόβλημα ομαδοποίησης και τη διατύπωση της ιδέας ότι ο αριθμητικός μέσος, παρότι βέλτιστος, δεν είναι ο καταλληλότερος αντιπρόσωπος των ομάδων κυρίως λόγω των ιδιαίτερων χαρακτηριστικών των δεδομένων του προβλήματος. Οι προτεινόμενες προσεγγίσεις είναι πρότυπες και αντιμετωπίζουν πιο αποτελεσματικά τα διάφορα ζητήματα που ανακύπτουν.

EXTENDED ABSTRACT IN ENGLISH

Kalogeratos, Argiris, O., S.
MSc, Computer Science Department, University of Ioannina, Greece. October, 2007.
Thesis Title: Methods for Clustering Documents.
Thesis Supervisor: Likas Aristeidis.

This thesis studies the problem of clustering documents (web documents or plain texts). Among the examined issues are the interesting global statistical properties of text and written languages, the technical and theoretical aspects of text preprocessing and filtering, the data models that are capable of representing the complexity of language as well as of efficient similarity calculations and memory allocation, the similarity measures and of course the clustering methods that can be used by a system to obtain the clustered documents.

All key components of the problem are studied and various *state-of-the-art* approaches are engaged to tackle several issues. Moreover, we attempt to approach some uncovered issues of literature, or review other, by spotting disadvantages of traditional techniques and mentioning some directions for further study.

The clustering algorithms that are examined in this dissertation are widely considered as general and established approaches for many clustering problems. In particular, two main categories are studied: a) the hierarchical agglomerative clustering (*HAC*) and b) the K-Means family of clustering algorithms.

The development of procedures that incorporate local information, such as *K-NN* neighborhoods, to a unsupervised clustering method is one of the major contributions of this work. This incorporation is not trivial for unsupervised learning problems, and to the best of our knowledge is not a well-studied idea in literature of clustering, for methods other than density-based. Two novel techniques have been developed and presented that are based on this latter concept. The first one concerns data preparation

before clustering and removes features from document models, considering the neighborhood of a document as an advisor for its term significance estimation. The neighborhood characteristics can limit or encourage term filtering, contrary to many filtering approaches that lack of a relevant threshold. This procedure indirectly leads in reduction of dimensionality of the problem.

The second technique concerns the main clustering procedure and aims at the separation of documents in a subspace of features that is learned dynamically during clustering. Specifically, this technique introduces the notion of Synthetic Cetroids – Representatives for cluster centers and modifies them appropriately using the general K-Means algorithm. The synthetic representatives can be constructed using a reference centroid, as well as a filtering scheme on it. Examples of reference centroids are the arithmetic mean, the medoid document or the medoid along with its *K-NN* neighborhood of co-clustered documents. We call the proposed technique K-Synthetic Centroids.

In the experimental evaluation, the clustering results using the proposed techniques show a clear improvement compared to traditional K-Means and HAC and are very encouraging for further improvements. Synthetic Centroids are also incorporated to the Global-K-Means algorithm, resulting in the Global-K-Synthetic Centroids method, a combination that presents the best results on all document collections that have been considered. Other K-means variants such as K-Medoids and K-Global-Medoids have also been examined and, despite having some useful properties, produce inferior results with respect to our approach.

In conclusion, this thesis contributes to the problem by putting forth the idea that although the arithmentic mean centroids are mathematically optimal as cluster representatives, their optimality is rather a drawback, because of the special characteristics of written language. The proposed methodologies for centroid definition and the corresponding algorithmic procedures developed in this thesis seem to cope efficiently with the various and complex difficulties of the document clustering problem.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

Με την ανάπτυξη και ευρεία χρήση των ηλεκτρονικών υπολογιστών ανέκυψαν ζητήματα αυτόματης οργάνωσης και διαχείρισης δεδομένων. Τα δεδομένα που παράγονται και διακινούνται ηλεκτρονικά από τον σύγχρονο άνθρωπο όχι μόνο αυξάνονται σε όγκο, αλλά μπορούμε να πούμε πως αυξάνεται και ο ίδιος ο ρυθμός παραγωγής τους.

Ο κυριότερος λόγος που συντελεί στην εμφάνιση του φαινομένου αυτού είναι πως τα ηλεκτρονικά μέσα επικοινωνίας και διαχείρισης πληροφορίας έχουν αποκτήσει κυρίαρχο ρόλο στις οικονομικά ανεπτυγμένες κοινωνίες. ο φάσμα των υπηρεσιών που παρέχονται στους χρήστες και των αναγκών που ικανοποιούν διευρύνεται καθιστώντας τους ηλεκτρονικούς υπολογιστές και το Διαδίκτυο από βασικό εργαλείο εργασίας μέχρι βασικό μέσο επικοινωνίας και ψυχαγωγίας. Ηλεκτρονικές εκδόσεις, ψηφιακές βιβλιοθήκες, ηλεκτρονικά βιβλία, μηνύματα ηλεκτρονικού ταχυδρομείου, άρθρα επικαιρότητας, ιστοχώροι συζητήσεων χρηστών, οι συμβατικές σελίδες Διαδικτύου και πρόσφατα ιστοχώροι διακίνησης πολυμεσικών δεδομένων, είναι μόνο μερικές από τις δραστηριότητες που απαιτούν τη διαχείριση μεγάλου όγκου δεδομένων.

Όλοι λοιπόν οι παράγοντες αυτοί συντελούν στην δημιουργία ενός περιβάλλοντος όπου ο χρήστης αδυνατεί να επιτελέσει, με τον παραδοσιακό χειρονακτικό τρόπο, ακόμα και τις πιο απλές εργασίες στα ηλεκτρονικά αρχεία που τον ενδιαφέρουν. Ένας ακόμα λόγος επιδείνωσης της κατάστασης είναι τα υψηλά επίπεδα διασύνδεσης των υπολογιστικών συστημάτων και των βάσεων δεδομένων με τον Παγκόσμιο Ιστό. Στα συστήματα μεγάλης κλίμακας (π.χ. συστήματα οργανισμών ή κόμβων διαδικτυακής αναζήτησης) ο έλεγχος και η διαχείριση των δεδομένων από τον άνθρωπο είναι διαδικασίες αφόρητες οικονομικά ή ακόμα και αδύνατες χρονικά.

Είναι έτσι, λογική η απαίτηση από πλευράς χρηστών οποιασδήποτε κατηγορίας - κατ' επέκταση η επιδίωξη από πλευράς ειδικών επιστημόνων- τέτοιες εργασίες να επιτελούνται αυτόματα ή ημιαυτόματα από τα υπολογιστικά συστήματα. Παρά το γεγονός ότι οι επιδόσεις των συστημάτων βελτιώνονται διαρκώς ώστε να μπορούν να χειρίζονται ταχύτερα μεγαλύτερους όγκους δεδομένων, δίχως αποτελεσματικές προσεγγίσεις σε επίπεδο επίλυσης των ίδιων των προβλημάτων η υπολογιστική ισχύς καθεαυτή (ταχύτητα επεξεργασίας) και οι επιμέρους επιδόσεις (π.χ. ταχύτητα δίσκων, μεταφοράς δεδομένων) είναι βέβαιο πως δε θα είναι ποτέ αρκετά για να καλύπτουν τις εκάστοτε διαμορφούμενες ανάγκες.

Η πληθώρα των ψηφιοποιημένων εγγράφων, όπως είπαμε, μπορεί να είναι η κύρια αιτία για τα προβλήματα που προέκυψαν στη διαχείρισή τους, όμως ως γεγονός ικανοποίησε και μια βασική προϋπόθεση για τη μελέτη και πρόοδο προς την επίλυση των προβλημάτων αυτών: την ύπαρξη αρκετού υλικού για επεξεργασία σε έτοιμη ηλεκτρονική μορφή. Επίσης, μία σπουδαία πρόκληση για την ανθρώπινη γνώση και επιστήμη είναι πως «τα δεδομένα μπορούν να δημιουργούν δεδομένα», δηλαδή η μελέτη φαινομένων μέσα από μεγάλους όγκους καταγεγραμμένης πληροφορίας μπορεί να δώσει λύσεις σε σημαντικά επιστημονικά προβλήματα. Τέτοια παραδείγματα είναι η ανάλυση ανθρώπινου γενετικού υλικού (*DNA, RNA*) για των εντοπισμό ύποπτων γονιδίων για διάφορες παθήσεις.

Η ιδέα της αυτόματης διαχείρισης κειμένων προέρχεται από τις αρχές της δεκαετίας του '60, όμως αξιόλογη ερευνητική δραστηριότητα έχει υπάρξει τα 15 τελευταία περίπου χρόνια. Υπάρχουν δύο διαφορετικές προσεγγίσεις: η μάθηση με επίβλεψη όπου διαθέτουμε ένα σύνολο εκπαίδευσης για το οποίο γνωρίζουμε την πραγματική κατηγορία κάθε αντικειμένου, βάσει του οποίου προσπαθούμε να προβλέψουμε την

κατηγορία αγνώστων δεδομένων, και η μάθηση χωρίς επίβλεψη η οποία δεν απαιτεί σύνολο εκπαίδευσης ούτε παρέμβαση από τον άνθρωπο. Στην εργασία αυτή θα ασχοληθούμε με την ομαδοποίηση κειμένων η οποία είναι μέθοδος μάθησης χωρίς επίβλεψη.

Όσων αφορά τα κείμενα, η βασική ανάγκη των χρηστών είναι η αναζήτηση πληροφορίας σε αρχεία και βάσεις δεδομένων, σε τοπικά ή απομακρυσμένα συστήματα. Εν γένει, ως διαδικασία αφορά την εύρεσης εγγράφων που ικανοποιούν μια σειρά κριτηρίων επιλογής του χρήστη, για παράδειγμα τη σχετικότητα με μια σειρά λεκτικών όρων ή φράσεων. Κατά την υλοποίηση ενός συστήματος προκύπτουν δύο σημαντικά ζητήματα. Η αποδοτική αναζήτηση απαιτεί την ύπαρξη ενός είδους οργάνωσης των δεδομένων βάσει κατάλληλων κριτηρίων (ευρετηριοποίηση). Το δεύτερο ζήτημα αφορά την εξασφάλιση της σχετικότητας των αποτελεσμάτων με τα ζητούμενα του χρήστη, το οποίο επιβάλλει τις περισσότερες φορές την παρέμβαση ανθρώπινης απόφασης για την επισύναψη χαρακτηριστικών στα αποθηκευμένα δεδομένα. Αν λοιπόν η αποτελεσματικότητα μπορεί να χρεωθεί ερευνητικά στην τεχνολογία υπολογιστικών συστημάτων και επικοινωνιών, το ζητούμενο της ορθότητας και της επεξηγηματικότητας των αποφάσεων (*decision reasoning*) θα πρέπει να λογίζεται ως στόχος και χώρος μελέτης πεδίων όπως η Τεχνητή Νοημοσύνη και συγκεκριμένα η Μηχανική Μάθηση.

Η ομαδοποίηση και η κατηγοριοποίηση αποτελούν προβλήματα ιδιαίτερης σημασίας για τον τομέα της Μηχανικής Μάθησης και είναι στοιχειώδεις λειτουργίες που μπορούν να εφαρμοστούν πάνω στα διαθέσιμα κείμενα. Ανήκουν σε μια γενικότερη κατηγορία ανάλυσης και διαχείρισης εγγράφων βάσει περιεχομένου (*content based document management tasks*). Η αυτόματη οργάνωση των εγγράφων μίας βάσης δεδομένων σε ομάδες με συνάφεια περιεχομένου (συχνά σε κάποιο ιεραρχικό σχήμα) αποτελεί ένα επίπεδο ευρετηριοποίησης το οποίο μπορεί να έχει παραχθεί χωρίς ή και με μερική ανθρώπινη επίβλεψη. Σε αυτό μπορεί στη συνέχεια να βασιστούν μια σειρά από άλλες εφαρμογές διαχείρισης εγγράφων. Αυτός είναι και ο λόγος που πολλές τεχνικές Μηχανικής Μάθησης έχουν ευρεία χρήση στο πεδίο της Διαχείρισης Δεδομένων.

1.1. Ορισμός Προβλήματος

Το πρόβλημα με το οποίο θα ασχολείται η εργασία είναι η ομαδοποίηση κειμένων: δίνεται ένα σύνολο κειμένων και απαιτείται ο αυτόματος διαχωρισμός τους σε ανεξάρτητες ομάδες, ώστε τα δεδομένα κάθε ομάδας να παρουσιάζουν θεματική συνοχή. Στη γενική μορφή του προβλήματος αφορά κάθε είδους κείμενα, αρκεί αυτά να παρέχονται σε ηλεκτρονική μορφή ώστε να αποτελούν είσοδο κάποιας μεθόδου. Στην εργασία θα μελετήσουμε μεθόδους ομαδοποίησης κειμένων τις οποίες θα εφαρμόσουμε σε έγγραφα φυσικής γλώσσας αλλά και μία ειδική κατηγορία κειμένων, τα υπερκείμενα (*HTML documents*).

Τα κείμενα συνήθως αναπαρίστανται με διανύσματα στον χώρο $\mathbb{R}^{|V|}$, όπου V το σύνολο των διαφορετικών λέξεων που εμφανίζονται σε μία συλλογή. Το σύνολο αυτό καλείται και λεξιλόγιο συλλογής και ορίζει τη διάσταση των δεδομένων του προβλήματος.

Τα βασικότερα προβλήματα που παρουσιάζονται στην ομαδοποίηση κειμένων είναι α) η μεγάλη διάσταση των δεδομένων, β) η σύνθετη εννοιολογική και νοηματική δομή τους (συμφραζόμενα, συντακτική δομή), γ) το συχνό φαινόμενο όπου κείμενα παρουσιάζουν θεματική συνάφεια με περισσότερες από μία ομάδες (δύσκολη η ταξινόμηση ακόμα και από ειδικούς [4]), δ) η δυσκολία προσδιορισμού σχήματος δομής ομάδων το οποίο αναζητούμε ως ομάδα, ε) αλλά και μία σειρά άλλων γλωσσικών ανωμαλιών που θα αναφέρουμε αναλυτικότερα στο Κεφάλαιο 2.

1.2. Σχετιζόμενα Επιστημονικά Πεδία και Σύγχρονες Εφαρμογές

Το πεδίο της Εξόρυξης Γνώσης από Κείμενα (*Text Mining*), έχει αρκετά κοινά στοιχεία με τις κλασικές μεθόδους Εξόρυξης Γνώσης από Δεδομένα (*Data Mining*). Πολλές από τις τεχνικές αυτές μπορούν να αναδείξουν την εσωτερική δομή των δεδομένων και να παράγουν συμπεράσματα συσχέτισης χρήσιμα για τον άνθρωπο. Συνεπώς μπορούν να καθορίσουν και αντίστοιχες μηχανικές διαδικασίες για την οργάνωση των δεδομένων.

Μια από τις πιο διαδεδομένες μορφές ηλεκτρονικών κειμένων που διακινούνται στο Διαδίκτυο είναι τα υπερκείμενα (*HTML documents*), τα οποία είναι δομημένα κείμενα που περιέχουν πληροφορία σε φυσική γλώσσα καθώς και ένα σύνολο στοιχείων προσδιορισμού ιδιοτήτων. Εφαρμόζοντας τεχνικές εξόρυξης στο Διαδίκτυο

προκύπτει το αντίστοιχο πεδίο (*Web Mining*). Μάλιστα σε γενικές γραμμές έχουμε τρεις υποκατηγορίες του πεδίου αυτού [2]: α) Εξόρυξη Δομών (*Web Structure Mining*), β) Εξόρυξη Χρήσης (*Web Usage Mining*), γ) Εξόρυξη Περιεχομένου (*Web Content Mining*).

Μερικές σύγχρονες εφαρμογές μεθόδων για την ομαδοποίηση εγγράφων είναι:

- Ομαδοποίηση με σκοπό την παρουσίαση ενός συνόλου εγγράφων με πιο οργανωμένο και κατανοητό τρόπο για το χρήστη μέσω της νοηματικής συσχέτισης τους.
- Ομαδοποίηση των εγγράφων μιας βάσης δεδομένων με σκοπό την ταχύτερη αναζήτηση εγγράφων σχετικών με ένα θέμα, εστιάζοντας σε υποσύνολα πληροφοριών (τις ομάδες συναφών κειμένων).
- Ομαδοποίηση ιστοσελίδων από Μηχανές Αναζήτησης Διαδικτύου [52], για την αποτελεσματικότερη αποθήκευση των δεδομένων και την ακριβέστερη ανάκτηση πληροφοριών (συνδυάζει τα προηγούμενα δύο).
- Φιλτράρισμα εγγράφων, τέτοια παραδείγματα είναι η διανομή ηλεκτρονικής αλληλογραφίας στα τμήματα μιας επιχείρησης ανάλογα με το αν υπάγεται το θέμα στις αρμοδιότητές του καθενός ή το φιλτράρισμα άρθρων τα οποία στέλνει ένα πρακτορείο ειδήσεων σε μία εφημερίδα ή ένα περιοδικό, ανάλογα με το χαρακτήρα του εντύπου.
- Αυτόματη παραγωγή μετα-δεδομένων, λέξεων κλειδιών κ.α (*automated metadata generation*).
- Αυτόματη δημιουργία ή εμπλουτισμός οντολογιών και αποσαφήνιση νοήματος λεκτικών όρων βάση συμφραζομένων σε κείμενα (*Word Sense Disambiguation*).
- Αυτόματη εξαγωγή περιλήψεων από κείμενα ή ομάδες κειμένων.

1.3. Ιστορική Αναδρομή της Αυτόματης Διαχείρισης και Οργάνωσης Κειμένων

Σαν πρόβλημα έχει διατυπωθεί εδώ και αρκετά χρόνια, παρόλα αυτά θεωρείται αιχμή τεχνολογικά καθώς η ύπαρξη τεχνικών μηχανικής οργάνωσης των δεδομένων είναι ζωτικής σημασίας ζήτημα για τον περιορισμό των επιπτώσεων της αύξησης των δεδομένων που αποθηκεύονται στα πληροφοριακά συστήματα.

Έως τα τέλη της δεκαετίας του ’80, η δημοφιλέστερη προσέγγιση στο πρόβλημα του διαμερισμού κειμένων σε ομάδες, βασιζόταν στην Μηχανική Γνώσης (*Knowledge Engineering – KE*). Αργότερα, στις αρχές του 1990 έγιναν οι πρώτες προσπάθειες να προσεγγιστεί το πρόβλημα του διαμερισμού των κειμένων με τεχνικές Μηχανικής Μάθησης (*Machine Learning - ML*).

Η ουσιώδης διαφορά ανάμεσα στις *ML* μεθόδους είναι η ύπαρξη ή μη επίβλεψης. Οι μέθοδοι μάθησης με επίβλεψη χρησιμοποιούν ένα βοηθητικό σύνολο δεδομένων το οποίο υποδεικνύει τα χαρακτηριστικά στοιχεία των κατηγοριών που επιθυμούμε να διαχωρίσουμε. Αντίθετα, οι μέθοδοι χωρίς επίβλεψη προσπαθούν να «ανακαλύψουν» εγγενείς νοηματικές ομοιότητες στα των κείμενα ώστε να παραχθεί μία διαμέρισή τους.

Πάντως, είναι ενδιαφέρον να αναφερθεί πως οι δύο προσεγγίσεις αλληλοσυμπληρώνονται. Σε αρκετές εφαρμογές χρησιμοποιούνται τεχνικές χωρίς επίβλεψη, κατά το στάδιο της προεπεξεργασίας δεδομένων, ώστε να εκτιμηθούν επιμέρους παράμετροι του προβλήματος, πριν τελικά εφαρμοστεί κάποια τεχνική με επίβλεψη. Ακόμα, είναι δυνατόν να υιοθετούμε μια προσέγγιση χωρίς επίβλεψη υποβοηθώντας παράλληλα την επίλυση με ένα σύνολο εκπαίδευσης.

1.4. Μηχανική Γνώσης

Οι τεχνικές αυτές στηρίζονται σε ένα σύνολο κανόνων κανονικής μορφής *DNF* (*Disjunctive Normal Form*) της μορφής:

if (*DNF formula*) **then** (*category*)

Συγκεκριμένα, κάθε κατηγορία περιγράφεται από ένα σύνολο τέτοιων κανόνων που ορίζουν τα επιθυμητά χαρακτηριστικά της. Στις λογικές συνθήκες μπορεί να ελέγχεται π.χ. η ύπαρξη λέξεων κλειδιών στο εξεταζόμενο κείμενο ή συνδυασμός τέτοιων συνθηκών με συζευκτικές και διαζευκτικές προτάσεις. Υπεύθυνοι για τη δημιουργία και ενημέρωση των κανόνων είναι οι μηχανικοί γνώσης (*knowledge engineers*) σε συνεργασία πάντα με κάποιους «ειδικούς» (*domain experts*) στο περιεχόμενο του συνόλου των εγγράφων που επιθυμούμε να ταξινομηθούν.

Το κύριο μειονέκτημα της προσέγγισης αυτής είναι γνωστό από το πεδίο των Έμπειρων Συστημάτων και ονομάζεται συμφόρηση απόκτησης γνώσης (*knowledge acquisition bottleneck*). Πέρα από το γεγονός ότι για την εφαρμογή και χρήση ενός τέτοιου συστήματος πάνω σε ένα πρόβλημα είναι απαραίτητη η συμβολή των ειδικών

περιοχής, οποιαδήποτε αλλαγή των παραμέτρων του προβλήματος (π.χ. η πρόσθεση μίας νέας κατηγορίας) απαιτεί την εκ νέου επέμβαση των ειδικών για την ενημέρωση των κανόνων.

if ((*wheat* **and** *farm*)	**or**
(*wheat* **and** *commodity*)	**or**
(*wheat* **and** *tones*)	**or**
(*bushels* **and** *export*)	**or**
(*wheat* **and** *winter* **and** ¬*soft***)**)	**then** WHEAT **else** ¬WHEAT

Σχήμα 1.1. Παράδειγμα σύνθετων λογικών εκφράσεων.

Πρόκειται λοιπόν, για συστήματα που αδυνατούν να γενικεύσουν τη γνώση που τους παρέχεται μέσω των κανόνων, ενώ η προσαρμογή τους σε νέα προβλήματα και η ενημέρωσή τους είναι διαδικασίες υψηλού κόστους.

1.5. Μηχανική Μάθηση

Οι μέθοδοι μηχανικής μάθησης έχουν το σαφές πλεονέκτημα ότι το πρόβλημα του προσδιορισμού των κανόνων είτε εξαλείφεται με μεθόδους μάθησης χωρίς επίβλεψη, είτε περιορίζεται σημαντικά. Στη χειρότερη περίπτωση, οι ειδικοί ταξινομούν χειρωνακτικά ένα σύνολο εγγράφων που αποτελούν παραδείγματα για την εκπαίδευση.

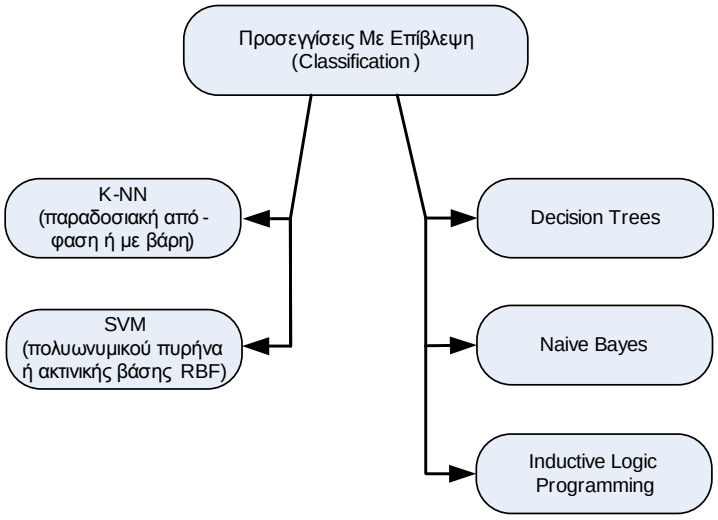
1.5.1. Κατηγοριοποίηση Κειμένων

Στην προσέγγιση της Μηχανικής Μάθησης [1], η οποία για τα κείμενα συναντάται και ως εντοπισμός θέματος (*topic spotting*), μία γενική επαγωγική διαδικασία κατασκευάζει αυτόματα έναν ταξινομητή για κάθε κατηγορία κάνοντας «παρατηρήσεις» πάνω στα χαρακτηριστικά τους. Τα κείμενα θα πρέπει να έχουν προηγουμένως ταξινομηθεί χειρωνακτικά από κάποιον ειδικό. Τελικά, ένα άγνωστο έγγραφο ταξινομείται βάση της συσχέτισής του με τους ταξινομητές που έχουν κατασκευαστεί.

Ορίζοντας αυστηρά το πρόβλημα κατηγοριοποίησης κειμένων θα λέγαμε πως είναι η διαδικασία με την οποία διαθέτοντας ένα σύνολο $D = \{d_1, \ldots, d_N\}$ αποτελούμενο από $|D| = N$ κείμενα και ένα σύνολο $K = \{\kappa_1, \ldots, \kappa_M\}$ με $|K| = M$ προκαθορισμένες κατηγορίες, θα πρέπει να αντιστοιχιστεί μία δυαδική τιμή σε κάθε ζεύγος (d_i, κ_j). Η

τιμή αυτή απαντά θετικά ή αρνητικά στο αν το κείμενο d_i κατηγοριοποιείται στην κατηγορία κ_j.

Το πρόβλημα μπορεί να διατυπωθεί και ως διαδικασία προσέγγισης μίας άγνωστης συνάρτησης $f : D \times K \rightarrow \{true, false\}$, η οποία θα αντιστοιχούσε ιδανικά τα κείμενα στις πραγματικές τους κατηγορίες. Η συνάρτηση αυτή προσεγγίζεται από τη συνάρτηση των ταξινομητών $f_c : D \times K \rightarrow \{true, false\}$ η οποία προσπαθεί να ελαχιστοποιήσει τις αποφάσεις που διαφωνούν με την *f*. Η εκμάθηση της συνάρτησης προσέγγισης γίνεται μέσω παραδειγμάτων που παρέχονται από ένα σύνολο εκπαίδευσης $T = \{(d_i, \kappa_i)\}$. Η απαίτηση για ύπαρξη συνόλου εκπαίδευσης κατατάσσει τις τεχνικές κατηγοριοποίησης στις τεχνικές μάθησης με επίβλεψη.



Σχήμα 1.2. Δημοφιλείς τεχνικές κατηγοριοποίησης οι οποίες έχουν εφαρμοστεί για τη μηχανική οργάνωση κειμένων.

Μια εκτενής σύγκριση μεθόδων κατηγοριοποίησης, όπως *Support Vector Machines* (*SVM*), *Neural Nets* [76], *Naive Bayes*, *Linear Least Square Fit* και *K-Nearest Neighbor* (*K-NN*) οι οποίες έχουν εφαρμοστεί σε κείμενα συναντάται στο [3]. Άλλες προσεγγίσεις αφορούν τα Δέντρα Αποφάσεων [30], τον *Inductive Logic Programming* [28][29][16], Συστήματα Κανόνων [26][27] και Στατιστική Ανάλυση [21][22][23].

1.5.2. Ομαδοποίηση Κειμένων

1.5.2.1. Βασικές Έννοιες

Οι τεχνικές ομαδοποίησης κατατάσσονται στις τεχνικές μάθησης χωρίς επίβλεψη. Ήρθαν αργότερα στο προσκήνιο, όταν τα προβλήματα (που δεν αφορούσαν μόνο ηλεκτρονικά έγγραφα) έθεταν στην πράξη επιπλέον περιορισμούς. Ένα προηγμένο σύστημα διαχείρισης είναι επιθυμητό να μην έχει την ανάγκη επέμβασης των «ειδικών περιοχής» για τον προσδιορισμό ή αποσαφήνιση του προβλήματος (κανόνες, επιλογές από ενδεχόμενα). Όταν η τάξη μεγέθους του όγκου πληροφορίας είναι οι εκατοντάδες χιλιάδες έγγραφα, με δυναμικές μεταβολές στις θεματικές ομάδες, τότε μία προσέγγιση με επίβλεψη θα ήταν ιδιαίτερα ασύμφορη από πολλές πλευρές.

Η ομαδοποίηση κειμένων έγκειται στον αυτόματο, μηχανικό διαμερισμό ενός συνόλου κειμένων σε επιμέρους ομάδες (*clusters*) βάσει του εννοιολογικού και θεματικού περιεχομένου τους. Ένας πιο αυστηρός ορισμός του προβλήματος, είναι πως πρόκειται για τη διαδικασία με την οποία κάθε κείμενο ενός συνόλου $|D| = N$ κειμένων, $D = \{d_1, \ldots, d_N\}$, αντιστοιχίζεται μοναδικά σε μία ομάδα από ένα σύνολο $|C| = M$ ομάδων, $C = \{c_1, \ldots, c_M\}$. Ο αριθμός M των ομάδων θεωρείται προκαθορισμένος και ιδανικά επιθυμούμε κάθε ομάδα του αποτελέσματος να αντιστοιχεί σε μία κατηγορία δεδομένων. Συνήθως μας αρκεί κάθε κατηγορία να περιέχεται σε μία μόνο ομάδα αποτελέσματος διότι η αντιστοιχία κειμένων σε κατηγορίες από τον άνθρωπο δεν είναι η μοναδική σωστή διαμέριση τους, έτσι η διαδικασία εκπαίδευσης μπορεί να ανακαλύψει σχέσεις οι οποίες συμπεραίνουν μία διαφορετική διαμέριση της συλλογής.

Για τη αναλυτική διάκριση και περιγραφή των αλγορίθμων ομαδοποίησης απαιτείται μία μακροσκελή συζήτηση [68] που ξεφεύγει από τα πλαίσια της εργασίας. Παρόλα αυτά στη βιβλιογραφία ο βασικός διαχωρισμός των αλγορίθμων ομαδοποίησης τους κατατάσσει σε δύο μεγάλες κατηγορίες:

- τους ιεραρχικούς (*hierarchical*), όπου παράγονται εμφωλευμένες ομάδες για το πρόβλημα προχωρώντας από το γενικό στο ειδικό (*top-down*), ή αντίστροφα (*bottom-up*), και
- τους διαμεριστικούς (*partitional* ή *non-hierarchical*) οι οποίοι προσπαθούν να βελτιώσουν επαναληπτικά την ποιότητα της ομαδοποίησης, βάσει κάποιων κριτηρίων ελέγχου.

- τους αλγόριθμους πυκνότητας (*density- based*) οι οποίοι αναγνωρίζουν ομάδες βάση της πυκνότητας αντικειμένων στο χώρο των δεδομένων χωρίς να είναι απαραίτητο να σχηματίζουν συγκεκριμένα σχήματα.

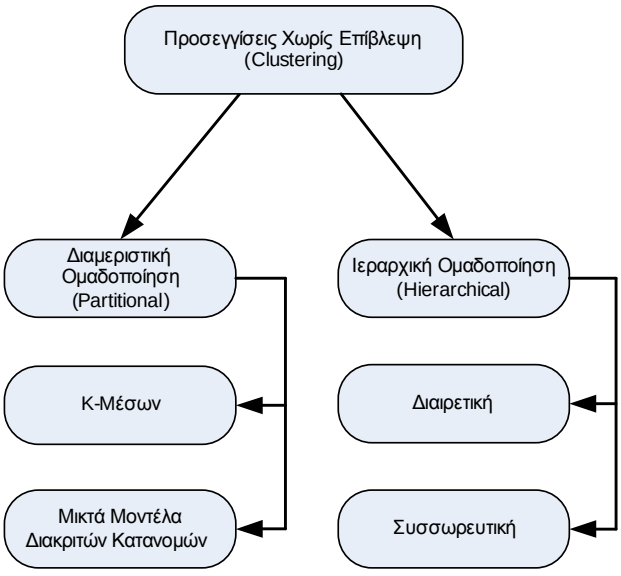
Ο τρόπος ανάθεσης δεδομένων στις ομάδες είναι επίσης ένα χαρακτηριστικό διαχωρισμού των αλγορίθμων, έτσι η μοναδικότητα της ανάθεσης κατατάσσει την προσέγγιση ως αυστηρή ομαδοποίηση (*hard clustering*). Μια άλλη εκδοχή είναι η ασαφής ομαδοποίηση (*fuzzy* ή *soft clustering*), στην οποία κάθε αντικείμενο συμμετέχει σε όλες τις ομάδες με μία πιθανότητα συμμετοχής $p \in [0,1]$ και αποτελεί γενίκευση της πρώτης περίπτωσης, όπου $p \in \{0, 1\}$. Τέλος, στην επικαλυπτόμενη ομαδοποίηση (*overlapping clustering*) δεν έχουμε ασαφή συμμετοχή στις ομάδες, αλλά επιτρέπεται η ανάθεση αντικειμένων σε περισσότερες από μία ομάδες.

Η ομαδοποίηση είναι μία στοιχειώδης διαχειριστική και οργανωτική λειτουργία που μπορεί να εφαρμοστεί πάνω στα διαθέσιμα δεδομένα. Ενσωματώνοντας πληροφορία σχετική με τον τύπο των αντικειμένων-δεδομένων του προβλήματος και στοχεύει στην ανακάλυψη δομών σε αυτά. Η πληροφορία καλείται συνήθως μέτρο ή συνάρτηση συσχέτισης (ομοιότητας ή συμπληρωματικά απόστασης), μέσω της οποίας ορίζονται σχέσεις μεταξύ ζευγών αντικειμένων, ζευγών ομάδων αλλά και σχέσεις αντικειμένου-ομάδας. Στο εξής θα αποκαλούμε τη συνάρτηση αυτή συνάρτηση ομοιότητας.

Στο υπό εξέταση πρόβλημα, η συνάρτηση ομοιότητας θα πρέπει να είναι κατάλληλα ορισμένη ώστε να μπορεί να απαντήσει στο κατά πόσο είναι νοηματικά συγγενές το περιεχόμενο δύο κειμένων. Είναι σημαντικό να παρατηρήσουμε πως η ποιότητά της καθορίζει σε πολύ μεγάλο βαθμό την ποιότητα των λύσεων που θα λάβουμε από οποιοδήποτε αλγόριθμο ομαδοποίησης, διότι παρέχει πληροφορία για το πρώτο επίπεδο συσχέτισης των δεδομένων που είναι οι αποστάσεις μεμονωμένων δεδομένων. Πάνω σε αυτό το στοιχειώδες επίπεδο σχέσεων βασίζονται οι περισσότεροι αλγόριθμοι για να δημιουργήσουν ένα ανώτερο επίπεδο αποφάσεων.

Υπάρχουν αρκετοί αλγόριθμοι ομαδοποίησης, με αρκετά διαφορετική φιλοσοφία προσέγγισης [10]. Στη σχετική βιβλιογραφία, επικρατούν γενικοί αλγόριθμοι οι οποίοι με παραλλαγές ή προσαρμογές καταφέρνουν να εφαρμοστούν και στο πρόβλημα αυτό

(Κ-Μέσων, Ιεραρχικός Συσσωρευτικός κ.α.). Επίσης, υπάρχουν και ειδικά σχεδιασμένοι μέθοδοι για το πρόβλημα της ομαδοποίησης κειμένων [5][8][47][55][56].



Σχήμα 1.4. Δημοφιλείς τεχνικές ομαδοποίησης οι οποίες έχουν εφαρμοστεί για τη μηχανική οργάνωση.

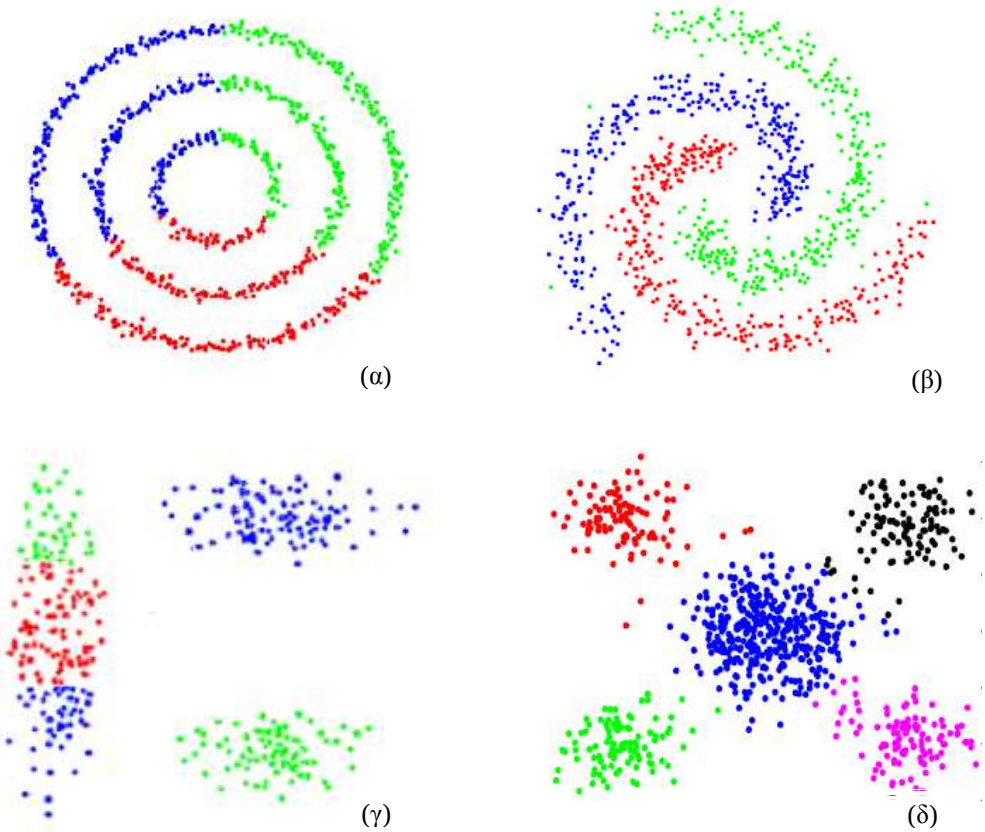
1.5.2.2. Ειδικά Θέματα

Η ομαδοποίηση γενικά είναι ένα δύσκολο πρόβλημα το οποίο φτάνει να έχει φιλοσοφικές προεκτάσεις. Δεν υπάρχει «συνταγή» για το ποια είναι η σωστή λύση για ένα σύνολο δεδομένων, επειδή αγνοούμε παντελώς τι περιέχουν αυτά. Το γεγονός ότι δε μπορεί να οριστεί αυστηρά και γενικά ταυτόχρονα το τί είναι «ομάδα», ιδιαίτερα στις περιπτώσεις όπου οι ομάδες δεν είναι διαχωρίσιμες, αποτελεί ένα βασικό θεωρητικό κενό της ομαδοποίησης.

Αν θέλαμε να δώσουμε έναν γενικό ορισμό της ομάδας δεδομένων, θα λέγαμε πως πρόκειται για ένα υποσύνολο των δεδομένων που παρουσιάζει (ακολουθεί ή συμμορφώνεται με) μία δομή στο χώρο η οποία προσδίδει εσωτερική συνοχή στην ομάδα σε σχέση με το περιβάλλον της που αποτελείται από τα υπόλοιπα δεδομένα του συνόλου. Τις περισσότερες φορές όμως δεν είναι σαφές ούτε το τι δομές αναζητούμε, αλλά ούτε και πώς θα εντοπίσουμε τι αξίζει να αναζητήσουμε. Υπάρχουν δύο ειδών αστοχίες που μπορεί να συμβούν και από διαφορετική αφετηρία να οδηγήσουν σε χαμηλής ποιότητας λύσεις:

- η επιλογή της μορφής των ομάδων (πόλωση προς συγκεκριμένες δομές),
- ο ορισμός της τάξης του μοντέλου (πολυπλοκότητα) η οποία στην ομαδοποίηση αφορά τον αριθμό των ομάδων που αναζητούμε.

Κάθε αλγόριθμος ομαδοποίησης προκρίνει το σχηματισμό ομάδων με συγκεκριμένη δομή, είτε αυστηρά (π.χ. ο K-Μέσων βρίσκει αποκλειστικά σφαιρικές ομάδες) είτε με ελαστικότητα (μία γκαουσιανή κατανομή μπορεί να βρει σφαίρες αλλά και ελλειπτικές δομές). Για το λόγο αυτό είναι χρήσιμη όποια πληροφορία μπορούμε να έχουμε σχετικά με το τι είναι πιθανό να υπάρχει στα δεδομένα, είτε για να δώσουμε παραπάνω πληροφορία στον αλγόριθμο που χρησιμοποιούμε ώστε να μπορέσει να την εκμεταλλευτεί, είτε να μας βοηθήσει να επιλέξουμε έναν καταλληλότερο αντικαταστάτη.



Σχήμα 1.3. Τέσσερις χαρακτηριστικές περιπτώσεις δομών ομάδων,
(α) δακτύλιοι χωρίς τομές, (β) σπειροειδείς, (γ) ελλείψεις, (δ) σφαίρες.

Στο Σχήμα 1.3 φαίνεται η σημαντικότητα των επιλογών που κάνουμε για την επίλυση των σχετικών προβλημάτων. Παράδειγμα λανθασμένης επιλογής αλγορίθμου-

δομής αναζήτησης είναι το (α), ενώ μία λάθος εκτίμηση της πολυπλοκότητας του προβλήματος που επιθυμούμε να λύσουμε είναι η (γ). Εκεί ο αλγόριθμος συμπεραίνει λανθασμένα τον διαχωρισμό μίας καλά δομημένης ομάδας σε τρεις μικρότερες.

Παρατηρεί κανείς ότι και ο ορισμός της πολυπλοκότητας και αυτό της εύρεσης των μορφών των ομάδων είναι συναφή προβλήματα τα οποία μπορούν να μελετηθούν μέσω μεθόδων ομαδοποίησης, διότι δεν απαιτούν καμία εκ των προτέρων γνώση για τα χαρακτηριστικά των κατηγοριών που αναζητούμε.

Ένα καθόλου απίθανο ενδεχόμενο είναι να διαθέτουμε ένα μεγάλο σύνολο κειμένων, για το οποίο έχουμε την πληροφορία μόνο ενός πολύ γενικού διαμερισμού, π.χ. σε 3-4 ομάδες. Τα κείμενα όμως διαχωρίζονται περαιτέρω σε πολλές υποκατηγορίες, πράγμα που σημαίνει πως θα εξυπηρετούσε να τα ομαδοποιήσουμε σε περισσότερες από τις γενικές ομάδες. Αυτό πετυχαίνεται λύνοντας επαναληπτικά το πρόβλημα αυξάνοντας τον αριθμό των ομάδων του τελικού διαμερισμού.

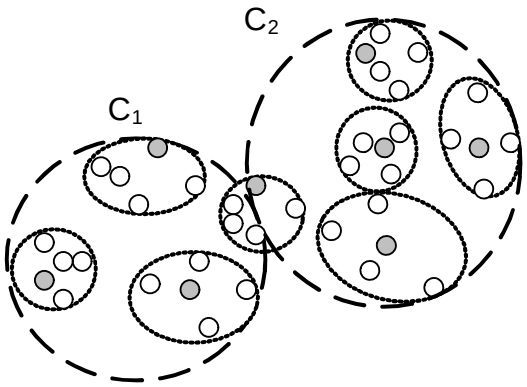
Στους ιεραρχικούς αλγόριθμους ομαδοποίησης παράγονται όλες οι λύσεις σε ένα ιεραρχικά εμφωλευμένο σχήμα. Παράλληλα ελέγχονται διάφοροι δείκτες οι οποίοι περιγράφουν την ποιότητα της ομαδοποίησης, έτσι ώστε να αποφανθούμε για μία καλή λύση και για τον αριθμό των ομάδων της. Ένα παράδειγμα ελέγχου της πολυπλοκότητας του μοντέλου είναι ο αλγόριθμος συσσώρευσης (αντίστοιχα διαμερισμού) να παύει την συνένωση ομάδων όταν αυτές παρουσιάζουν ομοιότητα κατώτερη ενός κατωφλίου [8].

Ο Milligan et.al [71] παρουσίασε την πρώτη ολοκληρωμένη πειραματική σύγκριση ποσοτήτων για την εκτίμηση του αριθμού των ομάδων η οποία αποτέλεσε σταθμό για την περαιτέρω διερεύνηση του προβλήματος. Στη συνέχεια υπήρξε μικρή εξέλιξη στο πρόβλημα [72][73], ενώ τα τελευταία χρόνια δύο προσεγγίσεις φαίνονται να κερδίζουν το ενδιαφέρον. Η πρώτη αφορά τη *Gap statistic* [74] και η δεύτερη τη *Stability-based* προσέγγιση [75].

1.5.2.3. Η Πληροφορία των Γειτονιών Κοντινότερων Δεδομένων

Οι γειτονιές μπορούν να θεωρούνται ένα επίπεδο οργάνωσης των δεδομένων πάνω από το επίπεδο των απομονωμένων σημειακών δεδομένων και ένα επίπεδο κάτω από τις δομές των ομάδων που αναζητούμε.

Η *K*-γειτονιά ενός αντικειμένου ορίζεται από τα *K* κοντινότερα αντικείμενα σε αυτό. Ως έννοια προέρχεται από τα προβλήματα κατηγοριοποίησης όπου έχει μελετηθεί εκτενώς. Έχοντας ένα σύνολο εκπαίδευσης, κάθε άγνωστο πρότυπο ταξινομείται στην κατηγορία την οποία υποδεικνύει η πλειοψηφία από τους *K*-κοντινότερους γείτονες του (*K-NN*) στο σύνολο εκπαίδευσης. Μια παραλλαγή της τεχνικής αυτής είναι η απόφαση να λαμβάνεται ζυγίζοντας την υπόδειξη κάθε αντικειμένου της γειτονιάς βάσει της απόστασης του από το άγνωστο πρότυπο, δίνοντας έμφαση στους κοντινότερους από τους *K* γείτονες. Το *K* είναι παράμετρος και παίζει γενικά σημαντικό ρόλο ανάλογα και των χαρακτηριστικών των δεδομένων.



Σχήμα 1.5. Η οργάνωσης των δεδομένων σε ένα υψηλό επίπεδο δύο ομάδων και σε ένα χαμηλότερο επίπεδο 4-*NN* για τα χρωματισμένα πρότυπα.

Τα προβλήματα κατηγοριοποίησης υποθέτουν ουσιαστικά τη συνέπεια σε επίπεδο γειτονιάς (*K-NN consistency*) και προσπαθούν να εξάγουν αποφάσεις αντλώντας πληροφορία από αυτό το επίπεδο οργάνωσης των δεδομένων. Από την άλλη πλευρά, αυτή η τοπική πληροφορία δεν έχει μελετηθεί εκτενώς στα προβλήματα ομαδοποίησης. Αν εξαιράισουμε τις μεθόδους που προσανατολίζονται πλήρως στο να αντλήσουν πληροφορία από την πυκνότητα των δεδομένων στο χώρο (*density-based*) και εν μέρει την ιεραρχική συσσωρευτική μέθοδο (*agglomerative*), η οποία στα αρχικά βήματα συνενώνει κοντινά αντικείμενα, συνήθως οι αλγόριθμοι ομαδοποίησης δεν εκμεταλλεύονται άμεσα την πληροφορία των γειτονιών αλλά ουσιαστικά υπεισέρχεται στη διαδικασία από τις ίδιες τις ιδιότητες των δεδομένων.

Αφού οι ομάδες ορίζονται ως συνεκτικά υποσύνολα δεδομένων και οι γειτονιές είναι ακόμα μικρότερα συνεκτικά υποσύνολα (έστω *K* πολύ μικρότερο από το μέγεθος του συνόλου δεδομένων, $K << N$), αναμένει κανείς πως οι περισσότερες γειτονιές θα

εμπεριέχονται ολόκληρες σε μία ομάδα του τελικού αποτελέσματος (Σχήμα 1.5). Μια ακόμα χρήσιμη παρατήρηση είναι πως η ύπαρξη ομάδων στα δεδομένα υπονοεί την ύπαρξη συνέπειας σε επίπεδο γειτονιών και αντίστροφα. Μάλιστα, το πόσο διαχωρισμένες είναι οι ομάδες είναι κατά κάποιο τρόπο ανάλογο της ποιότητας των γειτονιών, αφού η ύπαρξη τομών ορίζει και ασυνεπείς γειτονιές δεδομένων.

Στην εργασία αυτή θα μας απασχολήσει σε μεγάλο βαθμό το επίπεδο οργάνωσης των γειτονιών στα δεδομένα για την ομαδοποίηση κειμένων. Θα ενσωματώσουμε την πληροφορία αυτή στο φιλτράρισμα και στην ομαδοποίηση των κειμένων. Αυτό που αποτελεί πρόκληση είναι να χρησιμοποιήσουμε την τοπική πληροφορία σε μεθόδους οι οποίες την αγνοούν, όπως π.χ η οικογένεια αλγορίθμων K-Μέσων.

1.6. Δομή Συστημάτων Ομαδοποίησης Κειμένων

Κάθε τεχνική ομαδοποίησης μπορεί να αναλυθεί και να μελετηθεί ανεξάρτητα σε πέντε βασικούς άξονες τους οποίους θα μελετήσουμε στα επόμενα κεφάλαια ως ανεξάρτητες διαδικασίες αλλά και ως αλληλεπιδρώντα υποστοιχεία, Οι άξονες είναι:

1. το μοντέλο εξόρυξης δεδομένων από την συλλογή πληροφοριών (*information retrieval model*), το οποίο χειρίζεται την πηγαία μορφή δεδομένων ορίζοντας κατάλληλες διαδικασίες «άντλησης» πληροφορίας.
2. το μοντέλο αναπαράστασης δεδομένων (*data representation model*), το οποίο φέρει τα χαρακτηριστικά των φυσικών δεδομένων εισόδου.
3. το μέτρο ομοιότητας (*similarity measure*), το οποίο είναι ένας τρόπος συσχέτισης των δεδομένων.
4. το μοντέλο ομάδων (*cluster model*), που αφορά τα δομικά χαρακτηριστικά των επιθυμητών ομάδων.
5. τον αλγόριθμο ομαδοποίησης (*clustering technique*), ο οποίος διαμορφώνει τις ομάδες στηριζόμενος στα προηγούμενα.

Η διαδικασία εξαγωγής αποτελέσματος που ακολουθεί ένα σύστημα ξεκινά με την είσοδο των δεδομένων και τερματίζει έχοντας καταλήξει σε ένα διαμερισμό των κειμένων σε ομάδες. Τα επιμέρους στάδια είναι:

1. Στάδιο εξόρυξης πληροφορίας και προεπεξεργασίας κειμένων (*information retrieval and preprocessing*), όπου αφού αναγνωρίζεται η χρήσιμη πληροφορία

στα δεδομένα εισόδου, προεπεξεργάζεται και αναπαρίσταται στη μνήμη με κατάλληλες δομές δεδομένων (π.χ. διανύσματα, γραφήματα).

2. Βασικό στάδιο εφαρμογής αλγόριθμου ομαδοποίησης, το οποίο σχετίζεται με τους υπόλοιπους τρεις άξονες (3-5). Εκτελείται επαναληπτικά κάποιος αλγόριθμος ομαδοποίησης, προσαρμοσμένος στην αναπαράσταση που έχει επιλεγεί για τα δεδομένα.
3. Στάδιο βελτίωσης λύσης, όπου προαιρετικά εφαρμόζεται κάποια αλγοριθμική τεχνική, η οποία μπορεί να διαφέρει από αυτή του τρίτου σταδίου, επιχειρώντας να βελτιώσει το τελικό αποτέλεσμα. Το στάδιο αυτό μπορεί να εφαρμοστεί επαναληπτικά χρησιμοποιώντας τεχνικές διαφορετικών χαρακτηριστικών.

1.7. Οργάνωση της Διατριβής

Η διάθρωση της εργασίας έχει ως εξής: στο Κεφάλαιο 2 συζητούνται τα ιδιαίτερα χαρακτηριστικά των κειμένων φυσικής γλώσσας και παρουσιάζεται μία προτεινόμενη αλγοριθμική διαδικασία για την παραγωγή συνθετικών δεδομένων. Σε γενικές γραμμές η ροή της υπόλοιπης εργασίας συνδέεται άμεσα με την διαδοχή των σταδίων της ομαδοποίησης και των εμπλεκόμενων υποστοιχείων της Παραγράφου 1.5.

Στο Κεφάλαιο 3 παρουσιάζεται η διαδικασία προεπεξεργασίας των δεδομένων εισόδου και το φιλτράρισμα που βασίζεται στην τοπική πληροφορία των δεδομένων, το οποίο είναι μία τεχνική που αναπτύχθηκε στα πλαίσια της εργασίας. Στο Κεφάλαιο 4 παρουσιάζονται και αναλύονται δύο μοντέλα αναπαράστασης: το κλασσικό διανυσματικό που αναπαριστά μόνο της λέξεις ενός κειμένου και η προτεινόμενη γενικευμένη γραφική αναπαράσταση η οποία ενσωματώνει στις ακμές διάφορες σχέσεις μεταξύ των όρων. Στο Κεφάλαιο 5 γίνεται εκτενής αναφορά στα μέτρα ομοιότητας τα οποία χρησιμοποιούνται στα προβλήματα επεξεργασίας κειμένων, εστιάζοντας στις ιδιότητες που κρίνουν την καταλληλότητά τους για το πρόβλημα.

Στο Κεφάλαιο 6 παρουσιάζονται οι αλγοριθμικές τεχνικές ομαδοποίησης που θα εφαρμόσουμε καθώς και μία πρότυπη τεχνική ομαδοποίησης η οποία τροποποιεί τον γνωστό αλγόριθμο Κ-Μέσων. Η τεχνική αυτή ενσωματώνει μία δυναμική διαδικασία επιλογής χαρακτηριστικών στην βασική αλγοριθμική διαδικασία ομαδοποίησης των κειμένων, η οποία έχει ως στόχο τη μείωση της επίδρασης των όρων θορύβου ώστε να καταστεί πιο ευέλικτη η διαδικασία εκπαίδευσης.

Τέλος, στο Κεφάλαιο 7 παρουσιάζεται μία εκτενής πειραματική διαδικασία που αφορά το φιλτράρισμα, τα μοντέλα αναπαράστασης δεδομένων και τους αλγόριθμους ομαδοποίησης και στο 8^{ο} και τελευταίο Κεφάλαιο συζητούνται συνοπτικά τα αποτελέσματα και συμπεράσματα της εργασίας.

ΚΕΦΑΛΑΙΟ 2. ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΚΕΙΜΕΝΩΝ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Στο κεφάλαιο αυτό θα αναλύσουμε τις ιδιαιτερότητες των κειμένων φυσικής γλώσσας. Αν και είναι δυνατό να αντιμετωπίσουμε το πρόβλημα της ομαδοποίηση κειμένων χρήσει προσεγγίσεων γενικού σκοπού (αλγορίθμων, μοντέλων αναπαράστασης, μέτρων ομοιότητας), η ανάλυση όμως των ειδικών συνθηκών που επικρατούν στο πρόβλημα που επιθυμούμε να επιλύσουμε , μας επιτρέπει την καλύτερη κατανόηση:

- της συμπεριφοράς και απόδοσης των διαφόρων επιλογών τις οποίες κάνουμε για κάθε ένα από τα στάδια επίλυσης του προβλήματος (προεπεξεργασία, τεχνική μάθησης κ.α.), καθιστώντας δυνατή την ερμηνεία των πειραματικών παρατηρήσεων χωρίς να περιοριζόμαστε στην απλή καταγραφή τους.
- του ίδιου του προβλήματος σε βάθος, είναι η βάση η οποία ενδεχομένως να μας οδηγήσει σε μία αναζήτηση στρατηγικών και τεχνικών οι οποίες θα μπορούσαν να αποδώσουν καλύτερα, προσαρμόζοντας τις γενικές αρχές των τεχνικών επίλυσης στα προκείμενα ειδικά χαρακτηριστικά προβλήματος.

Ένας ακόμα στόχος που τίθεται είναι να παρουσιάσουμε ένα μοντέλο με το οποίο να είναι δυνατή η παραγωγή ρεαλιστικών συνθετικών δεδομένων για το πρόβλημα της ομαδοποίησης κειμένων. Τα υπαρκτά μοντέλα παράγουν ένα ενιαίο κείμενο ασυμπτωτικά μεγάλο. Στη δική μας περίπτωση το ζητούμενο είναι να παράγουμε δεδομένα ως ξεχωριστά κείμενα, τα οποία πέραν της ρεαλιστικής εσωτερικής τους στατιστικής δομής, θα πρέπει να επιτρέπουν τον σχηματίζουν ομάδες ρυθμιζόμενης συνοχής.

2.1. Βασικά Χαρακτηριστικά Φυσικών Γλωσσών

Οι φυσικές γλώσσες είναι περίπλοκοι κώδικες οι οποίοι είναι ικανοί να κωδικοποιήσουν και να μεταφέρουν μη τετριμμένη πληροφορία. Τέτοια παραδείγματα είναι η φυσική γλώσσα την οποία χρησιμοποιούν στην επικοινωνία τους οι άνθρωποι, αλλά και γλώσσες που προϋπήρχαν αυτής όπως το γενετικό υλικό. Στην περίπτωση αυτή με ένα σύστημα απλούστατων και περιορισμένων σε αριθμό συμβόλων κωδικοποιούνται τα χαρακτηριστικά των οργανισμών και οι «βιολογικές εντολές» για την αναπαραγωγή τους.

Η φυσική γλώσσα (εφεξής αναφερόμαστε στην ανθρώπινη) εξελίσσεται δυναμικά στο χρόνο (πλούτος λεξιλογίου, ευελιξία και ακρίβεια λόγου). Διατηρεί υψηλές δυνατότητες κωδικοποίησης πληροφορίας, αν και περιέχει ασύγκριτα περισσότερο «θόρυβο» στην πληροφορία που μεταφέρει σε σχέση με το γενετικό υλικό. Διαπιστώνει κανείς ότι τα δεδομένα που χειριζόμαστε δε μπορούν να αναπαρασταθούν με απλουστευτικά μοντέλα που αγνοούν την πολυπλοκότητα του ανθρώπινου λόγου. Ακόμα και όταν χρησιμοποιούμε αναντίστοιχα σχήματα αναπαράστασης, αυτό περισσότερο έχει να κάνει με την αδυναμία μηχανικής αποκωδικοποίησης της πολυπλοκότητας της φυσικής γλώσσας.

2.2. Προβλήματα στην Μηχανική Αποκωδικοποίηση της Φυσικής Γλώσσας

2.2.1. Η Μεγάλη Διάσταση του Προβλήματος

Πρώτο σημαντικό ζήτημα είναι η πολύ μεγάλη διάσταση τους, δηλαδή το μέγεθος του λεξιλογίου τους το οποίο είναι της τάξης των δεκάδων χιλιάδων όρων. Αυτή οφείλεται

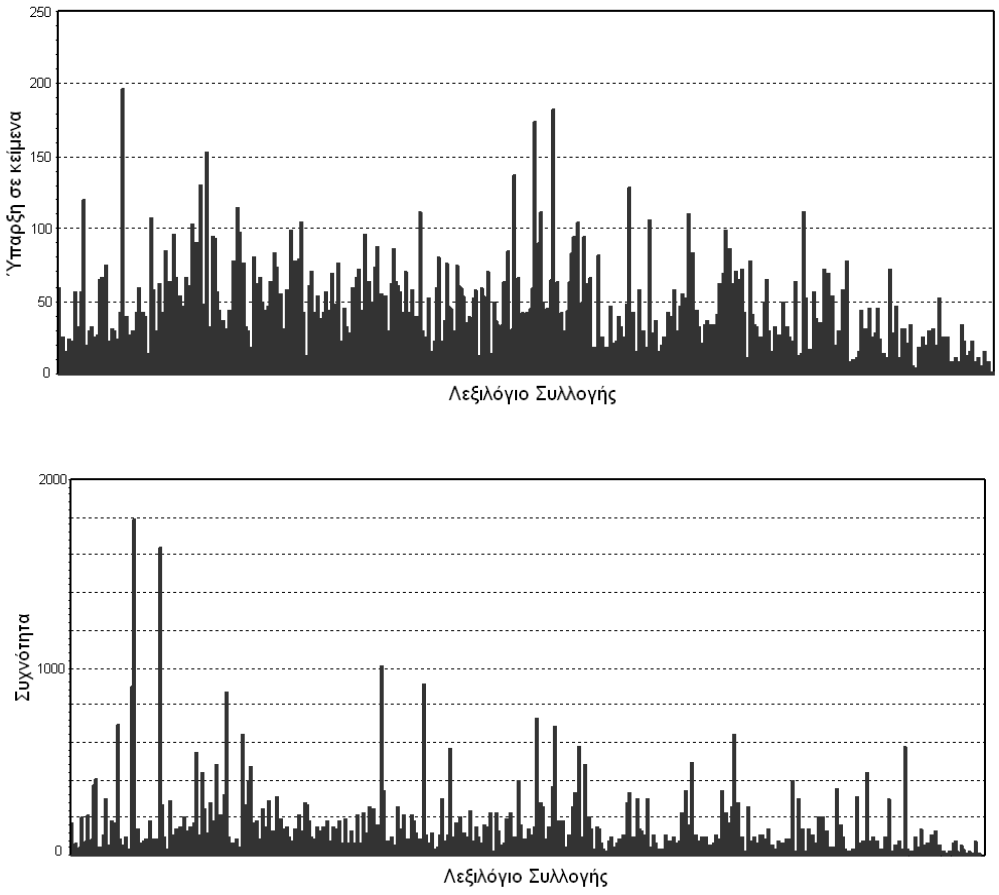
στα πολλά κείμενα που αποτελούν το σύνολο δεδομένων και στο μεγάλο λεξιλόγιο που διαθέτουν. Ο άνθρωπος από τη μία δεν καταφέρνει να κωδικοποιήσει σε πολύ πυκνές εκφραστικές δομές το νόημα που θέλει να μεταδώσει κι από την άλλη δε γράφει με στόχο να μπορεί ένα αυτόματο σύστημα να αναγνωρίσει το θέμα γραφής του. Υπάρχουν, για παράδειγμα, γραπτά τα οποία δεν έχουν απλά σκοπό την κωδικοποίηση της πληροφορίας, μπορεί να είναι διδακτικά, τεχνικής αναλυτικής περιγραφής, φιλοσοφικά, πολιτικά, θρησκευτικά κ.α.

Ακόμα, παρά τις πολλές διαστάσεις, τα δεδομένα δε μας παρέχουν πληροφορία για όλες αυτές. Τις περισσότερες φορές εξάγουμε το αυθαίρετο συμπέρασμα ότι ένα κείμενο δε σχετίζεται με έναν όρο όταν αυτός δεν παρουσιάζεται στο σώμα του. Σε μεγάλες συλλογές παρατηρείται κάθε κείμενο να αναφέρει το πολύ το 1% του λεξιλογίου ολόκληρης της συλλογής, το οποίο όπως είπαμε μπορεί να έχει χιλιάδες λέξεις. Κάποιες φορές δε, αν και ο αριθμός των κειμένων είναι μεγάλος, έχουμε λίγα δεδομένα αναλογικά με το λεξιλόγιο και τον αριθμό των ομάδων που αναζητούνται. Αυτό σημαίνει πολλαπλό πρόβλημα: πολλές διαστάσεις και ουσιαστικά ελλιπής πληροφορία για την πλειοψηφία των χαρακτηριστικών σε κάθε κείμενο. Στο Σχήμα 2.1 παρουσιάζονται τα ιστογράμματα για τη συλλογή *U*, αφού έχουμε αφαιρέσει κάποιες τετριμμένες λέξεις. Παρατηρούμε πως οι περισσότερες λέξεις εμφανίζονται σε λιγότερα από 50 κείμενα από τα 309 της συλλογής, ενώ και η αθροιστική συχνότητα εμφάνισης των περισσοτέρων σε όλη τη συλλογή είναι περιορισμένη.

Δεν είναι υπερβολικό να πούμε πως η ομαδοποίηση σε τόσο μεγάλη διάσταση μπορεί να είναι ακόμα και αδύνατη, λόγω της «κατάρας της μεγάλης διάστασης» [6]. Η αύξηση των χαρακτηριστικών στα δεδομένα αυξάνει εκθετικά τον αλγεβρικό χώρο του προβλήματος, με συνέπεια να απαιτείται αντίστοιχη αύξηση στον αριθμό των δεδομένων για να μπορεί να περιγραφεί ο χώρος αυτός και να αναζητηθούν δομές στην πληροφορία που περιέχει. Όμως σε πραγματικά προβλήματα δεν είναι δυνατόν να έχουμε πάντα μεγάλους όγκους διαθέσιμων δεδομένων, αλλά ούτε και την υπολογιστική ή χρονική ευχέρεια να τα επεξεργαστούμε.

Άμεση καταστροφική συνέπεια είναι ότι χάνεται η έννοια της πυκνότητας των δεδομένων στο χώρο και συνεπώς οι δομές ομάδων που αναζητούμε. Όταν περισσότερα χαρακτηριστικά συνυπολογίζονται στις αποστάσεις τότε αυτές γίνονται πιο ομοιόμορφες στο σύνολο δεδομένων. Για να αντιμετωπιστεί το πρόβλημα της

διάστασης, έχουν εφαρμοστεί κλασσικές τεχνικές μείωσης διάστασης αλλά και εξειδικευμένες τεχνικές για κείμενα.



Σχήμα 2.1. Επάνω: ιστόγραμμα αριθμού κειμένων που εμφανίζεται μία λέξη (συλλογή *U*, 10 ομάδες, 309 κείμενα, 9916 διαφορετικές λέξεις), κάτω: ιστόγραμμα αθροιστικών συχνοτήτων των λέξεων της ίδιας συλλογής.

2.2.2. Η Σύνθετη Νοηματική Δομή Περιεχομένου

Τα συμφραζόμενα σε ένα κείμενο έχουν πολύ μεγάλη σημασία για την μηχανική κωδικοποίηση και αναπαράσταση του νοήματός του. Το πεδίο επεξεργασίας φυσικής γλώσσας, *NLP*, και άλλοι ειδικοί στη Στατιστική Γλωσσική Ανάλυση (*Statistical Linguistics*) προσπαθούν να αντιμετωπίσουν αυτή τη διάσταση του προβλήματος.

Οι οντολογίες, ένα σύγχρονο πεδίο έρευνας, είναι συλλογές λεκτικών και φραστικών συσχετίσεων οργανωμένες σε ένα ιεραρχικό σχήμα που αλληλοκαθορίζουν το νοηματικό περιεχόμενό τους, ενώ από κοινού καθορίζουν τα ανώτερα νοηματικά

επίπεδα. Είναι ένα χρήσιμο εργαλείο για την ακριβέστερη μηχανική ανάλυση γλωσσικών δεδομένων και τη βελτίωση της ποιότητας των δεδομένων εισόδου.

Η σχέση που αναπτύχθηκε ανάμεσα στις τεχνικές μάθησης σε δεδομένα φυσικής γλώσσας με το πεδίο των οντολογιών είναι πρόσφατη και ως φαίνεται μπορεί να βοηθήσει και τα δύο πεδία. Οι πιο μοντέρνες προσεγγίσεις που στοχεύουν στην αυτόματη ή ημιαυτόματη κατασκευή οντολογιών βασίζονται στην ομαδοποίηση, η οποία χωρίς επίβλεψη μπορεί να ανακαλύψει σχέσεις στα γλωσσικά δεδομένα [69]. Η επίβλεψη στις μεθόδους ομαδοποίησης εδώ δεν έχει να κάνει με το πώς παράγεται μία λύση, αλλά με την επιβεβαίωση των αποτελεσμάτων της από τον άνθρωπο, ώστε αυτά να διαθέτουν την απαραίτητη ορθότητα.

2.2.3. Γλωσσικά Φαινόμενα

Τρία γλωσσικά φαινόμενα που εμφανίζονται σε κάθε γλώσσα, τα οποία μπορούν να κάνουν να αποτύχει οποιαδήποτε απλουστευτική μέθοδος είναι: η *πολυσημία*, η *ομωνυμία* και οι *σύνθετες φράσεις*. Η πολυσημία είναι το φαινόμενο όπου ένα όρος έχει τελείως διαφορετικό νόημα ανάλογα με τα συμφραζόμενα της εμφάνισής του σε ένα κείμενο. Έτσι, σε περίπτωση ύπαρξης περισσότερων από μία ομάδων στα δεδομένα, οι οποίες στηρίζονται σε τέτοιες λέξεις είναι μοιραίο να αποτύχει μία μέθοδος που ταυτίζει αλφαριθμητικά λέξεων. Παραδείγματα από τον χώρο της Πληροφορικής μπορεί να βρει κανείς πολλά: fork, pipe, disk, memory, bottleneck κ.α. έχουν ειδική σημασία στο πεδίο και τελείως διαφορετική στην καθημερινή χρήση τους.

Λέξη	Συνηθισμένη σημασία	Δευτερεύουσα σημασία
right	κατεύθυνση δεξιά	πολιτικός όρος
interest	ενδιαφέρον	τόκος δανείου
bank	τράπεζα	όχθη ποταμού
mine	κτητική αντωνυμία (μου)	νάρκη, ορυχείο
cold	χαμηλή θερμοκρασία	κρύωμα , ρίγος

Σχήμα 2.2. Χαρακτηριστικά παραδείγματα πολυσημίας.

Η ομωνυμία είναι κάθετο πρόβλημα, όπου περισσότεροι του ενός όρων καλύπτουν την ίδια έννοια ή βάσει των συμφραζομένων αποκτούν ταυτόσημο νόημα, π.χ. car, auto, vehicle ή street, avenue, highway. Ένα σχετικό πρόβλημα με την ομωνυμία είναι

διαφορετικές λέξεις να έχουν το ίδιο νοηματικό περιεχόμενο (*burstiness*) , π.χ. οι μάρκες αυτοκινήτων *VW, BMW* σε ένα κείμενο ενδέχεται να εννοούν το ίδιο πράγμα.

Οι σύνθετες φράσεις αφορούν όρους που αποτελούνται από περισσότερες από μία ανεξάρτητες λέξεις, π.χ. *Olympic Games, New York, city block, data mining, machine learning* κ.α. Το φαινόμενο είναι πολύ συχνό και αντιμετωπίζεται είτε με ανθρώπινη επέμβαση, δημιουργώντας μία λίστα σχετική με ένα θέμα, είτε μηχανικά, είτε συνδυάζοντας τα δύο προηγούμενα. Στην δεύτερη περίπτωση αφού συλλέγονται πληροφορίες για την πιθανότητα δύο διαδοχικές λέξεις να σχηματίζουν έναν σύνθετο όρο (π.χ. σε πολλά κείμενα συναντάται το *New York*, τα κείμενα αυτά μπορεί να είναι και διαφορετικά από αυτά που επιθυμούμε να οργανώσουμε) ελέγχονται τα συμφραζόμενα ώστε να αναγνωριστεί ή όχι ο σύνθετος όρος.

Και τα τρία παραπάνω προβλήματα είναι αντικείμενα του πεδίου *NLP* και αφορούν ιδιαίτερα το πρόβλημα Εννοιολογικού Καθορισμού Λέξεων (*Word Sense Disambiguation - WSD*). Μια προσέγγιση πολύ στενά συνδεδεμένη με την ομαδοποίηση κειμένων είναι να συλλέγονται δείγματα κειμένων (ή και ολόκληρα τα κείμενα) στα οποία περιέχεται ένας όρος και στη συνέχεια να ομαδοποιούνται ώστε να αναγνωρίζονται εννοιολογικά διαφορετικές χρήσεις του όρου [25]. Το σημαντικότερο πρόβλημα είναι αυτό που συζητήθηκε στο εισαγωγικό Κεφάλαιο, το ότι δε γνωρίζουμε πόσες διαφορετικές χρήσεις έχει ένας όρος ώστε να ορίσουμε τον αριθμό των ομάδων συμφραζομένων που αναζητούμε.

Πολλές προσεγγίσεις για την αντιμετώπισή των προβλημάτων που ανακύπτουν στα κείμενα χρησιμοποιούν το γνωστό εργαλείο *WordNet* για να προσδιορίσουν καλύτερα το εννοιολογικό περιεχόμενο των λέξεων στα κείμενα [1][10][18][25]. Μία πρόσφατη προσπάθεια να ταυτιστούν συνώνυμες και ομώνυμες λέξεις των κειμένων με μία προκαθορισμένη μορφή, δεν απέφερε σημαντική βελτίωση στα αποτελέσματα της ομαδοποίησης [20], παρότι δοκιμάστηκε και το λεκτικό αναγνωριστικό (*offset*) και το νοηματικό χαρακτηριστικό (*senseno*) που παρέχει το *WordNet*.

2.3. Η Αδυναμία Εφαρμογής Κλασικών Μεθόδων Ανάλυσης

Για να αναλύσουμε όμως τα ενδιαφέροντα χαρακτηριστικά των κειμένων θα πρέπει να χρησιμοποιήσουμε κάποια εργαλεία ανάλυσης και μελέτης πάνω στα δεδομένα. Εδώ όμως ερχόμαστε αντιμέτωποι με το μεγαλύτερο πρόβλημα που συναντά κανείς στο

πεδίο. Πέραν από τις συχνότητες των λέξεων (ως βάση αλλά και άλλων στατιστικών μεγεθών που αφορούν αυτές), και πιθανόν μίας ποσότητας πληροφορίας που αντλείται από την ακριβή παράθεσή τους στο κείμενο (φράσεις, προτάσεις, παράγραφοι) δεν είναι εύκολο να κωδικοποιήσουμε πολλά παραπάνω χαρακτηριστικά.

Δε μπορούμε επίσης να εφαρμόσουμε κλασσικές μαθηματικές μεθόδους για τον υπολογισμό ποσοτήτων, οι οποίες σε άλλα πεδία θεωρούνται στοιχειώδεις πληροφορίες που λαμβάνονται από τα δεδομένα. Ακόμα και σε μη συμβατικούς τύπους πληροφορίας όπως οι ψηφιακές εικόνες, λόγω της ύπαρξης διάταξης των εικονοστοιχείων στο δισδιάστατο χώρο μας επιτρέπεται κάτι τέτοιο (οι παράγωγοι λογίζονται ως διαφορές της συνάρτησης φωτεινότητας $I(i,j)$ για γειτονικά εικονοστοιχεία).

Αντίθετα στα κείμενα αδυνατούμε να υπολογίσουμε ολοκληρώματα και παραγώγους αφού η διάταξη της πληροφορίας δεν έχει αυστηρή σημασία. Η μελέτη που μπορούμε να εφαρμόσουμε περιορίζεται λοιπόν στη στατιστική ανάλυση της φυσικής γλώσσας και των κειμένων, η οποία μας επιτρέπει να προσεγγίσουμε περισσότερο μακροσκοπικά τη δομή αλλά και τους μηχανισμούς παραγωγής της.

2.4. Στατιστική Ανάλυση Κειμένων

2.4.1. Ο Νόμος του Zipf

Μία από τις πλέον γενικές στατιστικές ιδιότητές, η οποία εμπειρικά έχει επιβεβαιωθεί εδώ και πολλά χρόνια, είναι η συμμόρφωση του γραπτού Λόγου με την κατανομή *Zipf* [78]. Πρόκειται για μία διακριτή κατανομή πιθανοτήτων που οφείλεται αρχικά στον Kinsey Zipf, ενώ η γενίκευσή της στο νόμο που καλείται *Zipf-Mandelbrot* οφείλεται στον γνωστό μαθηματικό Benôit Mandelbrot.

Ο νόμος αυτός μπορεί να διατυπωθεί μαθηματικά με τη μορφή σχέσης μεταξύ της συχνότητας μίας λέξης σε ένα κείμενο και της θέσης της (*rank*), αν διατάξουμε όλο το λεξιλόγιο του κειμένου κατά φθίνουσα σειρά ως προς τις συχνότητες. Αν λοιπόν έχουμε ένα κείμενο με μήκος T (παράθεση του λεξιλογίου με επαναλήψεις όρων), τη θέση ενός όρου r στη διάταξη και την συχνότητα της $n(r)$ τότε ορίζουμε την κανονικοποιημένη συχνότητα μίας λέξης:

$$f(r) = \frac{n(r)}{T}.$$

Ο νόμος του *Zipf* με παραμέτρους β, c εκφράζεται ως εξής:

$$f(r) = x_r = \frac{c}{r^{\beta}} \Leftrightarrow$$

$$\log x_r = \log \frac{c}{r^{\beta}} \Leftrightarrow \log x_r = -\beta \log r + \log c,$$

όπου στην αρχική διατύπωσή του ο εκθέτης β και ο αριθμητής c είχαν τιμή τη μονάδα και η μορφή του νόμου ήταν: $f(r) \propto 1 / r$. Αυτή η διαπίστωση βασιζόταν στην παρατήρηση ότι σε ένα κείμενο μεγάλης έκτασης η δεύτερη συχνότερη λέξη έχει το ½ της συχνότητας της πρώτης, η τρίτη συχνότερη έχει το ⅓ της πρώτης κ.ο.κ. Τα πραγματικά δεδομένα δεν παρουσιάζουν ιδανική σύμπτωση με το στατιστικό μοντέλο, αυτός είναι και ο λόγος που παραμετροποιείται η κατανομή με τα β, c. Κάποιες μικρές αποκλίσεις σχετίζονται με τον λεξιλογικό πλούτο κάθε γλώσσας.

Η λογαριθμική μορφή δείχνει την ενδιαφέρουσα ιδιότητα πως υπάρχει γραμμική σχέση ανάμεσα στη συχνότητα εμφάνισης ενός όρου και τη θέση του στην φθίνουσα ακολουθία δειγμάτων. Η εύρεση των παραμέτρων του γραμμικού μοντέλου είναι ένα πρόβλημα εκτίμησης των παραμέτρων της ευθείας η οποία ταιριάζει καλύτερα στα δεδομένα (στη λογαριθμική κλίμακα). Με $E(\beta,c)$ συμβολίζουμε το τετραγωνικό σφάλμα ταιριάσματος της ευθείας με παραμέτρους β, c, δηλαδή το άθροισμα των τετραγωνικών αποστάσεων των σημείων από την ευθεία. Έτσι, με τη μέθοδο ελαχίστων τετραγώνων [77] οι παράμετροι εκτιμώνται ως εξής:

$$\log \hat{x}_r = -\beta \log r + \log c, \quad (b,c) = \arg\min_{\beta,c} \{E(\beta, c)\}$$

όπου με $\hat{x}_r$ συμβολίζεται η εκτιμώμενη συχνότητα για το r-οστό στοιχείο της διάταξης. Στο Παράρτημα παρατίθεται ο ακριβής υπολογισμός των παραμέτρων.

Διαπιστώνει κανείς πως η κατανομή χαρακτηρίζεται από «έλλειψη δικαιοσύνης», λίγα ενδεχόμενα έχουν υπερβολικά μεγάλη πιθανότητα εμφάνισης, αντίθετα με τη συντριπτική πλειοψηφία των οποίων η πιθανότητα εμφάνισης είναι πολύ μικρή. Σε τέτοιες συνθήκες ούτε ο μέσος, αλλά ούτε ο ενδιάμεσος μπορούν να περιγράψουν την κατανομή συχνοτήτων.

Ο νόμος αυτός είναι εμπειρικός και έχει επιβεβαιωθεί μόνο σε επίπεδο παρατηρήσεων για αρκετές γλώσσες. Στερείται έτσι αυστηρής θεωρητικής τεκμηρίωσης, μάλιστα οι θεωρητικές ερμηνείες που έχουν προταθεί στην βιβλιογραφία διαφωνούν αρκετά. Η πρώτη οφείλεται στον Simon ο οποίος προσομοίωσε τη

δυναμική της παραγωγής κειμένου με μία πολλαπλασιαστική στοχαστική διαδικασία [79], η οποία κατά την παραγωγή ενός μεγάλου κειμένου ασυμπτωτικά οδηγεί σε στατιστικά χαρακτηριστικά που συμμορφώνονται με την κατανομή *Zipf*. Μια δεύτερη προσέγγιση τυποποιήθηκε από τον Mandelbrot [80], αφού είχε ήδη προταθεί από τον Miller [81], και περιπλέκει ακόμα περισσότερο τα πράγματα. Συγκεκριμένα, η παρατήρηση που παρουσιάστηκε ήταν ότι μπορούμε να παράγουμε ένα κείμενο που ακολουθεί την ζητούμενη κατανομή, ακόμα και αν δημιουργούμε τυχαίες αλυσίδες χαρακτήρων που διαχωρίζονται από κενά. Μια ακόμα προσέγγιση, αφορά στην παραγωγή κειμένου μέσω στοχαστικών διαδικασιών *Markov* [84].

Αν παρατηρήσουμε, από γλωσσολογικής άποψης η πρώτη προσέγγιση δίνει εξήγηση στην δομή της φυσικής γλώσσας αφού θεωρεί ότι παράγεται από μία διαδικασία δυναμικής διαμόρφωσης περιεχομένου η οποία εξελίσσεται στο χρόνο, ενώ η δεύτερη αγνοεί κάθε δομή της φυσική γλώσσα και τη θεωρεί ένα απλό μοντέλο τυχαίας παραγωγής συμβόλων.

Η παρατήρηση της ισχύος του νόμου του *Zipf* δεν άργησε να επεκταθεί και σε άλλα πεδία. Το Διαδίκτυο είναι ένας χώρος όπου η κατανομή *Zipf* βοήθησε στην μοντελοποίηση των ραγδαία μεταβαλλόμενων συστημάτων [77][86][87]. Παραδείγματα είναι οι αιτήσεις πρόσβασης σε αρχεία, η τοπολογία και συνδεσιμότητα ανάμεσα σε κόμβους και σελίδες σε αυτό, το μέγεθος των πακέτων που διακινούνται.

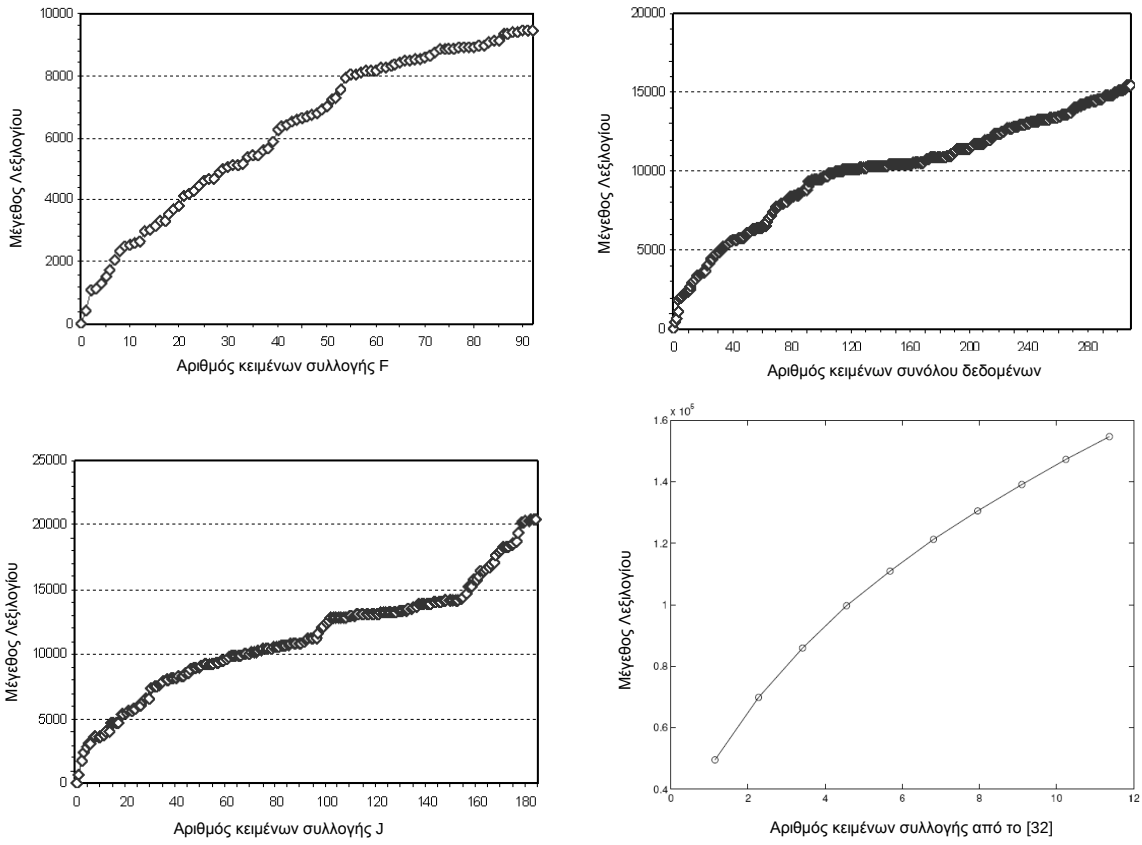
2.4.2. Ο Νόμος του Heap

Μία ακόμα σημαντική εμπειρική παρατήρηση είναι πως το λεξιλόγιο ενός κειμένου, οι διαφορετικές δηλαδή λέξεις που περιέχει, αυξάνει υπογραμμικά με το μέγεθος του. Αυτός ο νόμος καλείται νόμος του Heap [82] και έχει επίσης επιβεβαιωθεί πειραματικά [83]. Αν λοιπόν συμβολίσουμε το μέγεθος του λεξιλογίου ως V_t με t το μήκος του κειμένου, και ως α θεωρήσουμε μία παράμετρο στο [0, 1] που εκφράζει την πιθανότητα να εισάγεται ένας νέος όρος, τότε η προσέγγιση του νόμου εκφράζεται ως εξής:

$$\left|V_t\right| = a \cdot t^{\nu},\ \ \nu \in (0,1)$$

Στην περίπτωση που έχουμε εξετάσει ένα πολύ μεγάλο αριθμό $|D_t|$ ξεχωριστών κειμένων θα πρέπει να παρατηρούμε την παρακάτω σχέση:

$$\left|V_t\right| = O\left(\left|D_t\right|^{\nu}\right),\ \ \nu \in (0,\ 1).$$



Σχήμα 2.3. Η αύξηση του λεξιλογίου κατά την επεξεργασίας 4 συλλογών:.*F* (93 κείμενα, 4 ομάδες), *J* (183 κείμενα, 10 ομάδες), *U* (309 κείμενα, 10 ομάδες), συλλογή [32] (113716 κείμενα, 8 ομάδες).

Στο Σχήμα 2.3 παρουσιάζονται οι σχετικές παρατηρήσεις για τέσσερα σύνολα πραγματικών δεδομένων. Καθώς τα προεπεξεργαζόμαστε ένα-ένα παρατηρούμε πως και στα τρία σύνολα το λεξιλόγιο αυξάνεται γρηγορότερα από υπογραμμικά. Αυτό δε θα πρέπει να μας προβληματίζει, διότι οι συλλογές αποτελούνται από έχουμε γενικά λίγα κείμενα διαφορετικών θεμάτων, συνεπώς εν γένει πέρα από τις τετριμμένες λέξεις δεν υπάρχουν πολύ μεγάλες τομές λεξιλογίου, γεγονός που δεν επιτρέπει την πλήρη συμφωνία με τον εμπειρικό νόμο.

Συγκριτικά η εικόνα που μας δίνεται, μεταξύ ενός μικρού συνόλου 93 κειμένων και δύο μεγαλυτέρων δείχνει πως στα δύο μεγαλύτερα σύνολα έχουμε μία σαφή κάμψη μετά τα μέσα της επεξεργασίας, δηλαδή την σταδιακή μείωση του ρυθμού αύξησης του λεξιλογίου. Παρατηρήσεις για τον νόμο σε άλλα σύνολα μπορεί να βρει κανείς στο [32] απ’ όπου παραθέτουμε την αντίστοιχη γραφική παράσταση ενός πολύ μεγάλου

συνόλου από 113716 περιλήψεις κειμένων από 8 μεγάλες κατηγορίες, η οποία παρουσιάζει την αναφερόμενη γραμμική αύξηση λεξιλογίου.

Συμπερασματικά, ως φαινόμενο ο εμπειρικός νόμος του *Heap* μας δίνει μία εικόνα για την τάξη της πολυπλοκότητας και των απαιτήσεων σε μνήμη, των προβλημάτων μηχανικής επεξεργασίας κειμένων.

2.5. Το Μοντέλο του Simon για την Παραγωγή Κειμένων

Αυτό που προσπαθεί να ελέγξει το μοντέλο αυτό είναι ο τρόπος με τον οποίο εισάγονται νέες λέξεις στο παραγόμενο κείμενο. Έστω ότι βρισκόμαστε στο χρονικό σημείο t έχοντας παράγει ισάριθμες λέξεις για το κείμενο (έχει δηλαδή μήκος t), θεωρούμε το σύνολο $V_t = \{w_1, w_2, \ldots, w_k\}$ το οποίο περιέχει τους όρους που έχουν ήδη επιλεγεί τουλάχιστον μία φορά κατά την διαδικασία (τρέχον λεξικό του κειμένου), και L_t όλο το διαθέσιμο λεξικό μη χρησιμοποιημένων λέξεων. Δηλαδή ισχύει: $L_t = L_0 - V_t$ και $|L_t| > 0, \forall t$, ώστε να μπορούμε να εισάγουμε οποιαδήποτε στιγμή μία λέξη που δεν έχει προστεθεί στο κείμενο.

Με μία προκαθορισμένη πιθανότητα α, η οποία παραμένει σταθερή καθ' όλη τη διαδικασία παραγωγής, κάθε επόμενη στιγμή t+1 προσθέτουμε στο κείμενο μία λέξη $w \in L_t$ που δεν έχουμε συναντήσει. Με την συμπληρωματική της πιθανότητα 1-α εισάγεται μία λέξη από το λεξιλόγιο V_t που έχει ήδη χρησιμοποιηθεί. Η τυχαία επιλογή γίνεται διαλέγοντας ομοιόμορφα ανάμεσα σε t ενδεχόμενα (μήκος κειμένου), που αντιστοιχούν στις λέξεις που έχουν ήδη εισαχθεί στο κείμενο. Κατά αυτόν τον τρόπο δημιουργείται ανταγωνισμός στην επιλογή των λέξεων, αφού η πιθανότητα εμφάνισης μίας λέξης είναι συνάρτηση των προηγούμενων εμφανίσεων της. Μπορεί να δειχθεί πως η στοχαστική διαδικασία αυτή αν αφεθεί επί μακρόν οδηγεί σε χαρακτηριστικά της κατανομής *Zipf* με παράμετρο $\beta = 1 - \alpha$. Πειραματικά διαπιστώθηκε από τον Simon πως η παράμετρος που συμφωνούσε με πραγματικά δεδομένα είναι $\alpha = 0.01$.

Τρεις είναι οι βασικές αδυναμίες του μοντέλου του Simon. Πρώτον, δε μπορεί να εξηγήσει τις περιπτώσεις όπου $\beta > 1$, όπως παρατηρείται στην Ισπανική και Αγγλική γλώσσα, ενώ δε μπορεί να παράγει κείμενα με την ταχύτερη μείωση που παρουσιάζουν τα πραγματικά δεδομένα στις μικρές συχνότητες. Τέλος, δε συμμορφώνεται με τον εμπειρικό νόμο του *Heap* διότι λόγω της μη μεταβολής της παραμέτρου α με την οποία

επιλέγεται μία νέα λέξη, το λεξιλόγιο είναι κατά μέσο όρο $|V_t| = a \cdot t$ που υποδηλώνει γραμμική αύξηση με την πάροδο του χρόνου.

2.6. Το Μοντέλο των Zannete-Montemurro για την Παραγωγή Κειμένων

Το μοντέλο αυτό [85] κάνε μία σειρά τροποποιήσεων στο μοντέλο του Simon:

- Πρώτον, εισάγει μία παράμετρο *ν* που ρυθμίζει την επιλογή νέων όρων από το λεξικό L_t ώστε να μην είναι σταθερή η πιθανότητα εισαγωγής μίας λέξης, και να εξαρτάται από το χρόνο ώστε να επιτυγχάνεται υπογραμμική αύξηση λεξιλογίου. Η πιθανότητα αυτή ορίζεται ως: $a_t = a \cdot \nu \cdot t^{\nu-1}$, όπου *α* η αρχική σταθερή παράμετρος. Ο ορισμός αυτός βασίζεται στο ρυθμό με τον οποίο θα πρέπει να εισάγονται νέοι όροι βάσει του νόμου του *Heap* (προκύπτει παραγωγίζοντας τη $|V_t| = a \cdot t^{\nu}$). Η παράμετρος *ν* εξαρτάται: α) από τον τρόπο γραφής του κειμένου, το μέγεθος του λεξιλογίου που χρησιμοποιεί και το πόσο γρήγορα περνά από ένα νοηματικό σχήμα στο επόμενο, ενώ β) εξαρτάται ακόμα περισσότερο από τη γλώσσα στην οποία είναι γραμμένο το κείμενο (*language inflection*). Οι γλώσσες όπου έχουν πλουσιότερο λεξιλόγιο (ένας όρος δεν έχει γενικά πολλές σημασίες), παρουσιάζουν μεγαλύτερη ανάπτυξη λεξιλογίου σε ένα κείμενο ως προς το μήκος του. Συνεπώς, μπορούν να περιγραφούν με μικρότερες τιμές της παραμέτρου *ν* (γενικά: $\nu = [0.85, 0.95]$).
- Η δεύτερη τροποποίηση είναι να δίνεται ξεχωριστή προσοχή στις λέξεις που μόλις εισήχθηκαν στο κείμενο, αντίθετα με το αρχικό μοντέλο που απλά θεωρούσε την πιθανότητα επιλογής ανάλογη της συμμετοχής μιας λέξης στο κείμενο. Η λογική είναι πως η νέα λέξη δεν έχει «προλάβει» να παίξει στατιστικά ρόλο ώστε να έχει σημαντική πιθανότητα να επανεπιλεγεί, παρόλα αυτά καθορίζει σε μεγάλο βαθμό το τοπικό νόημα του κειμένου. Αυτό αναγκάζει κατά κάποιο τρόπο τον συγγραφέα να την ξαναχρησιμοποιήσει. Έτσι, η πιθανότητα επιλογής του όρου ορίζεται ανάλογη του $max\{n_i, \delta_i\}$, όπου δ_i επιλέγεται η εκθετική κατανομή ως αρχική ώθηση: $P(\delta_i) = \exp(-\delta_i / \delta_\mu) / \delta_\mu$.

Με τον τρόπο αυτό μειώνεται και η πόλωση της διαδικασίας ως προς το λεξιλόγιο το οποίο επιλέγεται στα αρχικά βήματα παραγωγής. Οι λέξεις με $n_i < \delta_i$ αποκτούν ένα

προβάδισμα για επανεπιλογή, ενώ όταν πλέον ισχύει $n_i > \delta_i$ δε θα έχουμε αλλοίωση της διαδικασίας για τις υψηλότερες συχνότητες. Προωθώντας τις λέξεις με χαμηλή συχνότητα επιτυγχάνεται και η γρηγορότερη κάμψη της κατανομής στις χαμηλές συχνότητες.

2.7. Μοντέλο Παραγωγής Συνθετικής Συλλογής Κειμένων με Δομή Ομάδων

Στην παράγραφο αυτή θα περιγράψουμε μία αφαιρετική προσέγγιση την οποία προτείνουμε για την παραγωγή συνθετικών κειμένων. Παρουσιάζουμε έναν πρότυπο αλγόριθμο για το σκοπό αυτό, ώστε να παράγονται κείμενα τα οποία μπορούν στη συνέχεια να χρησιμοποιούνται για την πειραματική εφαρμογή αλγορίθμων ομαδοποίησης και κατηγοριοποίησης κειμένων.

Θα χρησιμοποιήσουμε την στοχαστική διαδικασία των *Zannete-Montemurro* η οποία εξασφαλίζει τα επιθυμητά στατιστικά χαρακτηριστικά κάθε κειμένου χωριστά, σε συνδυασμό με ένα μοντέλο παραγωγής ομάδων. Η βασική τροποποίηση την οποία θα πρέπει να κάνουμε αφορά το γεγονός ότι εδώ δε μας αρκούν τα «ρεαλιστικά» στατιστικά χαρακτηριστικά, απαιτούμε να δημιουργήσουμε και «ρεαλιστική» δομή στο σύνολο δεδομένων το οποίο θα παράγουμε. Σε κάθε περίπτωση πάντως, η προσέγγιση αυτή θα μπορούσε να βελτιωθεί ύστερα από μία εκτενέστερη πειραματική μελέτη για την ορθότερη ρύθμισης των παραμέτρων της.

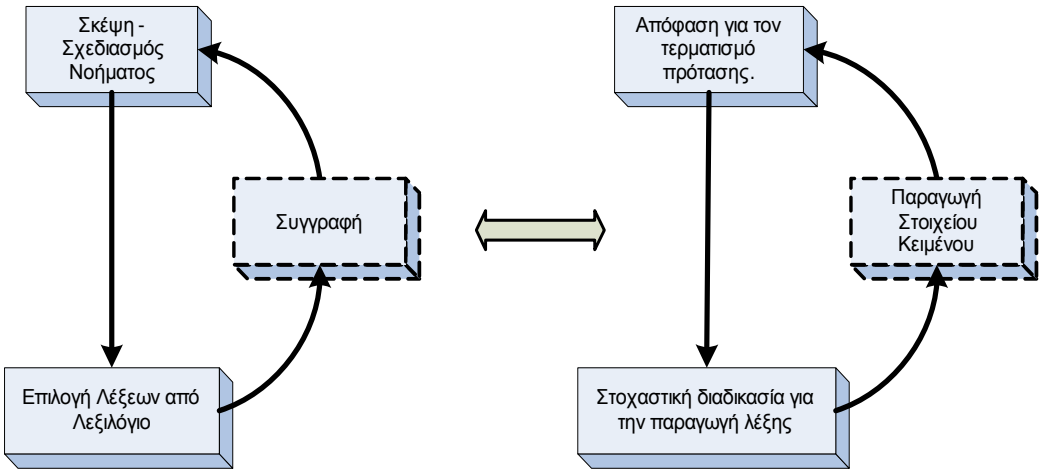
Τα κείμενα που παράγονται δεν έχουν, σαφώς, κανένα νόημα για τον άνθρωπο, παρόλα αυτά διατηρούν τα βασικά στατιστικά χαρακτηριστικά των πραγματικών κειμένων. Κατά βάση παράγεται φυσικό κείμενο, το οποίο επεκτείνεται εύκολα σε μορφές με πληροφορία ανάλογη των υπερκειμένων ή άλλων κειμένων με προσδιοριστικά σημαντικότητας.

2.7.1. Δημιουργία Κειμένου από τον Άνθρωπο

Ο άνθρωπος καθώς γράφει ένα κείμενο συνθέτει νοηματικά σύνολα λέξεων με έναν περισσότερο ιεραρχικό τρόπο. Επιλέγει λέξεις από το γνωστό του λεξιλόγιο, δημιουργεί φράσεις και προτάσεις οι οποίες αποτελούν στη συνέχει ένα κείμενο μεγαλύτερης έκτασης. Θεωρώντας ότι ένα κείμενο έχει ένα εν γένει προσδιορίσιμο θέμα, αναμένει κανείς να περιέχει λέξεις οι οποίες ανήκουν σε:

α) ένα ειδικό λεξιλόγιο το οποίο αφορά το βασικό θέμα του κειμένου,

β) ένα γενικότερο λεξιλόγιο το οποίο περιέχει τετριμμένους όρους (π.χ. άρθρα, προθέσεις, αντωνυμίες) καθώς επίσης και μη-τετριμμένους όρους οι οποίοι θα λέγαμε πως δεν είναι αρκετά ειδικοί αλλά βοηθούν τον συγγραφέα να συνθέσει λέξεις του πρώτου λεξιλογίου και να επεξηγήσει το ειδικό νόημα του κειμένου.



Σχήμα 2.4. Αριστερά: το υποθετικό μοντέλο σύνθεσης κειμένου από τον άνθρωπο, δεξιά: μία αφαιρετική προσέγγιση για την μηχανική προσομοίωση της διαδικασίας.

Αυτή είναι μία αφαιρετική διαδικασία που μπορεί να περιγράψει τη συγγραφή ενός κειμένου από τον άνθρωπο, την οποία θα προσομοιώσουμε μηχανικά. Πρόκειται για μία επαναληπτική διαδικασία η οποία δημιουργεί τμήματα κειμένου με τη βοήθεια μίας καλά ορισμένης στοχαστικής διαδικασίας. Η στοχαστική διαδικασία είναι υπεύθυνη για την τροφοδότηση της παραγωγικής διαδικασίας με λέξεις για το περιεχόμενο του κειμένου οι οποίες δεν προέρχονται απαραίτητα μόνο από ένα ειδικό θεματικό λεξικό.

2.7.2. Η Στοχαστική Διαδικασία για την Παραγωγή Όρων Κειμένου

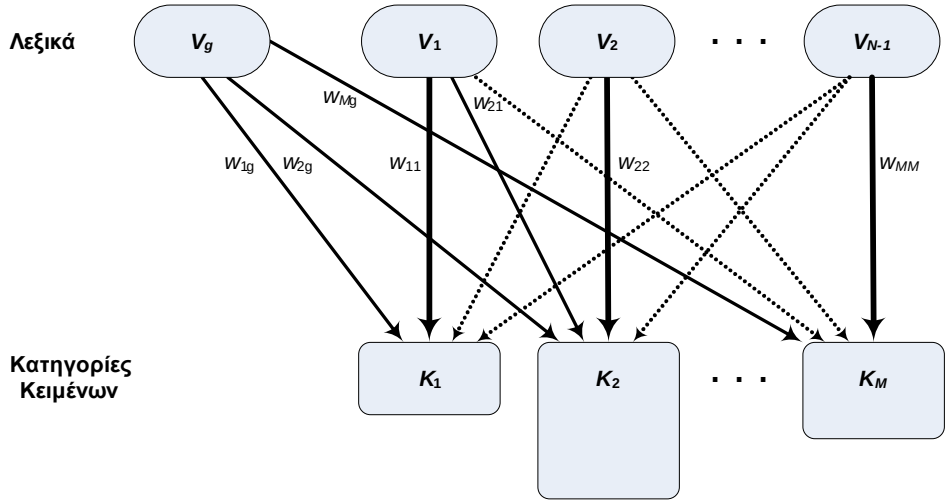
Θεωρούμε πως η συλλογή κειμένων παράγεται μέσω ενός μικτού μοντέλου, N συνιστωσών. Κάθε συνιστώσα παράγει διακριτά δείγματα από ένα ανεξάρτητο λεξιλόγιο V_i, i=1...N, και περιγράφεται από μία διαφορετική συνάρτηση πυκνότητας πιθανότητας (σ.π.π) P_i. Οι N αυτές συνιστώσες μπορούν να τροφοδοτήσουν οποιοδήποτε από τα κείμενα της συλλογής τους το ζητήσει, ώστε τελικά να παραχθεί ένα σύνολο δεδομένων από M κλάσεις διαφορετικών μεγεθών.

Η κατανομή όρων κάθε κατηγορίας εξαρτάται από το μηχανισμό μίξης των συνιστωσών παραγωγής λέξεων. Θέτουμε δύο βασικές προϋποθέσεις για τη συνέχεια, θα πρέπει να εξασφαλίζουμε πως α) η «δυσκολία» των δεδομένων θα μπορεί να καθορίζεται από τον χρήση, και β) πως αυτή θα μπορεί να έχει μη ομοιόμορφα χαρακτηριστικά. Ορίζοντας τη «δυσκολία», αυτή μπορεί να ερμηνευτεί στο χώρο του προβλήματος ομαδοποίησης, το οποίο σημαίνει την ύπαρξη τομών ανάμεσα στις διαφορετικές κλάσεις, στον χώρο των χαρακτηριστικών των δεδομένων.

Μια κλασσική προσέγγιση είναι να εισαχθεί θόρυβος σε ένα ιδανικά διαχωρίσιμο σχήμα ομάδων. Στην περίπτωσή μας ο θόρυβος μπορεί να μοντελοποιηθεί με μία σειρά επιλογών. Για να εισάγουμε ομοιόμορφο θόρυβο, υποθέτουμε πως μία από τις N συνιστώσες αποτελεί ένα γενικό λεξιλόγιο V_g, το οποίο συνεισφέρει σε κάθε κείμενο με «συνηθισμένες» λέξεις. Δεδομένου ότι όλα τα λεξικά είναι ανεξάρτητα μπορούμε να αυξήσουμε το επίπεδο του συνολικού θορύβου αυξάνοντας την επίδραση του λεξικού V_g στις κλάσεις. Έτσι, τα δεδομένα όλων των κλάσεων πλησιάζουν περισσότερο δυσκολεύοντας τον διαχωρισμό τους. Εναλλακτικά, θα μπορούσαμε να μειώσουμε σε μέγεθος του V_{g} εξαναγκάζοντας τα κείμενα σε δειγματοληψία κοινών λέξεων από ένα μικρότερο σύνολο.

Ακόμα όμως και σε αυτές τις περιπτώσεις, το αποτέλεσμα θα ήταν η ομοιόμορφη σύγχυση κάθε κλάσης με τις υπόλοιπες. Η απαίτηση για μη ομοιομορφία της δυσκολίας υποδεικνύει, τελικά, τη συμμετοχή όλων των λεξιλογίων στην παραγωγή όλων των κλάσεων, επιτρέποντας έτσι τον ορισμό της «δυσκολίας» ως: *ανά ζεύγος κλάσεων συντελεστή επικάλυψης*. Με άλλα λόγια το πόσο τα κείμενα δύο διαφορετικών κατηγοριών επιλέγουν λέξεις από κοινά σύνολα.

Στο Σχήμα 2.5 παρουσιάζεται η γραφική αναπαράσταση του μικτού μοντέλου παραγωγής κειμένων και εκδοχή την οποία έχουμε ρυθμίσει για την παραγωγή συνθετικών κειμένων M κατηγοριών. Τα βάρη συσχέτισης λεξιλογίου-κλάσης δεδομένων w_{ij} καθορίζουν την συνεισφορά του λεξικού V_j στα κείμενα της κατηγορίας K_i. Όπως διαφαίνεται από την ένταση των βελών, κάθε κατηγορία K_i θα περιέχει έγγραφα κυρίως αποτελούμενα από το ειδικό λεξιλόγιό V_i και από το κοινό λεξικό V_g. Το σχήμα αυτό καθιστά τις κλάσεις K_1, K_2 λιγότερο διαχωρίσιμες σε σχέση με οποιοδήποτε άλλο ζεύγος κατηγοριών.



Σχήμα 2.5. Μικτό μοντέλο παραγωγής κειμένων *M* κλάσεων από *N* λεξικά (οι ασθενείς ακμές υποδηλώνουν πολύ μικρές ή μηδενικές τιμές συνεισφοράς).

Μπορούμε να αναφέρουμε μερικούς ακόμα τρόπους αύξησης της δυσκολίας των δεδομένων χωρίς την εισαγωγή θορύβου:

- αύξηση των ειδικών λεξιλογίων V_i, αραιώνοντας κάθε ομάδα χωριστά,
- επιλέγοντας πιο ομοιόμορφα όρους από τα λεξικά, δηλαδή μειώνοντας την πιθανότητα σύμπτωσης όρων στα κείμενα ίδιας κατηγορίας,
- έχοντας περισσότερα ειδικά λεξικά από τις επιθυμητές κλάσεις, $M > N$, το οποίο και πάλι θα αυξάνει τη διάσταση του προβλήματος.

Παρατηρεί εύκολα κανείς πως αυτές είναι επιλογές των οποίων το αποτέλεσμα μπορεί να μοντελοποιηθεί από το συμπαγές μικτό μοντέλο που παρουσιάσαμε. Για τη συνέχεια επιλέγουμε ίσο αριθμό ειδικών συνιστωσών και παραγόμενων κατηγοριών και ένα κοινό λεξικό.

2.7.3. Συνιστώσες Παραγωγής Λέξεων

Μια δυνατότητα είναι να ορίσουμε τις συνιστώσες έτσι ώστε να παράγονται ομάδες με ιδιαίτερα σχήματα (π.χ. κυκλικές), όμως από τη θεωρητική και εμπειρική ανάλυση των κειμένων ως απαίτηση τίθεται μόνο η συμμόρφωση με την κατανομή *Zipf*.

Το μοντέλο των *Zannete-Montemurro* διαθέτει μία «μνήμη» για την παραγωγή ενός πολύ μεγάλου κειμένου και χρησιμοποιεί έναν μετρητή μήκους για το κείμενο που έχει παραχθεί (τη μεταβλητή *t*). Η ιδιαιτερότητα στην παραγωγή ενός συνόλου κειμένων με

δομή ομάδων, είναι πως τα ξεχωριστά κείμενα θα είναι γενικά μικρά και το λεξιλόγιό τους δε θα πρέπει να είναι τόσο πολωμένο από τις ήδη χρησιμοποιημένες λέξεις της ειδικής κατηγορίας (από την παραγωγή άλλων κειμένων της κατηγορίας) όσο τουλάχιστον θα ήταν αν παραγόταν ένα κείμενο με το μοντέλο *Zannete-Montemurro*. Αν μεγαλώναμε πολύ τα λεξικά, αυξάναμε την πιθανότητα επιλογής μίας νέας λέξης και προσπαθούσαμε να παράγουμε κείμενα μικρού ή μετρίου μεγέθους, τότε τα αποτελέσματα θα ήταν μη ρεαλιστικά. Θα παρατηρούσαμε ότι η αύξηση του λεξιλογίου σε ένα αρκούντως μεγάλο τέτοιο σύνολο δεδομένων θα είναι κατά πολύ μεγαλύτερη από γραμμική, τα κείμενα θα εμφάνιζαν όρους με πολύ περιορισμένες συχνότητες, και αντίστοιχες αθροιστικές συχνότητες, λέξεων για τη συλλογή. Το συνδυαστικό μοντέλο που χρησιμοποιούμε συνοπτικά έχει ως εξής:

- η παραγωγή κειμένων γίνεται ανά κατηγορία.
- το γενικό λεξιλόγιο διατηρεί τη μνήμη του καθ' όλη τη διάρκεια της διαδικασίας επιτρέποντας την εμφάνιση λέξεων υψηλού συχνοτικού περιεχομένου στα δεδομένα.
- κάθε συνιστώσα διαθέτει ξεχωριστή μνήμη για το ποιες λέξεις έχουν επιλεγεί από το υπό την ευθύνη της λεξιλόγιο, ενώ διατηρεί πληροφορία για τα δ_i όλων των όρων της που ορίζονται μία φορά.
- η παραγωγή εστιάζεται σε κάθε κείμενο χωριστά θεωρώντας τον μετρητή t ως το μήκος του κειμένου που έχει παραχθεί. Είναι ο συνδετικός κρίκος μεταξύ των διαφορετικών συνιστωσών και καθορίζει αν θα παράγουμε μία εντελώς νέα λέξη από οποιοδήποτε λεξικό.
- αφού ολοκληρώσουμε την παραγωγή ενός κειμένου και πριν την έναρξη του παραγωγής επομένου, η μνήμη κάθε ειδικού λεξικού χάνεται με εξαίρεση τα δ_i που είναι η αρχική ώθηση που δίνεται σε ένα υποσύνολο λέξεων. Αυτά μας βοηθούν να επαναρχικοποιηθεί η συνιστώσα με μία κοινή μικρή πόλωση.
- για την ρεαλιστική παραγωγή κειμένων μεταξύ της αρχικοποίησης ενός λεξικού και της παραγωγής όρων από αυτό παρεμβάλλεται μία διαδικασία προθέρμανσης.

Το μοντέλο των *Zannete-Montemurro*, όπως αναφέραμε, διαθέτει μία «μνήμη» για την παραγωγή του κειμένου, επομένως μπορεί να αμφιβάλλει κάποιος για το αν τελικά

με περισσότερες από μία τέτοιες διαδικασίες μπορούμε να παράγουμε δεδομένα τα οποία ακολουθούν μία *Zipf* ως ενιαίο σύνολο δεδομένων. Παρόλα αυτά μπορούμε να αντιληφθούμε διαισθητικά πως το πρόβλημα έγκειται στην ανεξαρτησία των λεξιλογίων, συνεπώς και της εσωτερικής μνήμης που διατηρεί τις επιλογές λέξεων από αυτά. Συνενώνοντας τα δεδομένα δύο ή περισσοτέρων διαδικασιών ανεξάρτητων λεξιλογίων (ας θεωρήσουμε ιδίων παραμέτρων) θα παίρναμε μία *Zipf* μορφή με μικρότερη αρνητική κλίση, κάτι που δεν είναι ασύμφωνο με την κατανομή αφού μοντελοποιείται από τις παραμέτρους α, β. Έτσι, η συνένωση δύο διατεταγμένων ιστογραμμάτων, θα έδινε μία μακρύτερη κατανομή, λόγω της ανεξαρτησίας των λεξιλογίων, με μικρότερη σαφώς κλίση χωρίς να παρατηρούσαμε αύξηση συχνοτήτων.

Αν θεωρήσουμε όμως πως έχουμε ένα (τουλάχιστον) κοινό λεξιλόγιο με κοινή μνήμη για την παραγωγή λέξεων, τότε μία τέτοια συνένωση θα αύξανε σίγουρα τις υψηλές συχνότητες στην κεφαλή της κατανομής (λέξεις που έχουν επιλεγεί σε κείμενα διαφορετικών κλάσεων), ενώ οι ανεξάρτητες λέξεις θα επιμήκυναν το λογαριθμικό ιστόγραμμα διατηρώντας περίπου την κλίση των *Zipf* κατανομών συχνοτήτων που είχαν οι δύο ξεχωριστές ομάδες δεδομένων.

Στον αλγόριθμο που προτείνεται, το γενικό λεξιλόγιο βοηθάει στην εμφάνιση των υψηλών συχνοτήτων στο ιστόγραμμα. Επίσης, η απόφαση για την παραγωγή μίας λέξης που δεν έχει εισαχθεί σε κάποιο κείμενο εξαρτάται μόνο από το μήκος του τρέχοντος κειμένου και είναι ιδιαίτερα κρίσιμη ως επιλογή. Αν ως μετρητή μήκους θεωρούσαμε τον αριθμό των λέξεων που μας έχει δώσει ένα λεξικό, αφού πρώτα το επιλέξουμε μέσω του μικτού μοντέλου, τα αποτελέσματα θα ήταν τελείως διαφορετικά διότι οι συνιστώσες δε θα συμπεριφέρονταν ως σύνολο. Κάθε μία χωριστά θα αύξανε το εύρος των λέξεων συνεισφοράς και θα οδηγούμαστε σε μία πιο πλατιά κατανομή. Οι πανομοιότυπες αρχικοποιήσεις με τα δ_i των ειδικών λεξικών ανάμεσα στην παραγωγή διαφορετικών κειμένων μπορούν να αναπαράγουν παρόμοιες συμπεριφορές στοχαστικά (τα κείμενα μίας κατηγορίας έχουν κοινή θεματική αλλά πολλές φορές και λεκτική αφετηρία), επιτρέποντας υψηλές συχνότητες σε λέξεις των ειδικών λεξικών.

Μια επιπλέον ρύθμιση είναι να διατηρούμε τα *δi* μόνο για το ειδικό λεξικό, κατά την παραγωγή μίας κατηγορίας κειμένων. Θα μπορούσε να γίνει κάτι τέτοιο, αν και πιθανόν να είναι επιθυμητό κάποιες κλάσεις να μοιράζονται όχι μόνο κοινό λεξικό αλλά και κοινούς πυρήνες πιο δομημένου περιεχομένου.

Τέλος, η προθέρμανση πειραματικά φάνηκε να βοηθάει αρκετά στο να μπορούμε να παίρνουμε όρους που δεν είναι πολωμένοι προς την αρχικοποίηση. Ουσιαστικά προσομοιώνουν την εγκεφαλική διεργασία που κάνει ο άνθρωπος πριν αρχίσει να γράφει, προετοιμάζοντας τους βασικούς νοηματικούς πυρήνες που θέλει να συνθέσει.

2.7.4. Αλγόριθμος Παραγωγής Συνθετικών Κειμένων

Ο ακόλουθος αλγόριθμος περιγράφει τη διαδικασία παραγωγής συνθετικών κειμένων όπως προτείνεται στην εργασία. Η παράμετρος *wg* έχει να κάνει με την εισαγωγή βάρους σε συγκεκριμένες προτάσεις, όταν επιθυμούμε να παράγουμε κείμενα με ετικέτες σημαντικότητας. Αν αποφασιστεί να επιβαρυνθεί κάποια πρόταση τότε θα πρέπει να επιλεγεί ένα αριθμητικό βάρος βάσει επιπρόσθετων επιλογών του χρήστη. Με δεδομένο πως γενικά οι λέξεις του ειδικού λεξικού συμβάλλουν περισσότερο στη διαμόρφωση του περιεχομένου ενός κειμένου τα βάρη θα αφορούν περισσότερο όρους σχετιζόμενους με το ειδικό θέμα. Οι διάφοροι όροι που επιβαρύνονται λόγω της ύπαρξής τους σε μία επιβαρυμένη πρόταση, χωρίς όμως να ανήκουν στο ειδικό λεξικό, παίζουν το ρόλο του θορύβου.

M	Ο αριθμός των κατηγοριών που θέλουμε να παράγουμε
N_i, i=1..M	Τα κείμενα που θα παραχθούν για κάθε κατηγορία
$\|V_i\|$, i=1..M+1	Το μέγεθος κάθε λεξικού (το (M+1)-ιοστό λεξικό είναι γενικό)
w_{ij}	Η πιθανότητα επιλογής μιας λέξης από το λεξικό j για κείμενο της κατηγορίας i
S, s	Ο μέγιστος και ελάχιστος αριθμός προτάσεων που μπορεί να περιέχει κάθε κείμενο
R, r	Ο μέγιστος και ελάχιστος αριθμός λέξεων ανά πρόταση
Wg	Το ποσοστό των προτάσεων που θα επιβαρύνονται περισσότερο
Ws	Ο αριθμός των προτάσεων προθέρμανσης
A	Παράμετρος της διαδικασίας *Zannete-Montemurro*
V	Παράμετρος της διαδικασίας *Zannete-Montemurro*
λ	Η παράμετρος λ της εκθετικής κατανομής (μ = 1 / λ)

Σχήμα 2.6. Παράμετροι εισόδου του αλγορίθμου παραγωγής συνθετικών κειμένων.

```
Διαδικασία Παραγωγής Συνθετικών Κειμένων διαφορετικών κατηγοριών
{
    Επέλεξε |V_i| λέξεις για κάθε λεξικό από ένα σύνολο διαθέσιμων λέξεων

    Αρχικοποίησε το κοινό λεξικό V_g και θεώρησέ το ως το λεξικό V_{M+1},
        αρχικοποίησε τις συχνότητες n_{(M+1)i} = 0 κάθε λέξης του λεξικού
        και τις δ_{(M+1)i} ~ Exp(λ) βάση της εκθετικής κατανομής

    Για κάθε κατηγορία K_i από τις M που παράγεται
    {   Αρχικοποίησε όλα τα ειδικά λεξικά V_k, k = 1,…M (όμοια με παραπάνω)

        Για κάθε κείμενο από τα N_i της κατηγορίας K_i
        {  Θέσε τον αριθμό των έτοιμων προτάσεων του κειμένου Q = 0
           Και των συνολικών λέξεων που έχουν επιλεγεί t = 0

           Επέλεξε τον αριθμό των προτάσεων q ~ U(s,S) (ομοιόμορφα)

           Για κάθε πρόταση από τις q
           {  Αν Q > ws της προθέρμανσης
                 Επέλεξε βάση της wg αν η πρόταση θα έχει επιπλέον βάρος και
                 στην περίπτωση αυτή ενημέρωσε το αρχείο εξόδου

              Επέλεξε τον αριθμό των λέξεων για την πρόταση t_q ~ U(R,r) (oομοιόμορφα)

              Για κάθε λέξη από τις t_q της πρότασης
              {   Επέλεξε λεξικό V_j βάση των w_ij
                  Βάση της πιθανότητας a_t = a·v·t^{v-1} αποφάσισε:
                  {  αν θα εισάγεις τυχαία μία νέα λέξη (επιλέγοντας ομοιόμορφα) ή
                     αν θα επιλέξεις ήδη εισηγμένη λέξη βάση των max{n_jk, δ_jk}
                                    των διαθέσιμων λέξεων από το λεξιλόγιο
                  }

                  Αύξησε τη συχνότητα της επιλεγμένης λέξης στο κείμενο

                  Αν Q > ws της προθέρμανσης
                     Γράψε στο αρχείο εξόδου τη λέξη

                  Θέσε t = t + 1
              }

              Αν Q > ws της προθέρμανσης
              {  Κλείσε το βάρος αν η πρόταση επιλέχθηκε να επιβαρυνθεί
                 Κλείσε την πρόταση στο αρχείο εξόδου αν έχει γραφεί
              }
           }
        }
    }
}
```

Σχήμα 2.7. Αλγόριθμος για την παραγωγή συνθετικής συλλογής κειμένων.

2.7.5. Δημιουργώντας Συνθετικά Δεδομένα

Με τον αλγόριθμο που παρουσιάστηκε δημιουργήσαμε, ενδεικτικά ως παράδειγμα εφαρμογής, 100 κείμενα ώστε να συγκρίνουμε τη διατεταγμένη λογαριθμικής κλίμακας κατανομή συχνοτήτων τους με τις αντίστοιχες των πραγματικών συλλογών *F*, *J*, *U*. Τα συνθετικά δεδομένα προέρχονται από 4 κατηγορίες των {30, 20, 25, 25} κειμένων. Ο αριθμός των λέξεων σε κάθε ειδικό λεξικό ορίστηκε {2000, 2000, 2000, 2000}. Για το δε κοινό λεξικό διαθέτουμε 2000 και 700 λέξεις, σε δύο διαφορετικές εκδόσεις

δεδομένων διατηρώντας κατά τα άλλα τις ίδιες παραμέτρους. Ορίσαμε τα διανύσματα πιθανοτήτων w_i για το μικτό μοντέλο επιλογής λεξιλογίου για κάθε ομάδα ως εξής:

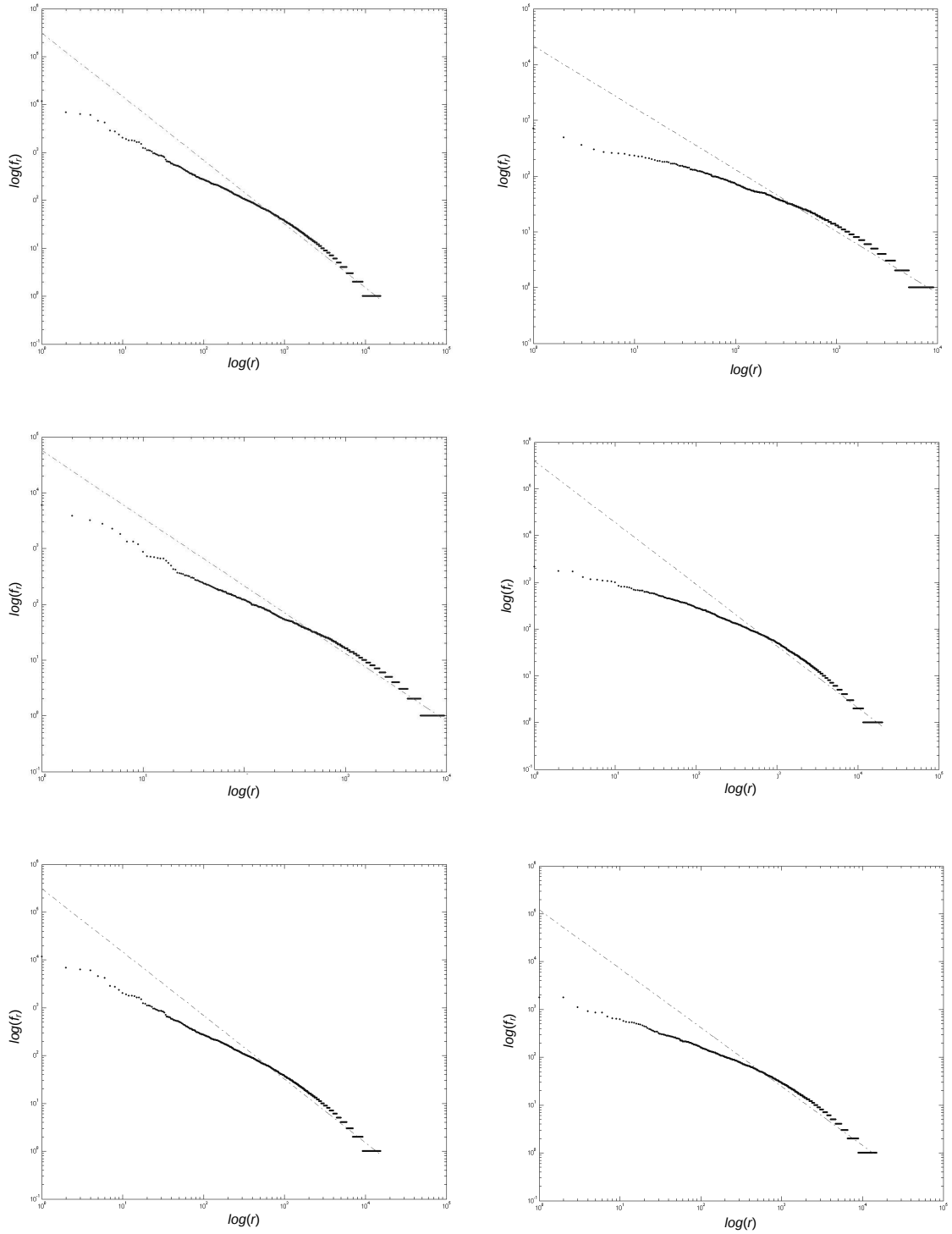
	V_1	V_2	V_3	V_4	V_g
K_1	**0.75**	0.04	0.01		0.20
K_2	0.02	**0.75**		0.03	0.20
K_3	0.02	0.01	**0.80**		0.17
K_4	0.05			**0.70**	0.25

Σχήμα 2.8. Οι πιθανότητες επιλογής όρων από τα διαφορετικά λεξικά για τις τέσσερις κατηγορίες κειμένων του παραδείγματος.

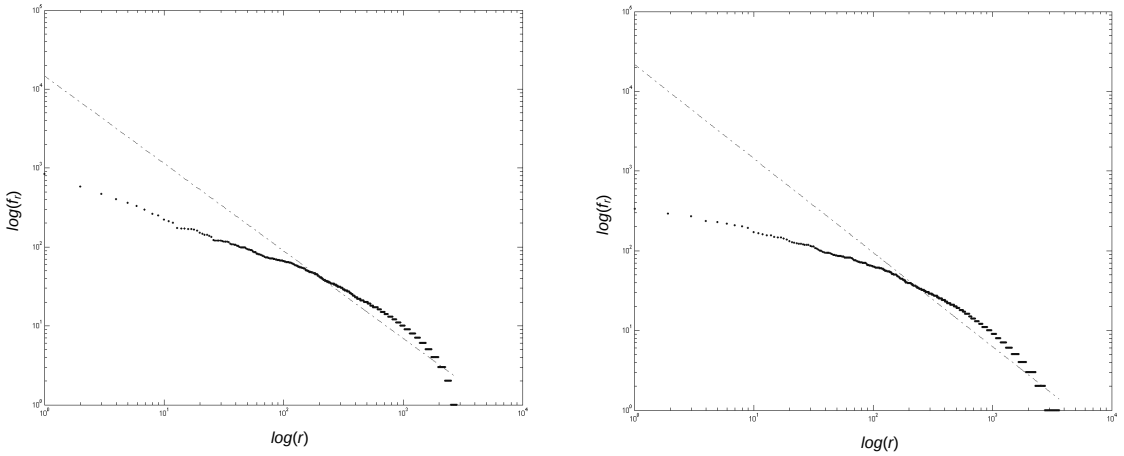
Επιλέγουμε ομοιόμορφα την παραγωγή 25 έως 100 προτάσεις για κάθε κείμενο, οι οποίες αποτελούνται από 4 έως 10 λέξεις. Επίσης, για κάθε κείμενο κάνουμε προθέρμανση 20 προτάσεων ανεξάρτητα από αυτές του κειμένου. Οι παράμετροι του μοντέλου *Zannete-Montemurro* τέθηκαν ($\alpha \cdot \nu$) = 0.3, β = 0.9 και το η μέση τιμή δ_μ = 4. Συνεπώς, λόγω των ομοιόμορφων επιλογών (εναλλακτικά: εκθετικές τ.μ.) το μέσο μήκος κειμένου είναι: $\overline{T} = \left[\left(S+s\right)/2\right]\cdot\left[\left(T+t\right)/2\right] = (100+25)\cdot(10+4)\,/\,4 = 438$ λέξεις.

Στο Σχήμα 2.10 φαίνεται το γραμμικό ταίριασμα ελαχίστων τετραγώνων με διακεκομμένη γραμμή. Παρατηρούμε να υπάρχει μία ασυμφωνία στις υψηλές συχνότητες όσων αφορά τα πραγματικά κείμενα. Αυτό εξηγείται αν λάβουμε υπόψη πως στα δεδομένα αυτά το λεξιλόγιο αυξανόταν περισσότερο από γραμμικά αποτρέποντας κάποιους όρους να αποκτήσουν μεγαλύτερες συχνότητες. Στα δεξιά κάθε γραφήματος φαίνεται η εκδοχή όπου έχουν αφαιρεθεί οι τετριμμένες λέξεις (προθέσεις κλπ.) χωρίς μετασχηματισμούς μορφολογικής ρίζας (*stemming*, βλ. Κεφάλαιο 3). Η προεπεξεργασία αυτή επηρεάζει εμφανώς τις υψηλές συχνότητες.

Τέλος, η συνθετική συλλογή, φαίνεται να διαθέτει στατιστικές ιδιότητες αντίστοιχες με αυτές πραγματικών κειμένων. Σε δύο διαφορετικές εκδόσεις ίδιων παραμέτρων μεταβάλλαμε μόνο το μέγεθος του γενικού λεξικού (2000 και 700 λέξεις) και παρατηρούμε την ίδια μεταβολή στο ιστόγραμμα, τη μείωση των υψηλών συχνοτήτων.



Σχήμα 2.9. Τα διατεταγμένα ιστογράμματα συχνοτήτων σε λογαριθμική κλίμακα για τις συλλογές (από επάνω): *F*, *J*, *U*. Αριστερή στήλη: ολόκληρη η συλλογή, δεξιά στήλη: μετά την αφαίρεση των τετριμμένων όρων.



Σχήμα 2.10. Τα διατεταγμένα ιστογράμματα συχνοτήτων σε λογαριθμική κλίμακα για τη συνθετική συλλογή. Αριστερά: κοινό λεξιλόγιο 2000 λέξεων, δεξιά: κοινό λεξιλόγιο 700 λέξεων.

Αυτό που πρέπει να τονιστεί είναι πως η προτεινόμενη διαδικασία μπορεί να δώσει χρήσιμα δεδομένα, απαιτείται όμως προσεκτική επιλογή των παραμέτρων ώστε να προκύπτουν ρεαλιστικά χαρακτηριστικά. Είναι δυνατόν να ρυθμίσουμε κατάλληλα το μοντέλο παρατηρώντας την αύξηση στο λεξιλόγιο και τη συμφωνία με τον εμπειρικό νόμο του *Zipf*, έχοντας υπόψη την συμπεριφορά που θα πρέπει να έχουν τα πραγματικά δεδομένα. Μπορούμε ακόμα να συμβουλευόμαστε πραγματικά δεδομένα όπως πράξαμε στη δική μας ανάλυση, ώστε να καταφέρουμε να παραμετροποιήσουμε καλύτερα το μοντέλου.

ΚΕΦΑΛΑΙΟ 3. ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΩΝ

3.1. Φάσεις Προεπεξεργασίας

Ένα κείμενο για να μπορεί να εξεταστεί από οποιοδήποτε αλγόριθμο ομαδοποίησης, θα πρέπει προηγουμένως να έχει υποβληθεί σε μια διαδικασία προεπεξεργασίας και μετασχηματισμού. Για την καλύτερη κατανόηση μπορούμε να θεωρήσουμε πως τα κείμενα είναι διανύσματα από χαρακτηριστικά περιεχομένου. Στη κλασική προσέγγιση κάθε διάσταση ορίζει μία ποσότητα ανάλογη της συχνότητας μίας λέξης. Ως δεύτερος στόχος τίθεται η μείωση της τάξης πολυπλοκότητας του προβλήματος. Γενικά, αυτό μπορεί να επιτελεστεί ανεξάρτητα από την ίδια τη διαδικασία ομαδοποίησης (*off-line*).

Από τις πρώτες προσπάθειες των ερευνητών στην επεξεργασία κειμένων άρχισαν να διατυπώνονται θεωρητικές και πειραματικές παρατηρήσεις, πως η ποιότητα της λύσης ίσως επηρεάζεται περισσότερο από την ποιότητα των δεδομένων που παρέχουμε στις κύριες διαδικασίες επίλυσης και λιγότερο από τις ίδιες τις αλγοριθμικές τεχνικές. Έτσι, αποκτά ιδιαίτερο ενδιαφέρον το φιλτράρισμα και το μοντέλο ορισμού βαρών στα χαρακτηριστικά του κειμένου.

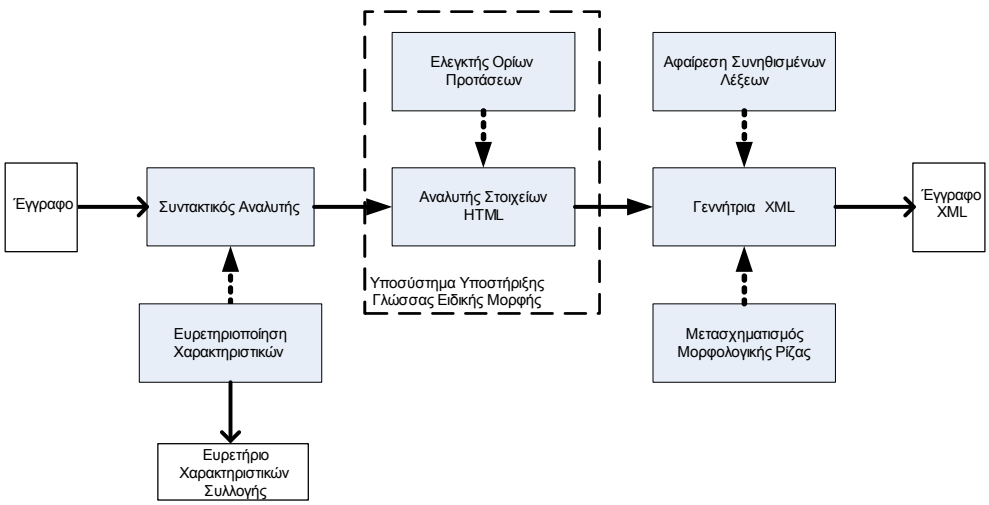
Εδώ πλέον μπορούν να προσαρμοστούν με ευκολία διάφορα εργαλεία της στατιστικής ανάλυσης (π.χ. *Bayesian Theory*), αφού υπολογίζονται εύκολα οι δεσμευμένες και οι από κοινού πιθανότητες ύπαρξης όρων σε κατηγορίες κειμένων. Για το ζήτημα αυτό υπάρχει μεγάλος όγκος βιβλιογραφίας, ενώ οι περισσότερες προσεγγίσεις αφορούν τη μάθηση με επίβλεψη.

Σχεδιάζοντας ένα πραγματικό σύστημα, θα πρέπει να κάνουμε κάποιες επιλογές που αφορούν τη μορφή εξόδου της προεπεξεργασμένης πληροφορίας. Μία επιλογή είναι να εκτελεστεί αμέσως μετά κάποιος αλγόριθμος ομαδοποίησης, με μειονέκτημα την επανάληψη της διαδικασίας προεπεξεργασίας σε κάθε ξεχωριστή εκτέλεση. Μια ακόμα επιλογή είναι να παράγουμε μία προεπεξεργασμένη συλλογή δεδομένων. Για παράδειγμα αν επιλεγεί η διανυσματική αναπαράσταση των κειμένων τότε μπορούμε να δημιουργήσουμε ένα αρχείο εξόδου με τα διανύσματα των κειμένων. Οι επιλογές αυτές σίγουρα περιορίζουν το σύστημα.

Στην υλοποίησή μας η διαδικασία προεπεξεργασίας χωρίζεται σε δύο μέρη:

- τη βασική προεπεξεργασία, που κάνει στοιχειώδεις παρεμβάσεις και μετασχηματισμούς στα δεδομένα. Εκτελείται μία φορά εκτός διαδικασίας (*off-line*),
- και την κατά την εκτέλεση προεπεξεργασία που εφαρμόζει συγκεκριμένες επιλογές του χρήστη ειδικά για τη λύση που πρόκειται να παραχθεί. Η φάση αυτή επιτελείται πάνω στα δεδομένα που παράγει η βασική προεπεξεργασία.

3.2. Προεπεξεργασία εκτός Διαδικασίας (*offline*)



Σχήμα 3.1. Η προεπεξεργασία εκτός διαδικασίας ενός εγγράφου.

Οι κυριότερες ενέργειες οι οποίες επιτελούνται από την πρώτη φάση είναι οι ακόλουθες:

- εντοπισμός ωφέλιμης πληροφορίας στο υπερκείμενο (*content spotting*),
- αφαίρεση συνηθισμένων λέξεων (*stopwords removal*),
- μετασχηματισμός μορφολογικής ρίζας (*stemming process*),

3.2.1. Εντοπισμός Τμημάτων Ωφέλιμης Πληροφορίας Υπερκειμένου

Εδώ θα παρουσιάσουμε τα ζητήματα που προκύπτουν κατά την προεπεξεργασία των υπερκειμένων σε μορφή *HTML*, και πως αυτά αντιμετωπίζονται από το σύστημα προεπεξεργασίας που υλοποιήσαμε. Για κείμενα με ετικέτες ιδιοτήτων άλλης κατηγορίας (οικογένεια *XML*) είναι πολύ εύκολη η προσαρμογή της διαδικασίας.

Όσων αφορά τα απλά κείμενα, δεν απαιτούν συντακτικούς κανόνες για την ανάγνωσή τους επομένως ο μηχανισμός εντοπισμού ωφέλιμης πληροφορίας παίζει το ρόλο του λεκτικού αναλυτή, αναγνωρίζοντας μία-μία τις λεκτικές μονάδες του κειμένου.

3.2.1.1. Δομή Υπερκειμένων

Η γλώσσα HTML (*Hypertext Markup Language*) περιγράφεται με εμφανίσιμους χαρακτήρες ASCII οι οποίοι αποθηκεύονται σε αρχεία κειμένου. Είναι δομημένη και βασίζεται κυρίως στην απόδοση ιδιοτήτων σε τμήματα κειμένου μέσω ετικετών (*tags*). Οι ετικέτες αυτές συντάσσονται με τη μορφή `<HTML_TAG>` και παίζουν το ρόλο των εντολών, αφού είναι καθορισμένες από την γλώσσα (ή από τον χρήστη αν είναι επεκτάσιμες ετικέτες). Τα αρχεία *HTML* προβάλλονται στο χρήστη μέσω ειδικών προγραμμάτων που ονομάζονται φυλλομετρητές.

Για τους σκοπούς της εργασίας, στόχος είναι να αντλήσουμε πληροφορία από τα βασικά χαρακτηριστικά που συναντά κανείς σε αυτή τη μορφή εγγράφων. Αυτό γίνεται από έναν κατάλληλο συντακτικό αναλυτή (*parser*) και ένα σύνολο χαρακτηριστικών που έχει καθορίσει ο χρήστης ως σημαντικά.

```
<HTML>
      <HEAD><TITLE> Document’s title </TITLE></HEAD>
      <BODY> Document’s body text and information. </BODY>
</HTML>
```

Σχήμα 3.2. Γενική δομή ιστοσελίδων.

Στοιχείο	Περιγραφή	Ύπαρξη
TITLE	Τίτλος κειμένου	Προαιρετική
META – KEYWORDS	Λέξεις κλειδιά του κειμένου	Προαιρετική
META – DESCRIPTION	Σύντομη περιγραφή περιεχομένων	Προαιρετική
META – AUTHOR	Συγγραφέας κειμένου	Προαιρετική

Σχήμα 3.3. Στοιχεία στο τμήμα της κεφαλίδας HEAD.

3.2.1.2. Οι ιδιαίτερες Δυσκολίες της Εξόρυξης Δεδομένων (*Web Mining*)

Ανορθόδοξη Συντακτική Δομή Υπερκειμένου

Στην πράξη ο κώδικας των σελίδων διαφέρει πολύ από τα εγχειρίδια της *HTML*, διότι το τελικό ζητούμενο είναι να φαίνεται μία σελίδα στον φυλλομετρητή. Αυτό δε σημαίνει όμως ότι δεν υπάρχουν συντακτικά σφάλματα. Ο συντακτικός αναλυτής που χρησιμοποιήσαμε κάνει μία αξιοπρεπή προσπάθεια να αντιμετωπίσει τέτοιου είδους προβλήματα.

Εντοπισμός Ορίων Προτάσεων

Ένα επιπλέον πρόβλημα είναι η εύρεση των ορίων των προτάσεων στις ιστοσελίδες. Η παράληψη των σημείων στίξης, κυρίως της τελείας (‘.’), αποτελεί συχνό φαινόμενο. Αυτό συμβαίνει γιατί η κατανόηση του περιεχομένου στηρίζεται σε μεγάλο βαθμό στην οπτική διαμόρφωση τους και στην ανθρώπινη αντίληψη. Είναι εξαιρετικά σημαντικό να ανακτούμε τα πραγματικά όρια των προτάσεων στην περίπτωση που η γειτνίαση των λέξεων θεωρείται χαρακτηριστικό των κειμένων.

Μια αξιόπιστη γενικά λύση είναι να λαμβάνουμε υπόψη τα οπτικά αποτελέσματα των ετικετών διαμόρφωσης της *HTML* και να υποθέτουμε το τέλος των προτάσεων βάσει κάποιων ευρετικών κανόνων. Για παράδειγμα, μπορούμε να θεωρήσουμε ότι η

εμφάνιση της ετικέτας που προκαλεί αλλαγής γραμμής
 ή παραγράφου <P>, υπονοούν το νοηματικό τέλος της αμέσως προηγούμενης πρότασης.

Να συμπληρώσουμε πως και στα φυσικά κείμενα μπορεί να βρει κανείς ασυνέπειες στα σημεία στίξης. Τέτοια παραδείγματα είναι οι επικεφαλίδες κεφαλαίων και παραγράφων. Οι λύσεις είναι και πάλι ευρετικές: συνήθως απέχουν μία γραμμή από το κυρίως κείμενο (διπλή αλλαγή γραμμής: "/n/n"), ή αναγνωρίζεται το αρχικό κεφαλαίο γράμμα της αμέσως επόμενης πρότασης.

3.2.2. Μετατροπή σε Χαμηλή Γραφή

Με την απλή αυτή διαδικασία μετατρέπονται όλοι οι χαρακτήρες του κειμένου, σε χαμηλή γραφή κατά την ανάγνωσή τους από το αρχείο. Παράλληλα, αγνοούνται όλοι οι μη αλφαριθμητικοί όροι, π.χ. αριθμοί, ημερομηνίες, σύμβολα όπως '<', '>', '=' κ.α. Οι σύνθετες λέξεις οι οποίες περιέχουν τους χαρακτήρες '−' ή '_', συνενώνονται σε έναν όρο αγνοώντας τους χαρακτήρες αυτούς (π.χ. "*intra-cluster*" → "*intracluster*"). Το ίδιο συμβαίνει και με τις εμφανίσεις άλλων συμβόλων μέσα σε λέξεις.

3.2.3. Μετασχηματισμός Μορφολογικής Ρίζας

Αν παρατηρήσουμε, σε ένα κείμενο είναι συχνό μία λέξη να εμφανίζεται με εννοιολογικά ισοδύναμες εκφράσεις (κλίσεις, χρόνοι, παράγωγες λέξεις κ.α). Για παράδειγμα οι λέξεις *player, players, play, playing, played* υπονοούν την ίδια πληροφορία για τα υποκείμενα ή αντικείμενα με τα οποία συσχετίζονται σε ένα σημείο του κειμένου. Έτσι, αντιμετωπίζουμε τα παραπάνω ως μια λέξη, την *play*, η οποία είναι η μορφολογική ρίζα τους (*stem*).

Επίσης με τη διαδικασία αυτή, η εσωτερική δομή κάθε κειμένου γίνεται πιο συμπαγής. Αν μας ενδιαφέρουν οι γειτνιάσεις όρων τότε η λέξη *play* θα κληρονομήσει τους γείτονες των υπόλοιπων λέξεων με αποτέλεσμα να αναδεικνύονται πιο ισχυρές συνδέσεις μεταξύ των λέξεων του κειμένου.

Ο μετασχηματισμός αυτός μπορεί να χαρακτηριστεί και ως διαδικασία μείωσης διάστασης για τα κείμενα. Η επιλογή της εφαρμογής του στα δεδομένα εισόδου έχει να κάνει με ένα ακόμα ζήτημα. Οι διαφορετικές εκδοχές ενός όρου δεν είναι κενές πληροφορίας. Σαφώς και προσαρμόζουν την έννοια ρίζας, π.χ. την play, σε ένα

νοηματικό σχήμα. Το πρόβλημα είναι ότι είναι πολύ δύσκολο να εκμεταλλευτούμε μηχανικά τις λεπτές αυτές νοηματικές διαφοροποιήσεις. Εδώ και πάλι υπεισέρχονται μέθοδοι επεξεργασίας και ανάλυσης φυσικής γλώσσας (*NLP*) οι οποίες βασίζονται κατά πολύ σε λεξικά-οντολογίες, γραμματικές και αναλυτικούς συντακτικούς κανόνες.

Για το μετασχηματισμό μορφολογικής ρίζας χρησιμοποιήθηκε ο αλγόριθμος του *Porter* [33], ο πλέον διαδεδομένος αλγόριθμος για το σκοπό αυτό, ενώ μία σύγκριση άλλων προσεγγίσεων της βιβλιογραφία συναντά κανείς στο [34].

Λέξη	Μορφ/κή ρίζα		Λέξη	Μορφ/κή ρίζα
transmission	transmit		society	societ
independent	independ		surface	surfac
exercises	exercis		conceptual	concept
identification	identif		journalist	journal
characteristic	characterist		transaction	transact
instance	instanc		andvances	advance
classification	classif		preferable	prefer
released	releas		prohibitive	prohibit
machine	machin		arraqngement	arrang
embodyment	embodi		January	januari
durable	durabl		manually	manual
consider	consider		maintenance	mainten
summary	summari		specifically	specif
diameter	diamet		islandic	island
economists	economi		elementary	element
qualitative	qualiti		filing	file
obtain	obtain		abbreviation	abbrevi
development	develop		conducted	conduct
prescribes	prescribe		chemical	chemic
energy	energi		replacing	replac
convincing	convinc		constraction	construct

Σχήμα 3.4. Παραδείγματα εφαρμογής του μορφολογικού μετασχηματισμού.

3.2.3. Αφαίρεση Συνηθισμένων Λέξεων

Έχει παρατηρηθεί πως οι 10 συχνότερες λέξεις της αγγλικής γλώσσας αποτελούν το 20-30% των λεκτικών μονάδων σε ένα κείμενο [35]. Ανάμεσά τους οι λέξεις {is, the, to, for, and, it…} που συναντάμε σχεδόν σε κάθε πρόταση. Αυτές οι λέξεις αφαιρούνται διότι έχουν βοηθητικό ρόλο, ώστε να συνδέονται συντακτικά οι έννοιες που ο άνθρωπος θέλει να παραθέσει. Η ποσότητα των όρων που αποφασίζουμε πως δε μεταφέρουν πληροφορία επηρεάζει και την ποιότητα των δεδομένων που θα αναπαραστήσουμε. Γι’ αυτό και υπάρχουν «γενικά» σύνολα λέξεων που προτείνονται

για το σκοπό αυτό, τα οποία περιέχουν διαφορετικό αριθμό λέξεων (συνήθως ανάμεσα σε 200 – 600 λέξεις).

Στη βιβλιογραφία το σύνολο που προτείνει ο Fox [37] θεωρείται ως κλασσικό, υπάρχουν επίσης πολλά ακόμα σύνολα που είναι πιο εξειδικευμένα σε θεματικές κατηγορίες, για παράδειγμα για ιατρικές συλλογές εγγράφων. Από την άλλη πλευρά, υπάρχουν και προσεγγίσεις αυτόματου εντοπισμού των μη πληροφοριακών όρων στα κείμενα εισόδου. Για την ακρίβεια πέραν από τους εντελώς τετριμμένους όρους εντοπίζονται λέξεις που έχουν μη πληροφοριακή στατιστική συμπεριφορά στα δεδομένα και με τον ορισμό κατωφλίων αφαιρούνται [36].

Ο έλεγχος για τον εντοπισμό των συνηθισμένων λέξεων γίνεται μετά την διαδικασία *stemming*. Με αυτόν τον τρόπο καθορίζουμε τις μετασχηματισμένες λέξεις τις οποίες θέλουμε να αποκόψουμε χωρίς να απαιτείται ο προσδιορισμός όλων των μορφών που μπορεί να συναντώνται. Αν δε γνωρίζουμε για κάθε λέξη που επιθυμούμε να αφαιρέσουμε τη μετασχηματισμένη μορφή της (ρίζα), μπορούμε να δημιουργήσουμε ένα αρχείο το οποίο να περιέχει ένα σύνολο λέξεων ή παραγώγων. Ο μετασχηματισμός του αρχείου αυτού δίνει το τελικό σύνολο λέξεων αποκοπής.

Στην υλοποίησή μας, οι λέξεις αυτές εισάγονται στη διαδικασία μετασχηματισμού μέσω ενός αρχείου εισόδου. Για την δημιουργία του αρχείου χρησιμοποιήσαμε τρία διαφορετικά αρχεία λέξεων που διατίθενται στο Διαδίκτυο. Αφού τα συνενώσαμε δημιουργήσαμε ένα λεξικό συνηθισμένων λέξεων με περίπου 570 όρους.

I, you,…	mine,yours	have,had,…	me, him, …
how	overall	meanwhile	still
hopefully	rather	much	then
however	really	namely	this, that, …
indeed	seem	need	thus
instead	reardless	next	together
last	serious	nevertheless	too
less	one, two…	often	unfortunately
like	several	once	usefull
ltd	someone	otherwise	upon
mainly	soon	quite	to, for, off,
may	so	particularly	up, down…
moreover	first, second…	whether	use

Σχήμα 3.5. Παραδείγματα συνηθισμένων λέξεων.

Αν μας ενδιαφέρει η τοπολογική διάταξη των λέξεων στο κείμενο, τότε μια ουσιώδης παρατήρηση είναι πως δύο μη γειτονικές λέξεις της πρότασης στην αρχική μορφή του κειμένου, μπορεί να καθίστανται γειτονικές ύστερα από την αφαίρεση συνηθισμένων λέξεων που παρεμβάλλονται ανάμεσά τους. Αυτό είναι σύμφωνο τις περισσότερες φορές με το νόημα του κειμένου. Για παράδειγμα η φράση: "*visit Greece and its islands*" κωδικοποιείται ορθά νοηματικά ως: "*visit – Greece – islands*".Αυτή είναι μία σημαντική διαφορά ανάμεσα σε αυτή τη διαδικασία και την διαδικασία φιλτραρίσματος. Η τελευταία, έπεται της αφαίρεσης συνηθισμένων λέξεων και η αποκοπή ενδιάμεσων όρων δεν επηρεάζει τη σχέση γειτνίασης των όρων που παραμένουν στο κείμενο. Για παράδειγμα, αν φανταστούμε μη συνηθισμένους όρους ανάμεσα στις λέξεις "*Greece*" και "*islands*" τότε η αφαίρεσή τους κατά το φιλτράρισμα δε σημαίνει την άμεση νοηματική σύνδεση των λέξεων που απομένουν.

3.2.4. Αποτέλεσμα Βασικής Προεπεξεργασίας

Το τέλος της πρώτης φάσης διαδικασίας δημιουργεί ένα νέο προεπεξεργασμένο σύνολο δεδομένων το οποίο αποτελείται από κείμενα που μοιάζουν με την *XML* μορφή και περιέχουν τις λέξεις διατηρώντας την διάταξη στο αρχικό κείμενο. Τα κείμενα υπόκεινται μόνο στους γενικούς μετασχηματισμούς που δεν επιθυμούμε να παραμετροποιηθούν στις ακόλουθες εκτελέσεις της διαδικασίας ομαδοποίησης.

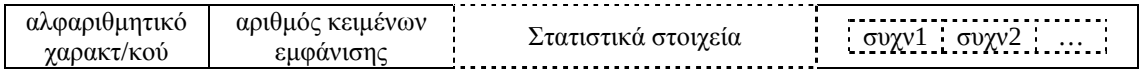
Στα νέα κείμενα, θα υπάρχουν μόνο οι ετικέτες που αλλάζουν τη σημαντικότητα των λέξεων, θα έχουν αφαιρεθεί τμήματα των εγγράφων όπως *script* κώδικας, σχόλια κ.α. Στο παρακάτω παράδειγμα φαίνεται πως ύστερα από το κλείσιμο της ετικέτας `TITLE` ο προεπεξεργαστής εντόπισε και διόρθωσε το τέλος της πρότασης, βάσει της υπόθεσης πως ο τίτλος του κειμένου είναι κάτι ανεξάρτητο από το υπόλοιπο κείμενο.

```
<DOCUMENT>
      <TITLE> Document's title </TITLE>  .
      <BODY>  Document's body text and information. </BODY>
</DOCUMENT>
```

Σχήμα 3.6. Γενική αφαιρετική δομή μετασχηματισμένου κειμένου σε *XML*.

Στην περίπτωση όπου η είσοδος είναι απλό κείμενο παρακάμπτεται το υποσύστημα συντακτικής ανάλυσης υπερκειμένου και συνεπώς το κείμενο *XML* θα περιέχει μόνο τις ετικέτες DOCUMENT, έναρξης και λήξης.

Το υποσύστημα της ευρετηριοποίησης συγκεντρώνει στοιχεία καθ' όλη τη διάρκεια της πρώτης φάσης προεπεξεργασίας. Διατηρεί έναν πίνακα κατακερματισμού ως ευρετήριο αποθηκεύοντας τα χαρακτηριστικά που έχει συναντήσει και είναι χρήσιμα για την συνέχεια (για τις ακμές που θα χρησιμοποιηθούν στο μοντέλο με τα γραφήματα μία ακμή $a \rightarrow b$ μετατρέπεται σε αλφαριθμητικό της μορφής "*a–b*"). Κάθε φορά που συναντάται ένα χαρακτηριστικό σε κάποιο κείμενο, ελέγχεται η ύπαρξή του στο ευρετήριο και είτε προστίθεται ως εγγραφή, είτε ενημερώνεται η υπάρχουσα. Τελικά, παράγεται ένα ακόμα αρχείο, το οποίο περιέχει πλειάδες της μορφής:

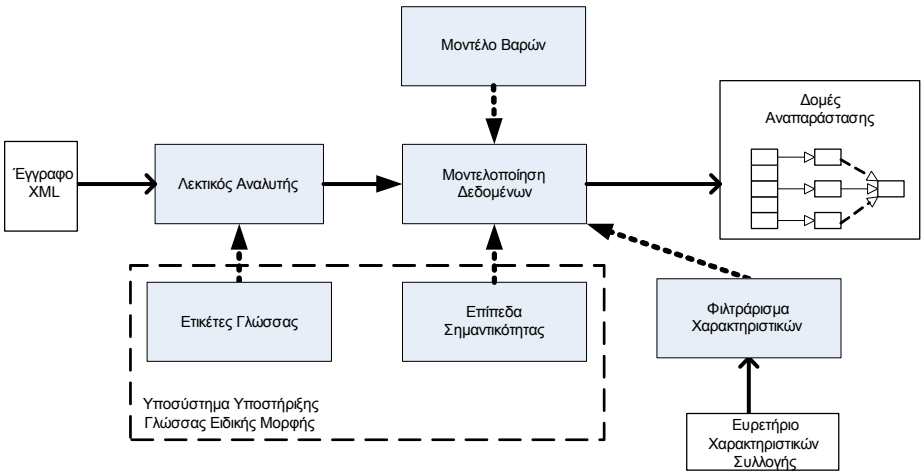


Σχήμα 3.7. Εγγραφή ευρετηρίου συλλογής.

Προαιρετικά μπορούμε να κρατάμε όποιες άλλες στατιστικές πληροφορίες επιθυμούμε. Είναι χρήσιμο κάποιες φορές να αποθηκεύουμε ένα διάνυσμα με τις συχνότητες εμφάνισης των όρων σε κάθε κείμενο.

3.3. Προεπεξεργασία κατά την Εκτέλεση

Πλέον ζητείται από τον χρήστη να παραχθεί μία λύση για το πρόβλημα που αφορά τα συγκεκριμένα δεδομένα. Ο προεπεξεργαστής εδώ έχει έναν πολύ απλούστερο ρόλο. Διαβάζει απευθείας τις *XML* μορφές των κειμένων αναγνωρίζει απλώς τις ανεξάρτητες λεκτικές μονάδες και τις ετικέτες και τροφοδοτεί τη διαδικασία μοντελοποίησης με τις αντίστοιχες πληροφορίες. Η διαδικασία επιταχύνεται ιδιαίτερα, επειδή έχουμε κρατήσει τις μορφολογικές ρίζες των χρήσιμων όρων, ενώ διατηρεί και τη γενικότητά τους επιτρέποντας στον χρήστη να αλλάξει διάφορες παραμέτρους χωρίς να επιβάλλεται νέα προεπεξεργασία.



Σχήμα 3.8. Η δεύτερη φάση προεπεξεργασίας κατά την εκτέλεση.

3.3.1. Μοντέλο Βαρών

Στο σημείο αυτό, θα πρέπει να αποφασιστεί ποια είναι η σημαντικότητα ενός χαρακτηριστικού για κάθε κείμενο στο οποίο παρουσιάζεται. Δύο είναι οι βασικές θεωρήσεις για τις τεχνικές ανάθεσης βαρών, α) ως αναφορά όλη τη συλλογή η υπόθεση της μονοτονικότητας [38] αναφέρει: «μία λέξη που εμφανίζεται σε πολλά κείμενα δε θα πρέπει να θεωρείται σημαντικότερη για τη συλλογή από μία άλλη που εμφανίζεται σε λίγα» και β) γα κάθε κείμενο χωριστά: «η υψηλή συχνότητα είναι ένδειξη υψηλής σημαντικότητας». Το πρώτο ονομάζεται και διακριτική δύναμη (*discrimination power*), δηλαδή το πόσο μπορεί να θεωρηθεί ιδιαίτερο χαρακτηριστικό μερίδας κειμένων. Οι παραδοχές που γίνονται ώστε να τυποποιηθεί το πρόβλημα είναι:

1. η κατανομές των λέξεων σε σχετικά κείμενα είναι ανεξάρτητες, η ίδια υπόθεση γίνεται και για τα λιγότερο σχετικά κείμενα,
2. οι εμφανίσεις των λέξεων σε ένα κείμενο είναι ανεξάρτητες (*Bayesian* προσέγγιση των τυχαίων μεταβλητών των όρων),
3. η σημαντικότητα ενός όρου σε ένα κείμενο θα πρέπει να είναι ανάλογη των εμφανίσεών του σε αυτό,
4. για τη συλλογή κειμένων, ως πληροφορία θα πρέπει να λογίζεται η παρουσία ενός όρου αλλά και η απουσία του,
5. τα κείμενα, παρά την ανισότητα μεγέθους, θα πρέπει να αντιμετωπίζονται με δικαιοσύνη κατά τον υπολογισμό της συνάρτησης συσχέτισης.

Τα σημεία 3, 4 υπονοούν ότι το βάρος w_{ij} του όρου j στο κείμενο d_i θα καθορίζεται από: α) την εσωτερική σημαντικότητα h_{ij} η οποία εξαρτάται από τη συχνότητά του σε αυτό και β) την εξωτερική-συνολική g_j σημαντικότητα που αφορά τη συνολική συμπεριφορά του στο σύνολο δεδομένων. Τυπικά αυτό εκφράζεται ως:

$$w_{ij}\left(h_{ij}, g_j, n_i\right) = h_{ij} \cdot g_j \cdot c_i ,$$

όπου c_i ένας παράγοντας κανονικοποίησης για το κείμενο d_i. Η υπόθεση ανεξαρτησίας των λέξεων στο εσωτερικό του κειμένου περιορίζει τη σχετική έρευνα στις τεχνικές υπολογισμού της εξωτερικής συνιστώσας. Επίσης, υπάρχουν προσεγγίσεις με επίβλεψη οι οποίες προσπαθούν να «μάθουν τα καλύτερα βάρη» για τις λέξεις της συλλογής, αλλά τα συμπεράσματά τους δεν έχουν γενικευτική δυνατότητα [39].

3.3.1.1. Σχήματα Εσωτερικών Βαρών

Το απλούστερο σχήμα είναι τα δυαδικά βάρη (*boolean weighting – BW*) που καταγράφουν την ύπαρξη του όρου σε ένα κείμενο με βάρος τη μονάδα, διαφορετικά θεωρούν μηδενικό βάρος. Η άμεση γενίκευσή του είναι τα συχνοτικά βάρη $h_{ij} = f_{ij}$ (*term frequency weighting – TF*).

Το επόμενο βήμα αφορά έγγραφα τα οποία έχουν ετικέτες προσδιορισμού ιδιοτήτων (*term significance weighting – TS*), π.χ. τα υπερκείμενα. Αυτή είναι μία επιπλέον πληροφορία όπου αναθέτοντας τιμές στις ετικέτες αυτές μπορούμε να αυξάνουμε ή να μειώνουμε τη σημαντικότητα σε διαφορετικά τμήματα του κειμένου. Αντίστοιχα με το συχνοτικό βάρος, το βάρος που μπορεί να προκύψει από την πληροφορία των ετικετών είναι το άθροισμα της σημαντικότητας κάθε εμφάνισης k, $k=1,\ldots f_{ij}$, που ισούται με τη σπουδαιότητα του σημείου $\mathrm{sgn}_{ij}^{(k)}$ του κειμένου που παρουσιάζεται:

$$h_{ij} = \sum_{k=1}^{f_{ij}} \mathrm{sgn}_{ij}^{(k)}$$

Αυτές οι προσεγγίσεις μπορούν να εφαρμοστούν και χωρίς την ύπαρξη εξωτερικού βάρους, δηλαδή ορίζοντας $w_{ij} = h_{ij}$. Συνολικά, για τα σχήματα εσωτερικών βαρών:

$$h_{ij} = \begin{cases} \sum_{k=1}^{f_{ij}} \mathrm{sgn}_{ij}^{(k)}, & \textit{significance weighting} \\ f_{ij}, & \textit{frequency weighting} \\ I\left(j \in d_i\right), & \textit{boolean weighting} \end{cases}$$

3.3.1.2. Σχήματα Εξωτερικών Βαρών

Τα σχήματα αυτά λαμβάνουν υπόψη τη συχνότητα ενός χαρακτηριστικού σε όλη τη συλλογή κειμένων, προσπαθώντας να ενισχύσουν τις λέξεις ισχυρής διακριτικής δύναμης. Η πιο δημοφιλής συνολική προσέγγιση εσωτερικού και εξωτερικού βάρους είναι η $TF \cdot IDF$ (*term frequency* · *inverse document frequency*), η οποία θεωρεί το βάρος w_{ij} ανάλογο της συχνότητας του όρου στο κείμενο και αντιστρόφως ανάλογο με την συνολική παρουσία του στη συλλογή:

$$g_j = IDF_j = \log_\beta\left(\frac{N}{n_j}\right),$$

όπου n_j το σύνολο των εγγράφων που παρουσιάζεται ο όρος j και N το σύνολο των εγγράφων της συλλογής. Παρατηρούμε πως ισχύει $N / n_j \geq 1$ και συνεπώς $IDF_j \in [0, \log_\beta \mathrm{N}]$. Ο λογάριθμος επιφέρει μια μη γραμμική ανάθεση βαρών, επίσης μπορούμε να θεωρήσουμε τη βάση του λογαρίθμου β ως παράμετρο. Συνήθως θέτουμε $\beta = 10$, ενώ μικρότερες τιμές επεμβαίνουν πιο δραστικά στη σημαντικότητα των όρων.

Υπάρχουν πολλά ακόμα σχήματα τα οποία ανήκουν σε αυτή την οικογένεια. Όλα υπολογίζουν μία λογαριθμική ποσότητα αναλογίας μεταξύ υποσυνόλων της συλλογής. Η $TF \cdot IDF$ έχει παρουσιάσει τη μεγαλύτερη ανθεκτικότητα σε συγκριτικές μελέτες όπως η [39], δεν είναι τυχαίο ότι διατυπώθηκε και χρησιμοποιείται από το 1968 [40]! Στο [41] μπορεί να βρει κάποιος μια σύντομη περιγραφή κάποιων γνωστών σχημάτων ανάθεσης βαρών.

3.3.1.3. Επίπεδα Σημαντικότητας σε Υπερκείμενα

Η οπτική παρουσίαση του κειμένου υπονοεί τη διαφορετική σημαντικότητα κάποιων τμημάτων του. Ορίσαμε την συνάρτηση εκτίμησης σημαντικότητας h_{ij} η οποία υπολογίζει ένα βάρος για τον όρο j στο κείμενο d_i συναρτήσει της ποιότητας των

εμφανίσεων κάθε χαρακτηριστικού. Έχουμε θεωρήσει 5 κύρια επίπεδα σημαντικότητας που φαίνονται παρακάτω και κάποια υποεπίπεδα που βρίσκονται ανάμεσα στο επίπεδο *H* και *VH* και αφορούν τις επικεφαλίδες των επιμέρους τμημάτων *H1*…*H6*.

Επίπεδο	Σημαντικότητα
NONE (N)	0
LOW (L)	1
MEDIUM (M)	2
HIGH (H)	4
VERY_HIGH (VH)	8

Σχήμα 3.9. Επίπεδα σημαντικότητας για τα τμήματα υπερκειμένων.

Τμήμα Κειμένου	Επίπεδο
Τίτλος	VH
Meta–δεδομένα: Λέξεις κλειδιά	VH
Meta–δεδομένα: Περιγραφή	H
Συγγραφέας	-
Επικεφαλίδες Παραγράφων 1…6	M…VH
Ετικέτες κελιών πινάκων	M
Ετικέτες εικόνων	M
Κείμενο σώματος	L
Τονισμένο κείμενο	M
Υπερσύνδεσμοι (*hyperlinks*)	H

Σχήμα 3.10. Γενική πολιτική απόδοσης σημαντικότητας στα τμήματα κειμένου.

Ετικέτες *HTML*			
H6	STRIKE	FONT	CAPTION
H5	DFN	B	TH
H4	CITE	I	LIST
H3	ADDRESS	U	IMG
H2	ABBR	BIG	INPUT
H1	A	EM	OPTION
TITLE	LABEL	STRONG ή S	SELECT
META	LEGEND	BR	CAPTION

Σχήμα 3.11. Ετικέτες τις οποίες λαμβάνουμε υπόψη για να αναθέσουμε βαρύτητα, να αφαιρέσουμε τμήματα κειμένου, ή να αναγνωρίζουμε το τέλος των προτάσεων.

Μία άξια λόγου παρατήρηση, είναι το ότι διαμορφώνουμε τους συνδέσμους σε ένα αλφαριθμητικό. Για παράδειγμα ένας υπερσύνδεσμος της μορφής: '`http://www.cs.uoi.gr/~akaloger`' θα μετασχηματιστεί σε ένα ενιαίο αλφαριθμητικό: '`wwwcsuoigrakaloger`'.

3.3.1.4. Κανονικοποίηση Βαρών

Η κανονικοποίηση των βαρών έχει παρατηρηθεί ως απαραίτητη στη βιβλιογραφία. Ουσιαστικά, επιβάλλει τη μετέπειτα σύγκριση των κειμένων πάνω στην αναλογία του περιεχομένου τους (σημαντικότητα όρων), αντιμετωπίζοντας τα κείμενα διαφορετικού μεγέθους με δικαιοσύνη. Αν θεωρήσουμε τα κείμενα d_i ως διανύσματα, η κανονικοποίηση μπορεί να είναι ως προς τη νορμα-1 ή τη νόρμα-2, κάτι που εξαρτάται από τη συνάρτηση συσχέτισης που επιλέγεται. Για παράδειγμα στην πιο διαδεδομένη συνολική προσέγγιση, όπου επιλέγεται η κανονικοποιημένη $TF \cdot IDF$ βάσει της νόρμας-2, για τον παράγοντα c_i του κειμένου d_i ισχύει:

$$\left\| d_i^{(tf \cdot idf)} \right\|_2 = \left[\sum_{j=1}^{|V|} \left(f_{ij} \cdot \log_\beta \left(N/n_j \right) \cdot c_i \right)^2 \right]^{1/2} = 1 \Leftrightarrow c_i = \left[\sum_{j=1}^{|V|} f_{ij} \cdot \log_\beta \left(N/n_j \right) \right]^{-1/2}$$

3.4. Μείωση Διάστασης

Η μείωση της διάστασης των κειμένων αφορά τη μείωση του λεξιλογίου της συλλογής και κάθε κειμένου χωριστά. Παρουσιάστηκαν ήδη κάποια ζητήματα της προεπεξεργασίας που μπορούν να θεωρούνται και τεχνικές μείωσης λεξιλογίου, όπως η αφαίρεση συνηθισμένων λέξεων και ο μετασχηματισμός μορφολογικής ρίζας.

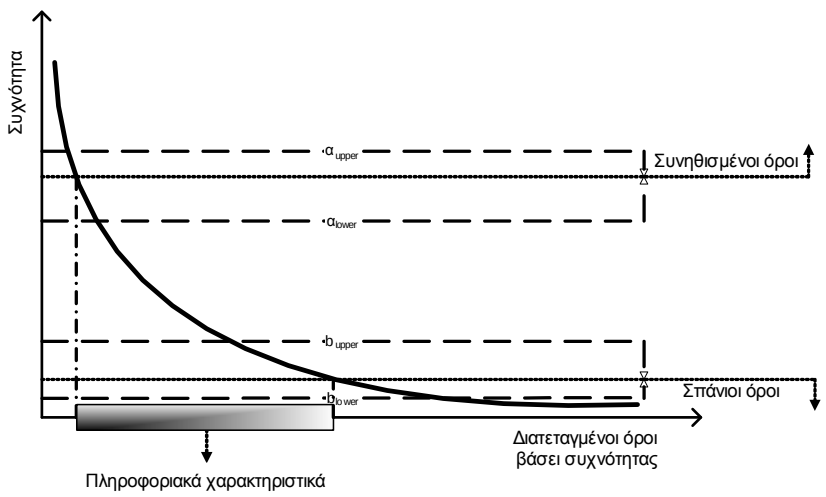
Για τη μείωση της διάστασης κειμένων φυσικής γλώσσας έχουν χρησιμοποιηθεί κλασσικές τεχνικές, όπως *Principal Component Analysis* (*PCA*) η οποία προβάλει τα δεδομένα σε ένα χώρο μικρότερης διάστασης στον οποίο παρουσιάζουν μεγάλη μεταβλητότητα, *Multidimensional Scaling* (*MDS*) η οποία προσπαθεί να μειώσει δραστικά τις διαστάσεις των δεδομένων διατηρώντας τις σχετικές αποστάσεις μεταξύ των κειμένων οι οποίες συνθέτουν την πληροφορία της δομής τους, *Latent Semantic Indexing* (*LSI*) [88] αλλά και πιο πρόσφατες όπως η *Locality Preserving Projection* (*LPP*) [19]. Οι μέθοδοι αυτές βασίζονται στην γραμμική άλγεβρα και προσπαθούν να προβάλλουν τα δεδομένα σε χώρους μικρότερης διάστασης διατηρώντας τα χαρακτηριστικά των δεδομένων στον αρχικό χώρο.

Δύο ακόμα ενδιαφέρουσες μέθοδοι μείωσης διάστασης για κείμενα είναι η [18], η οποία αναλύει τα κείμενα χρησιμοποιώντας κανόνες συσχέτισης (*association rule analysis and discovery*) και αφού εντοπίσει τα *frequent itemsets* που αποτελούνται από

συχνές ομάδες λέξεων που εμφανίζονται από κοινού σε αρκετά κείμενα, φιλτράρει τους κανόνες αυτούς. Στο [17] βρίσκει κανείς μία διαφορετική προσέγγιση όπου ύστερα από μία ομαδοποίηση σε M ομάδες, τα δεδομένα προβάλλονται ως διανύσματα στο χώρο $\mathbb{R}^M$, ορίζοντας κάθε διάσταση ως την απόσταση από την διαμορφωμένη ομάδα του πρώτου αποτελέσματος. Ύστερα, επαναλαμβάνεται ένα ακόμα στάδιο ομαδοποίησης για να πάρουμε το τελικό αποτέλεσμα.

Εδώ περνάμε σε μεθόδους που αναζητούν ποσότητες οι οποίες να είναι σε θέση να εκφράσουν την πληροφορία που περιέχει ένα χαρακτηριστικό. Η μείωση συνήθως πετυχαίνεται θέτοντας κατώφλια πάνω στις ποσότητες αυτές.

Πειραματικά αποτελέσματα και μία κατανοητή παρουσίαση πάνω στις γνωστές μεθόδους της βιβλιογραφίας που θα παρουσιάσουμε συνοπτικά και εδώ, μπορεί να βρει κανείς στο [42]. Μια σημαντική διαφορά ανάμεσα στην μείωση διάστασης σε άλλα προβλήματα και σε αυτό των κειμένων είναι πως ανάμεσα στα σπάνια χαρακτηριστικά των κειμένων θεωρείται πως υπάρχουν αρκετοί «σχετικά πληροφοριακοί» όροι οι οποίοι αναδεικνύουν μη τετριμμένες σχέσεις ανάμεσα στα λίγα κείμενα που περιέχουν τα σπανιότερα χαρακτηριστικά του συνόλου δεδομένων. Έτσι, δεν είναι προφανές πως μπορούμε να ξεκινήσουμε τυφλά από τα χαρακτηριστικά την αποκοπή.



Σχήμα 3.12. Η διαδικασία μείωσης διάστασης ως σύνολο κατωφλίων.

Στο Σχήμα 3.12 απεικονίζονται τα βασικά χαρακτηριστικά του προβλήματος. Στο διατεταγμένο ιστόγραμμα συχνοτήτων, φαίνεται η αποκοπή των συνηθισμένων και των

σπάνιων λέξεων που θεωρούνται μη-πληροφοριακοί. Από τους περιορισμούς αυτούς προκύπτουν τα χρήσιμα χαρακτηριστικά για τα προβλήματα επεξεργασίας κειμένων. Επίσης, φαίνεται η αβεβαιότητα των κατωφλίων αποκοπής.

3.4.1. Μείωση Διάστασης με Επίβλεψη

3.4.1.1. Γενική Προσέγγιση

Η περισσότερη βιβλιογραφία αναφέρεται σε μεθόδους όπου διαθέτουμε σύνολο εκπαίδευσης και μπορούμε να το μελετήσουμε στατιστικά με κριτήρια θεωρίας της πληροφορίας (*information theoretic*). Συμβολίζουμε, έστω, τον εξεταζόμενο όρο t την κατηγορία κειμένων C, και έναν δείκτη ποιότητας $\Delta(t, C)$ για την κατηγορία αυτή. Τότε οι συνολικές ποσότητες που παραδοσιακά υπολογίζονται για όλη τη συλλογή είναι:

$$\Delta_{\max}(t) = \max_i \{\Delta(t, C_i)\}, \text{ και } \Delta_{avg}(t) = \sum_{i=1}^{M} [P(C_i) \cdot \Delta(t, C_i)].$$

Ο τρεις βασικότερες ποσότητες που χρησιμοποιούνται για το σκοπό αυτό είναι [42]: ο χ^2 που εκφράζει την εξάρτηση των τ.μ. της κατηγορίας και του όρου, ο *IG* (*information gain*) που υπολογίζει το κέρδος σε *bit* πληροφορίας για την εύρεση της κατηγορίας του κειμένου όταν είναι γνωστή η παρουσία ή η απουσία του όρου t σε αυτό και τέλος ο *MI* (*mutual information*) που υπολογίζει το πόσο η γνώση της συμπεριφοράς της τ.μ. μίας λέξης μπορεί να καθορίζει την κατηγορία ενός κειμένου που τν περιέχει. Πειραματικά, στο [42] εμφανίζονται ως καλύτερες επιλογές αυτή του *IG* και του χ^2, ενώ το *MI* δε θεωρείται κατάλληλη επιλογή διότι πολώνει τις αποφάσεις ως προς τους όρους χαμηλότερου συχνοτικού περιεχομένου.

3.4.1.2. Κατωφλίωση Ισχύος

Αν και δε θα μας απασχολήσουν οι τεχνικές με επίβλεψη, αναφέρουμε αυτή επειδή έχει μια σχέση σε επίπεδο ιδέας με την προτεινόμενη τεχνική φιλτραρίσματος που θα περιγράψουμε στη συνέχεια. Ως ισχύς ορίζεται το πόσο έντονα εμφανίζεται ένας όρος σε κοντινά κείμενα. Η ιδέα αυτή παρουσιάζεται στο [57] και έχει εφαρμοστεί σε προβλήματα κατηγοριοποίησης κειμένων [58][59]. Αφού τεθεί ένα κατώφλι πάνω στην συνάρτηση συσχέτισης, που εκφράζει το πότε θα πρέπει να θεωρούνται «συναφή» δύο έγγραφα, εντοπίζονται κοντινά ζεύγη (d_k, d_l). Η ισχύς του όρου t υπολογίζεται τελικά

βάσει της δεσμευμένης πιθανότητας ως: $ts(t) = P(t \in d_k \mid t \in d_l)$. Ο στόχος της τεχνικής είναι να βρει τη σχετική σημαντικότητα ανάμεσα σε κείμενα σχετικού περιεχομένου.

3.4.2. Μείωση Διάστασης χωρίς Επίβλεψη

Στο πρόβλημα της ομαδοποίησης δε δίνεται σύνολο εκπαίδευσης, συνεπώς εφαρμόζονται πιο απλές λύσεις οι οποίες και εδώ σχετίζονται με κατώφλια σε διάφορες ποσότητες.

3.4.2.1. Κατωφλίωση Μεγέθους Μοντέλου Κειμένου

Πρόκειται για την πιο επιθετική πρακτική αποκοπής όρων, η οποία βασίζεται μόνο στην διατήρηση των καλύτερων λέξεων από κάθε κείμενο. Θέτοντας ένα κατώφλι *mf* στα μοντέλα (*model filtering*), π.χ. 50 λέξεις, κρατάμε τις *mf* λέξεις με τα καλύτερα βάρη από κάθε κείμενο. Ως βάρος όρου μπορεί να θεωρηθεί οποιαδήποτε ποσότητα που προκύπτει από τα αντίστοιχα μοντέλα που ήδη παρουσιάστηκαν.

Μία παρατήρηση είναι πως με την προσέγγιση αυτή είναι πιθανό να μπορούμε να αφαιρέσουμε πολύ μεγάλο μέρος της πληροφορίας, διατηρώντας τα βασικά χαρακτηριστικά του νοήματος των κειμένων, το οποίο γενικά περιέχεται στα πιο έντονα χαρακτηριστικά του. Με άλλα λόγια γνωρίζουμε πως θα χαθεί πληροφορία αλλά ελπίζουμε πως αυτό δε θα αποβεί καταστροφικό για το πρόβλημα. Από την άλλη πλευρά δε λαμβάνεται καθόλου υπόψη το αρχικό μέγεθος του κειμένου.

Μία τέτοια προσέγγιση θα μπορούσε να βοηθήσει και στην περίπτωση που έχουμε ένα ιδιαίτερα μεγάλο σύνολο δεδομένων. Να παρατηρήσουμε πάντως, πως για να εκμεταλλευτούμε το γεγονός των αραιών αναπαραστάσεων απαιτούνται κατάλληλες δομές δεδομένων, αλλιώς η πολυπλοκότητα παραμένει υψηλή.

3.4.2.2. Κατωφλίωση Συμμετοχής στα Κείμενα της Συλλογής

Μία συχνά χρησιμοποιούμενη τεχνική είναι το κατώφλι πάνω στις συχνότητες των χαρακτηριστικών σε όλο το σύνολο δεδομένων (*document frequency threshold*), για παράδειγμα αν θέταμε *dft* = 1 τότε θα αφαιρούνταν οι λέξεις που εμφανίζονται σε ένα μόνο κείμενο, ανεξαρτήτως της συχνότητας τους στο κείμενο αυτό. Πάνω στον ίδιο δείκτη μπορεί να βασιστεί η αποκοπή λέξεων που εμφανίζονται σε περισσότερα από *dft*

κείμενα, δηλαδή οι λέξεις που εμφανίζονται σε ένα πολύ μεγάλο αριθμό κειμένων της συλλογής.

Χρησιμοποιήσαμε την επιλογή *dft* = 1 ως βασική κατά το πειραματικό στάδιο. Η αφαίρεση λέξεων που εμφανίζονται σε ένα μόνο κείμενο δε επηρεάζει αρνητικά τους υπολογισμούς ομοιότητας μεταξύ των κειμένων. Αντίθετα ενισχύει το ποσοστό περιεχομένου που μεταφέρουν οι όροι οι οποίοι εμφανίζονται και σε άλλα κείμενα.

Γενικά πάντως, αν και είναι υπολογιστικά ο απλούστερος τρόπος, με γραμμική πολυπλοκότητα στο μέγεθος των δεδομένων, μειώνει αρκετά το λεξιλόγιο για τους λόγους που περιγράψαμε και στο Κεφάλαιο 2. Το κατώφλι δε συνηθίζεται να ξεπερνά το 2 επειδή οι σπάνιοι όροι μεταφέρουν αρκετή μη τετριμμένη πληροφορία.

3.4.2.3. Κατωφλίωση Συμμετοχής σε Γειτονιά (*K-NN* Φιλτράρισμα)

Στην παράγραφο αυτή θα παρουσιάσουμε μια τεχνική μείωσης της διάστασης των δεδομένων, την οποία προτείνουμε ως βάση για περαιτέρω έρευνα. Από τις παραπάνω τεχνικές θα λέγαμε πως σχετίζεται περισσότερο με την κατωφλίωση ισχύος διότι αναζητά πληροφορία στις γειτονιές των δεδομένων, αν και αυτή εφαρμόζεται σε προβλήματα εκπαίδευσης με επίβλεψη. Ο αναγνώστης που έχει υπόψη του βασικές έννοιες επεξεργασίας εικόνας, θα παρατηρήσει ομοιότητες με μεθόδους που ορίζουν παράθυρα από εικονοστοιχεία πάνω στις εικόνες και μελετούν την τοπική πληροφορία (π.χ. εξομάλυνση με φίλτρα *Gauss*).

Όπως είδαμε, οι περισσότερες στατιστικές μέθοδοι δεν αποκόπτουν αβίαστα όρους και προσπαθούν να ποσοτικοποιήσουν την ποιότητα τους βάσει παρατηρήσεων πάνω σε όλη τη συλλογή. Στην τεχνική που προτείνουμε, η βασική θεώρηση είναι πως ο χώρος του προβλήματος, αν και πολύ μεγάλος σε διαστάσεις, μπορεί να περιγραφεί καλύτερα από την τοπική πληροφορία παρά από την συνολική που αφορά όλο το σύνολο δεδομένων. Στις χιλιάδες των διαστάσεων, τα δεδομένα είναι πολύ αραιά και τις περισσότερες φορές κάθε αντικείμενο έχει πληροφορία σε ένα πολύ μικρό ποσοστό των διαστάσεων αυτών. Για τον ίδιο λόγο εξηγήσαμε πως μπορούμε να δούμε το πρόβλημα ως πρόβλημα τιμών που λείπουν (*missing values*).

Το ερώτημα τίθεται ως εξής: είναι πάντα αποτελεσματικό να ψάχνουμε στατιστικές δομές σε ένα τέτοιο περιβάλλον, ή θα μπορούσαμε να εκμεταλλευτούμε και άλλου

είδους πληροφορία; Τα κοντινά κείμενα μπορούν να παρέχουν καλύτερη πληροφορία για τα τοπικά χαρακτηριστικά του χώρου.

Κάθε ξεχωριστό θέμα της συλλογής μπορεί να περιγραφεί από ένα σύνολο λέξεων οι οποίες σίγουρα εμφανίζονται σε πολύ περισσότερα από ένα κείμενα της συλλογής. Σύμφωνα με το μοντέλο παραγωγής σύνθετων κειμένων του Κεφαλαίου 2, μπορούμε να αναλύσουμε αντίστροφα το πρόβλημα: θεωρούμε πως υπάρχει ένα ιδεατό σύνολο χαρακτηριστικών για κάθε θεματική κατηγορία (αντίστοιχο του λεξικού), και όλα τα σύνολα μαζί ορίζουν τις διαστάσεις του πραγματικού προβλήματος. Από ένα σύνολο επιλέγουν συχνότερα λέξεις τα κείμενα που σχετίζονται με μία κατηγορία. Τα κείμενα με άλλα λόγια, προκύπτουν από δειγματοληψία του αρχικού χώρου.

Κατ' επέκταση ένα πλήθος «κοντινών» κειμένων αποστάσεις περιγράφει καλύτερα τα τοπικά χαρακτηριστικά του χώρου, επειδή ουσιαστικά αποτελείται από πυκνότερα δείγματα πάνω σε έναν μικρό υποχώρο του συνολικού χώρου. Είναι γνωστό ότι ο κοντινότερος 1-*NN* και ο μακρινότερος γείτονας είναι ιδιαίτερα ευαίσθητες στο θόρυβο πληροφορίες. Έτσι, αν λάβουμε μία μεγαλύτερη γειτονιά, έστω *K*, ως αναφορά τότε αυξάνουμε την πιθανότητα να περιλαμβάνει επί των πλείστον κείμενα από μία κατηγορία (*K-NN* συνέπεια). Θεωρούμε τα κατώφλια που παρουσιάζονται στο Σχήμα 3.13 και καλύπτουν αρκετές δυνατότητες ρυθμίσεων για τη διαδικασία.

`dft`	Ο ελάχιστος αριθμός κειμένων στον οποίο θέλουμε να υπάρχει ένας όρος ώστε να θεωρείται χρήσιμος (*least document frequency*)
`LFinD`	Η ελάχιστη (*least frequency in a document*) συχνότητα σε ένα κείμενο, κάτω από την οποία θα «αμφισβητείται» η σημαντικότητα του όρου για το συγκεκριμένο κείμενο
`Ki`	Το μέγεθος της γειτονιάς γύρω από ένα κείμενο d_i, την οποία θα συμβουλευόμαστε ώστε να λαμβάνουμε αποφάσεις για την αποκοπή ή μη των αμφισβητούμενων όρων του κειμένου
`LNNF`	Ορίζει την απαίτηση να υπάρχει ένας όρος σε περισσότερα από *LNNF* από τα κείμενα της γειτονιάς (*least NN frequency*)
`LFinNN`	Η συχνότητα που πρέπει να έχει ένας όρος σε έναν γείτονα ώστε να θεωρούμε πως αυτός «υπάρχει» και παίζει ρόλο σε αυτόν
`LFSinD`	Όταν έχουμε κείμενα με ετικέτες είναι η σημαντικότητα πάνω από την οποία οι όροι ενός κειμένου δεν αποκόπτονται ποτέ (*least favored significance in document*)

Σχήμα 3.13. Οι παράμετροι ρύθμισης της τεχνικής φιλτραρίσματος που στηρίζεται στη δομή των δεδομένων σε επίπεδο *K-NN* γειτονιάς.

Η παρακάτω διαδικασία περιγράφει όλη τη ρουτίνα φιλτραρίσματος όρων που χρησιμοποιήσαμε στην πειραματική διαδικασία. Αρχικά γίνεται φιλτράρισμα κατωφλίωσης συμμετοχής στη συλλογή, υπολογίζονται οι ομοιότητες μεταξύ των

κειμένων ώστε να βρεθούν οι κοντινότεροι γείτονες και στη συνέχεια εξετάζονται οι όροι που «αμφισβητούνται» από την παράμετρο `LfinD`. Τα κατώφλια αυτά ορίστηκαν για να προσδοθεί μία γενικότητα και να μπορούν να καλυφθούν διάφορες επιλογές χρήστη για φιλτράρισμα. Στη συνέχεια ακολουθεί ο ψευδοκώδικας της υλοποίησης.

Η αυστηρότητα του φιλτραρίσματος είναι δυνατόν να ρυθμιστεί με μία σειρά από διαφορετικούς τρόπους. Για παράδειγμα, αυξάνοντας την `LfinNN` παράμετρο απαιτούμε ποιοτικότερες εμφανίσεις του αμφισβητούμενου όρου στους γείτονες του κειμένου. Αντίστοιχα, έχουμε τη δυνατότητα να απαιτήσουμε να βρεθεί ο όρος σε τουλάχιστον `LNNF` γείτονες πριν αποφασίσουμε πως χαρακτηρίζει μία γειτονιά.

Να σημειωθεί πως, αν η επιλογή μας είναι να στρέψουμε την αναζήτησή σε γειτονιές κοντινότερων γειτόνων (*nearest neighbors*), θα πρέπει να έχουμε κατά νου πως η σχέση αυτή δεν είναι συμμετρική μεταξύ των κειμένων: αν $d_j \in NN_i$ τότε δε συνεπάγεται πως $d_i \in NN_j$. Αυτή η παρατήρηση αφορά το γεγονός πως σε κάθε πέρασμα των κειμένων κάθε κείμενο συμβουλεύεται τη γειτονιά του, η οποία δε συμπίπτει με τις γειτονιές των γειτόνων του. Έτσι, αφού αποφασίσουμε πως ένα χαρακτηριστικό έχει πληροφορία για το κείμενο, επειδή περιέχεται σε κάποιο γείτονα, υπάρχει το ενδεχόμενο ο γείτονας να αποφασίσει την αποκοπή του χαρακτηριστικού αυτού από το περιεχόμενό του. Για να καταλήξει σε συνεπείς αποφάσεις, ο αλγόριθμος θα πρέπει να επαναλαμβάνεται έως ότου να μην αποκοπούν χαρακτηριστικά από τα κείμενα, ή ο αριθμός αυτός να είναι τόσο μικρός ώστε να μην αξίζει να συνεχίσουμε. Η ορθότητα του πρόωρου σταματήματος δικαιολογείται απόλυτα διότι ο αριθμός των χαρακτηριστικών που αποκόπτονται φθίνει στο πέρας των επαναλήψεων.

Η επιλογή των *NN* γειτόνων δίνει ευελιξία ιδιαίτερα όταν ορίζεται μικρό *K* σε σχέση με το μέγεθος συλλογής |*D*|, ώστε κάθε κείμενο να έχει την ελευθερία να επιλέγει τη γειτονιά που θα συμβουλεύεται. Μια ακόμα παρατήρηση αφορά τις ομοιότητες, οι οποίες υπολογίζονται μόνο μία φορά βάση όλων των χαρακτηριστικών των κειμένων (εξαιρούνται αυτά που δεν πληρούν το *dft* κριτήριο) με συνέπεια να μην μεταβάλλονται οι γειτονιές κατά τη διαδικασία.

```
Διαδικασία Φιλτραρίσματος Συλλογής Κειμένων
{
    Μοντελοποίησε τα N κείμενα της συλλογής (το πρώτο στάδιο προεπεξεργασίας),
       παράλληλα συγκέντρωσε δεδομένα για τους όρους (ευρετηριοποίηση χαρακτ/κών)

    Εφάρμοσε φιλτράρισμα με κατωφλίωση συμμετοχής στα κείμενα της συλλογής,
       θέσε παράμετρο dft = 1, και δημιούργησε νέα συλλογή με d_i κείμενα

    Υπολόγισε τον πίνακα ομοιοτήτων A = {sim(i,j)}, i,j = 1,…,N
       θεωρώντας μία κατάλληλη συνάρτηση ομοιότητας sim()

    Για κάθε κείμενο d_i της συλλογής, i = 1,…,N
    {   Εντόπισε τους K_i κοντινότερους γείτονες του κειμένου
       Αποθήκευσέ τους στο σύνολο NN_i = {nn_ik}, k = 1,…,K_i
    }

    Όσο αφαιρούνται χαρακτηριστικά από τα κείμενα,
        [ή στο προηγούμενο βήμα αφαιρέθηκαν R > P όροι],
    {
        Για κάθε κείμενο d_i της συλλογής, i = 1,…,N
        {   Για κάθε χαρακτηριστικό j του d_i (f_ij ≠ 0), j = 1,…,|d_i|
            {   Αν  f_ij < LFinD && s_ij < LFSinD       /* αμφισβητούμε το χαρακ/κο j */
                {   m = 0
                    Για κάθε κείμενο n = nn_ik της γειτονιάς του d_i, k = 1,…,K_i
                    {   Αν  f_nj > LFinNN
                           m = m + 1

                        Αν  m >= LNNF || m + (k-m) < LNNF
                           Σταμάτα την αναζήτηση στους γείτονες
                    }
                                              /* βρέθηκαν ποιοτικές εμφανίσεις σε */
                     Αν  m > LNNF             /* αρκετούς γείτονες */
                        Πρόσθεσε το χαρακτηριστικό j στο φιλτραρισμένο κείμενο d_i'
                }
                 Αλλιώς  Πρόσθεσε το χαρακτηριστικό j στο φιλτραρισμένο κείμενο d_i'
            }
        }

        Μέτρησε τα χαρακτηριστικά R που αποκόπηκαν στη συλλογή
        Θεώρησε την φιλτραρισμένη συλλογή ως νέα συλλογή d_i = d_i'
    }

    Αν ΣΥΝΥΠΟΛΟΓΙΣΜΟΣ_ΑΚΜΩΝ_ΣΤΟ_ΜΟΝΤΕΛΟ == 1
       Πρόσθεσε όλες τις ακμές των αρχικών κειμένων που ενώνουν όρους που
                οι οποίοι παρέμειναν ύστερα από το φιλτράρισμα
}
```

Σχήμα 3.14. Ψευδοκώδικας για τη προτεινόμενη τεχνική μείωσης διάστασης χωρίς επίβλεψη.

Η τεχνική αυτή φαίνεται να είναι αποτελεσματική στο να αποκόπτει αρκετή πληροφορία προσπαθώντας παράλληλα να διατηρεί ή και να ενισχύει τη σχέση ομοιότητας με τα γειτονικά κείμενα (κοντινά δείγματα του χώρου αρχικής διάστασης). Το γεγονός αυτό έχει ιδιαίτερη σημασία για την βελτίωση της ποιότητας της ομαδοποίησης. Βέβαια, η ποιότητα της μείωσης διάστασης που επιτυγχάνει εξαρτάται σε μεγάλο βαθμό από το θόρυβο που υπάρχει στις γειτονιές, δηλαδή αν το *K* ορίζει γειτονιές που περιέχουν αντικείμενα μίας κατηγορίας. Ούτως ή άλλως όμως, από τον παράγοντα αυτό εξαρτώνται έμμεσα και οι περισσότεροι αλγόριθμοι ομαδοποίησης.

Το επιθυμητό είναι να ρυθμίζουμε τη διαδικασία με τρόπο ο οποίος να «σέβεται» τα δεδομένα. Κάτι τέτοιο μπορεί να γίνει ορίζοντας ένα αρκετά μεγάλο $K = N/q$ ώστε να αποκόψουμε την πιο προφανή θορυβώδη πληροφορία. Για παράδειγμα ορίζοντας $q = M$ (τον αριθμό των ομάδων), ή q = μία εικασία για την μικρότερη ομάδα των δεδομένων. Ύστερα, αφού τα δεδομένα θα έχουν πιο ξεκάθαρη δομή μπορεί να ακολουθήσει ένα ακόμα στάδιο όπου επαναϋπολογίζονται οι ομοιότητες και ορίζεται μικρότερο K (δηλαδή μεγαλύτερο q).

Ζητήματα τα οποία μπορούν να μελετηθούν πειραματικά για τη βελτίωση της τεχνικής είναι η συμβολή των αμοιβαίων γειτόνων (*mutual neighbors*), ο συνδυασμός της με τεχνικές στατιστικής ανάλυσης, και ο ορισμός του K_i ο οποίος μπορεί να εξαρτάται από την ποιότητα της γειτονιάς (θεωρήσαμε για κάθε αντικείμενο ίδιο $K_i = K$). Είναι ενδιαφέρον να εξεταστεί η εφαρμογή της τεχνικής μετά από ένα στάδιο ομαδοποίησης, όπου έχουμε περισσότερη πληροφορία για της «πραγματικές» ομάδες. Οι γείτονες σε αυτή την περίπτωση μπορούν να επιλέγονται μέσα από την ομάδα στην οποία συμμετέχει τελικά ένα κείμενο.

Θα μπορούσε ακόμα, να χρησιμοποιηθεί ένα διαφορετικό μέτρο ομοιότητας. Ένα τέτοιο παράδειγμα είναι η ομοιότητα βάση των κοινών κοντινότερων γειτόνων (*shared nearest neighbors similarity*).

ΚΕΦΑΛΑΙΟ 4. ΜΟΝΤΕΛΑ ΑΝΑΠΑΡΑΣΤΑΣΗΣ ΚΕΙΜΕΝΩΝ

Στη φυσική γλώσσα η παράθεση των λέξεων αποσκοπεί στη σύνθεση νοήματος υψηλότερου επιπέδου από αυτό των μεμονωμένων λέξεων. Είναι σημαντικό να αναπαραστήσουμε την οντότητα «κείμενο» σε μια πιο συμπαγή και αυστηρή μορφή. Η συντριπτική πλειοψηφία των άρθρων της βιβλιογραφίας και των μελετών που γίνονται πάνω σε κείμενα θεωρεί ως χρήσιμη πληροφορία τις συχνότητες των λέξεων ενός κειμένου (*single-term analysis*).

Στο Κεφάλαιο αυτό θα μελετήσουμε σε βάθος την αναπαράσταση των κειμένων με γραφήματα τα οποία εκμεταλλεύονται την εσωτερική δομή των κειμένων. Προτείνεται μία αφαιρετική αναπαράσταση γραφημάτων, η οποία αναπαριστά τα κείμενα ως σύνολα βασικής πληροφορίας, που είναι οι όροι του, και σύνολα σχέσεων πάνω σε αυτούς.

4.1. Διανυσματικό Μοντέλο

Οι περισσότερες μέθοδοι ομαδοποίησης εγγράφων που χρησιμοποιούνται στην πράξη σήμερα βασίζονται στο διανυσματικό μοντέλο (*vector-space model*) [43][44][45][46]. Είναι επίσης το κυρίαρχο μοντέλο σε πολλά από τα προβλήματα που σχετίζονται με τα έγγραφα και έχουμε αναφέρει μέχρι στιγμής, αλλά και γενικά στα περισσότερα προβλήματα αναπαράστασης δεδομένων.

Στο διανυσματικό μοντέλο το πρόβλημα αναπαράστασης περιορίζεται στον ορισμό της πληροφορίας που θα περιγράφει κάθε διάσταση. Έτσι, αν V το λεξιλόγιο μίας συλλογής, το μοντέλο αυτό αναπαριστά ένα κείμενο ως διάνυσμα του χώρου $\mathbb{R}^{|V|}$:

$$d_i = \{w_{ij}\},\ j = 1,...,|V|,$$

όπου σε κάθε διάσταση j αναφέρει το βάρος της αντίστοιχης λέξης. Το βάρος αυτό προκύπτει συνήθως επιλέγοντας ένα σχήμα ανάθεσης βαρών από αυτά που περιγράφονται στο Κεφάλαιο 3. Η αναπαράσταση αυτή καλείται συχνά *bag-of-words*, για προφανείς λόγους και συνήθως χρησιμοποιείται σε συνδυασμό με το σχήμα *TF* · *IDF*. Η κανονικοποίηση όπως εξηγήσαμε θεωρείται απαραίτητη σε κάθε περίπτωση, εδώ γίνεται ως προς τη νόρμα-2. Τελικά, το διάνυσμα αναπαράστασης του κειμένου d_i παίρνει τη παρακάτω μορφή με τη βοήθεια του παράγοντα κανονικοποίησης c_i:

$$d_i^{(tf\cdot idf)} = c_i \cdot \left(f_{i1} \cdot \log_\beta\left(\frac{N}{n_1}\right), ..., f_{i|V|} \cdot \log_\beta\left(\frac{N}{n_{|V|}}\right)\right), \qquad c_i = \left[\sum_{j=1}^{|V|} f_{ij} \cdot \log_\beta\left(\frac{N}{n_j}\right)\right]^{-1/2}$$

Τα διανύσματα αυτά είναι ιδιαίτερα αραιά δυσκολεύοντας την επεξεργασία τους όταν το πρόβλημα αφορά πολλά έγγραφα. Με κατάλληλες δομές δεδομένων είναι δυνατόν να τα διατηρούμε πιο αποδοτικά στη μνήμη και κυρίως να κάνουμε ταχύτερους υπολογισμούς.

Η ομοιότητα μεταξύ των κειμένων μπορεί να υπολογιστεί με διάφορα μέτρα πάνω σε διανύσματα. Παραδείγματα τέτοιων μέτρων ομοιότητας είναι η συνημιτονοειδής ομοιότητα (*cosine measure*), ή Ευκλείδεια ομοιότητα (*Euclidian measure*), κ.α. Διαπιστώνει κανείς ότι, ανεξαρτήτου μέτρου, τίθεται ένας αναπόφευκτος περιορισμός που αφορά την ανάλυση ως προς έναν όρο (*single–term analysis*). Αυτό σημαίνει ότι η ομοιότητα δύο κειμένων περιορίζεται στον έλεγχο ύπαρξης κοινού λεξιλογίου.

Το πρόβλημα αυτό μπορεί να επεξηγηθεί με ένα απλό παράδειγμα: δεδομένου του λεξιλογίου ενός κειμένου και των συχνοτήτων εμφάνισης , θα μπορούσε να παραχθεί ο ίδιος αντιπρόσωπος-διάνυσμα, για κάθε κείμενο παραγόμενο από μία τυχαία παράθεση των όρων του αρχικού κειμένου. Συνεπώς, η αναπαράσταση αυτή αγνοεί το βασικό δομικό χαρακτηριστικό των ανθρώπινων κειμένων, που δεν είναι άλλο από την παράθεση νοηματικών σχημάτων.

Παρόλα αυτά το διανυσματικό μοντέλο έχει αρκετά ενδιαφέρουσες ιδιότητες:

- το κυριότερο είναι πως επιτρέπει την απευθείας εφαρμογή μαθηματικών πράξεων πάνω στα δεδομένα, όπως ο υπολογισμός εσωτερικού γινομένου, νορμών κ.α., και μέσω αυτών αυστηρότερη ανάλυση των διαφόρων τεχνικών μάθησης στα δεδομένα. Επίσης μπορούν να εφαρμοστούν άμεσα γενικοί μέθοδοι μάθησης για διανυσματικά δεδομένα, χωρίς την απαίτηση καποιας τροποποίησης.
- σε αντίθεση με τις πιο σύνθετες προσεγγίσεις των γραφημάτων, απαιτεί ένα απλούστερο στάδιο προεπεξεργασίας διότι αρκεί να καταγραφεί η συχνότητα των λεκτικών μονάδων σε ένα κείμενο.
- γενικά θεωρείται πετυχημένο μοντέλο, διότι παρουσιάζει αποτελέσματα που καλύπτουν από πλευράς αναπαράστασης τα προβλήματα επεξεργασίας κειμένων.

Ολοκληρώνοντας την παρουσίαση του διανυσματικού μοντέλου, θα λέγαμε, βάσει της διαίσθησης αλλά και της λογικής, πως τίθεται ως απαίτηση η καλύτερη αναπαράσταση των κειμένων. Μάλιστα, το πρόβλημα δεν εντοπίζεται στην μαθηματική μορφή του μοντέλου άλλα στην *bag-of-words* προσέγγιση που αντλεί πληροφορία μόνο από τους όρους ενός κειμένου.

4.2. Μοντέλα Γραφημάτων

Όπως αντιλαμβάνεται ο άνθρωπος κατά την ανάγνωση ενός κειμένου, στις λέξεις περιέχεται το εννοιολογικό περιεχόμενο το οποίο συνδέεται μέσω της παράθεσης των λέξεων. Η σύνδεση αυτή διαμορφώνει το νοηματικό περιεχόμενο του κειμένου. Ένα γράφημα G είναι μία σύνθετη μορφή δεδομένων η οποία αποτελείται από κόμβους και

συνδέσεις μεταξύ αυτών. Λόγω της δυνατότητας περιγραφής σύνθετων σχέσεων μεταξύ δεδομένων τα γραφήματα είναι κατάλληλα για την αναπαράσταση κειμένων.

Στη βιβλιογραφία, για το ζήτημα της ενσωμάτωσης πληροφορίας από την παράθεση των λέξεων των κειμένων υπάρχει μικρό σχετικά πλήθος εργασιών. Ένας λόγος είναι ότι έχουν διατυπωθεί αμφιβολίες για το αν μπορούσε η πληροφορία παράθεσης των λέξεων να βοηθήσει στα προβλήματα μηχανικής μάθησης [16]. Για την ακρίβεια διατυπώθηκε η άποψη πως η πληροφορία αυτή δεν είναι κρίσιμης σημασίας και συνεπώς με απλούστερα μοντέλα μπορούμε να πετύχουμε παρόμοια αποτελέσματα.

Σε άλλες εργασίες ρητά διατυπώθηκε πως η σχέση μεταξύ των λέξεων στις φράσεις ενός κειμένου μπορεί να οδηγήσει σε μία πιο ακριβή αναπαράσταση του [15][48], και να βοηθήσει τους αλγόριθμους εκπαίδευσης στην παραγωγή ποιοτικότερων λύσεων. Σε κάποιες εργασίες προτείνονται μοντέλα κατευθυνόμενων γραφημάτων [54][48][49] [50][67][51], ή μέτρα ομοιότητας πάνω σε αυτά, υποστηρίζοντας την άποψη υπέρ ενός πλουσιότερου μοντέλου αναπαράστασης. Στο [54] το κείμενο αναπαρίσταται με ένα κατευθυνόμενο γράφημα συνδέσεων μεταξύ των όρων του, όπως αυτοί συναντώνται κατά την ανάγνωση του, στο [47] προτείνεται μία προσέγγιση αναπαράστασης βάσει των φράσεων, και στο [18] η χρησιμότητα της πληροφορίας των συνδέσεων των λέξεων υπονοείται αντλώντας τα *frequent itemsets* των όρων του κειμένου (συχνές από κοινού εμφανίσεις λέξεων).

Στη βιβλιογραφία της επεξεργασίας φυσικής γλώσσας, είναι συχνό με από την ανάλυση των φράσεων μέσω γραμματικής και συντακτικής ανάλυσης και τον προσδιορισμό του ρόλου κάθε όρου (π.χ. επίθετο, ρήμα, αντικείμενο κ.λ.π.) προκύπτουν κατευθυνόμενα γραφήματα. Παρατηρεί κανείς, πως υπάρχει εμφανής διάσταση των προσεγγίσεων με αυτό που περιγράψαμε παραπάνω, από τη μία η εμμονή στο απλό διανυσματικό μοντέλο με βασική πληροφορία τις λέξεις και από την άλλη πολύπλοκα γραφήματα με πολλαπλές ακμές μέσα σε μία μόνο φράση.

Μια πιθανή εξήγηση είναι το ίδιο το πρόβλημα εφαρμογής. Κατά την μηχανική οργάνωση κειμένων υπάρχει αρκετή πληροφορία στους πολλούς όρους που διαθέτει κάθε κείμενο. Αντίθετα, η σημασιολογική ανάλυση σε βάθος εφαρμόζεται συνήθως σε προβλήματα όπως η ανάλυση ερωτήσεων (*query systems, query answering*), όπου το επεξεργαζόμενο κείμενο αποτελείται από τις λίγες λέξεις ενός ερωτήματος χρήστη προς ένα σύστημα. Αντλείται λοιπόν ότι είναι δυνατό από τα δεδομένα, ενώ λόγω της

συντακτικής ανάλυσης διατηρούνται ακόμα και οι τετριμμένες λέξεις. Μία χαρακτηριστική προσέγγιση σε τέτοια προβλήματα είναι η επέκταση των όρων του ερωτήματος (*query term expansion*), όπου για τους μη επουσιώδεις όρους του αναζητούνται π.χ. συνώνυμα τα οποία επεκτείνουν το ερώτημα.

4.3. Κατευθυνόμενο Γράφημα Μονοπατιών-Φράσεων

Στην προσέγγιση αυτή [47][52], προτείνεται η χρήση ενός τύπου κατευθυνόμενου γραφήματος τον οποίο αποκαλούν *Document Index Graph*, το οποίο διατηρεί όλα τα μονοπάτια, όλων των κειμένων της συλλογής. Τα μονοπάτια λέξεων αναπαριστούν τις προτάσεις ή γενικά τις φράσεις που αποτελούνται από διαδοχικούς όρους. Το γράφημα αυτό περιέχει έναν κόμβο για κάθε διαφορετικό όρο της συλλογής, στον οποίο αποθηκεύει πληροφορίες για το ποια κείμενα περιέχουν τον όρο. Αντίστοιχα, χειρίζεται και τις ακμές που είναι οι συνδέσεις διαδοχικών όρων σε αυτά.

Κάθε λέξη που αναγνωρίζεται από το λεκτικό και συντακτικό αναλυτή, εισέρχεται για ευρετηριοποίηση στο γράφημα. Ελέγχεται αν υπάρχει ήδη εγγραφή, αν όχι προστίθεται, αν ναι ενημερώνεται για το κείμενο στο οποίο υπάρχει ο όρος. Ταυτόχρονα, ενημερώνονται οι ομοιότητα μεταξύ των κειμένων.

4.4. Γενικευμένο Γράφημα Σχέσεων για την Αναπαράσταση Κειμένων

Θα διατυπώσουμε μία σειρά ορισμών οι οποίοι μας επιτρέπουν να περιγράψουμε το γενικευμένο γράφημα αναπαράστασης κειμένων. Με t συμβολίζουμε έναν όρο, R ένα σύνολο σχέσεων ανάμεσα σε όρους και Σ ένα σύνολο από όρους.

Ορισμός 4.1. *Νοηματική σχέση* μεταξύ των όρων t_i, t_j τύπου $r(t_i, t_j) \in R$, καλείται μία μηχανικά αναγνωρίσιμη σχέση μεταξύ των δύο όρων t_i, t_j ενός κειμένου, η οποία αποτελεί χαρακτηριστικό της αναπαράστασής του. Καλούμε, επίσης, το ζεύγος όρων t_i, t_j *νοηματικά σχετιζόμενο ζεύγος βάσει της σχέσης* r.

Ορισμός 4.2. *Νοηματικά σχετιζόμενο σύνολο όρων* $\Sigma_r(t)$, *με όρο* t, καλείται το σύνολο το οποίο αποτελείται από όλους τους όρους ενός κειμένου που σχετίζονται με τον όρο t βάσει της σχέσης $r(t, t_j) \in R$, $j = 1, \ldots, |\Sigma_r(t)|$.

Ορισμός 4.3. *Κλάση γραφημάτων* για την αναπαράσταση κειμένων καλούμε την οικογένεια γραφημάτων που προκύπτει από τον ορισμό ενός συνόλου σχέσεων *R*. Συμβολικά γράφουμε: GC_R.

Ο παρακάτω γενικός ορισμός μπορεί να συμπεριλάβει από απλές σχέσεις μέχρι συνθετότερες που θα προέκυπταν από τη εφαρμογή γραμματικής, συντακτικής και σημασιολογικής ανάλυσης. Αρκεί, το σύνολο *R* να περιέχει καλώς ορισμένες σχέσεις οι οποίες να είναι δυνατόν να αναγνωριστούν μηχανικά από κατάλληλες διαδικασίες εξόρυξης δεδομένων από τα κείμενα.

Ένα γράφημα σχέσεων *G* μπορεί να οριστεί ως $G = \{T, E, W, R\}$, όπου:

T: ένα σύνολο κόμβων $T = \{ t_i \}$, όπου κάθε κόμβος αντιπροσωπεύει μοναδικά μία λέξη του κειμένου.

E: ένα σύνολο ακμών $E = \{ e_{ij}^{(r)} \}$, με $e_{ij}^{(r)} = (t_i, t_j, r)$ μία τριάδα στοιχείων, που δηλώνει την ύπαρξη σχέσης $r(t_i, t_j)$ και αντίστοιχα ακμής που αναπαριστά τη σχέση αυτή στο μοντέλο γραφήματος.

W: ένα σύνολο βαρών $W = \{w_j\}$, $j = 1, \ldots, |G|$, που εκφράζουν την σημαντικότητα των στοιχείων του γραφήματος. Το σύνολο *W* περιέχει ένα βάρος για κάθε στοιχείο γραφήματος του *G*, δηλαδή $|W| = |G| = |T| + |E|$.

Στη συνέχεια θα αναφερόμαστε στις σχέσεις που έχουν οριστεί για το γράφημα G_i ως R_i, ομοίως για T_i, E_i, W_i.Παρατηρούμε, πως η *κλάση* ενός τέτοιου γραφήματος ορίζεται από το σύνολο R_i.

Το σύνολο *W* έχει μέγεθος όσο τα στοιχεία του γραφήματος *G* και όχι όσο τα στοιχεία όλων των κειμένων. Τα w_j καθορίζονται από μία συνάρτηση ανάθεσης $s(\cdot)$, και δε θα πρέπει να συγχέονται με τα σχήματα βαρών του προηγούμενου κεφαλαίου. Ορίζουμε λοιπόν τη βοηθητική συνάρτηση $s(g)$ η οποία καθορίζει τη σημαντικότητα κάθε στοιχείου γραφήματος (g: κόμβος ή ακμή) σε όλα τα γραφήματα των κειμένων, $w_{ij} = s(g_{ij})$. Η συνάρτηση αυτή αφορά όλο το σύνολο δεδομένων, αντίθετα το σύνολο *R* μπορεί να είναι ειδικό για κάθε κείμενο, με τη σύμβαση πως η ίδια σχέση *r* σε δύο διαφορετικά κείμενα θα αντιμετωπίζεται με τον ίδιο τρόπο από τη συνάρτηση ανάθεσης βάρους $s(\cdot)$ ώστε να έχουμε συγκρίσιμες ποσότητες σημαντικότητας. Η συνάρτηση αυτή παρέχει ένα επίπεδο ελευθερίας στον καθορισμό των βαρών *W*, ώστε

να επιτρέπεται ο διαφορετικός τρόπος υπολογισμού κάθε βάρους ανάλογα με το αν πρόκειται για λέξη ή σχέση. Για παράδειγμα, σε μία συγκεκριμένη σχέση μπορεί να θέλουμε να δίνουμε βάρος με έναν ειδικά ορισμένο τρόπο.

Ορισμός 4.4. Υπεργράφημα (*hypergraph*) $HG = \{T_{HG}, E_{HG}, W_{HG}, R_{HG}\}$ είναι το γράφημα το οποίο προκύπτει από την (συν)ένωση των *n* κειμένων ενός συνόλου. Το γράφημα αυτό περιγράφεται ως:

$$HG = \bigcup_{i=1}^{n} G_i ,$$

$$T_{HG} = \bigcup_{i=1}^{n} T_i ,\ E_{HG} = \bigcup_{i=1}^{n} E_i ,\ R_{HG} = \bigcup_{i=1}^{n} R_i$$

$$W_{HG} = \{w_{(HG)k}\},\ k = 1, ..., \left|T_{HG}\right| + \left|E_{HG}\right|,$$

$$w_{(HG)k} = c \cdot \sum_{i=1}^{n} \sum_{j=1}^{|W_i|} I(\delta_{ij} == k) \cdot w_{ik}.$$

Το νέο διάνυσμα βαρών W_{HG} έχει για κάθε διαφορετικό όρο του λεξιλογίου το άθροισμα των βαρών που παρουσιάζει αυτός σε όλα τα γραφήματα του συνόλου. Η συνάρτηση $I(\cdot)$ έχει τιμή 1 ή 0 ανάλογα με το αν ισχύει η έκφραση εισόδου. Όταν $c = 1$, το γράφημα αυτό είναι γράφημα αθροιστικών βαρών και όταν $c = 1/n$ είναι γράφημα μέσου όρου βαρών. Επίσης, η συνάρτηση $s(\cdot)$ παραμένει ως έχει.

Ο υπολογισμός του υπεργραφήματος είναι μία πράξη που μπορούμε να εφαρμόσουμε σε οποιοδήποτε (υπο)σύνολο κειμένων. Σε αλγοριθμικό επίπεδο τα υπεργραφήματα θα μας φανούν χρήσιμα στην αναπαράσταση των κέντρων των ομάδων όπου ουσιαστικά θα είναι η συνένωση όλων των κειμένων μίας ομάδας.

Αν εξετάσουμε καλύτερα την τελευταία σχέση για τα βάρη του υπεργραφήματος, αντιλαμβανόμαστε πως αυτό που πετυχαίνεται είναι η αντιστοίχιση των διαστάσεων των W_i στο μεγαλύτερο χώρο $\mathbb{R}^{(|T_{HG}|+|E_{HG}|)}$ που ανήκει το W_{HG}, μέσω των δεικτών δ. Το αποτέλεσμα αυτής της θεώρησης είναι πως παρά το ότι πρόκειται για αφαιρετική δομή αναπαράστασης, τα μοντέλα μπορούν να θεωρούνται διανύσματα που προκύπτουν από την 1-1 αντιστοιχία-παράθεση των χαρακτηριστικών τους στις διαστάσεις του χώρου $\mathbb{R}^{(|T_{HG}|+|E_{HG}|)}$. Ο χώρος αυτός είναι ο χώρος χαρακτηριστικών (*feature space*) ολόκληρης της συλλογής.

Θα εξηγήσουμε πως αυτό γίνεται μέσω γραφοθεωρητικών διαδικασιών και δε χρειάζεται σε καμία περίπτωση η δημιουργία διανυσμάτων στον χώρου $\mathbb{R}^{(|T_{HG}|+|E_{HG}|)}$. Έτσι, μπορούμε να χρησιμοποιούμε απευθείας διανυσματικές έννοιες, συμβολισμούς και πράξεις, θεωρώντας πως παρέχεται ένα λογικό επίπεδο προσομοίωσης των πράξεων αυτών στο χώρο των γραφημάτων. Στη συνέχεια θα παρουσιάσουμε γραφήματα που περιορίζουν το γενικό αυτό μοντέλο.

4.5. Κατευθυνόμενο Γράφημα Απλών Γειτνιάσεων

Στην προσέγγιση αυτή [54] το κείμενο αναπαρίσταται μέσω ενός κατευθυνόμενου γραφήματος, (*Directed Index Graph - DIG*). Θα δείξουμε πως η προσέγγιση αυτή αποτελεί ειδική περίπτωση του γενικευμένου γραφήματος της Παραγράφου 4.4.

Συγκεκριμένα, θεωρούμε το σύνολο σχέσεων $R = \{r_{dn}\}$ το οποίο περιέχει μία μόνο σχέση, αυτή της κατευθυνόμενης γειτνίασης r_{dn} (*directed neighbor*). Οι κατευθύνσεις κωδικοποιούν τη ροή του λόγου στο κείμενο και υποθέτουν ότι: $(t_i \rightarrow t_j) \neq (t_j \rightarrow t_i)$.

Ορισμός 4.5. *Κατευθυνόμενη γειτνίαση* r_{dn} καλείται η σχέση μεταξύ δύο όρων t_i, t_j, η οποία απορρέει από την διαδοχική παράθεση τους στο κείμενο, με τον t_i να προηγείται του t_j. Συμβολικά γράφουμε: $(t_i \rightarrow t_j)$.

Η σχέση αυτή αναγνωρίζεται εύκολα κατά την μηχανική ανάγνωση του κειμένου. Κατ' επέκταση των ορισμών της Παραγράφου 4.4, το σύνολο Σr_{dn} (t) είναι η *γειτονία* του όρου t στο αναφερόμενο κείμενο, περιέχοντας όλους τους όρους με τους οποίους ο όρος t σχετίζεται με σχέση κατευθυνόμενης γειτνίασης.

4.6. Γράφημα Απλών Γειτνιάσεων χωρίς Κατευθύνσεις

Το μοντέλο αυτό προτείνεται στην παρούσα εργασία, θεωρεί νοηματικά ισοδύναμες τις σχέσεις: $(t_i \rightarrow t_j) \equiv (t_i \leftarrow t_j)$. Αναφέρουμε ένα παράδειγμα χωρίς να υποστηρίζουμε ότι είναι το πλέον γενικό. Ας θεωρήσουμε τα κείμενα: "*the car is fast*", "this is a fast car" για τα οποία το κατευθυνόμενο μοντέλο θα έδινε ("*car*" → "*fast*"), ("*fast*" → "*car*"). Παρατηρούμε πως η πραγματική νοηματική σχέση περιλαμβάνει και τις δύο

αναπαραστάσεις, η απάντηση στην ερώτηση πώς χαρακτηρίζεται το ουσιαστικό "*car*" στα δύο κείμενα είναι "*fast*".

Η βασική παρατήρηση, είναι πως δύο λέξεις που συναντώνται μαζί θα έπρεπε να αναπαρίστανται ως μη διατεταγμένο ζεύγος από κοινού παρατηρήσεων. Θεωρούμε μη-κατευθυνόμενο γράφημα και το σύνολο $R = \{r_{undn}\}$ περιέχει μόνο μία σχέση, αυτή της μη-κατευθυνόμενης γειτνίασης r_{undn} (*undirected neighbor*).

Ορισμός 4.6. *Μη-κατευθυνόμενη γειτνίαση* r_{undn} καλείται η συμμετρική σχέση μεταξύ δύο όρων $r(t_i, t_j,)$, η οποία απορρέει από την διαδοχική παράθεση τους στο κείμενο. Συμβολικά γράφουμε: $(t_i - t_j)$.

Η σχέση αυτή αναγνωρίζεται όμοια με την κατευθυνόμενη γειτνίαση, το σύνολο $\Sigma r_{undn}(t)$, είναι η γειτονία του όρου t στο αναφερόμενο κείμενο και περιέχει όλους τους όρους με τους οποίους σχετίζεται ο όρος βάσει της σχέσης r_{undn}.

4.7. Σύνθετες Κλάσεις Γραφημάτων

Στην Παράγραφο 4.4 προτείναμε μία αφαιρετική δομή γραφημάτων η οποία μπορεί να αναπαραστήσει τα κείμενα ως σύνολα βασικής πληροφορίας που είναι οι όροι του, και σύνολα σχέσεων πάνω σε αυτούς. Δίνεται συνεπώς, η δυνατότητα να οριστούν απλές σχέσεις, όπως αυτές που είδαμε στις Παραγράφους 4.5 και 4.6, αλλά και πιο περίπλοκες για προβλήματα ειδικών απαιτήσεων.

Ένα παράδειγμα γραφήματος με περισσότερες σχέσεις είναι το γράφημα που επιτρέπει κατευθυνόμενες και μη-κατευθυνόμενες γειτνιάσεις. Αν χρήσει μίας γραμματικής ήταν δυνατό να διακρίνουμε την κατεύθυνση της νοηματικής σχέσης μεταξύ ζευγών όρων, θα μπορούσαμε να αναπαραστήσουμε αυστηρότερα και το περιεχόμενο του κειμένου. Το εν λόγω γράφημα περιγράφεται από το $R = \{r_{undn}, r_{dn}\}$.

4.8. Η Πολυπλοκότητα Χειρισμού των Μοντέλων

4.8.1. Το Διανυσματικό Μοντέλο

Για να κατανοήσουμε την ουσιαστική διαφορά ανάμεσα στην υλοποίηση που έγινε στα πλαίσια της εργασίας και των παραδοσιακών διανυσματικών θα αναφέρουμε συνοπτικά πώς θα υλοποιούσαμε το διανυσματικό μοντέλο αποδοτικά βάσει της βιβλιογραφίας [32].

Θα δημιουργούσαμε έναν ενιαίο πίνακα κατακερματισμού όρων για όλη τη συλλογή εγγράφων, αναθέτοντας ένα μοναδικό ακέραιο αναγνωριστικό σε κάθε διαφορετικό όρο της συλλογής. Για κάθε όρο θα αποθηκεύαμε επίσης τη συχνότητα εμφάνισής του σε κάθε κείμενο (ή γενικότερα το βάρος του), σε ένα διάνυσμα μήκους $|D|$. Στη συνέχεια για κάθε κείμενο θα δημιουργούσαμε ένα πυκνό διάνυσμα, το οποίο σε κάθε διάσταση θα περιείχε το ακέραιο αναγνωριστικό μίας λέξης του κειμένου (ή εναλλακτικά έναν δείκτη στον πίνακα κατακερματισμού). Έτσι, θα αντιμετωπιζόταν το φαινόμενο των αραιών τιμών στα διανύσματα συχνοτήτων.

Αν επιχειρούσαμε να προσθέσουμε και την πληροφορία των ακμών στον ίδιο σχεδιασμό δομών (δεν υπάρχει πρόταση στη βιβλιογραφία), τότε η διάσχιση ενός κειμένου θα παρήγαγε μία αυθαίρετη διάταξη από κόμβους και ακμές. Δε θα μπορούσαμε να εκμεταλλευτούμε αλγοριθμικά το γεγονός ότι η ύπαρξη των ακμών σε ένα γράφημα εξαρτάται από την ύπαρξη των αντίστοιχων κορυφών που ενώνουν. Για τη σύγκριση δύο κειμένων G_1, G_2, θα έπρεπε να ελέγχουμε την ύπαρξη κάθε στοιχείου του G_1 στο G_2 διάνυσμα. Επίσης, ο έλεγχος ύπαρξης ενός στοιχείου γραφήματος σε ένα κείμενο θα γινόταν βάση του ενιαίου πίνακα κατακερματισμού, και θα αφορούσε τον έλεγχο της συχνότητας του ζητούμενου στοιχείου στο αντίστοιχο διάνυσμα του ευρετηρίου. Το κόστος της πράξης αυτής είναι σταθερό.

Η προσέγγιση αυτή απαιτεί να έχουμε όλη τη συλλογή εξαρχής διαθέσιμη για να μπορέσουμε να δημιουργήσουμε τον ενιαίο πίνακα κατακερματισμού (ευρετήριο). Επίσης, η επέκτασή της σε γραφήματα ουσιαστικά θα δημιουργούσε ένα μεγαλύτερο αραιό διάνυσμα με διαστάσεις τα μοναδικά αναγνωριστικά κόμβων και ακμών.

4.8.2. Το Γενικευμένο Μοντέλο Γραφημάτων

4.8.2.1. Ανάλυση Δομών Δεδομένων Μοντέλου

Κάνοντας κάποιες επιλογές στις δομές δεδομένων που θα χρησιμοποιήσουμε μπορούμε να βελτιώσουμε την απόδοση διαφόρων λειτουργιών πάνω στα γραφήματα των κειμένων. Στη δική μας υλοποίηση αποθηκεύουμε τους όρους κάθε κειμένου σε έναν Πίνακα Κατακερματισμού Όρων (*ΠΚΟ*), ο οποίος ρυθμίζεται σε μέγεθος της τάξης των $O(|T|)$ κάδων παρέχοντας τη δυνατότητα γραμμικής διάσχισης όλων των όρων-κόμβων του γραφήματος και σταθερού χρόνου αναζήτησης λέξης στο κείμενο $O(1)$.

Παράλληλα, οι σχέσεις και οι σχετιζόμενοι όροι του συνόλου $\Sigma_R(t)$ (η γειτονιά στις σχέσεις απλών γειτνιάσεων) ανάμεσα στον όρο t και σε άλλους του ίδιου κειμένου διατηρούνται ως ακμές σε έναν Πίνακα Κατακερματισμού Ακμών (*ΠΚΑ*(t)), ο οποίος αποθηκεύεται εσωτερικά σε κάθε κόμβο του *ΠΚΟ*. Οι διαφορετικές σχέσεις πάνω στο ίδιο ζεύγος όρων μπορούν να διατηρούνται ως διαφορετικές ετικέτες στις ακμές προκαθορισμένου αριθμού (σταθερός χρόνος για έλεγχο ετικέτας σε ακμή). Η δυνατότητα ρύθμισής του μεγέθους του *ΠΚΑ*(t) για κάθε όρο χωριστά, βάσει της συνδεσιμότητας του δηλαδή O(αριθμού ακμών του t στο κείμενο), επιτρέπει γραμμική διάσχιση και των ακμών που αφορούν έναν όρο και είναι αποθηκευμένες σε αυτόν (π.χ. η ακμή στα κατευθυνόμενα γραφήματα αποθηκεύεται στον όρο αφετηρία), και σταθερό χρόνο αναζήτησης ακμής σε αυτή. Τέλος, ο *ΠΚΑ*(t) μπορεί να μη δεσμεύεται καν, σε περίπτωση που ο κόμβος t δεν έχει συνδέσεις ή οι σχετιζόμενοι όροι αποκοπούν κατά το φιλτράρισμα.

Να σημειώσουμε πως αν επιλεγεί το μοντέλο μη-κατευθυνόμενων γραφημάτων γίνεται μία σύμβαση ώστε κάθε σχέση $r(t_i, t_j)$ να αποθηκεύεται ως εγγραφή-ακμή μόνο στη γειτονιά του όρου που είναι λεξικογραφικά μεγαλύτερος. Το γεγονός αυτό λαμβάνεται υπόψη όταν αναζητείται μία ακμή σε ένα κείμενο, και έτσι περιορίζεται στη γειτονιά που θα μπορούσε να υπάρχει η ακμή.

Η διάσχιση ολόκληρου του γραφήματος αναπαράστασης είναι γραμμική ως προς το μέγεθος των στοιχείων γραφήματος $O(|T|+|E|)$ (Σχήμα 4.1), με το $|E|$ να εξαρτάται σε μικρό βαθμό από το πλήθος των διαφορετικών σχέσεων που έχουμε ορίσει $|R|$ λόγω της ενσωμάτωσης πολλών ετικετών σε μία ακμή.

```
Διαδικασία Διάσχισης Μοντέλου Γραφήματος Κειμένου (di)
{
    Για κάθε κάδο του κειμένου di->ΠΚΟ[b], b = 1, …,αριθμός_κάδων(ΠΚΟ)
    {   L = ΠΚΟ[b]->λίστα_όρων_κάδου

        Όσο L != NULL,                          /* διέτρεξε την λίστα κάδου όρων */
        {   Επεξεργασία όρου L->αλφαριθμητικό_όρου

                                    /* εύκολη μετατροπή μοντέλου σε διανυσματικό */
            Αν ΣΥΝΥΠΟΛΟΓΙΣΜΟΣ_ΑΚΜΩΝ_ΣΤΟ_ΜΟΝΤΕΛΟ == 1 && L->αριθμός_σχέσεων > 0
            {   Για κάθε κάδο (L->ΠΚΣ)[j], j = 1, …,αριθμός_κάδων(L->ΠΚΣ)
                {   Μ = (L->ΠΚΣ)[j]->λίστα_σχέσεων_του_όρου

                    Όσο Μ != NULL,              /* διέτρεξε την λίστα κάδου ακμών */
                    {  Επεξεργασία ακμής Μ->αλφαριθμητικό_ακμής και ετικετών της
                       Μ = Μ->επόμενος_σχετιζόμενος_όρος
                    }
                }
            }

            L = L->επόμενος_όρος
        }
    }
}
```

Σχήμα 4.1. Ψευδοκώδικας γραμμικής διάσχισης και επεξεργασίας του γραφήματος ενός κειμένου.

Η πορεία διάσχισης ενός κειμένου μπορεί να περιγραφεί από την ακολουθία στοιχείων: (*όρος*1, {*ακμές όρου*1}, …, *όρος*|*T*|, {*ακμές όρου*|*T*|}). Η διάσχιση αυτή είναι ακριβώς ίδια με τη διάσχιση ενός πίνακα γειτνίασης (2Δ πίνακα, ή πίνακα λιστών) που συνήθως αναπαριστά στη μνήμη του υπολογιστή ένα γράφημα. Κατά την σύγκριση δύο γραφημάτων στις εντολές επεξεργασίας του Σχήματος 4.1 εκτελείται μία αναζήτηση στο δεύτερο γράφημα, για το στοιχείο γραφήματος που συναντάμε στο πρώτο. Όπως εξηγήσαμε, η αναζήτηση οποιουδήποτε στοιχείου γραφήματος είναι πράξη σταθερού χρόνου στην υλοποίησή μας. Συνεπώς και η από κοινού επεξεργασία δύο γραφημάτων, η οποία απαιτεί εκατέρωθεν ταυτοποίηση όρων και ακμών, είναι μία γραμμικού χρόνου διαδικασία στον αριθμό των στοιχείων γραφήματος του πρώτου κειμένου (κείμενο αναφοράς), $O(|T_1|+|E_1|)$.

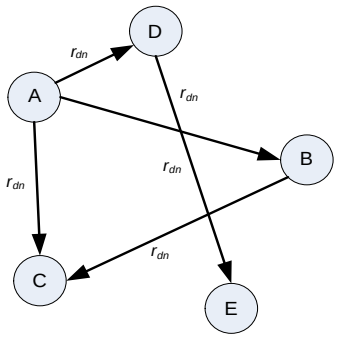
Το κέρδος σε σχέση με την απλή επέκταση της υλοποίησης της παρ. 4.8.1 είναι πως λόγω της δομημένης διάσχισης του γραφήματος, με τη μη ταυτοποίηση όρου *t* στο δεύτερο γράφημα παραλείπεται και ο έλεγχος ύπαρξης των ακμών του *t* και η διαδικασία προχωρά στην εξέταση του επόμενου όρου του κειμένου. Αυτό μας συμβουλεύει να ελέγχουμε το μέγεθος των δύο εξεταζόμενων γραφημάτων και να θεωρούμε ως εξωτερικό γράφημα αναφοράς της διαδικασίας (πρώτο γράφημα) αυτό με το μικρότερο αριθμό στοιχείων. Με τον τρόπο αυτό μειώνουμε τον αριθμό των

αναζητήσεων στο δεύτερο γράφημα, όπου μαζί με την ξεχωριστή ρύθμιση των μεγεθών των πινάκων κατακερματισμού παίζουν σημαντικό ρόλο στη αποδοτικότερη εκτέλεση ρουτινών που αφορούν συνθετικά γραφήματα, που είναι συνενώσεις γραφημάτων κειμένων και περιέχουν πολλά στοιχεία γραφήματος (π.χ. το υπεργράφημα G_{HG}).

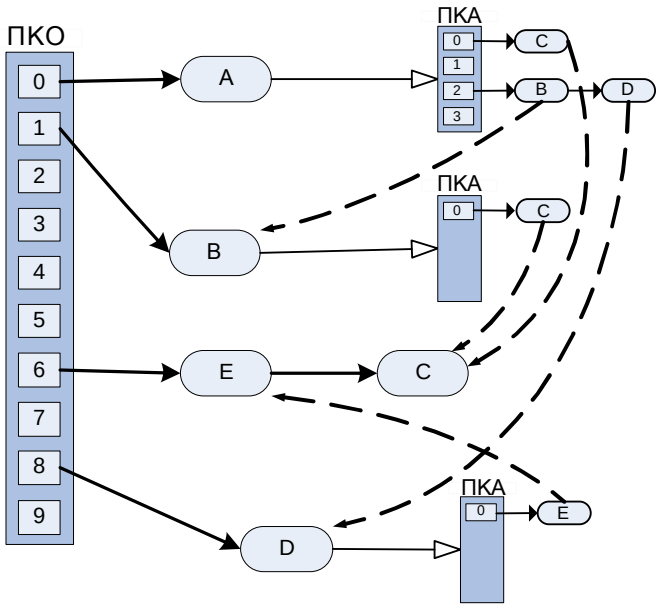
Επίσης, δεν απαιτείται η ύπαρξη ενιαίου ευρετηρίου σε περίπτωση που δε διατίθεται ολόκληρη η συλλογή εξαρχής, διότι μπορούν να διατηρούνται ακόμα και τα αλφαριθμητικά σε κάθε *ΠΚΟ* ανεξάρτητα. (Στην προεπεξεργασία το ευρετήριο που αναφέραμε στο Κεφάλαιο 3 χρησιμοποιείται για τον υπολογισμό των *idf*(*t*) ο οποίος είναι αδύνατος έτσι κι αλλιώς στην περίπτωση που δεν έχουμε εξαρχής ολόκληρο το σύνολο δεδομένων).

4.8.2.2. Ένα Παράδειγμα Αναπαράστασης

Ας υποθέσουμε ότι έχουμε ένα κείμενο 5 όρων το οποίο αναπαρίσταται ως γράφημα κατευθυνόμενης απλής γειτνίασης. Επίσης θεωρούμε πίνακες κατακερματισμού *ΠΚΟ* δέκα κάδων, και *ΠΚΑ* ρυθμιζόμενου μεγέθους περίπου στον αριθμό των ακμών μίας γειτονιάς (σημείωση: οι τιμές κατακερματισμού είναι υποθετικές). Τα Σχήματα 4.2 και 4.3 παρουσιάζουν την αφαιρετική δομή και την εσωτερική αναπαράσταση στη μνήμη του υπολογιστή αντίστοιχα.



Σχήμα 4.2. Αναπαράστασης κειμένου με γράφημα κατευθυνόμενης γειτνίασης, $R = \{r_{dn}\}$.



Σχήμα 4.3. Η εσωτερική αναπαράσταση του κειμένου στη μνήμη.

4.8.3. Εκτίμηση Υπολογιστικού Κόστους Βασικών Πράξεων στα Μοντέλα

Μπορεί κάποιος να ισχυριστεί πως ασυμπτωτικά έχουμε το ίδιο αποτέλεσμα με την υλοποίηση 4.8.1, όμως θυμίζουμε πως ένα γράφημα $|T|$ κόμβων περιέχει το πολύ $|T|^2$ ακμές. Δηλαδή, η διάσχισή του κοστίζει το πολύ $O(|T|+|T|^2) = O(|T|^2)$. Είναι δύσκολο να συναντήσουμε τη χειρότερη περίπτωση στα κείμενα, διότι αυτό θα σήμαινε πως κάθε όρος έχει σύνδεση με κάθε άλλο στο κείμενο, όμως παρόλα αυτά μπορούμε να κάνουμε μία υπόθεση εργασίας για να εκτιμήσουμε το υπολογιστικό κέρδος της αποθήκευσης των ακμών στους κόμβους, όπως προτείνουμε στην εργασία.

Ως αναφορά την πράξη σύγκρισης των γραφημάτων, μπορούμε να επιτύχουμε μία βελτίωση του υπολογιστικού κόστους της. Είναι από τις πλέον κρίσιμες διαδικασίες διότι εφαρμόζεται σε κάθε ζεύγος του συνόλου δεδομένων αρχικά ώστε να δημιουργηθεί ο $|D| \times |D|$ πίνακας ομοιοτήτων.

Παρατηρείται, λοιπόν, πως οι ακμές των γραφημάτων έχουν τάξη $O(\alpha \cdot |T|) = O(|T|)$. Η τιμή του α μπορεί να εκτιμηθεί από το γεγονός ότι μία λέξη συμμετέχει εν γένει σε δύο ακμές σε κάθε εμφάνισή της, αυτή με την προηγούμενη και αυτή με την επόμενη λέξη του κειμένου (εξαίρεση αποτελούν η αρχική και τελική λέξη μίας πρότασης), δηλαδή ισχύει $\alpha > 1$, ενώ το $\alpha \cdot |T|$ φράσσεται άνω από το μήκος του

κειμένου (το μήκος είναι το άθροισμα συχνοτήτων των όρων του). Αν αφαιρούνται οι ακμές της συλλογής που εμφανίζονται σ ένα μόνο κείμενο τότε παρατηρείται πως α < 1, λόγω της αυστηρότητας της άμεσης γειτνίασης, παρόλο που σε ένα μεγάλο σύνολο δεδομένων δε θα παρατηρείται έντονα κάτι τέτοιο. Επίσης, η μέση τομή λεξιλογίου λ των δύο κειμένων, σπάνια ξεπερνά το 25% (συνήθως είναι λιγότερο λόγω των διαφορετικών θεμάτων των κατηγοριών). Έτσι, συμπεραίνουμε πως με την αναπαράσταση του κειμένου ως γράφημα κάνουμε την καλύτερη δυνατή προσπάθεια και περιορίζουμε το κόστος σε $O(\lambda \cdot |T|)$, με κέρδος το σταθερό όρο $\lambda/\alpha < 1$.

Για παράδειγμα, θεωρώντας πως οι παρατηρήσεις μας έχουν μία δόση γενικότητας, τότε: αν $\alpha = 2$ και $\lambda = 0.25$, ο όρος γίνεται $\lambda/\alpha = 0.125$. Αν και ασυμπτωτικά ασήμαντο, σε ένα πραγματικό σύστημα αυτό σημαίνει πως η βασική πράξη της σύγκρισης δύο κειμένων θα εκτελείται σε περίπου 8 φορές λιγότερο υπολογιστικό κόστος. Σε περιπτώσεις όπου το σύνολο δεδομένων περιέχει χιλιάδες έγγραφα μία βελτίωση της τάξης αυτής επιφέρει μεγάλη οικονομία χρόνου.

4.9. Διανυσματοποίηση Μοντέλων Αναπαράστασης Γραφημάτων

Αναφερόμενοι στα μοντέλα γραφημάτων, παρουσιάζεται ένα σημαντικό πρόβλημα που αφορά τον αυστηρό μαθηματικό χειρισμό τους. Για παράδειγμα, πώς θα οριστεί η νόρμα-2 $\|G\|_2$ ενός γραφήματος ώστε να κανονικοποιήσουμε τα βάρη του; Αν η κανονικοποίηση αφορούσε μόνο τους όρους του κειμένου θα αντιμετωπίζαμε το γράφημα ως διάνυσμα και διατρέχοντάς το θα κάναμε τους απαραίτητους υπολογισμούς για τη νόρμα. Η επέκταση της πράξης σε ολόκληρο το γράφημα, ακμών και κορυφών, είναι προφανής διότι και πάλι θα διατρέχαμε το γράφημα συμπεριλαμβάνοντας και τις ακμές στους υπολογισμούς.

Οι πράξεις σύγκρισης περιεχομένου (ομοιότητας) δύο γραφημάτων εμπεριέχουν πράξεις ανάμεσα στα βάρη των κοινών στοιχείων γραφήματος (πολλαπλασιαστικά ή προσθετικά). Είναι εύκολο να παρατηρήσει κανείς πως η διαδικασία γραμμικής διάσχισης ενός γραφήματος και ταυτοποίησης στοιχείων του με ένα δεύτερο γράφημα είναι εντελώς ισοδύναμη με την επεξεργασία διανυσμάτων κατά διάσταση.

Θεωρούμε μία αυθαίρετη συνάρτηση πάνω σε μοντέλα γραφημάτων $F_G(i)$ η οποία επιστρέφει το i-οστό στοιχείο γραφήματος, $i = 1, \ldots, |G|$, θεωρώντας μία πλήρη

διάταξή τους, ώστε η $s(F_G(i))$ να επιστρέφει το βάρος του *i*-οστού στοιχείου του γραφήματος. Ακόμα, θεωρούμε την προφανή συνάρτηση πάνω σε διανύσματα $F_v(i) = v_i$ που επιστρέφει απευθείας την τιμή της *i*-οστής διάστασης του διανύσματος *v*, και *dim*(*v*) τη διάσταση του χώρου του *v*, τότε οι παρακάτω υπολογισμοί δείχνουν την ισοδυναμία υπολογισμών:

η νόρμα διανύσματος: $\|v\|_2 = \left(\sum_{i=1}^{\dim(v)} \left(F_v(i)\right)^2 \right)^{1/2}$

η «νόρμα» γραφήματος: $\|G\|_2 = \left(\sum_{i=1}^{|G|} \left(s\left(F_G(i)\right)\right)^2 \right)^{1/2}$

Η νόρμα είναι ένα χαρακτηριστικό παράδειγμα σύνθετου υπολογισμού πάνω στις διαστάσεις του μοντέλου δεδομένων, με τον ίδιο τρόπο είναι δυνατόν να υπολογιστούν ποσότητες ανάμεσα σε ζεύγη γραφημάτων, όπως το εσωτερικό γινόμενο $G_1 \cdot G_2$, με τη βοήθεια μίας συνάρτησης αναζήτησης στοιχείου σε γράφημα, έστω *search* $(G, F_G(i))$. Θα παίρναμε λοιπόν:

«εσωτερικό γινόμενο» γραφήματος: $\left(\sum_{i=1}^{|G|} s\left(F_{G_1}(i)\right) \cdot s\left(search(G_2, F_{G_1}(i))\right) \right)^{1/2}$

Εδώ φαίνεται πως μία διανυσματική πράξη, η οποία προϋποθέτει την ύπαρξη διανυσμάτων στον ίδιο χώρο, μπορεί να γίνει απευθείας στον χώρο της κλάσης των γραφημάτων GC_R. Λόγω της συνάρτησης *search*(·) και της $F_G(\cdot)$ η εκ των προτέρων γνώση του μεγέθους του χώρου $\mathbb{R}^{(|T_{HG}|+|E_{HG}|)}$ δεν απαιτείται. Συνεπώς, το ενιαίο ευρετήριο δεν είναι απαραίτητο για να αναθέτει μοναδικά αναγνωριστικά στα διαφορετικά στοιχεία της συλλογής, αν και μπορεί να αντικαταστήσει τις συγκρίσεις αλφαριθμητικών με συγκρίσεις αναγνωριστικών.

Η πράξη του εσωτερικού γινομένου είναι πολύ σημαντική επειδή ισοδυναμεί με τον υπολογισμό μίας μορφής ομοιότητας ανάμεσα σε δύο διανύσματα κειμένων όταν αυτά είναι κανονικοποιημένα (*cosine measure*). Το ίδιο πετυχαίνεται και στην προτεινόμενη προσέγγιση χρήσει του γενικευμένου μοντέλου αναπαράστασης με γραφήματα.

Συμπεράσματα για τα Μοντέλα Αναπαράστασης

Η ουσιαστική συμβολή του ορισμού που δώσαμε για το γενικευμένο γράφημα αναπαράστασης κειμένων είναι πως ορίζεται ένα λογικό επίπεδο χειρισμού των γραφικών αναπαραστάσεων. Για μία εφαρμογή, η πληροφορία των ακμών-σχέσεων θα πρέπει να αντληθεί από τα κείμενα με διαδικασίες εξόρυξης δεδομένων. Οι όροι αντιπροσωπεύονται μοναδικά από έναν κόμβο και οι σχέσεις αποτελούν ετικέτες των ακμών στο γράφημα αναπαράστασης. Αυτές μπορούν να συνυπάρχουν ως χαρακτηριστικά μίας ακμής, δηλ. στην υλοποίηση των σχετικών δομών δεδομένων να επιτρέπεται η ανάθεση περισσότερων από μία ετικέτες στις ακμές. Τελικά, στην περίπτωση του μη κατευθυνόμενου γραφήματος απλής γειτνίασης που χρησιμοποιείται στην εργασία, απαιτούνται τόσες ακμές όσες και τα διαφορετικά μη διατεταγμένα ζεύγη όρων που εμφανίζονται στο κείμενο.

Η προσέγγιση των κατευθυνόμενων μονοπατιών γειτνίασης αποτελεί γενίκευση της κατευθυνόμενης απλής γειτνίασης, αν δεχτούμε ότι η γειτνίαση δύο λέξεων είναι ένα μονοπάτι μήκους ένα. Παρουσιάζει το πλεονέκτημα πως μπορεί να εκμεταλλευτεί την πληροφορία των ακμών χωριστά αλλά και ως ενιαίο σύνολο χαρακτηριστικών του κειμένου, όμως το κόστος είναι να διατηρεί ακμές ανάλογες του μήκους του κειμένου. Επίσης, στερείται ανοχής στη διαφορετικότητα με την οποία μπορεί να εμφανίζεται η ίδια φράση σε διαφορετικά κείμενα, π.χ. η φράση *"A→B→C→D"* δεν αναγνωρίζεται σωστά όταν συναντάται ως *"C→D→B→A"*. Δεν είναι καθόλου δύσκολο να σκεφτούμε τέτοια παραδείγματα, διότι πηγάζουν από την ελευθερία του γραπτού Λόγου.

Φράση Κειμένου	Μοντέλο Κατευθυνόμενων Μονοπατιών	Μοντέλο Μη-Κατευθυνόμενων Απλών Γειτνιάσεων
The kid took the ball and went to play…	kid → ball → play	kid − ball − play
The kid went to play with the ball…	kid → play → ball	kid − play − ball

Σχήμα 4.4. Διαφορές στη μοντελοποίηση ίδιων νοηματικά φράσεων από τα εξεταζόμενα μοντέλα αναπαράστασης κειμένων.

Κάναμε ιδιαίτερη αναφορά στο πρόβλημα της υψηλής διάστασης των κειμένων και την ανάγκη μείωσης των χαρακτηριστικών τους. Η αφαίρεση μη-πληροφοριακών όρων από το κείμενο καταστρέφει σε μεγάλο βαθμό τις δομές των φράσεων που επιθυμούμε να αντλήσουμε από τα κείμενα. Θυμίζουμε πως η αφαίρεση ενός τετριμμένου όρου

(π.χ. ένα άρθρο) δημιουργεί ακμή μεταξύ των εκατέρωθεν όρων, αντίθετα για μη-τετριμμένο όρο δεν πράττουμε κάτι τέτοιο γιατί δεν είναι προφανής η ορθότητα του.

Η διατήρηση της ακριβούς πληροφορίας για τις φράσεις του κειμένου και οι τεχνικές μείωσης διάστασης είναι γενικά ασύμβατες προσεγγίσεις. Ειδικά σε περιπτώσεις μεγάλων συλλογών που πιθανόν επιθυμούμε να μειώσουμε δραστικά τον όγκο των μοντέλων η διατήρηση των μονοπατιών θα ήταν δύσκολη.

Με άλλα λόγια, η πληροφορία που ιδανικά θα μπορούσαμε να αντλήσουμε από τις φράσεις είναι σημαντική αλλά είναι δύσκολο να τη μεταφέρουμε στο χώρο των χαρακτηριστικών απομνημονεύοντας την ακριβή διαδοχή των λέξεων. Η αναπαράσταση αυτή επιβαρύνει περισσότερο τις τεχνικές μάθησης: α) αυξάνοντας την πολυπλοκότητα και τη διάσταση των μοντέλων με πολλές κατευθυνόμενες ακμές μονοπατιών, β) περιορίζοντας την εφαρμογή τεχνικών μείωσης του λεξιλογίου, γ) μη καταφέρνοντας την ευέλικτη αναπαράσταση των φράσεων του κειμένου λόγω αυστηρότητας διαδοχής των όρων. Αυτές βέβαια είναι μία σειρά λογικών παρατηρήσεων οι οποίες περισσότερο σκιαγραφούν το τι θα πρέπει να εξετάσει πειραματικά κάποιος για να ισχυριστεί την ακαταλληλότητα της προσέγγισης και όχι έναν πλήρως τεκμηριωμένο ισχυρισμό.

Αυτό που υποστηρίζεται και προτείνεται στην παρούσα εργασία είναι πως:

- η αποσύνθεση των χαρακτηριστικών ενός κειμένου σε στοιχειώδη και ανεξάρτητα χαρακτηριστικά μπορεί να δώσει πολύ μεγαλύτερη ευελιξία στο μοντέλο αναπαράστασης ώστε να μην εξαρτάται από την ακριβή διάταξη των όρων στις φράσεις. Η κατεύθυνση των ακμών είναι μία υπόθεση που επίσης στερεί από το μοντέλο τη δυνατότητα ανάδειξης πραγματικών ομοιοτήτων μεταξύ των κειμένων.
- η δυνατότητα άμεσου μαθηματικού χειρισμού των μοντέλων ως διανυσματικές αναπαραστάσεις είναι ακόμα μία ευκολία, ενώ για την περίπτωση της ομοιότητας σε γράφημα μονοπατιών θα έπρεπε να οριστεί μία πιο πολύπλοκη διαδικασία για φράσεις ο οποίος να ταυτίζει και υποφράσεις.

Το πρώτο σημείο, περιορίζει αρκετά και το μέγεθος των μοντέλων αναπαράστασης. Επίσης, σε περιπτώσεις που τίθενται περιορισμοί από το πρόβλημα (χρόνου, μεγέθους συλλογής, πολυπλοκότητας) δίνεται η δυνατότητα να εφαρμοστούν ακόμα και δραστικό φιλτράρισμα χωρίς να αναιρείται η φιλοσοφία του μοντέλου.

4.10. Διαχείριση Περιεχομένου – Βαρών των Γραφημάτων

4.10.1. Ο Διαχωρισμός του Περιεχομένου Σχέσεων και 'Όρων Κειμένου

Κατά την δημιουργία του γραφήματος αναπαράστασης, παρουσιάζεται ένα πρόβλημα που έχει ως αιτία τους διαφορετικούς τύπους πληροφορίας που χειριζόμαστε: τους όρους και τις ακμές. Οι όροι αυτόνομα είναι ένα σύνολο πληροφορίας στο οποίο μπορούμε να εφαρμόσουμε οποιοδήποτε σχήμα βαρών και να προκύψει ένα συνεπές σχήμα πληροφορίας, το ίδιο ισχύει μεμονωμένα και για τις ακμές.

Η συνέπεια αφορά το ότι κάθε ένα από τα στοιχειώδη χαρακτηριστικά είναι ίδιας φύσης-κατηγορίας, άρα και οι παρατηρήσεις πάνω σε αυτά (π.χ. συχνότητες) είναι συγκρίσιμες. Οι κατηγορίες στη γενική περίπτωση είναι: α) όλες οι λέξεις αποτελούν την κύρια κατηγορία πληροφορίας, β) κάθε διαφορετικού τύπου σχέση $r \in R_{HG}$ που ορίζεται στο σύνολο γραφημάτων της συλλογής ορίζει και μία ξεχωριστή κατηγορία χαρακτηριστικών. Θα περιοριστούμε στην περίπτωση μίας σχέσης, της μη-κατευθυνόμενης γειτνίασης $R_{HG} = \{r_{undn}\}$. Να διευκρινίσουμε πως εδώ πλέον χρησιμοποιούμε το γράφημα G ως διάνυσμα χαρακτηριστικών. Ομοίως θεωρείστε και για την τις περιπτώσεις των υπεργραφημάτων G_{HG}.

Η συνέπεια που συμπεράναμε για τις ακμές όταν αυτές θεωρούνται χωριστά, αντίστοιχα και για τους όρους, δεν είναι προφανής και στην περίπτωση που έχουμε ένα γράφημα με όρους και σχέσεις (ακμές). Τα ερωτήματα που τίθεται είναι:

- ποια είναι η διαφορά στον συνυπολογισμό της συχνότητας f_e μίας ακμής, συγκριτικά με τον συνυπολογισμό της συχνότητας $f_t = f_e$ ενός όρου, στη διαμόρφωση του περιεχομένου του κειμένου.
- πόση συνολική συνεισφορά θα έχει κάθε κατηγορία πληροφορίας στη διαμόρφωση του περιεχομένου.

Όσων αφορά το πρώτο σημείο, είναι ο κύριος λόγος που ορίστηκε η συνάρτηση $s(g_{ij})$ μαζί με το μοντέλο γραφημάτων. Για να μπορεί να διαχωρίσει το σχήμα ανάθεσης βαρών στα διαφορετικής κατηγορίας στοιχεία του γραφήματος. Έτσι, το $W_i = \{w_{ij}\}$ ενός κειμένου, δεν εξαρτάται απευθείας από τα σχήματα βαρύτητας. Η συνάρτηση υπολογίζει ειδικά το βάρος κάθε κατηγορίας στοιχείου γραφήματος (προς αποφυγή σύγχυσης χρησιμοποιούμε απευθείας την $s(g)$). Για το δεύτερο σημείο, μία σχετική

παρατήρηση συναντά κάποιος στο [52] όπου πειραματικά διαπιστώνεται πως η αναλογία του ποσοστού που βασίζεται η συνάρτηση ομοιότητας στα μονοπάτια και στους όρους του κειμένου παίζει σημαντικό ρόλο στην ποιότητα της ομαδοποίησης.

Θα πρέπει λοιπόν, να ορίζεται μία αναλογία συνεισφοράς περιεχομένου κάθε κατηγορίας στοιχείων γραφήματος στο συνολικό περιεχόμενο. Αν για παράδειγμα θεωρήσουμε ισονομία στις κατηγορίες, το περιεχόμενο ενός κειμένου ορίζεται ως το άθροισμα των κανονικοποιημένων βαρών ως προς κάποια νόρμα διανυσμάτων $\|\cdot\|_p$:

$$Cnt^{(p)}(G_i) = \frac{1}{\|G_i\|_p}\left(\sum_{j=1}^{|G_i|} s(g_{ij})^p\right)^{1/p} = 1.$$

Αν πάλι δεν επιθυμούμε ισονομία χρησιμοποιούμε την παράμετρο μίξης περιεχομένου ακμών *blending factor* (*bf*), $bf \in [0,1]$. Θεωρούμε πως «σπάμε» το διάνυσμα χαρακτηριστικών σε δύο διανύσματα, αυτό που περιέχει τα βάρη των κόμβων του γραφήματος $G^{(t)} = \{g^{(t)}\}$ και αυτό των ακμών $G^{(e)} = \{g^{(e)}\}$. Η αποσύνθεση αυτή παρατηρήστε, ότι μπορεί να γίνει μόνο στον χώρο των χαρακτηριστικών (δε μπορούμε να απομονώσουμε μία ακμή από τους κόμβους που συνδέει). Συνεπώς, το περιεχόμενο ενός κειμένου υπολογίζεται ως εξής:

$$Cnt^{(p)}(G_i) = (1-bf)\cdot\left[\frac{1}{\|G_i^{(t)}\|_p}\left(\sum_{j=1}^{|G_i^{(t)}|} s(g_{ij}^{(t)})^p\right)^{1/p}\right] + bf\cdot\left[\frac{1}{\|G_i^{(e)}\|_p}\left(\sum_{j=1}^{|G_i^{(e)}|} s(g_{ij}^{(e)})^p\right)^{1/p}\right] = 1.$$

Η τελευταία σχέση υποδηλώνει πως η κανονικοποίηση στην περίπτωση αυτή θα γίνει χωριστά για τα βάρη ακμών και κορυφών και το αποτέλεσμα της μίξης είναι και πάλι περιεχόμενο ίσο με τη μονάδα, το οποίο είναι βασική απαίτηση.

Θα μελετήσουμε στο πειραματικό κομμάτι της εργασίας την επίδραση του περιεχομένου των ακμών στην ομαδοποίηση. Κατά την υλοποίηση, η μίξη αυτή ενσωματώθηκε στη συνάρτηση ομοιότητας ώστε να είναι δυνατό να μεταβάλλεται ο *bf* κατά την εκτέλεση. Όταν ο παράγοντας $bf = 0$ μηδενίζουμε την επιρροή των ακμών στο περιεχόμενο του κειμένου, αντίστοιχα για τους όρους όταν $bf = 1$. Θα φανεί πως οι ακμές μπορούν να βοηθήσουν το πρόβλημα της ομαδοποίησης, αλλά η μεταβολή του *bf* μπορεί να μεταβάλλει αρκετά και τις λύσεις. Θα δείξουμε επίσης μία ευρετική τεχνική χωρίς επίβλεψη που μπορεί να βρει μία καλή τιμή, αλλά όχι βέλτιστη, για την παράμετρο μίξης.

4.10.2. Σχήματα Βαρών για τα Στοιχεία Γραφήματος

Στο προηγούμενο κεφάλαιο περιγράφηκαν δημοφιλή σχήματα τα οποία μπορούν να εφαρμοστούν σε οποιαδήποτε αναπαράσταση. Το *boolean weighting* μοντέλο αναφέρεται στα [49][50][67], ενώ στο [54] χρησιμοποιήθηκαν και τα *TF*, *TS*. Στην εργασία αυτή, πέρα από τα πειράματα που αφορούν τις συναρτήσεις ομοιότητας και τα μοντέλα αναπαράστασης όπου χρησιμοποιούμε όλες τις εκδοχές, χρησιμοποιήσαμε την $h_{ij} \cdot IDF$ προσέγγιση (βλ. Παράγραφο 3.3.1). Με τον τρόπο αυτό καθορίζουμε το σύνολο W_i για το γράφημα G_i. Εξαίρεση εφαρμογής του *IDF* είναι φυσικά το λογικό σχήμα (*boolean weighting*) που δε μπορούμε να ορίσουμε βάρη διάφορα των {0, 1}.

Πειραματικά παρατηρήθηκε πως ο *IDF* όρος δε βοηθούσε την διαδικασία μάθησης όταν εφαρμοζόταν και στις ακμές (για τα εσωτερικά σχήματα *TF* και *TS*). Ο λόγος είναι πως μία ακμή εμφανίζεται σπανιότερα από τους όρους που συνδέει και συνεπώς θα έχει υπερβολικά μεγάλες τιμές *idf*. Επίσης, η ανάθεση βαρών στις ακμές με τον τρόπο που αντιμετωπίζουμε τις λέξεις είναι μία απλουστευτική προσέγγιση διότι οι συχνότητες των ακμών που δημιουργούνται ανάμεσα σε όρους χαμηλού περιεχομένου (οι οποίοι περιέχουν αρκετό θόρυβο) μπορεί να είναι αθροιστικά περισσότερες από αυτές των αντιπροσωπευτικών όρων. Για να αντιμετωπίσουμε τα δύο ζητήματα αυτά επιλέξαμε την ανάθεση βάρους μέσω του μικρότερου βάρους από τους συνδεόμενους όρους, η συνάρτηση $s(g)$ ορίστηκε ως:

$$s(g_{ij}) = h_{ij} \cdot \left[1 + \ln\left(1 + \min(w_i, w_j)\right)\right],$$

όπου w_i, w_j τα βάρη των συνδεόμενων λέξεων. Ο φυσικός λογάριθμος επιλέχθηκε (αντί μίας βάσης π.χ. του 10) ώστε να δίνεται μεγαλύτερη έμφαση στους συνδεόμενους όρους, ενώ το ότι παίζει ρόλο μόνο ο μικρότερος συνδεόμενος όρος της ακμής σημαίνει πως οι όροι με μικρά βάρη θα δημιουργούν μικρής σημασίας ακμές, ακόμα και αν τυχαίνει να συνδέονται με συχνές λέξεις, με αποτέλεσμα οι συχνότεροι-σημαντικότεροι όροι να βγαίνουν περισσότερο κερδισμένοι από την ύπαρξη ακμών.

ΚΕΦΑΛΑΙΟ 5. ΜΕΤΡΑ ΟΜΟΙΟΤΗΤΑΣ

5.1. Ορισμός Συνάρτησης Ομοιότητας

Έχοντας δύο αντικείμενα *x, y* ενός δοθέντος χώρου, το μέτρο ομοιότητας ή συνάρτηση ομοιότητας *sim*(*x, y*) αποτιμά την ομοιότητα τους σε μία αριθμητική ποσότητα. Η έννοια της ομοιότητας είναι συμπληρωματική της απόστασης *dist*(*x, y*), που εκφράζει την ανομοιότητα των αντικειμένων. Όταν ισχύει $sim(x, y) \in [0, 1]$ εύκολα μετατρέπουμε τη συνάρτηση ομοιότητας σε συνάρτηση απόστασης στο ίδιο διάστημα (και αντίστροφα): $dist(x, y) = 1 - sim(x, y)$. Επίσης, αν θεωρήσουμε ένα ακόμα αντικείμενο *z*, είναι επιθυμητό να ισχύουν οι παρακάτω ιδιότητες ώστε να χαρακτηρίσουμε μία συνάρτηση απόστασης μετρική (*metrics*):

1. $dist(x, y) = 0 \Leftrightarrow x = y$,
2. $dist(x, y) = dist(y, x)$,
3. $dist(x, y) + dist(y, z) \geq dist(x, z)$.

Η τελευταία σχέση ονομάζεται τριγωνική ανισότητα και είναι δυνατόν να επιταχύνει τη λήψη αποφάσεων σε κάποια προβλήματα, συμπεραίνοντας σχέσεις μεταξύ των αντικειμένων χωρίς να υπολογίζεται αναλυτικά η απόστασή τους.

Θεωρείται όμως αρκετά αυστηρή για αρκετά προβλήματα, γι' αυτό συνήθως στεκόμαστε στις δύο πρώτες. Στο [63] δίνεται ένας γενικός ορισμός για τα μέτρα ομοιότητας σύμφωνα με τη Θεωρία της Πληροφορίας:

$$sim(x, y) = \frac{x \cap y}{x \cup y}.$$

Η σχέση αυτή εκφράζει την ομοιότητα ως «ποσοστό» όμοιων χαρακτηριστικών, για παράδειγμα στα κείμενα αφαιρετικά σημαίνει ποσοστό κοινού νοηματικού περιεχομένου. Πολλές συναρτήσεις ομοιότητας μπορούν να οριστούν πάνω σε αυτή την γενική αρχή, όμως ο ακριβής υπολογισμός που υλοποιούν παίζει σημαντικό ρόλο στο αν τελικά ένα μέτρο είναι κατάλληλο για ένα πρόβλημα προβλήματα.

Πριν από όλα, είναι απαραίτητο να εντοπίσουμε τα χαρακτηριστικά ενός «κατάλληλου» μέτρου ομοιότητας και να αποφασίσουμε για μία σειρά επιλογών:

- αν και υπάρχουν μέτρα απόστασης-ομοιότητας τα οποία έχουν επιδείξει μια γενικότητα που τους επιτρέπει να χρησιμοποιούνται σε πολλά πεδία (π.χ. Ευκλείδεια απόσταση), η καταλληλότητα έχει να κάνει με τα ιδιαίτερα χαρακτηριστικά του προβλήματος που επιθυμούμε να λύσουμε.
- η βασική απαίτηση από μία συνάρτηση είναι να επιδεικνύει ανοχή σε μετασχηματισμούς που είναι πιθανοί στα δεδομένα του προβλήματος, αλλά δεν είναι σημαντικοί για αυτό. Τέτοιους μετασχηματισμούς θα δούμε στη συνέχεια.
- η κανονικοποίηση των δεδομένων μπορεί να επηρεάσει θετικά ή αρνητικά την ποιότητα ενός μέτρου ομοιότητας.

Η κανονικοποίηση είναι απαραίτητη στο πρόβλημα που αναλύουμε, διότι ο μετασχηματισμός στα δεδομένα που παρατηρείται συχνά αλλά δεν είναι πληροφοριακός για τον υπολογισμό της ομοιότητας δύο κειμένων, είναι αυτός της διαφοράς μεγέθους δύο κειμένων ίδιου περιεχομένου. Θέλουμε δηλαδή $sim(d_1, \lambda \cdot d_1) = 1$, με λ ένα σταθερό όρο.

Η Ευκλείδεια απόσταση ήταν η συνάρτηση που χρησιμοποιήθηκε πρώτη στην ομαδοποίηση κειμένων, ενώ συναντάται ακόμα σε σύγχρονες εργασίες [31]. Τελευταία, η συνημιτονοειδής ομοιότητα (*cosine similarity*) έχει κερδίσει έδαφος και αποτελεί σταθερή επιλογή πολλών συστημάτων, χωρίς να εκλείπουν και οι προστάσεις για νέα μέτρα [54][8] και νέες φιλοσοφίες υπολογισμού ομοιότητας.

5.2. Η Οικογένεια Αποστάσεων *Minkowski*

Οι αποστάσεις αυτές είναι ευρύτατα χρησιμοποιούμενες σε διάφορα προβλήματα όπου τα αντικείμενα είναι διανύσματα του $\mathbb{R}^k$. Για κείμενα d_1, d_2 και χρησιμοποιώντας μία παράμετρο p ορίζεται ένα άπειρο πλήθος αποστάσεων στο [0, ∞):

$$L_p(d_1, d_2) = \left(\sum_{i=1}^{|k|} \left| d_{1i} - d_{2i} \right|^p \right)^{1/p} \in [0, \inf).$$

Οι συνηθέστερες επιλογές του είναι: α) απόσταση *Hamming* με $p = 1$, β) *Ευκλείδεια* απόσταση με $p = 2$, γ) απόσταση *Tschebyshev* με $p = \infty$.

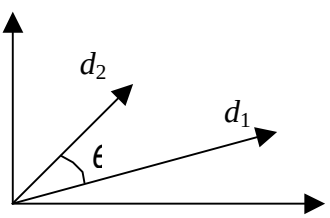
Υπάρχουν πολλοί τρόποι να μετατρέψει κανείς μία συνάρτηση απόστασης σε συνάρτηση ομοιότητας. Αρκεί μέσω μίας έκφρασης να απεικονίσουμε μία τιμή: [0, ∞) → [0, 1], το οποίο γίνεται χρησιμοποιώντας μία μονοτονικά φθίνουσα συνάρτηση για την απεικόνιση. Μια γνωστή επιλογή είναι η $sim(d_1,d_2) = 1 / (1 + L_2(d_1, d_2))$, ενώ μία πιο κατάλληλη για κείμενα [12] είναι η $sim^{(E)}(d_1, d_2) = exp(-L_2(d_1, d_2)^2)$.

5.3. Συνημιτονοειδής Ομοιότητα / Συσχέτιση

Ίσως το πιο δημοφιλές μέτρο στο χώρο της ομαδοποίησης κειμένων είναι αυτό της συνημιτονοειδούς ομοιότητας ή συσχέτισης (*cosine similarity*, *correlation*). Αυτή υπολογίζεται από την:

$$sim^{(\cos)}(d_1, d_2) = \cos\left(\theta(d_1, d_2)\right) = \frac{d_1^T d_2}{\|d_1\|_2 \cdot \|d_2\|_2} = \left(\frac{d_1^T}{\|d_1\|_2} \right) \cdot \left(\frac{d_2}{\|d_2\|_2} \right).$$

Όπως παρατηρεί κανείς υπολογίζεται το εσωτερικό γινόμενο των κανονικοποιημένων κειμένων-διανυσμάτων, γεγονός που την καθιστά ανεξάρτητη από το μήκος των κειμένων. Με κανονικοποίηση όλων των διανυσμάτων στην υπερσφαίρα με ακτίνα ένα, αποφεύγονται ακριβές αριθμητικές πράξεις υπολογισμού των νορμών.



Σχήμα 5.1. Η γωνία που σχηματίζεται μεταξύ δύο διανυσμάτων του χώρου $\mathbb{R}^2$.

Η τιμή της συνάρτησης εξαρτάται μόνο από την κλίση ανάμεσα στα διανύσματα d_1, d_2. Αν η κατανομή του λεξιλογίου είναι ανάλογη σε δύο κείμενα, τότε θα έχουν ίδια διεύθυνση και η ομοιότητά τους (εσωτερικό γινόμενο) θα είναι ίση με τη μονάδα. Η ιδιότητα αυτή καθιστά τη συνάρτηση κατάλληλη για τα κείμενα που μπορεί να έχουν διαφορετικά μεγέθη λαμβάνοντας υπόψη την αναλογία περιεχομένου.

5.4. Extended Jaccard ομοιότητα

Πρόκειται για την επέκταση του *Jaccard coefficient* (ή *Tanimoto coefficient*) για διανύσματα με δυαδικά χαρακτηριστικά:

$$J = \frac{f_{11}}{f_{11} + f_{01} + f_{10}},$$

όπου f_{kl} ο αριθμός των φορών που συμπίπτουν οι τιμές k, l σε αντίστοιχα *bits* στα διανύσματα. Η επέκταση για διακριτά, μη αρνητικά χαρακτηριστικά δίνεται από την:

$$sim^{(J)}(d_1, d_2) = \frac{d_1^T d_2}{\|d_1\|_2^2 + \|d_2\|_2^2 - d_1^T d_2}.$$

Η συνάρτηση αυτή, όπως θα δούμε, διαθέτει χαρακτηριστικά ανάμεσα σε αυτά της Ευκλείδειας απόστασης και της συνημιτονοειδούς ομοιότητας.

Ένα μέτρο που προκύπτει από την πρόσθεση του εσωτερικού γινομένου $d_1^{\mathrm{T}} d_2$ στον αριθμητή και τον παρονομαστή του $sim^{(J)}$ είναι ο *Dice coefficient*:

$$sim^{(D)}(d_1, d_2) = \frac{2 \cdot d_1^T d_2}{\|d_1\|_2^2 + \|d_2\|_2^2}.$$

Η συμπεριφορά του είναι γενικά πολύ κοντά σε αυτή του *Jaccard* μέτρου. Παρόμοιο μέτρο είναι και ο *Simple Matching Coefficient* ο οποίος λαμβάνει υπόψη και τον παράγοντα f_{00}:

$$SM = \frac{f_{11}}{f_{11} + f_{01} + f_{10} + f_{00}}.$$

5.5. Ομοιότητα Κοινών Κοντινότερων Γειτόνων

Είναι από τις πιο σύγχρονες προσεγγίσεις, η οποία δεν προτείνει μία μαθηματική έκφραση για τον υπολογισμό της ομοιότητας αλλά μία διαφορετική φιλοσοφία στο ζήτημα της ομοιότητας δύο αντικειμένων. Η υπόθεση στην οποία στηρίζεται είναι πως:

«*δύο αντικείμενα έχουν αρκετή ομοιότητα όταν έχουν μεγάλο αριθμό κοινών κοντινότερων γειτόνων, ακόμα και αν δεν προκύπτει κάτι τέτοιο από τον άμεσο υπολογισμό της ομοιότητας των δύο αντικειμένων μέσω, ενός μέτρου ομοιότητας*».

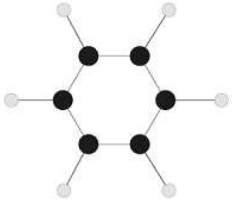
Η προσέγγιση των κοινών κοντινότερων γειτόνων (*shared nearest neighbors similarity*) έχει την ίδια βάση με αυτή που παρουσιάσαμε στο Κεφάλαιο 2 για το φιλτράρισμα με την *K-NN* προσέγγιση. Προσπαθεί δηλαδή, να εκτιμήσει την τοπική πληροφορία-δομή που παρουσιάζουν τα δεδομένα.

Ο υπολογισμός είναι απλός: αρχικά υπολογίζουμε τους *K-NN* γείτονες κάθε αντικειμένου, και υποθέτουμε μηδενική ομοιότητα για αντικείμενα που δεν αλληλοπεριέχονται στις κοντινότερες γειτονιές τους. Διαφορετικά, για να εκτιμήσουμε την ομοιότητα δύο αντικειμένων *x*, *y* υπολογίζουμε κάποια ποσότητα που εμπλέκει τον αριθμό των κοινών αντικειμένων που βρίσκονται στις γειτονιές των *x*, *y* [68].

5.6. Υπολογισμός Ομοιότητας σε Σύνθετα Αντικείμενα-Γραφήματα

5.6.1. Γενικές Προσεγγίσεις

Επειδή μας απασχόλησαν τα γραφήματα καλό είναι να αναφέρουμε πως στη γενική περίπτωση όπου δεν έχουμε μοναδικές ετικέτες κόμβων (π.χ. μόρια ανόργανων ενώσεων όπου πολλοί άνθρακες μπορεί να συμμετέχουν χωρίς να μπορούμε να τους ξεχωρίσουμε ή μόρια DNA, RNA κ.α.) το να μετρηθεί η ομοιότητα μεταξύ δύο γραφημάτων δεν είναι καθόλου εύκολη υπόθεση.



Σχήμα 5.2. Χημικό μόριο βενζίνης C_6H_6.

Ουσιαστικά το πρόβλημα του υπολογισμού της ομοιότητας γίνεται *NP*-πλήρες λόγω της συνδυαστικής πολυπλοκότητας, και αποκτά χαρακτηριστικά παρόμοια με αυτά της εύρεσης ισομορφισμού σε γραφήματα. Οι λύσεις για τον υπολογισμό του

μέγιστου κοινού υπογραφήματος δύο γραφημάτων βασίζονται στην εύρεση της μέγιστης κλίκας (*maximum clique detection*) ή την οπισθοδρόμηση (*backtracking*). Οι αλγόριθμοι είναι απλοί στη λογική, είναι όμως εξαιρετικά ακριβοί υπολογιστικά. Για παράδειγμα, η πολυπλοκότητα χρόνου στη χείριστη περίπτωση για δύο γραφήματα με n και m κόμβους αντίστοιχα είναι $O((n \cdot m)^n)$.

Για την περίπτωση γραφημάτων χωρίς ή με μη-μοναδικές ετικέτες κόμβων έχουν αναπτυχθεί τεχνικές όπως:

- εξόρυξη υπο-δομών (*substructure mining*) στην οποία δημιουργείται ένα σύνολο σύνθετων χαρακτηριστικών που θεωρούνται πληροφοριακά για το σύνολο δεδομένων, και ορίζουν το χώρο των χαρακτηριστικών του προβλήματος. Ύστερα, δημιουργείται ένα νέο σύνολο δεδομένων από διανύσματα, καθένα από τα οποία έχει τις συχνότητες εμφάνισης των χαρακτηριστικών του χώρου σε ένα αντικείμενο. Η εκπαίδευση γίνεται στο νέο σύνολο αυτό [66].
- *edit-distance-based* ομοιότητα στην οποία αφού οριστούν μία σειρά από βασικές πράξεις τροποποίησης των δεδομένων (π.χ. στα γραφήματα τέτοιες πράξεις θα μπορούσαν να είναι η εισαγωγή / διαγραφή ακμής κ.α.), για να υπολογιστεί η ομοιότητα μεταξύ δύο σύνθετων αντικειμένων προσεγγίζεται το πόσες έγκυρες πράξεις πρέπει να γίνουν στο ζεύγος αντικειμένων ώστε να διαφέρουν πλήρως [64][65].

Η πρώτη από τις τεχνικές μοιάζει με την προσέγγιση που προτείνουμε περί γενικευμένου γραφήματος αναπαράστασης. Οι διαφορές είναι ότι πρώτον στα κείμενα έχουμε μοναδικές ετικέτες και δεύτερον ότι θέσαμε ως προϋπόθεση μία σχέση να μπορεί να αναγνωριστεί μηχανικά μέσω μίας διαδικασίας εξόρυξης δεδομένων, συνεπώς ο χώρος των χαρακτηριστικός προκύπτει απευθείας από τα χαρακτηριστικά που ο χρήστης έχει ορίσει ως χρήσιμα.

Αυτές κάποιες προσεγγίσεις γα το πρόβλημα, η υλοποίησή των οποίων μπορεί να αφορά πολύ πιο εξειδικευμένα εργαλεία όπως οι πυρήνες (*kernels*).

5.6.2. Συνελικτικοί Πυρήνες για Προβλήματα Ανάλυσης Φυσικής Γλώσσας

Στην παράγραφο αυτή θα παρουσιάσουμε πολύ συνοπτικά πώς αντιμετωπίζεται το πρόβλημα χειρισμού σύνθετων αντικειμένων με προσεγγίσεις που έχουν εφαρμοστεί κατά κόρων και στο πεδίο *NLP*.

Οι συνελικτικοί πυρήνες (*convolution kernels*) [60] περιγράφουν τον τρόπο με τον οποίο είναι δυνατόν να δημιουργηθούν πυρήνες (*kernels*) για διακριτές δομές αντικειμένων, όπως τα δέντρα, τα αλφαριθμητικά και τα γραφήματα. Οι πυρήνες μπορούν να θεωρούνται συναρτήσεις ομοιότητας μεταξύ ζευγών αντικειμένων οι οποίες αν και χειρίζονται απευθείας τα πραγματικά αντικείμενα, δίνουν ισοδύναμα αποτελέσματα με το να προβάλλαμε τα αντικείμενα σε έναν άλλο χώρο και ύστερα να υπολογίζαμε την ομοιότητα τους μέσω της πράξης του εσωτερικού γινομένου.

Όπως είπαμε, λειτουργούν ως συντόμευση αποφεύγοντας την ίδια την προβολή των αντικειμένων. Διατηρούν την αρχική αναπαράσταση βάσει της οποίας υπολογίζουν τις συναρτήσεις πυρήνα (*kernel functions*) ως εσωτερικά γινόμενα μεταξύ ζευγών αντικειμένων. Οι συναρτήσεις πυρήνα με τη σειρά τους βασίζονται σε υποπυρήνες (*sub-kernels*) που χειρίζονται τα υπομέρη των σύνθετων αντικειμένων.

Αποφεύγοντας την μακροσκελή παρουσίαση των μεθόδων αυτών, θεωρείται πιο χρήσιμη η αναφορά ενός παραδείγματος. Ο πυρήνας για τη σύγκριση δύο σύνθετων δενδρικών αντικειμένων *x*, *y* (*tree kernel*) μπορεί να εκφραστεί ως [61]:

$$\left(x \in S_o\right) \xrightarrow{\phi} \left(\phi(x) \in S_f\right),$$
$$\left(y \in S_o\right) \xrightarrow{\phi} \left(\phi(y) \in S_f\right),$$

$$K(x, y) = \left\langle \phi(x) \cdot \phi(y) \right\rangle = \sum_{i=1}^{M} K_i(x, y) = \sum_{i=1}^{M} \phi_i(y) \cdot \phi_i(y),$$

όπου K_i οι υποπυρήνες, M ο συνολικός αριθμός υποδένδρων (γενικά υποστοιχείων) από τα οποία αποτελούνται τα αντικείμενα. Ο πυρήνας τελικά υπολογίζει το εσωτερικό γινόμενο των συχνοτήτων εμφάνισης $\phi_i(\cdot)$ των υποδένδρων στα δένδρα, αθροίζοντας τους υποπυρήνες K_i. Ο ίδιος πυρήνας μπορεί να χρησιμοποιηθεί και για αλφαριθμητικά. Για κείμενα μία σχετική εργασία είναι η [62].

5.6.3. Ομοιότητα Γραφημάτων Κειμένων

5.6.3.1. Γραφοθεωρητική Ομοιότητα Μεγίστου Κοινού Υπογραφήματος

Όπως αναφέραμε οι λέξεις αποτελούν τις μοναδικές ετικέτες του γραφήματος. Τη γενική διαδικασία εύρεσης ομοιότητας σε γραφήματα την ονομάζουμε ταίριασμα γράφου (*graph matching*) και λόγω αυτού του χαρακτηριστικού είναι δυνατόν να γίνει σε O(|T| + |E|) χρόνο όπως περιγράψαμε στο Κεφάλαιο 4.

Μια συνάρτηση ομοιότητας η οποία είναι σύμφωνη με την προσέγγιση του [63] και υπολογίζεται απευθείας πάνω στο γράφημα ενός κειμένου προτείνεται στο [51], με τη διαφορά ότι διαιρεί την τομή των γραφημάτων με το μέγιστο από τα αντικείμενα. Αυτό το μέτρο έχει χρησιμοποιηθεί στο πρόβλημα της ομαδοποίησης [49][50] και της κατηγοριοποίησης κειμένων *K-NN* [67]. Ως μετρήσιμο μέγεθος ομοιότητας ορίζεται το ποσοστό τομής των γραφημάτων, το οποίο υπολογίζεται μέσω του μέγιστου κοινού υπογραφήματος G_{mcs} (*maximal common subgraph*) των γραφημάτων G_1, G_2 (ο συμβολισμός εδώ είναι πιο κατανοητός όταν αναφέρεται στα γραφήματα και όχι στα διανύσματα). Ως συνάρτηση ομοιότητας έχουμε:

$$sim^{(G)}(G_1, G_2) = \frac{\left|mcs\left(G_1, G_2\right)\right|}{\max(|G_1|, |G_2|)},$$

όπου $mcs(\cdot,\cdot)$ η συνάρτηση εύρεσης του μέγιστου κοινού υπογραφήματος που παίρνει μέγιστη τιμή το μέγεθος του μικρότερου από τα δύο γραφήματα και $|G|$ το μέγεθος ενός γραφήματος.

Θεώρημα 5.1: Για γραφήματα G_1, G_2, G_3, και αν θεωρήσουμε $dist^{(G)} = 1 - sim^{(G)}$ ισχύουν οι παρακάτω ιδιότητες:

1. $0 \leq dist^{(G)}(G_1, G_2) \leq 1$,
2. $dist^{(G)}(G_1, G_2) = 0 \Leftrightarrow G_1 = G_2$, (είναι ισομορφικά γραφήματα),
3. $dist^{(G)}(G_1, G_2) = dist^{(G)}(G_2, G_1)$,
4. $dist^{(G)}(G_1, G_2) + dist^{(G)}(G_2, G_3) \geq dist^{(G)}(G_1, G_3)$.

Η απόδειξη του θεωρήματος, αλλά και μία εκτεταμένη συζήτηση πάνω στη συγκεκριμένη μετρική, μπορεί να αναζητηθεί στην εργασία [51].

Παρατηρώντας τη συνάρτηση αυτή αντιλαμβανόμαστε ένα σημαντικό μειονέκτημα, ότι ουσιαστικά δεν ενσωματώνει καμία πληροφορία βαρών στα στοιχεία του γραφήματος (*boolean term weighting*). Στις εργασίες [49][50][67] εμφανίζεται να δουλεύει καλά, αλλά στην πειραματική διαδικασία οι συγγραφείς χρησιμοποιούσαν αντιπροσωπευτικά γραφήματα 10 έως 80 κόμβων περίπου εφαρμόζοντας κατωφλίωση μεγέθους μοντέλων. Επίσης, παρατηρήθηκε πως μετά από την αύξηση των γραφημάτων πέρα από έναν αριθμό κόμβων (ή αντίστοιχα αύξηση των κοντινότερων γειτόνων), η απόδοση έπεφτε αρκετά.

Θεωρούμε πως ο λόγος που συνέβαινε κάτι τέτοιο είναι η έλλειψη βαρών στα γραφήματα. Όπως εξηγήσαμε στο Κεφάλαιο 2, πολλή από την πληροφορία των κειμένων είναι ουσιαστικά θόρυβος ο οποίος μπορεί να βρίσκεται σε χαρακτηριστικά είτε χαμηλού είτε υψηλού συχνοτικού περιεχομένου. Όσο επιλέγουμε ένα μικρό γράφημα αντιπρόσωπο από τις υψηλές συχνότητες του κειμένου, το να μην έχουμε βάρη μπορεί να παίζει και θετικό ρόλο, διότι μειώνει την επιρροή τυχόν συνηθισμένων όρων με μεγάλες συχνότητες. Σε ένα τέτοιο μικρό αντιπροσωπευτικό γράφημα οι περισσότεροι όροι είναι πράγματι «αντιπροσωπευτικοί», δηλαδή για την αναλογία:

$$ratio^{(b)} = \frac{\#representative\ \ features}{\#noisy\ \ features}$$

του αριθμού των αντιπροσωπευτικών όρων του κειμένου προς αυτή των όρων θορύβου ισχύει $ratio^{(b)} > 1$.

Όταν όμως αυξάνεται το μέγεθος των γραφημάτων, η συνάρτηση ομοιότητας δε διαθέτει πλέον κριτήρια και εξισώνει το ταίριασμα ενός ζεύγους όρων πολύ μικρής συχνότητας με ένα άλλο μεγάλης συχνότητας. Έτσι, θα λέγαμε πως ο παραπάνω δείκτης πέφτει πολύ κάτω της μονάδας, $ratio^{(b)} < 1$. Ο ισχυρισμός αυτός ως φαινόμενο τεκμηριώνεται και βάσει της εμπειρικής κατανομής *Zipf*.

5.6.3.2. Γραφοθεωρητική Ομοιότητα Βάσει του Περιεχομένου του Μεγίστου Κοινού Υπογραφήματος

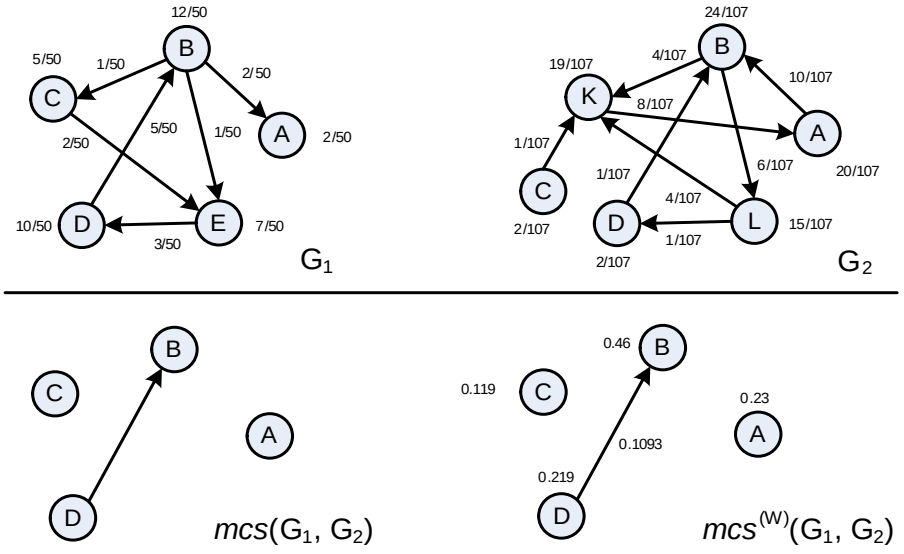
Η επέκταση της γραφοθεωρητικής συνάρτησης ομοιότητας παρουσιάστηκε στο [54] ως ομοιότητα μέγιστου κοινού περιεχομένου (*maximum common content similarity*), με μορφή που συμφωνεί απόλυτα με την προσέγγιση του [63]. Αναθέτοντας βάρη στο γράφημα, σύμφωνα με κάποιο σχήμα βαρών, η συνάρτηση καταφέρνει να έχει

καλύτερες επιδόσεις. Συγκεκριμένα αυτή υπολογίζει ως ομοιότητα το περιεχόμενο (δηλ. το άθροισμα βαρών) του μέγιστου κοινού υπογραφήματος προς το αντίστοιχο συνολικό περιεχόμενο της ένωσής τους. Δηλαδή έχουμε:

$$sim^{(GW)}(G_1, G_2) = \frac{\left\| mcs^W\left(G_1, G_2\right)\right\|_1}{\|G_1\|_1 + \|G_2\|_1} = \frac{1}{2} \cdot \left\| mcs^W\left(G_1, G_2\right)\right\|_1,$$

όπου $mcs^W(\cdot,\cdot)$ η συνάρτηση που βρίσκει το μέγιστο κοινό υπογράφημα και αθροίζει τα βάρη των στοιχείων γραφήματος και από τα δύο κείμενα. Αν τέλος, κανονικοποιήσουμε τα γραφήματα ως προς νόρμα L_1 τότε προκύπτει η τελευταία μορφή με $0 \leq \|mcs^W(\cdot, \cdot)\|_1 \leq 2$.

Στην έκφραση αυτή εφαρμόζουμε απευθείας και τη θεώρηση περί διανυσματοποίησης των γραφημάτων. Έτσι, το $mcs^W(G_1, G_2)$ μπορεί να θεωρείται ένα διάνυσμα στο χώρο της ένωσης των χαρακτηριστικών των γραφημάτων (ή στο χώρο της ένωσης όλων των χαρακτηριστικών της συλλογής) το οποίο προσθέτει τα βάρη των κοινών μη-μηδενικών διαστάσεων τους. Έτσι, υπολογίζεται εύκολα η νόρμα-1.



Σχήμα 5.3. Μέγιστο κοινό υπογράφημα κατευθυνόμενων γραφημάτων G_1, G_2, κάτω αριστερά: αγνοώντας τα βάρη, κάτω δεξιά: με κανονικοποιημένα βάρη.

Στο Σχήμα 5.3 παρουσιάζεται ένα παράδειγμα υπολογισμού του μέγιστου κοινού υπογραφήματος για τις δύο εκδοχές της γραφοθεωρητικής συνάρτησης: με βάρη και χωρίς. Είναι εμφανές πως τα περιεκτικά σε βάρος στοιχεία θα παίζουν μεγαλύτερο

ρόλο κατά τον υπολογισμό της πρώτης ομοιότητας. Η σχέση περιεχομένου για ένα γράφημα δίνεται πλέον από την:

$$ratio^{(W)} = \frac{content\left(representative\ \ features\right)}{content\left(noisy\ \ features\right)}.$$

Αν ως περιεχόμενο γραφήματος θεωρήσουμε το άθροισμα των βαρών των στοιχείων του, βλέπουμε πως δυσκολότερα θα συμβεί $ratio^{(W)} < 1$, ακόμα και όταν αυξήσουμε αρκετά το μέγεθος του αντιπροσωπευτικού γραφήματος.

Έχοντας τα δύο υπό σύγκριση γραφήματα G_1 και G_2, ο υπολογισμός του μεγέθους $|mcs(G_1, G_2)|$ μπορεί να γίνει με κόστος $O(min\{|G_1|, |G_1|\})$ χρησιμοποιώντας τον αλγόριθμο διάσχισης του Κεφαλαίου 4. Ο υπολογισμός υλοποιείται μετρώντας τα κοινά στοιχεία (αντίστοιχα αθροίζοντας τα βάρη τους για το $\|mcs(G_1, G_2)\|_1$) στα σημεία του ψευδοκώδικα «επεξεργασίας όρου» και «επεξεργασίας ακμής».

5.7. Ανάλυση των Κυριότερων Μέτρων Ομοιότητας για Κείμενα

Στην παράγραφο αυτή θα αναλύσουμε τις κυριότερες συναρτήσεων ομοιότητας, όπως η Ευκλείδεια, η συνημιτονοειδής, η *extended Jaccard* και η γραφοθεωρητική με βάρη. Η τελευταία δεν είναι από τα διαδεδομένα μέτρα, αλλά στη βάση του υπολογισμού της, που είναι το ποσοστό της τομής περιεχομένου δύο κειμένων, στηρίζονται πολλές συναρτήσεις ομοιότητας.

Αρχικά θα βασιστούμε στο ακόλουθο παράδειγμα (Σχήματα 5.4 και 5.5) [12]: θεωρούμε δυο σημεία του 2Δ χώρου, $x = (3, 1)^T$ και $y = (1, 2)^T$ τα οποία σημειώνονται με ‘x’ στο Σχήμα 5.4. Αυτό που εμφανίζεται στο σχήμα είναι οι επιφάνειες ίσης ομοιότητας (*iso-similarity surfaces*) όπου αναπαριστώνται ως συνεχείς γραμμές για τιμές ομοιότητας 0.25, 0.5, 0.75. Επίσης, με διακεκομμένη γραμμή εμφανίζονται τα σημεία του χώρου που ισαπέχουν από τα *x*, *y*. Κατά μήκος μίας συνεχούς γραμμής η ομοιότητα παραμένει σταθερή και βάσει των μεταβολών ομοιότητας στο χώρο παρατηρούμε πως:

- ίση Ευκλείδεια ομοιότητα μεταξύ ενός αντικειμένου *z* και σημείων του χώρου λαμβάνουμε σε ομόκεντρες υπερσφαίρες με κέντρο το αντικείμενο *z*. Η ομοιότητα αυτή αγνοεί τους μετασχηματισμούς μετατόπισης (*translations*), για παράδειγμα αν κρατήσουμε σταθερό το *x* και περιστρέψουμε το *y* γύρω από το *x* διατηρώντας την ακτίνα περιστροφής ίση με την απόστασή τους, τότε η

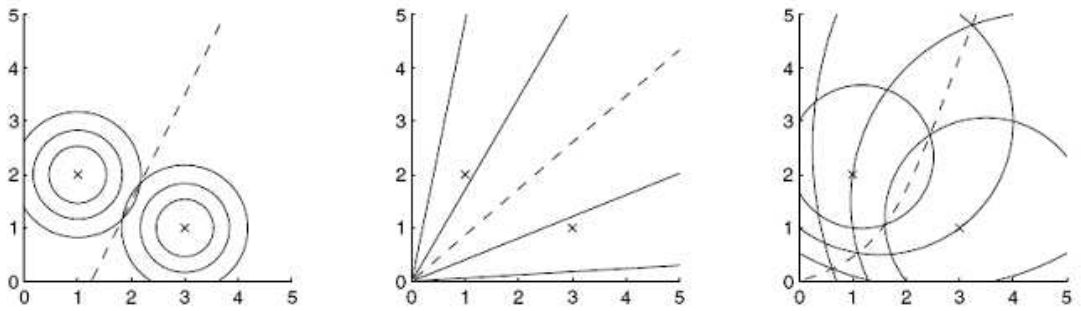
μεταξύ τους ομοιότητα θα παραμείνει σταθερή. Αντίθετα αν δεν κανονικοποιηθούν τα διανύσματα *x*, *y* η συνάρτηση είναι ευαίσθητη στις μεταβολές της κλίμακας τους, δηλαδή $sim^{(E)}(x, y) \neq sim^{(E)}(x, \lambda \cdot y)$, $\lambda \in \mathbb{R}$.

Το μόνο σημείο του χώρου με ομοιότητα ίση με τη μονάδα είναι το ίδιο το αντικείμενο *z* ενώ σε όλες τις υπόλοιπες πεπερασμένες συντεταγμένες του χώρου η ομοιότητα είναι διάφορη του 0. Αυτό συντελεί σε ένα αρκετά αρνητικό φαινόμενο, στο να δημιουργούνται πυκνοί πίνακες ομοιότητας.

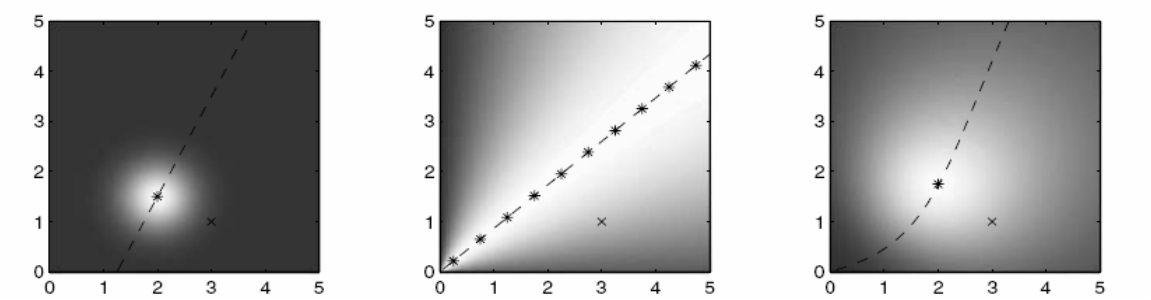
- ίση συνημιτονοειδή ομοιότητα λαμβάνουμε σε υπερκώνους που έχουν την κορυφή τους στην αρχή των αξόνων και άξονα που συμπίπτει με το διάνυσμα του z. Το μέτρο αυτό από τη φύση του αγνοεί την κλίμακα των διανυσμάτων και ενδιαφέρεται μόνο για τη διεύθυνσή τους. Η κανονικοποίηση γίνεται μόνο για λόγους μείωσης των πράξεων κατά τον υπολογισμό των ομοιοτήτων. Από την άλλη πλευρά είναι ευαίσθητο στους μετασχηματισμούς μετατόπισης.

  Τα σημεία του χώρου που παρουσιάζουν ομοιότητα ίση με τη μονάδα είναι όλα αυτά που έχουν την ίδια διεύθυνση με το αντικείμενο (στα κείμενα αυτό σημαίνει ίδια σχετική κατανομή λεξιλογίου) και ομοιότητα ίση με μηδέν τα σημεία που βρίσκονται στο κάθετο επίπεδο στο διάνυσμα του αντικειμένου.
- για την *extended Jaccard* παρατηρούμε πως συνδυάζει τα χαρακτηριστικά και των δύο προηγούμενων μέτρων ομοιότητας. Οι ίσες ομοιότητες λαμβάνονται πάνω σε υπερσφαίρες οι οποίες όμως δεν έχουν ως κέντρο το αντικείμενο αναφοράς, αλλά ούτε και είναι ομόκεντρες. Όσο η ομοιότητα τείνει στο μηδέν το μέτρο αυτό πλησιάζει σε συμπεριφορά τη συνημιτονοειδή ομοιότητας, ενώ όταν η ομοιότητα τείνει στη μονάδα πλησιέστερη συμπεριφορά είναι αυτή της Ευκλείδειας ομοιότητας.

Στο Σχήμα 5.5 παρουσιάζεται για το ίδιο παράδειγμα η μεταβολή της ομοιότητας στο χώρο. Στα φωτεινά σημεία του χώρου η ομοιότητα είναι μεγάλη. Με ‘*’ σημειώνεται το σημείο πάνω στην γραμμή ίσης ομοιότητας (γενικά επιφάνεια) για τα *x*, *y* το οποίο είναι πιο κοντά στο *x*.



Σχήμα 5.4. Επιφάνειες ίσης ομοιότητας για $x = (3, 1)^T$ και $y = (1, 2)^T$, (πηγή: [12]). Από αριστερά: Ευκλείδεια, συνημιτονοειδής και *extended Jaccard*.



Σχήμα 5.5. Μεταβολή ομοιότητας των $x = (3, 1)^T$ και $y = (1, 2)^T$ στον χώρο, (πηγή: [12]). Από αριστερά: Ευκλείδεια, συνημιτονοειδής, *extended Jaccard*.

Η συνημιτονοειδής ομοιότητα είναι σίγουρα μία καλή επιλογή για προβλήματα επεξεργασίας κειμένων, ενώ και η *extended Jaccard* με κανονικοποιημένα διανύσματα κειμένων παρουσιάζει επίσης κατάλληλη συμπεριφορά.

Από τις γραφοθεωρητικές προσεγγίσεις θα ασχοληθούμε με αυτή που εισάγει βάρη στο γράφημα αναπαράστασης. Αν κανονικοποιούσαμε τα γραφήματα των κειμένων και δημιουργούσαμε τα αντίστοιχα σχήματα ίσης ομοιότητας για τη συνάρτηση αυτή, θα παρατηρούσαμε πως δε θα παρουσίαζε ευαισθησία στην κλίμακα. Δύο κείμενα με ίδια σχετική κατανομή περιεχομένου στους ίδιους όρους και ακμές θα είχαν ομοιότητα ίση με τη μονάδα. Επίσης, μηδενικές ομοιότητες υπάρχουν όταν τα κείμενα δεν έχουν καθόλου τομή σε επίπεδο γραφήματος. Άρα δεν παρουσιάζουν το πρόβλημα των πυκνών πινάκων ομοιότητας (στοιχείο θορύβου) που απαιτούν την εφαρμογή κατωφλίων για το μηδενισμό των χαμηλών τιμών ομοιότητας.

Ακόμα, οι επιφάνειες ομοιότητας θα ήταν παρόμοιες με της Ευκλείδειας, πάνω σε υπερσφαίρες με τη διαφορά ότι όσο θα διέφεραν δύο κείμενα σε περιεχόμενο, θα

παρουσίαζε γραμμική αύξηση της συνάρτησης ομοιότητας τους. Η ακτίνα θα αυξανόταν επίσης γραμμικά ως προς την ανομοιότητα. Στην Ευκλείδεια η αντίστοιχη αύξηση είναι υπερβολική (βάσει της προσέγγισης: $sim^{(E)}(d_1, d_2) = exp(-L_2(d_1, d_2)^2)$.

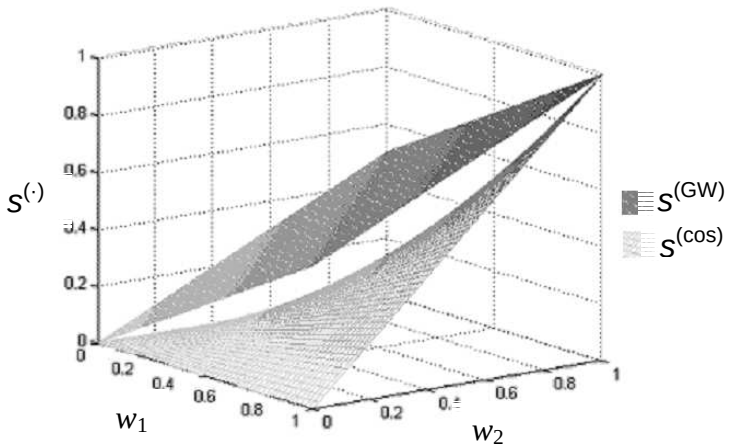
Για την σύγκριση της συμπεριφοράς της με αυτή της συνημιτονοειδούς ομοιότητας μπορούμε να δούμε κάτι επίσης σημαντικό. Θεωρώντας δύο κανονικοποιημένα κείμενα, τα δύο μέτρα γίνονται:

$$sim^{(GW)}(G_1, G_2) = \frac{1}{2} \cdot \left\| mcs^W \left(G_1, G_2\right) \right\|_1,$$

$$sim^{(\cos)}(G_1, G_2) = G_1^T G_2$$

Αυτές οι εκφράσεις δείχνουν πως και τα δύο μέτρα προσθέτουν «στοιχειώδεις» ομοιότητες που απορρέουν από κοινά ζεύγη στοιχείων γραφήματος. Αν για ένα ζεύγος κοινών χαρακτηριστικών συμβολίσουμε τα βάρη w_1, w_2 , το πρώτο μέτρο υπολογίζει τη στοιχειώδη ομοιότητα προσθετικά: $s^{(GW)}(w_1, w_2) = w_1 + w_2$, και η συνημιτονοειδής ομοιότητα πολλαπλασιαστικά: $s^{(\cos)}(w_1, w_2) = w_1 \cdot w_2$.

Περιορίζοντας τη μελέτη στις στοιχειώδεις ομοιότητες παρουσιάζουμε την παρακάτω 3Δ αναπαράσταση της συνάρτησης στοιχειώδους ομοιότητας που απορρέει από τα βάρη $w_1, w_2 \in [0, 1]$ και υπολογίζεται από τις δύο συναρτήσεις $s^{(GW)}$, $s^{(\cos)}$.



Σχήμα 5.6. Στοιχειώδεις ομοιότητες βάσει των συναρτήσεων $s^{(GW)}$ και $s^{(\cos)}$. Τα βάρη w_1, w_2 στους άξονες *x*, *y* στον z η τιμή των συναρτήσεων.

Η στοιχειώδης ομοιότητα της συνημιτονοειδούς συνάρτησης έχει μία «κεκλιμένη σαμαρωτή» μορφή η οποία δίνει μικρότερες τιμές για ζεύγη μικρών βαρών αλλά και ζεύγη όπου το ένα από τα δύο βάρη είναι μικρό. Από την άλλη πλευρά η

γραφοθεωρητική ομοιότητα δίνει ένα κεκλιμένο επίπεδο χωρίς να «καταπιέζει» τις τιμές της για τα ζεύγη μικρών βαρών. Αυτό πηγάζει από το ότι το άθροισμα βαρών είναι γραμμική συνάρτηση προς αυτά, αντίθετα με τον πολλαπλασιασμό τους.

Θα πρέπει όμως να αναλύσουμε τι ρόλο παίζουν στη συνολική ομοιότητα των κειμένων οι συμπεριφορές αυτές. Βάσει όσων έχουμε αναφέρει, στα κείμενα λίγοι όροι έχουν μεγάλες συχνότητες και οι πολλοί χαμηλές. Επίσης, οι πιο συχνές λέξεις ενός κειμένου είναι από στατιστικής άποψης πιο ασφαλή χαρακτηριστικά (οι συνηθισμένες λέξεις έχουν αφαιρεθεί). Άρα:

- η συνημιτονοειδής ομοιότητα υπολογίζοντας πολλαπλασιαστικά τις στοιχειώδεις ομοιότητες, περιορίζει συνολικά το ποσοστό ομοιότητας που πηγάζει από όρους χαμηλού συχνοτικού περιεχομένου στα κείμενα. Ακόμα η κανονικοποίηση των κειμένων με τη $\|\cdot\|_2$ ανακατανέμει εσωτερικά το περιεχόμενο του κειμένου δίνοντας έμφαση στους συχνότερους όρους. Έτσι γίνεται ακόμα πιο δύσκολο να παρατηρήσουμε $ratio^{(\cos)} < 1$.
- η γραφοθεωρητική αντίθετα, λειτουργώντας προσθετικά, είναι γραμμική στον υπολογισμό της στοιχειώδους ομοιότητας και αναμένει κανείς να επηρεάζεται περισσότερο από τους όρους χαμηλής συχνότητας. Η κανονικοποίηση με την $\|\cdot\|_1$ αφήνει αναλλοίωτη την εσωτερική κατανομή περιεχομένου στα κείμενα και συνεπώς επιτρέπει στα χαρακτηριστικά θορύβου να παίζουν αθροιζόμενα μεγαλύτερο ρόλο.

Παρατηρήσεις πάνω στη συμπεριφορά αυτή θα έχουμε τη δυνατότητα να παρουσιάσουμε και στην πειραματική μελέτη της εργασίας, στο Κεφάλαιο 7.

ΚΕΦΑΛΑΙΟ 6. ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ ΚΕΙΜΕΝΩΝ

Αναφέραμε στο εισαγωγικό κεφάλαιο, ότι οι τεχνικές ομαδοποίησης ανήκουν στην κατηγορία της μάθησης χωρίς επίβλεψη: μας δίνεται ένα σύνολο δεδομένων χωρίς να γνωρίζουμε την κατηγορία κάθε προτύπου και μας ζητείται να διαχωρίσουμε το σύνολο αυτό σε *M* ξεχωριστές ομάδες (βάσει της αυστηρής ομαδοποίησης). Ο αντικειμενικός στόχος είναι κείμενα που συνυπάρχουν στην ίδια ομάδα να παρουσιάζουν μεταξύ τους μεγαλύτερη συνάφεια περιεχομένου σε σχέση με την συνάφεια μεταξύ κειμένων διαφορετικών ομάδων.

Στην πράξη αυτό που προσπαθούν να κάνουν αρκετοί αλγόριθμοι είναι να μεγιστοποιήσουν την ομοιότητα μεταξύ των ομάδων (ισοδύναμα να ελαχιστοποιήσουν τη διασπορά των ομάδων) αναμένοντας κάτι τέτοιο να προκαλέσει και μείωση της ομοιότητας μεταξύ διαφορετικών ομάδων. Η ποσότητα που μεγιστοποιείται ή ελαχιστοποιείται καλείται αντικειμενική συνάρτηση. Το σφάλμα ορίζεται με διαφορετικό τρόπο για κάθε διαφορετική επιλογή συνάρτησης ομοιότητας, για παράδειγμα για την Ευκλείδεια απόσταση θεωρούμε το άθροισμα των τετραγωνικών

αποστάσεων των προτύπων από το κέντρο της ομάδας στην οποία ανήκουν (αν ορίζεται) αλλιώς από όλα τα αντικείμενα της ομάδας.

Οι δύο κύριοι αλγόριθμοι που θα μας απασχολήσουν είναι:

- ο ιεραρχικός αλγόριθμος συσσωρευτικής ομαδοποίησης (*Hierarchical Agglomerative*),
- η οικογένεια αλγορίθμων Κ-Μέσων (*K-Means*).

Διάφορες παραλλαγές των αλγορίθμων αυτών έχουν εφαρμοστεί σε πολλά προβλήματα, γι' αυτό και θεωρούνται κατά κάποιο τρόπο γενικού σκοπού. Σε ότι αφορά τα κείμενα, αποτελούν τους αλγόριθμους που επικρατούν στη βιβλιογραφία.

6.1. Μοντέλα Ομάδων

Το μοντέλο ομάδων καθορίζει το πώς «αντιλαμβάνεται» ένας αλγόριθμος ομαδοποίησης την ίδια την έννοια της ομάδας και κατ' επέκταση πώς την περιγράφει ως σύνθετη δομή. Η περιγραφή των ομάδων παίζει σημαντικό ρόλο στον χειρισμό τους, όταν δηλαδή κατά την αλγοριθμική διαδικασία απαιτείται κάποιου είδους συσχέτιση δεδομένων σε ανώτερο επίπεδο από αυτό των μεμονωμένων αντικειμένων. Τέτοιες συσχετίσεις είναι ο υπολογισμός αποστάσεων μεταξύ ομάδων ή ομάδων και αντικειμένων, η εκτίμηση της ποιότητας των ομάδων κ.α.

6.1.1. Παραμετρική Ομαδοποίηση

Η βασική διαφοροποίηση στα μοντέλα ομάδων έγκειται στο αν το μοντέλο είναι παραμετρικό ή όχι. Στις παραμετρικές τεχνικές ομαδοποίησης (*parametric* / *prototype-based clustering*) για κάθε ομάδα αντικειμένων ορίζεται ένας αντιπρόσωπος. Αναλόγως της συνολικής προσέγγισης με την οποία ομαδοποιούμε τα δεδομένα ο αντιπρόσωπος αυτός μπορεί να διαφέρει. Έτσι, αν για παράδειγμα εκπαιδεύουμε ένα μικτό μοντέλο κατανομών τότε ο αντιπρόσωπος θα είναι μία κατανομή με συγκεκριμένες παραμέτρους, ενώ αν εφαρμόζουμε μία πιο παραδοσιακή μέθοδο αυστηρής ομαδοποίησης ο αντιπρόσωπος μπορεί να είναι ένα υποσύνολο της πληροφορίας της ομάδας.

Στην πιο δημοφιλή προσέγγιση ο αντιπρόσωπος ονομάζεται *κέντρο* ή *κεντροειδές* της ομάδας (*cluster's centroid*). Το κέντρο αυτό μπορεί να είναι ένα από τα πρότυπα

της ομάδας ή όταν επιτρέπεται από το πρόβλημα να είναι ο αλγεβρικός μέσος (*cluster's mean*) που δεν αντιστοιχεί σε κάποιο πραγματικό πρότυπο.

Σε κάθε περίπτωση, ο ορισμός του αντιπροσώπου στις παραμετρικές τεχνικές είναι άμεσα συνυφασμένος με την ίδια τη διαδικασία εκπαίδευσης, γι' αυτό συνήθως αναφέρονται ως τεχνικές παραμετρικής ομαδοποίησης. Η ομοιότητα ανάμεσα σε ένα αντικείμενο και έναν αντιπρόσωπο εκφράζει τη σχέση του αντικειμένου με την αντίστοιχη ομάδα. Βάσει των σχέσεων αυτών τα αντικείμενα επιλέγουν να ενταχθούν στις ομάδες με τον κοντινότερο προς αυτά αντιπρόσωπο.

Ο στόχος των σχετικών τεχνικών είναι να οριστούν οι κατάλληλοι αντιπρόσωποι c_k, k: 1, ..., M, των ομάδων ώστε να περιγράφουν και τη λύση του ίδιου του προβλήματος ομαδοποίησης. Η διαδικασία εκπαίδευσης τελειώνει και το παραδοτέο σύστημα αποτελείται από τις M περιγραφές ομάδων.

6.1.2. Μη-παραμετρική Ομαδοποίηση

Στις *μη-παραμετρικές* τεχνικές (*non-parametric clustering*) κάθε ομάδα περιγράφεται από τα ίδια τα στοιχεία τα οποία ανήκουν σε αυτή. Στον υπολογισμό της απόστασης ενός προτύπου από μία ομάδα λαμβάνονται υπόψη όλα τα πρότυπα που ανήκουν ήδη σε αυτή ή κατά περίπτωση μέρος αυτών (π.χ. *single-link agglomerative*). Το παραδοτέο σε αυτή την περίπτωση είναι η αντιστοιχία πρότυπο-ομάδα, για τα δεδομένα εισόδου.

6.2. Ομαδοποίηση με τον Αλγόριθμο K-Μέσων

Ο αλγόριθμος K-Μέσων είναι ίσως ο πιο δημοφιλής αλγόριθμος ομαδοποίησης και ανήκει στην κατηγορία των διαμεριστικών αλγορίθμων αυστηρής ομαδοποίησης. Στην κατηγορία αυτή ο αριθμός των ομάδων που αναζητάμε είναι γνωστός εξ' αρχής και παραμένει σταθερός κατά τη διάρκεια της εκπαίδευσης.

Αρχικά ορίζεται μία διαμέριση K = M ομάδων την οποία στη συνέχεια προσπαθεί να βελτιώσει επαναληπτικά βάσει κάποιου κριτηρίου. Εδώ θα αναφέρουμε τις βασικότερες εκδοχές του για το πρόβλημα των κειμένων. Οι ομάδες περιγράφονται από αντιπροσώπους οι οποίοι είναι τα κεντροειδή τους, ενώ κριτήριο για την επιλογή της ομάδας στην οποία θα ενταχθεί ένα αντικείμενο είναι η μέγιστη εγγύτητα με το κέντρο της. Η είσοδος είναι ίδια για όλες τις εκδοχές του αλγορίθμου.

K	Ο αριθμός των ομάδων που αναζητούμε
$D=\{d_i\}, i=1,\ldots,N$	Ο αριθμός των κειμένων της συλλογής
DE	Κριτήριο τερματισμού, διαφορά λύσεων δύο διαφορετικών βημάτων
MAX_ITER	Κριτήριο τερματισμού , μέγιστος αριθμός επαναλήψεων

Σχήμα 6.1. Είσοδος των αλγορίθμων της οικογένειας K-Μέσων.

6.2.1. Βασικός Αλγόριθμος K-Μέσων

```
    Αλγόριθμος K-Μέσων ομαδικής ενημέρωσης κέντρων
   {
(1)    Αρχικοποίησε τα K κέντρα των ομάδων

(2)    Επανέλαβε,
(3)    {   Ανέθεσε κάθε κείμενο di, i=1,…,N στην ομάδα με το κοντινότερο κέντρο
(4)        Ενημέρωσε τα κέντρα των ομάδων
(5)    } μέχρι να ικανοποιηθεί κάποιο κριτήριο τερματισμού (DE ή MAX_ITER)
    }
```

Σχήμα 6.2. Ψευδοκώδικας αλγορίθμου K-Μέσων ομαδικής ενημέρωσης κέντρων.

Ο αλγόριθμος αυτός είναι γρήγορος, απλός στην κατανόηση και την υλοποίηση. Το χαρακτηριστικό αυτής της εκδοχής είναι πως τα κέντρα ενημερώνονται ύστερα από την ανάθεση όλων των δεδομένων στις κοντινότερες ομάδες (*batch centroid update*).

Το βασικό μειονέκτημα του είναι πως, αν και συγκλίνει μονότονα και γρήγορα, καταλήγει ντετερμινιστικά σε ένα τοπικό ελάχιστο, εξαρτώμενος αποκλειστικά από την αρχικοποίηση των κέντρων. Μάλιστα η αρχικοποίηση επηρεάζει σε πολύ μεγάλο βαθμό και, ανάλογα με τα δεδομένα, μπορεί να δώσει λύσεις που διαφέρουν αρκετά σε ποιότητα. Το γεγονός αυτό μας αναγκάζει να εφαρμόζουμε ειδικές διαδικασίες ώστε να μετριάσουμε την εξάρτηση από τον παράγοντα «αρχικοποίηση».

6.2.2. K-Μέσων Αυξητικής Ομαδοποίησης

Μια παραλλαγή του βασικού αλγορίθμου η οποία λειτουργεί αυξητικά (*online / incremental K-Means*) έχει προταθεί στη βιβλιογραφία για το πρόβλημα των κειμένων και αναφέρεται πως συμπεριφέρεται καλύτερα από την ομαδική ενημέρωση των κέντρων [10][70].

Αφού αρχικοποιήσουμε τα κεντροειδή και τις ομάδες, αφήνουμε ένα-ένα τα κείμενα να ενταχθούν στην κοντινότερη ομάδα ενημερώνοντας κανένα ή δύο κέντρα

(κανένα: αν παραμείνει στην ίδια ομάδα, δύο: αν αλλάξει ομάδα). Τα δεδομένα ομαδοποιούνται ξανά και ξανά ενημερώνοντας τα κέντρα (βήματα 2 έως 6). Το πέρασμα όλων των δεδομένων από τη διαδικασία εκπαίδευσης καλείται εποχή.

```
    Αλγόριθμος Κ-Μέσων αυξητικής ομαδοποίησης
   {
(1)    Αρχικοποίησε τα Κ κέντρα των ομάδων
(2)    Ανέθεσε κάθε κείμενο di, i=1,…,N στην ομάδα με το κοντινότερο κέντρο

(3)    Επανέλαβε,
(4)    {   Για κάθε κείμενο di, i=1,…,N,
(5)        {   Ανέθεσε το κείμενο di στην ομάδα με το κοντινότερο κέντρο
(6)            Ενημέρωσε το κέντρο της νικήτριας ομάδας
(7)        }
(8)    } μέχρι να μην συμβούν αλλαγές στα κέντρα
   }
```

Σχήμα 6.3. Ψευδοκώδικας αυξητικού αλγόριθμου Κ-Μέσων.

Στην κλασσική έκδοση ο αλγόριθμος προσπαθεί να βελτιώσει τη λύση του προηγούμενου βήματος, το οποίο βελτίωσε το προηγούμενο κ.ο.κ. και τελικά προκύπτει εξάρτηση από την αρχική διαμέριση των δεδομένων, η οποία είναι συχνά τυχαία. Στην αυξητική έκδοση η διαδικασία εξαρτάται από τη σειρά επεξεργασίας των δεδομένων από τον αλγόριθμο, γι' αυτό και μεταξύ εποχών μεριμνούμε για την τροποποίηση της σειράς αυτής. Ο αλγόριθμος συγκλίνει ντετερμινιστικά σε τοπικό ελάχιστο μόνο σε περίπτωση που σε διαφορετικές εκτελέσεις διατηρήσουμε ίδια τη σειρά επεξεργασίας των δεδομένων κάθε εποχής της εκπαίδευσης.

Η σύγκλιση είναι σαφώς βραδύτερη και επιτυγχάνεται κάνοντας μικρότερα βήματα προόδου κάθε φορά που συναντάται ένα νέο για την εποχή κείμενο, σε σχέση με τον κλασσικό αλγόριθμο ομαδικής ενημέρωσης. Αυτό επιτρέπει στα κεντροειδή να μαθαίνουν με μεγαλύτερη ασφάλεια και σταθερότητα από τα δεδομένα. Από την άλλη πλευρά ο αλγόριθμος χάνει την συνολική εικόνα με την οποία αντιμετωπίζει τα δεδομένα προσπαθώντας να διαχειριστεί την είσοδο ενός μόνο στοιχείου τη φορά.

Ένα σημαντικό πλεονέκτημα της τεχνικής αυτής είναι ότι μπορούμε να χρησιμοποιήσουμε διαφορετικές αντικειμενικές συναρτήσεις [68], πράγμα που δεν είναι εύκολο να κάνουμε στην παραδοσιακή μορφή του αλγορίθμου και να δείξουμε τη σύγκλιση. Αν η στοιχειώδης μετακίνηση βελτιώνει την αντικειμενική συνάρτηση που

έχουμε ορίσει τότε εξασφαλίζεται εύκολα ότι ο αλγόριθμος θα σταματήσει μόνο όταν δε μπορεί να βελτιώσει άλλο την συνάρτηση αυτή. Η σύγκλιση είναι προφανής.

6.2.3. Κ-Μέσων Διαιρετικής Ομαδοποίησης

Μία ακόμα παραλλαγή της οικογένειας αυτής είναι ο αλγόριθμος Κ-Μέσων διαιρετικής ομαδοποίησης. Κατ' εξαίρεση περιλαμβάνεται στους διαμεριστικούς αλγορίθμους λόγω τη σχέσης του με τον Κ-Μέσων. Ουσιαστικά είναι ένας διαιρετικός ιεραρχικός αλγόριθμος από πάνω προς τα κάτω (από το γενικό στο ειδικό) ο οποίος συνδυάζει σε κάθε βήμα του κάποια από τα προτερήματα του αλγορίθμου Κ-Μέσων. Έχει αναφερθεί πως μπορεί να αποτελέσει καλύτερη επιλογή για την ομαδοποίηση κειμένων από τον παραδοσιακή έκδοση του αλγορίθμου.

```
    Αλγόριθμος Κ-Μέσων διαιρετικής ομαδοποίησης
   {
(1)    Τοποθέτησε όλα τα κείμενα σε μία μοναδική ομάδα
(2)    Θέσε K = 1

(3)    Επανέλαβε,
(4)    {   Επέλεξε μία ομάδα από αυτές που αποτελούν την τρέχουσα λύση
(5)        Για έναν αριθμό δοκιμών, test=1,…,TESTS,
(6)           Διάσπασε την επιλεγμένη ομάδα χρησιμοποιώντας τον Κ-Μέσων με Κ=2
(7)
(8)        Επέλεξε από τις δοκιμές τη διάσπαση που έδωσε το μικρότερο σφάλμα
(9)        Εφάρμοσέ τη στη λύση k και θέσε Κ = Κ+1
(10)   } μέχρι να προκύψουν Κ = Μ ομάδες
    }
```

Σχήμα 6.4. Ψευδοκώδικας διαιρετικού αλγόριθμου Κ-Μέσων.

Επειδή ο αλγόριθμος αυτός επικεντρώνεται στη διάσπαση μίας ομάδας σε κάθε βήμα, παρουσιάζει το πρόβλημα ότι δε βρίσκει το τοπικό ακρότατο της αντικειμενικής συνάρτησης όταν τερματίζει. Επίσης, οι αποφάσεις που λαμβάνει είναι τελικές, δηλαδή ένα αντικείμενο που ακολούθησε ένα κλαδί του δενδρογράμματος λύσεων στον κόμβο *Λ* δε θα περάσει ποτέ στο άλλο υποδένδρο που δημιουργήθηκε στον *Λ*. Γι' αυτό χρησιμοποιείται πολύ συχνά ως αρχικοποίηση για μία τελική εφαρμογή του παραδοσιακού Κ-Μέσων, ώστε να λάβουμε ντετερμινιστικά το αντίστοιχο τοπικό ελάχιστο.

Για την επιλογή της ομάδας που θα διασπαστεί υπάρχουν αρκετά κριτήρια. Ένα είναι να εντοπιστεί η ομάδα με το μεγαλύτερο σφάλμα, η οποία υποδεικνύει ότι περιέχει αρκετά ανόμοια αντικείμενα και θα πρέπει να διαμεριστεί. Μία άλλη επιλογή είναι να επιλεγεί η μεγαλύτερη ομάδα, η οποία προσπαθεί να παράγει λύσεις με ομάδες ίδιας τάξης μεγέθους. Στη βιβλιογραφία αναφέρεται ότι και οι δύο λύσεις δίνουν παρόμοια αποτελέσματα.

Ο διαιρετικός αλγόριθμος έχει προταθεί και ως *Principal Direction Divisive Partitional Algorithm* [7] ή *PCA Partitioning Algorithm* [18], διαφέροντας από τον προαναφερόμενο διαιρετικό στο πως επιλέγεται η διαίρεση μίας ομάδας. Αντί να δοκιμάζονται διάφορες διασπάσεις μέσω του K-Μέσων, επιλέγεται η ομάδα με το μεγαλύτερο σφάλμα η οποία διασπάται σε δύο μέρη που ορίζονται από το υπερεπίπεδο που περνά από τον αριθμητικό μέσο της ομάδας και είναι κάθετο στην κατεύθυνση της μέγιστης μεταβλητότητας των διανυσμάτων της ομάδας (η πρώτη ιδιοτιμή του πίνακα συμμεταβλητότητας ενός δείγματος δεδομένων από την ομάδα). Κατά τα λοιπά ο αλγόριθμος είναι πανομοιότυπος με τον διαιρετικό που περιγράφηκε.

6.3. Ο Αλγόριθμος K-Μέσων ως Πρόβλημα

6.3.1. Ομοιότητες Αλγορίθμου K-Μέσων με Μεθόδους Βελτιστοποίησης

Οι αλγόριθμοι ομαδοποίησης μπορούν να αναλυθούν ως προβλήματα βελτιστοποίησης. Η αδυναμία να βρούμε την καλύτερη δυνατή λύση του προβλήματος, λόγω της *NP*-πληρότητας και του μεγάλου όγκου δεδομένων, μας εξωθεί σε λύσεις που σταδιακά βελτιώνουν την ποιότητα της ομαδοποίησης, βάσει πάντα κάποιων κριτηρίων.

Η εύρεση νέων κριτηρίων, ιδιαίτερα για τον αλγόριθμο K-Μέσων, είναι ένα πολύ μοντέρνο ερευνητικό πρόβλημα [71][14]. Και αναφέρουμε συγκεκριμένα για τον αλγόριθμο αυτό διότι έχει μία πολύ χαρακτηριστική δομή η οποία είναι απόλυτα συνυφασμένη με μία κατηγορία μεθόδων βελτιστοποίησης, αυτή της απότομης καθόδου (*gradient descent*). Στην προσέγγιση αυτή εφαρμόζεται ακριβώς η ίδια γενική διαδικασία:

α) αρχικά θεωρούμε μία λύση στο πρόβλημα βελτιστοποίησης η οποίο καθορίζεται από μία συνάρτηση την οποία προσπαθούμε να ελαχιστοποιήσουμε

β) υπολογίζεται η μεταβολή στη λύση η οποία βελτιστοποιεί την αντικειμενική συνάρτηση,

γ) ενημερώνεται η λύση,

δ) επαναλαμβάνουμε μέχρι να βρεθούμε σε τοπικό ακρότατο ή οι μεταβολή στη λύση να είναι μικρότερη ενός κατωφλίου *ε*.

Με άλλα λόγια βελτιώνουμε μία αρχική λύση σε διακριτά βήματα όπως και στον Κ-Μέσων, όπου κάθε φορά επιλέγουμε την καλύτερη από τις δυνατές μεταβολές στη λύση του προβλήματος.

Μία διαφορά είναι πως με τον *gradient descent* γενικά αναζητούμε ακρότατα μίας συνάρτησης στο συνεχή χώρο, οπότε και ο υπολογισμός της καλύτερης μεταβολής στη λύση εμπλέκει τον υπολογισμό της πρώτης παραγώγου της συνάρτησης στο τρέχον σημείο. Επίσης, ο *gradient descent* διαθέτει μία παράμετρο: το βήμα της μεταβολής πάνω στην καλύτερη κατεύθυνση.

Αντίθετα στον Κ-Μέσων η βελτιστοποίηση μπορεί να γίνει μέσω διακριτών μεταθέσεων των δεδομένων ανάμεσα στις ομάδες, κάθε μία από τις οποίες είναι στην «καλύτερη κατεύθυνση» για τη βελτίωση της λύσης. Κατά κάποιο τρόπο το βήμα του *gradient descent* αντιστοιχεί στο πώς θα ενημερώσουμε τα κεντροειδή, δηλαδή με ομαδική ενημέρωση ή για κάθε ένα δεδομένο χωριστά.

Αποδεικνύεται, κυρίως με μία σειρά εμπειρικών διαπιστώσεων, πως όντως οι μεταβολές της λύσης που κάνει ο αλγόριθμος Κ-Μέσων είναι στην καλύτερη κατεύθυνση για τη βελτίωση της λύσης. Αυτό εξασφαλίζει τη μονότονη σύγκλιση σε τοπικό ελάχιστο της αντικειμενικής συνάρτησης, αρκεί:

- σε κάθε βήμα η ανάθεση των κειμένων να γίνεται στην ομάδα με το κοντινότερο κεντροειδές,
- η ενημέρωση του κεντροειδούς κάθε ομάδας να είναι τέτοια ώστε να ελαχιστοποιείται το εσωτερικό σφάλμα (αν πρόκειται για τον αυξητικό τότε αυτό αφορά την ενημέρωση της νέας και της παλιάς ομάδας του κειμένου).

Αν παρατηρήσουμε στο πρώτο βήμα μειώνεται το σφάλμα με την «βέλτιστη» ανάθεση κάθε αντικειμένου, ενώ στο δεύτερο επίσης μειώνεται το σφάλμα με την «βέλτιστη» περιγραφή κάθε ομάδας χρήσει κατάλληλου αντιπροσώπου. Αυτή είναι και

η εμπειρική παρατήρηση που τεκμηριώνει τη σύγκλιση. Παρόλα αυτά ο αλγόριθμος δεν καταλήγει σε ολικό ελάχιστο διότι α) εξαρτόμαστε πάντα από την αρχικοποίηση, και β) ταυτόχρονα κάθε αντικείμενο προσπαθεί να βελτιώσει τη θέση του ανάμεσα στις ομάδες, με αποτέλεσμα να μην υπάρχει ένα «συνολικά βέλτιστο σχέδιο» για το πώς τελικά θα επιτευχθεί η καλύτερη δυνατή συνολική λύση (*global optimum*).

6.3.2. Υπολογισμός Κεντροειδών για Μονότονη Σύγκλιση

Ο υπολογισμός των κέντρων γίνεται με τέτοιο τρόπο ώστε να βελτιστοποιείται η αντικειμενική συνάρτηση. Επειδή αυτή αφορά όλες τις ανεξάρτητες ομάδες, η βελτιστοποίηση της συνιστώσας που αφορά μία μεμονωμένη ομάδα οδηγεί και στην ελαχιστοποίηση της αντικειμενικής συνάρτησης.

Η μονότονη σύγκλιση απαιτεί ανά πάσα στιγμή το κέντρο να αναπαριστά την ομάδα με τέτοιο τρόπο ώστε να ελαχιστοποιεί το εσωτερικό της σφάλμα ή αντίστοιχα να μεγιστοποιεί την εσωτερική ομοιότητα των αντικειμένων με αυτό.

6.3.2.1. Ευκλείδεια Απόσταση

Για την Ευκλείδεια απόσταση χρησιμοποιείται το άθροισμα των τετραγωνικών σφαλμάτων (τετραγωνικών διαφορών των αντικειμένων από το κέντρο της ομάδας):

$$SSE_i = \sum_{d \in C_i} \left(m_i - d \right)^2, \qquad SSE = \sum_{i=1}^{M} \sum_{d_j \in C_i} \left(m_i - d_j \right)^2,$$

$$c_i = m_i^* = \arg\max_{m_i} \left\{ \sum_{d \in C_i} sim^{(E)} \left(m_i, d \right) \right\} = \frac{1}{|C_i|} \sum_{d_j \in C_i} d_j .$$

Είτε παραγωγίζοντας με μερικές παραγώγους την πρώτη έκφραση, είτε κάθε συνιστώσα της SSE_i ως προς m_i και θέτοντας ίσο με μηδέν, αποδεικνύεται πως τα κέντρα που ελαχιστοποιούν αυτή τη συνάρτηση είναι οι αριθμητικοί μέσοι των διανυσμάτων όλης της ομάδας.

6.3.2.2. Συνημιτονοειδής Ομοιότητα

Γενικά, αυτό που προσπαθεί να κάνει η πλειοψηφία των αλγορίθμων ομαδοποίησης είναι να μεγιστοποιήσει την ανά ζεύγος ομοιότητα των αντικειμένων κάθε ομάδας και εμμέσως να ελαχιστοποιήσει και την ομοιότητα μεταξύ διαφορετικών ομάδων. Αν

έχουμε κανονικοποιημένα τα διανύσματα των κειμένων τότε η αντικειμενική συνάρτηση ορίζεται ως:

$$Objective_i = \sum_{d_j \in C_i} d_j^T \frac{m_i}{\|m_i\|_2}, \qquad m_i = \frac{1}{|C_i|} \sum_{d_j \in C_i} d_j ,$$

όπου με m_i συμβολίζεται ο μέσος όρος των διανυσμάτων της ομάδας και είναι το βέλτιστο κέντρο. Με άλλα λόγια όταν χρησιμοποιούμε τη συνημιτονοειδή συνάρτηση ως μέτρο ομοιότητας τότε προκύπτει πως αρκεί κατά την ομαδοποίηση να αυξάνεται η ομοιότητα των κειμένων από τον αριθμητικό μέσο της ομάδας τους. Να σημειώσουμε πως το m_i δεν εξασφαλίζεται ότι θα είναι κανονικοποιημένο αν τα επιμέρους διανύσματα που αθροίζονται είναι κανονικοποιημένα. Έτσι, μπορούμε να ορίσουμε το κανονικοποιημένο κεντροειδές της ομάδας C_i και να επαναδιατυπώσουμε την αντικειμενική συνάρτηση:

$$Objective_i = \sum_{d_j \in C_i} d_j^T c_i, \qquad c_i = \frac{m_i}{\|m_i\|_2}.$$

Για τη συνημιτονοειδή ομοιότητα αποδεικνύεται βάσει της ανισότητας Cauchy-Schwarz πως το κανονικοποιημένο κεντροειδές c_i είναι το εγγύτερο μοναδιαίο διάνυσμα σε όλα κανονικοποιημένα διανύσματα των κειμένων. Για κάθε κανονικοποιημένο διάνυσμα x στον χώρο του προβλήματος, και κάθε ομάδα C_i ισχύει:

$$\sum_{d_j \in C_i} d_j^T c_i \geq \sum_{d_j \in C_i} d_j^T x.$$

Τελικά, αθροίζοντας τη συνοχή όλων των ομάδων προκύπτει η Ολική Συνοχή (*Total Cohesion - TC*):

$$TC_i = \sum_{d_j \in C_i} sim^{(\cos)}\left(m_i, d_j\right), \qquad TC = \sum_{i=1}^{M} \sum_{d_j \in C_i} sim^{(\cos)}\left(m_i, d_j\right)$$

$$m_i^* = \arg\max_{m_i} \left\{ \sum_{d_j \in C_i} sim^{(\cos)}\left(m_i, d_j\right) \right\} = \frac{1}{|C_i|} \sum_{d_j \in C_i} d_j .$$

Η συνάρτηση αυτή παρουσιάζει κάποια ενδιαφέροντα χαρακτηριστικά, για μια ομάδα μπορούμε να γράψουμε:

$$TC_i = \sum_{d_j \in C_i} \frac{d_j^T c_i}{\|d_j\|_2 \cdot \|c_i\|_2} = \left(\sum_{d_j \in C_i} d_j^T \right) c_i = \left(|C_i| \cdot m_i^T \right) c_i =$$

$$= \left(|C_i| \cdot \left\| m_i^T \right\|_2 \cdot c_i^T\right) c_i = \left(|C_i| \cdot \left\| m_i^T \right\|_2\right) \|c_i\|_2 = \left(|C_i| \cdot \left\| m_i^T \right\|_2\right) = \left\| \sum_{d_j \in C_i} d_j^T \right\|_2 .$$

Το αποτέλεσμα αυτό σημαίνει πως για να υπολογίσουμε τη συνοχή μίας ομάδας, δηλαδή την αντικειμενική συνάρτηση για την ομάδα, αρκεί να υπολογίσουμε τη νόρμα του αθροίσματος όλων των κανονικοποιημένων κειμένων στην ομάδα.

Μια ενδιαφέρουσα παρατήρηση βασίζεται στα εξής:

1) $d_j^T m_i = d_j^T \left(\frac{1}{|C_i|} \cdot \sum_{d_k \in C_i} d_k \right) = \frac{1}{|C_i|} \cdot \sum_{d_k \in C_i} sim^{(cos)}(d_j, d_k)$,

2) $\|m_i\| = \left[\left(\frac{1}{|C_i|} \sum_{d_j \in C_i} d_j \right)^T \left(\frac{1}{|C_i|} \sum_{d_k \in C_i} d_k \right) \right]^{1/2} = \left[\frac{1}{|C_i|^2} \sum_{d_j \in C_i} \sum_{d_k \in C_i} d_j^T d_k \right]^{1/2} =$

$$= \left(\frac{1}{|C_i|^2} \cdot \sum_{d_k \in C_i} \sum_{d_l \in C_i} sim^{(cos)}(d_j, d_k) \right)^{1/2} .$$

Από την ομοιότητα ενός κειμένου με το κεντροειδές μίας ομάδας βλέπουμε ότι:

$$sim^{(cos)}\left(d_j, c_i\right) = d_j^T m_i \cdot \frac{1}{\|m_i\|},$$

- εξαρτάται από το εσωτερικό γινόμενο του αριθμητικού μέσου με το διάνυσμα του κειμένου, δηλαδή από τη μέση ομοιότητα του κειμένου με τα κείμενα της ομάδας,
- και από έναν πολλαπλασιαστικό παράγοντα 1 / $\|m_i\|$, ο οποίος επειδή όλα τα κείμενα είναι κανονικοποιημένα και ισχύει: $\|m_i\|_2 \leq 1$, είναι μεγαλύτερος της μονάδας και παίζει ένα ρόλο «παράγοντα ενίσχυσης».

Όπως φαίνεται από τη δεύτερη σχέση, η νόρμα-2 ενέχει τον υπολογισμό της μέσης ομοιότητας των κειμένων ανά ζεύγη, της ομάδας. Άρα όσο καλύτερη η ποιότητα της ομάδας, σύμφωνα με το κριτήριο αυτό, τόσο μικρότερος ο παράγοντας ενίσχυσης, ενώ αντίθετα μία ομάδα με μέτρια χαρακτηριστικά επιδιώκει την βελτίωσή της έχοντας μεγάλη τιμή ενίσχυσης.

Αυτά που παρουσιάστηκαν για την συνημιτονοειδή ομοιότητα δείχνουν και τα ιδιαίτερα χαρακτηριστικά που την καθιστούν από τα πλέον κατάλληλα μέτρα ομοιότητας για το πρόβλημα της ομαδοποίησης κειμένων.

6.3.3. Υπολογισμός Ενδιάμεσων Κειμένων για Κεντροειδή

Αφού γνωρίζουμε τα έγγραφα τα οποία περιλαμβάνει μία ομάδα, μπορούμε να θεωρήσουμε ως κέντρο της το ενδιάμεσο κείμενο, το οποίο έχει τη μικρότερη απόσταση από όλα τα κείμενα της ομάδας (*medoid*). Έτσι, ο υπολογισμός του κέντρου μπορεί να γίνει με τον ακόλουθο τρόπο:

$$m_i = \arg\max_{\forall d_j \in C_i} \left(\sum_{d_k \in C_i} sim(d_j, d_k) \right).$$

Ο ενδιάμεσος όπως είναι φυσικό δεν είναι η βέλτιστη επιλογή για τη μεγιστοποίηση της συνοχής μιας ομάδας. Αυτή η ιδιότητα είναι προνόμιο το οποίο έχουν μόνο συγκεκριμένες εκφράσεις κέντρων που εξαρτώνται από τη συνάρτηση ομοιότητας και την αντικειμενική συνάρτηση. Κατά τη διαδικασία ομαδοποίησης και έχοντας επιλέξει τους ενδιάμεσους για κεντροειδή των ομάδων, παρατηρούμε να αυξάνεται η συνοχή των ομάδων σε κάθε βήμα, και είτε η διαδικασία τερματίζει σε τοπικό ακρότατο, είτε παρατηρείται μείωση της συνοχής μετά από μία ενημέρωση κέντρων. Η μείωση αυτή έχει να κάνει με το εξής φαινόμενο:

- όλα τα κείμενα επιλέγουν την κοντινότερη ομάδα βάσει του ενδιαμέσου της αυξάνοντας την αντικειμενική συνάρτηση (αυξάνεται η συνοχή),
- όταν όμως επαναϋπολογίζεται ο ενδιάμεσος κάθε είναι δυνατό η συνοχή να μειωθεί.

Τις περιπτώσεις αυτές μπορούμε να τις αναγνωρίσουμε παρακολουθώντας σε κάθε βήμα την αντικειμενική συνάρτηση συνοχής της λύσης. Όταν διαπιστωθεί μείωση τότε διακόπτουμε τη διαδικασία μάθησης και αναφέρουμε την αμέσως προηγούμενη λύση η οποία ήταν και η καλύτερή που βρέθηκε.

Το πλεονέκτημα της επιλογής αυτής είναι πως ο ενδιάμεσος είναι ένα κείμενο κάποιας κατηγορίας και όχι κάποιος συγκερασμός των χαρακτηριστικών της ομάδας. Επίσης βοηθάει στην μετρίαση της επιρροής ακραίων κειμένων. Μπορεί μεν να περιγράφει ελλιπώς την ομάδα, όμως την αναγκάζει να πάρει πιο ξεκάθαρες αποφάσεις για την κατηγορία κειμένων στην οποία θα δώσει μεγαλύτερη βαρύτητα. Το πρόβλημα παρόλα αυτά παραμένει, ένα κείμενο είναι ελλιπής αναπαράσταση της ομάδας και κατ' επέκταση μίας ολόκληρης κατηγορίας, έτσι η απόδοση του εξαρτάται από το αν υπάρχουν κείμενα κάθε κατηγορίας που περιέχουν τα βασικά χαρακτηριστικά της. Η

αδυναμία αυτή ισχύει επειδή κάθε κείμενο έχει ένα σύνολο χαρακτηριστικών κατά πολύ μικρότερο από αυτό της συλλογής δεδομένων (και της ομάδας στην οποία βρίσκεται). Όταν υπάρχουν αρκετά αντικείμενα, λίγες ακραίες περιπτώσεις και γενικά διαχωρίσιμες ομάδες τότε είναι πιθανότερο να υπάρχουν κείμενα ικανά να περιγράψουν τις κατηγορίες.

Μία λύση στο πρόβλημα αυτό θα ήταν να υπολογίζαμε το διάνυσμα το οποίο σε κάθε διάστασή του θα είχε το ενδιάμεσο βάρος ενός χαρακτηριστικού που εμφανίζεται στην ομάδα. Ο υπολογισμός αυτός, όμως, είναι ιδιαίτερα ακριβός. Αν τα διανύσματα είναι διάστασης r τότε απαιτείται $O(r \cdot n_i^2)$ χρόνος, όπου επειδή γενικά $r >> n_i$ τελικά θα ξεπερνά κατά πολύ τους n^3 υπολογισμούς.

```
    Αλγόριθμος Κ-Ενδιαμέσων ομαδικής ενημέρωσης
   {
(1)    Αρχικοποίησε τα Κ κεντροειδή των ομάδων

(2)    Επανέλαβε,
(3)    {   Αποθήκευσε τη τρέχουσα λύση ως Λs-1 του βήματος s-1
(4)        Ανέθεσε κάθε κείμενο di, i=1,…,N στην ομάδα με το κοντινότερο κεντροειδές
(5)        Εντόπισε τους ενδιάμεσους των ομάδων και θέσε αυτούς ως κεντροειδή
(6)        Εκτίμησε την ποιότητα της ομαδοποίησης βάσει της συνοχής
                των ομάδων TCs και των ενδιαμέσων

(7)        Αν η συνοχή μειώθηκε στο τελευταίο βήμα, TCs < TCs-1
                Τερμάτισε τη διαδικασία και επέστρεψε τη λύση Λs-1
(8)    } μέχρι να μην συμβούν αλλαγές στα κεντροειδή
   }
```

Σχήμα 6.5. Ψευδοκώδικας αλγορίθμου Κ-Ενδιαμέσων ομαδικής ενημέρωσης.

Οι υπολογισμοί που απαιτούνται για την εύρεση του ενδιαμέσου έχουν μεγαλύτερο κόστος από τον υπολογισμό του αριθμητικού μέσου. Ο δεύτερος υπολογίζεται ύστερα από κάθε βήμα ανάθεσης των κειμένων απλά προσθέτοντας τα διανύσματα που εισήλθαν σε μία ομάδα και αφαιρώντας αυτά που αποχώρησαν. Αντίθετα η εύρεση του ενδιαμέσου της ομάδας C_i, αν υλοποιηθεί απλά, απαιτεί $O(n_i^2)$ υπολογισμούς στο πλήθος των κειμένων της ομάδας.

6.4. Αρχικοποίηση Κέντρων

Η αρχικοποίηση των κέντρων είναι πολύ σημαντική για λόγους που εξηγήσαμε. Υπάρχουν δύο κατηγορίες τεχνικών αρχικοποίησης:

- απλές τεχνικές επιλογής στοιχείων από τα δεδομένα χωρίς την εκτέλεση αλγορίθμου μάθησης,
- σύνθετες τεχνικές οι οποίες περιλαμβάνουν και αλγορίθμους μάθησης.

Για την αρχικοποίηση του αλγορίθμου Κ-Μέσων θα ασχοληθούμε με τεχνικές της πρώτης κατηγορίας, ενώ αυτές της δεύτερης έχουν μεγάλη σχέση με τις τεχνικές βελτίωσης λύσεων στις οποίες δε θα επεκταθούμε στην εργασία αυτή. Θα πρέπει να έχουμε κατά νου πως αξίζει να δαπανήσουμε λίγο χρόνο για τον ορθότερο ορισμό των αρχικών κέντρων.

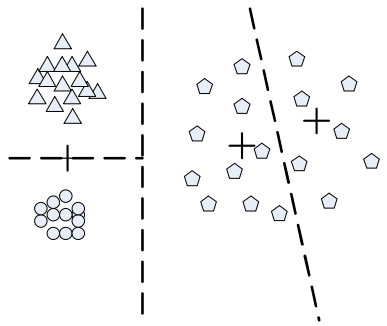
6.4.1. Ομοιόμορφη Αρχικοποίηση Κέντρων

Για τον ορισμό των αρχικών κέντρων επιλέγουμε ομοιόμορφα M στοιχεία από τα δεδομένα τα οποία θεωρούμε ως αρχικά κέντρα. Γενικά ο αλγόριθμος εκτελείται αρκετές φορές και λαμβάνουμε την καλύτερη διαμέριση βάση του κριτηρίου που βελτιστοποιείται. Αν πάλι γνωρίζουμε τις πραγματικές κατηγορίες των κειμένων και θέλουμε να έχουμε μία εικόνα για το σύνολο των λύσεων, υπολογίζουμε το μέσο όρο διαφόρων μέτρων εκτίμησης της ποιότητας της ομαδοποίησης.

Η τυχαία αρχικοποίηση έχει αρκετούς περιορισμούς που εξαρτώνται κυρίως από τα δεδομένα. Μία λύση είναι να χρησιμοποιούνται σύνθετες τεχνικές αρχικοποίησης, είτε εφαρμόζοντας έναν ιεραρχικό αλγόριθμο ομαδοποίησης, είτε τον Κ-Μέσων, σε ένα δείγμα που παίρνουμε από τα δεδομένα. Για να σκεφτεί κάποιος τρόπους αρχικοποίησης πρέπει να γνωρίζει τι επηρεάζει τον αλγόριθμο κατά την εκτέλεσή του.

Ο αλγόριθμος Κ-Μέσων δυσκολεύεται να εντοπίσει μη σφαιρικές ομάδες, ομάδες με διαφορετικές πυκνότητες και ομάδες διαφορετικών μεγεθών. Το βασικό στοιχείο είναι πως αν η αρχικοποίηση αναθέσει για παράδειγμα δύο κείμενα μίας κατηγορίας σε περισσότερες από μία ομάδες ως αρχικά κέντρα, τότε είναι πολύ πιθανό η κατηγορία αυτή να περιγράφεται επίσης από δύο ομάδες στην τελική λύση, ενώ αντίστοιχα τουλάχιστον μία άλλη ομάδα θα περιέχει στοιχεία από δύο κατηγορίες. Όσο αυξάνονται οι ομάδες του προβλήματος τόσο πιθανότερο είναι το φαινόμενο αυτό.

Πάντως σε κάποιες περιπτώσεις τα προβλήματα αυτά ξεπερνιούνται χωρίς πολύ κόπο. Για παράδειγμα, αν το πρόβλημα δεν απαιτεί να περιέχονται όλα τα δεδομένα μίας κατηγορίας σε μία ομάδα, αλλά επιθυμεί κάθε ομάδα του αποτελέσματος να αποτελείται από αντικείμενα μίας κατηγορίας, τότε μπορούμε να ομαδοποιήσουμε τα αντικείμενα σε περισσότερες από *M* ομάδες.



Σχήμα 6.6. Εσφαλμένος διαχωρισμός τριών ομάδων λόγω κακής αρχικοποίησης (διαφορετικές πυκνότητες ομάδων).

6.4.2. Ομοιόμορφη Αρχικοποίηση Ομάδων

Στην εκδοχή αυτή ανατίθεται κάθε αντικείμενο σε από τις *M* ομάδες. Χρησιμοποιούμε την αρχικοποίηση αυτή στο πειραματικό στάδιο, όχι για να μελετήσουμε την καταλληλότητά της, αλλά για να εξετάσουμε τη συμπεριφορά των τεχνικών ομαδοποίησης όταν οι αρχικές ομάδες είναι ακατάλληλες.

Η διαφορά στην περίπτωση της ομοιόμορφης αρχικοποίησης των ομάδων είναι πως όσο οι κατηγορίες έχουν ίδιο μέγεθος τόσο πιθανότερο είναι οι ομάδες να μην έχουν κυρίαρχα χαρακτηριστικά που ταιριάζουν με κάποια από τις κατηγορίες που αναζητούμε. Αν πάλι οι κατηγορίες έχουν μεγάλες διαφορές μεγεθών, είναι πιθανόν όλες οι ομάδες να περιγράφουν χαρακτηριστικά των πολυπληθών κατηγοριών.

Μετά την αρχικοποίηση των ομάδων το κεντροειδές μέσου όρου αντιπροσωπεύει βέλτιστα την ομάδα έχοντας μεγάλη σύγχυση περιεχομένου διαφορετικών κατηγοριών. Έτσι, κάθε αντικείμενο δεν έχει ξεκάθαρες επιλογές για το ποια ομάδα έχει κείμενα μεγαλύτερης συνάφειας και ο αλγόριθμος σταματά σε ένα μη αξιοπρεπές τοπικό ελάχιστο. Μπορούμε να αντιληφθούμε πως ακόμα και αν ομαδοποιήσουμε τα δεδομένα σε $K > M$ ομάδες, δεν είναι εύκολο να πάρουμε ομοιογενείς ομάδες. Οι συνθήκες που

δημιουργούνται με την ομοιόμορφη αρχικοποίηση κάνουν εξαιρετικά δύσκολο το πρόβλημα για τον K-Μέσων, περιπτώσεις που μας ενδιαφέρουν να μελετήσουμε.

Αντίθετα, όταν επιλέγουμε αντικείμενα ως αρχικά κέντρα, τα υπόλοιπα δεδομένα επιλέγουν με βάση τη συνάφεια περιεχομένου με τα *M* κείμενα. Έτσι είναι αρκετά πιθανό μετά την ανάθεση όλων των δεδομένων στις ομάδες και τον υπολογισμό του κεντροειδούς να έχουμε μερικές ομοιογενείς ομάδες.

6.4.3. Αρχικοποίηση Βάσει των M-μακρινότερων Αντικειμένων

Μια άλλη τεχνική αρχικοποίησης είναι να εντοπίσουμε ένα σύνολο *M* αντικειμένων τα οποία απέχουν αρκετά μεταξύ τους. Η τεχνική καλείται αρχικοποίηση με τα *M*-μακρινότερα αντικείμενα και στόχο έχει να καλύψει αρκετά τον χώρο των δεδομένων, ώστε τα κέντρα να μην είναι μόνο τυχαία αλλά και καλά διαχωρισμένα. Πιο συγκεκριμένα δε βρίσκονται τα *M*-μακρινότερα αντικείμενα της συλλογής κειμένων λόγω υψηλού υπολογιστικού κόστους. Ξεκινώντας από το πρώτο αντικείμενο, αναζητείται το αντικείμενο το οποίο βρίσκεται μακρύτερα από αυτό. Ύστερα, το επόμενο κέντρο θα πρέπει να απέχει τη μεγαλύτερη δυνατή απόσταση από τα προηγούμενα δύο κ.ο.κ.

Παρότι γενικά είναι καλύτερη προσέγγιση από την τυχαία αρχικοποίηση, έχει δύο μειονεκτήματα. Το μεν πρώτο αφορά την πολυπλοκότητα των υπολογισμών, σε σχέση με την απλότητα της τυχαίας επιλογής. Το δεύτερο και σοβαρότερο είναι πως αν τα δεδομένα περιέχουν ακραία αντικείμενα (*outliers*) τότε τείνει να τα επιλέξει ως αρχικά κέντρα, διότι αυτά είναι πιθανό να βρίσκονται αρκετά μακριά από τα «φυσιολογικά» δεδομένα. Στην πράξη συνήθως η αναζήτηση γίνεται σε ένα δείγμα δεδομένων το οποίο εξασφαλίζει ότι οι πυκνές περιοχές είναι πιθανότερο να αντιπροσωπεύονται στο δείγμα, σε σχέση με τις αραιές, οπότε περιορίζεται και η επιρροή των τελευταίων στην αρχικοποίηση.

Δύο εναλλακτικές που μπορούμε να εφαρμόσουμε για την επιλογή του πρώτου αντικειμένου είναι να επιλέξουμε:

- τον αριθμητικό μέσο (αν βέβαια μπορεί να υπολογιστεί),
- το ενδιάμεσο αντικείμενο όπως συζητήθηκε, ή
- την τυχαία ομοιόμορφη επιλογή ενός αντικειμένου από τη συλλογή.

Επειδή όπως είναι εμφανές τα αντικείμενα επιλέγονται κατά κάποιο τρόπο, από την περιφέρεια του συνόλου, η πρώτη επιλογή προσπαθεί να εισάγει ένα μεσαίο αντικείμενο ως πρώτο στοιχείο για να καλύψει και την περιοχή αυτή.

```
    Αλγόριθμος Εύρεσης Μ-μακρινότερων αντικειμένων στα δεδομένα
    {
(1)    Θεώρησε το σύνολο Mf = {di) που αποτελείται από ένα αντικείμενο
(2)    Θέσε K = 1

(3)    Επανέλαβε,
              /* αυτό με τη μικρότερη μέση ομοιότητα */
(4)    {   Εντόπισε το αντικείμενο που βρίσκεται μακρύτερα από τα αντικείμενα του Mf
(5)        Πρόσθεσέ το στο σύνολο Mf και θέσε K = K+1
(6)   } μέχρι να προκύψει |Mf| = M
    }
```

Σχήμα 6.7. Ψευδοκώδικας για την εύρεση των *M*-μακρινότερων αρχικών κέντρων.

Η μεθοδολογία αυτή πειραματικά έδειξε πολύ καλά χαρακτηριστικά και τη συστήνουμε για το πρόβλημα. Μπορεί να εφαρμοστεί κι αυτή επαναληπτικά αν επιθυμούμε να αναζητήσουμε τυχόν καλύτερες αρχικοποιήσεις.

6.5. Γενικευμένος Αλγόριθμος K-Μέσων

Η τεχνική αυτή (*Global K-Means*), προσπαθεί με έναν αυξητικό τρόπο να εισάγει βέλτιστα μία νέα ομάδα, σε μία λύση μικρότερης τάξης [72]. Πιο αναλυτικά, έχοντας τα *N* έγγραφα και επιθυμώντας τον διαχωρισμό τους σε *M* ομάδες, η διαδικασία μπορεί να περιγραφεί ακολούθως: υπολογίζουμε το βέλτιστο κέντρο θεωρώντας ότι όλα τα έγγραφα ανήκουν σε μία ομάδα (k = 1), ελέγχοντας όλα τα υποψήφια έγγραφα στο ρόλο αυτό και εκτελώντας τον αλγόριθμο K-Μέσων κάθε φορά. Για να λάβουμε τη λύση για το πρόβλημα της επόμενης τάξης (k = 2) αναζητούμε ένα πρότυπο, το οποίο μαζί με το κέντρο του προηγούμενου σταδίου θα δίνουν τη βέλτιστη λύση (τη λύση με το μικρότερο σφάλμα) για την τάξη αυτή κ.ο.κ. Στη γενική περίπτωση:

- αν υποθέσουμε ότι το σύνολο $\left(c_1^*(k-1), ..., c_{k-1}^*(k-1)\right)$ ορίζει τη βέλτιστη λύση για το πρόβλημα της k-1 τάξης,
- υπολογίζονται όλες οι πιθανές λύσεις του προβλήματος k ομάδων, με αρχικές συνθήκες $\left(c_1^*(k-1), ..., c_k^*(k-1) = d_i\right)$ με i=1, …N, εφαρμόζοντας τοπική αναζήτηση με τον K-Μέσων,

- η καλύτερη λύση που προκύπτει αποτελεί και τη βέλτιστη για το πρόβλημα k ομάδων $\left(c_1^*(k),\ c_2^*(k),\ ...,\ c_{k-1}^*(k),\ c_k^*(k)\right)$.
- Επανάληψη των βημάτων έως ότου $k = M$.

Μπορούμε να αναφέρουμε κάποιες σημαντικές παρατηρήσεις για τη μέθοδο αυτή. Πριν από όλα για να αποφευχθεί η σύγχυση των συμβόλων, οι αρχικές συνθήκες του κάθε σταδίου επίλυσης δεν είναι μέρος της βέλτιστης λύσης του επόμενου βήματος, γενικά δηλαδή: $c_t^*(k\text{-}1) \neq c_t^*(k)$, αφού εκτελώντας τον Κ-Μέσων τα κέντρα μεταβάλλονται.

Είναι εμφανές ότι σε κάθε βήμα γίνονται κατά προσέγγιση N υπολογισμοί λύσεων εκτελώντας τον αλγόριθμο Κ-Μέσων, μία για το κάθε υποψήφιο κέντρο της νέας ομάδας. Στην περίπτωση που ως κέντρα θεωρούμε το ενδιάμεσο αντικείμενο κάθε ομάδας, τότε υποψήφια για να αποτελέσουν το αρχικό κέντρο της νέας ομάδας είναι όλα τα έγγραφα τα οποία δεν αποτελούν μέρος της λύσης μικρότερης τάξης. Ασφαλώς και πρόκειται για υπολογιστικά άπληστη και αρκετά βαριά μέθοδο, όμως υπάρχουν αρκετές τροποποιήσεις που μπορούν να γίνουν και να την επιταχύνουν αρκετά. Ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο [72], όπου οι συγγραφείς προτείνουν και έναν τρόπο να επιταχυνθεί η αλγοριθμική διαδικασία.

Δεν είναι δύσκολο να διαπιστώσει κανείς ότι ο Γενικευμένος Κ-Μέσων είναι τουλάχιστον όσο αποτελεσματικός είναι ο Κ-Μέσων τυχαίας αρχικοποίησης. Η λύση που παράγεται είναι κοντά στην βέλτιστη, το οποίο βασίζεται στο συλλογισμό ότι η βέλτιστη λύση με k ομάδες μπορεί να προκύψει εφαρμόζοντας τοπική αναζήτηση με τον παραδοσιακό Κ-Μέσων, ξεκινώντας από μία αρχική κατάσταση όπου:

- τα k-1 κέντρα είναι τοποθετημένα στις βέλτιστες θέσεις για το πρόβλημα της αντίστοιχης τάξης,
- το εναπομείναν k-οστό κέντρο τοποθετείται στην κατάλληλη θέση η οποία ανακαλύπτεται βάση της λύσης μικρότερης τάξης.

Σε κάθε περίπτωση η μέθοδος είναι ανεξάρτητη των όποιων συνθηκών αρχικοποίησης και παράγει την ίδια λύση σε κάθε εφαρμογή της, όταν δε μεταβάλλεται το σύνολο δεδομένων. Αυτό είναι ιδιαιτέρως χρήσιμο διότι εξασφαλίζοντας ότι ο αλγόριθμος δίνει συγκεκριμένο αποτέλεσμα, μπορούμε να συγκρίνουμε την απόδοση

διαφόρων επιλογών ρύθμισης που αφορούν το μοντέλο αναπαράστασης, τις συναρτήσεις ομοιότητας κ.α.

Κατά την επίλυση του προβλήματος *M* ομάδων έχουμε επιλύσει το πρόβλημα για κάθε $k < M$, χαρακτηριστικό το οποίο μπορούμε να εκμεταλλευτούμε ώστε εφαρμόζοντας κάποιο κριτήριο διακοπής της διαδικασίας να εκτιμήσουμε τον αριθμού ομάδων που υπάρχουν στα δεδομένα.

6.6. Μια άλλη Προσέγγιση για τον Υπολογισμό του Κέντρου Ομάδας

Αναφέραμε τους αναλυτικούς υπολογισμούς των κεντροειδών που οδηγούν σε μονότονη σύγκλιση του αλγορίθμου. Κάθε αναλυτικό κέντρο αποτελεί το διάνυσμα το οποίο μεγιστοποιεί την ομοιότητα των κειμένων στο εσωτερικό μίας ομάδας.

Στην παράγραφο αυτή θα συζητηθεί η καταλληλότητα της βέλτιστης αναπαράστασης των χαρακτηριστικών μίας ομάδας με το μέσο διάνυσμα ως κεντροειδές. Επίσης θα παρουσιάσουμε μία προσέγγιση την οποία προτείνουμε για το πρόβλημα και μπορεί να εφαρμοστεί ως τροποποίηση του αλγορίθμου K-Μέσων αλλά και του συσσωρευτικού αλγορίθμου ομαδοποίησης.

6.6.1. Το Βέλτιστο Κέντρο ως Βέλτιστος Αντιπρόσωπος

6.6.1.1. Το Πρόβλημα του Θορύβου

Στην παρούσα εργασία περιγράφοντας κάθε στάδιο της επίλυσης του προβλήματος, από την επεξεργασία έως εδώ, έχουμε θίξει πολλά ζητήματα που αφορούν τη φύση των κειμένων ως δεδομένα υψηλής διάστασης με σύνθετη νοηματική δομή αλλά και πολλά περιττά χαρακτηριστικά. Το ζήτημα που ανακύπτει εδώ είναι το αν ο αριθμητικός μέσος είναι μία καλή επιλογή για τα κεντροειδή των ομάδων.

Αν το δούμε ως πρόβλημα βελτιστοποίησης της αντικειμενικής συνάρτησης, τότε κάποιος θα έλεγε με βεβαιότητα πως η επιλογή αυτή αποτελεί τον καλύτερο δυνατό αντιπρόσωπο των κειμένων μίας ομάδας διότι κωδικοποιεί τη σχετική κατανομή όλων των χαρακτηριστικών του (διεύθυνση διανύσματος). Από μαθηματικής άποψης, λοιπόν, δε μπορούμε να αμφισβητήσουμε τον αριθμητικό μέσο. Όμως, το συμπέρασμα αυτό αφήνει στην άκρη δύο βασικά ζητήματα που αφορούν τα προβλήματα οργάνωσης κειμένων φυσικής γλώσσας:

1. Η αντικειμενική συνάρτηση είναι μία έκφραση-κριτήριο που εκτιμά χωρίς γνώση των πραγματικών κατηγοριών των δεδομένων την ποιότητα της ομαδοποίησης. Επίσης, η συνάρτηση αυτή δεν αμφισβητεί τα δεδομένα, θεωρεί ότι τα κείμενα περιέχουν χρήσιμα χαρακτηριστικά και αντιλαμβάνεται τη σημαντικότητα τους μέσω της συχνότητας τους σε ένα κείμενο ή μία ομάδα.
2. Γνωρίζουμε πως τα χαρακτηριστικά αυτά δεν είναι συμπαγείς πληροφορίες για το μηχανικό χειρισμό τους, πολλές φορές ούτε για τον άνθρωπο ο οποίος αναγκάζεται να ξεφυλλίσει ένα κείμενο-άρθρο πριν συμπεράνει το τι θέμα έχει.

Το ότι άλλωστε αφιερώνεται πολύς χρόνος για τη σωστή επιλογή των χαρακτηριστικών που θα παρέχουμε στη βασική αλγοριθμική διαδικασία κατά την προεπεξεργασία συμπληρώνει και τα δύο παραπάνω σημεία.

Διατυπώνοντας ξεκάθαρα το γενικότερο συλλογισμό και το πού εντοπίζουμε το ενδιαφέρον μας, θα λέγαμε ότι:

- δεν αμφισβητείται το μαθηματικά βέλτιστο κεντροειδές σε καμία περίπτωση.
- αυτό που αμφισβητείται είναι η ποιότητα της πληροφορίας που περιέχουν οι αναπαραστάσεις των κειμένων, κατ' επέκταση και οι αντιπρόσωποι των ομάδων.
- ο ισχυρισμός αυτός επεκτείνεται, όποιες κι αν είναι οι μηχανικές τεχνικές άντλησης πληροφορίας (*IR*), φιλτραρίσματος, μοντέλο αναπαράστασης δεδομένων, που έχουμε επιλέξει για τα κείμενα.
- συνεπώς, η προσκόλληση στη βέλτιστη αναπαράσταση μίας ομάδας η οποία βασίζεται στη μη-βέλτιστη μηχανική κωδικοποίηση των χαρακτηριστικών των κειμένων τίθεται τουλάχιστον υπό συζήτηση.

Από την άλλη πλευρά ο ισχυρισμός περί συζητήσιμης ποιότητας δεδομένων, μεταφέρει το πρόβλημα στο επίπεδο της εξαγωγής των χαρακτηριστικών, θεωρώντας ότι το βασικό αλγοριθμικό στάδιο κάνει ότι είναι δυνατόν βάσει των δεδομένων που του παρέχονται. Πάντα όμως θα αντιμετωπίζεται κάποιου είδους δυσκολία στην μηχανική αποκωδικοποίηση των κειμένων και στην επιλογή χαρακτηριστικών.

Μια βασική διαπίστωση, που μπορεί να φανεί χρήσιμη και στη συνέχεια, είναι ότι: με τις παραδοσιακές μεθόδους δε μπορούμε να παράγουμε περισσότερη πληροφορία

από αυτή που μας παρέχεται από τα δεδομένα, μπορούμε όμως να επιλέξουμε μικρότερο μέρος από αυτή. Αυτό αφορά κάθε κείμενο χωριστά, όπου ουσιαστικά απαλείφονται χαρακτηριστικά τους, αλλά κατ' επέκταση αφορά και τα κέντρα των ομάδων τα οποία ορίζονται βάσει των χαρακτηριστικών των κειμένων της ομάδας. Μπορεί επίσης να καταλάβει κανείς πως αν σε κάθε κείμενο μας απασχολεί το γεγονός ότι μπορεί να περιέχει όρους θορύβου, τότε σε επίπεδο ομάδας ο θόρυβος που προέρχεται από πολλά κείμενα υψηλής διάστασης μπορεί να αποτελεί ακόμα μεγαλύτερο κίνδυνο για το πρόβλημα.

6.6.1.2. Αριθμητικά και Τεχνικά Προβλήματα

Ο αριθμητικός μέσος παρουσιάζει κάποια προβλήματα στο συγκεκριμένο πρόβλημα:

1) Πρώτον «συσσωρεύει» πολλή πληροφορία από τα χαρακτηριστικά της ομάδας. Υπό άλλες συνθήκες, όπου τα διανύσματα των κειμένων δεν ήταν αραιά και περιείχαν πληροφορία για τα περισσότερα από τα διαφορετικά χαρακτηριστικά της συλλογής, το φαινόμενο αυτό δε θα μας απασχολούσε ιδιαίτερα. Μαζί με τον θόρυβο κάθε κείμενο θα περιείχε έντονο συχνοτικό περιεχόμενο σε όλες τις λέξεις με τις οποίες έχει σχέση το θέμα του. Έτσι, αν χρησιμοποιούσαμε την ομοιότητα συνημίτονου η κλίση του κεντροειδούς με το διάνυσμα ενός μη σχετικού αντικειμένου της ομάδας θα μεγάλωνε, και αυτό πιθανόν να ανάγκαζε το αντικείμενο να μεταφερθεί σε άλλη ομάδα.

   Όμως, η συσσώρευση στη συγκεκριμένη περίπτωση είναι ιδιαίτερα προβληματική, διότι τα αντικείμενα δεν έχουν ξεκάθαρη πληροφορία για κάθε όρο της συλλογής με τον οποίο σχετίζονται, παρά μόνο για αυτούς που εμφανίζονται σε καθένα χωριστά. Αυτό σημαίνει πως λόγω της μεγάλης διάστασης η μέση ομοιότητα των αντικειμένων της ίδιας κατηγορίας θα είναι μικρή, σε σχέση με την πραγματική ομοιότητα την οποία θα μπορούσε να συμπεράνει ο άνθρωπος. Συνεπώς, το κεντροειδές συγκεντρώνοντας χαρακτηριστικά αυξάνει την πιθανότητα να ισχύσει $ratio < 1$, όπως ορίστηκε στο Κεφάλαιο 5. Το φαινόμενο αυτό εντείνεται στην περίπτωση όπου το σύνολο δεδομένων περιέχει λίγα κείμενα.

2) Παρουσιάζεται το πρόβλημα της αυτό-ομοιότητας (*self-similarity*) κατά την επιλογή της κοντινότερης ομάδας. Το πρόβλημα αυτό είναι άμεσα συνδεδεμένο με το σημείο (1). Αφορά γενικά κάθε μέτρο ομοιότητας, αλλά εδώ θα περιγράψουμε την περίπτωση της συνημιτονοειδούς ομοιότητας.

Ο υπολογισμός της ομοιότητας ενός κειμένου με τον αριθμητικό μέσο της ομάδας στην οποία βρίσκεται περιλαμβάνει και την ομοιότητα του κειμένου με τον εαυτό του:

$$d_j^T c_i = d_j^T \frac{m_i}{\|m_i\|} = d_j^T \frac{1}{\|m_i\|} \cdot \sum_{d_k \in C_i} d_k = \frac{1}{\|m_i\|} \cdot d_j^T d_j + \frac{1}{\|m_i\|} \cdot \sum_{d_k \in C_i} d_j^T d_k \ .$$

Αν όπως εξηγήσαμε στο σημείο (1), η μέση ομοιότητα ανάμεσα στα κείμενα της ίδιας ομάδας είναι μικρή τότε και η ομοιότητα ενός κειμένου προς την ομάδα που ανήκει (επειδή ανατέθηκε εκεί αρχικά ή βρέθηκε κατά την διαδικασία) θα είναι επίσης μικρή (δεύτερος όρος της σχέσης). Όμοια συμπεραίνουμε και για την ομοιότητα του κειμένου με τις υπόλοιπες ομάδες.

Κατά αυτόν τον τρόπο, ο πρώτος όρος που αφορά την ομοιότητα ενός κειμένου με τον εαυτό του, μπορεί να είναι πολύ μεγαλύτερος του δεύτερου με αποτέλεσμα να αποφασίζει από μόνος του την παραμονή του αντικειμένου στην ομάδα που ήδη βρίσκεται. Γενικεύοντας την παρατήρηση αυτή σε όλο το σύνολο δεδομένων τότε στην περίπτωση που έχουμε μία μέτρια αρχικοποίηση (με μικρή εσωτερική ομοιότητα στις ομάδες) υπάρχει ο κίνδυνος να πάρουμε μία κακή λύση. Μάλιστα αναμένεται να γίνονται πολύ λίγες μεταθέσεις αντικειμένων μεταξύ των ομάδων. Μία σχετική συζήτηση μπορεί να αναζητηθεί στην βιβλιογραφία [13].

Επίσης, η έκφραση αυτή δικαιολογεί γιατί πολλές φορές οι αλγόριθμοι της οικογένειας K-Μέσων δημιουργούν ομάδες συγχωνευμένων κατηγοριών.

Συχνά για να ξεπεραστούν παρόμοια προβλήματα εφαρμόζονται μέθοδοι δραστικού φιλτραρίσματος σε κάθε κείμενο, ευελπιστώντας να διατηρηθεί ένα μικρό αλλά αντιπροσωπευτικό σύνολο χαρακτηριστικών. Είναι προτιμότερο δηλαδή να χάσουμε αρκετή πληροφορία αποσκοπώντας στη μείωση της επίδρασης του θορύβου. Ουσιαστικά, έτσι γίνεται πιο αραιός ο πίνακας ομοιοτήτων των κειμένων.

Επίσης έχουν προταθεί διάφορες τεχνικές στη βιβλιογραφία οι οποίες προσπαθούν να λύσουν το πρόβλημα αυτό. Στο [13] αντί να γίνεται η ανάθεση όλων των κειμένων στις νέες ομάδες τους, δημιουργούνται αλυσίδες αναθέσεων οι οποίες βελτιώνουν τη λύση. Στο [31] παρεμβάλλεται ένα βήμα μετά την ανάθεση των κειμένων στις ομάδες, όπου κάθε κείμενο «ρωτάει» έναν αριθμό κοντινότερων γειτόνων της συλλογής για την ομάδα στην οποία βρίσκονται. Βάση της πλειοψηφίας των προτάσεων αυτών κάθε κείμενο μπορεί να μεταφερθεί σε μία άλλη ομάδα από αυτή του κοντινότερου κεντροειδούς. Αναγκαίο είναι επίσης να δημιουργηθούν αλυσίδες μεταθέσεων οι οποίες να εξασφαλίζουν τη βελτίωση της ομαδοποίησης.

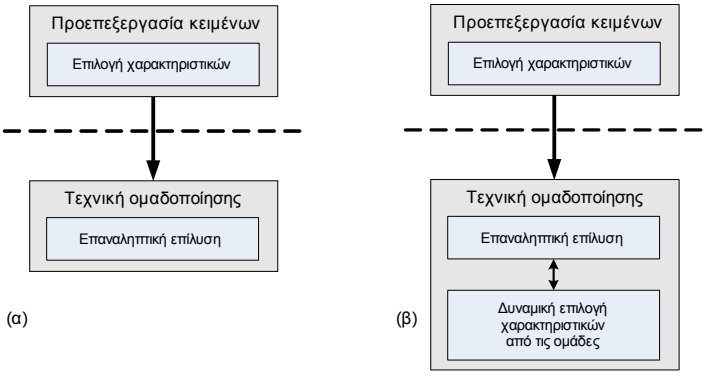
6.6.2. Τεχνικές για τον Υπολογισμό Καλύτερων Αντιπροσώπων Ομάδων

Η συμβολή της εργασίας αυτής σε ότι αφορά το βασικό στάδιο ομαδοποίησης των κειμένων στηρίζεται στην ιδέα ότι μπορούμε να πάρουμε ένα υποσύνολο της πληροφορίας από κάθε ομάδα ως αντιπρόσωπό της. Οι αλγόριθμοι ομαδοποίησης και η διαδικασία επιλογής χαρακτηριστικών είναι παραδοσιακά δύο ανεξάρτητα στάδια της λύσης ενός προβλήματος. Το δεύτερο επιτελείται κατά τη διάρκεια της προεπεξεργασίας, ενώ το πρώτο λύνει επαναληπτικά το πρόβλημα. Έχουμε επίσης αναφέρει πως υπάρχουν προσεγγίσεις οι οποίες πρώτα ομαδοποιούν τα δεδομένα, ύστερα εξάγουν συμπεράσματα από τις ομάδες, αναθέτουν καλύτερα βάρη στους όρους των κειμένων και τέλος προσπαθούν να παράγουν την τελική διαμέριση.

Η προσέγγιση που προτείνεται αφορά στην ενσωμάτωση μίας διαδικασίας επιλογής χαρακτηριστικών στην βασική διαδικασία επίλυσης του προβλήματος με σκοπό τον καταλληλότερο ορισμό των αντιπροσωπευτικών κέντρων των ομάδων. Τα κέντρα αυτά τα ονομάζουμε συνθετικά, διότι δεν προκύπτουν από τον αριθμητικό μέσο των αντικειμένων. Η κλασσική προεπεξεργασία και επιλογή χαρακτηριστικών φυσικά και θα παραμείνει ως ανεξάρτητο στάδιο επεξεργασίας των δεδομένων εισόδου.

Η λογική είναι πως ενώ αρχικά η επιλογή χαρακτηριστικών δεν έχει πληροφορία για τα κατηγορίες και δε μπορεί να επιλέξει πάντα με επιτυχία χαρακτηριστικά, κατά τη διάρκεια της επίλυσης δημιουργείται επιπλέον πληροφορία για τη σχετικότητα των κειμένων, η οποία δεν είναι άλλη από την τρέχουσα ομαδοποίηση. Επειδή ακριβώς τα κριτήρια εκτίμησης της σημαντικότητας των χαρακτηριστικών αλλάζουν κατά την

εξέλιξη της λύσης, η διαδικασία επιλογής χαρακτηριστικών λαμβάνει δυναμικά αποφάσεις για την επιλογή των σημαντικών χαρακτηριστικών κάθε ομάδας.



Σχήμα 6.8. (α) Ανεξάρτητα στάδια επιλογής χαρακτηριστικών και επίλυσης, (β) Η προτεινόμενη προσέγγιση της δυναμικής επιλογής χαρακτηριστικών.

Η τεχνική έχει μία διαφορετική φιλοσοφία επίλυσης του προβλήματος επιτρέποντας τη συνεργασία δύο μηχανισμών επεξεργασίας των δεδομένων, και παρουσιάζει ένα σύνολο πλεονεκτημάτων:

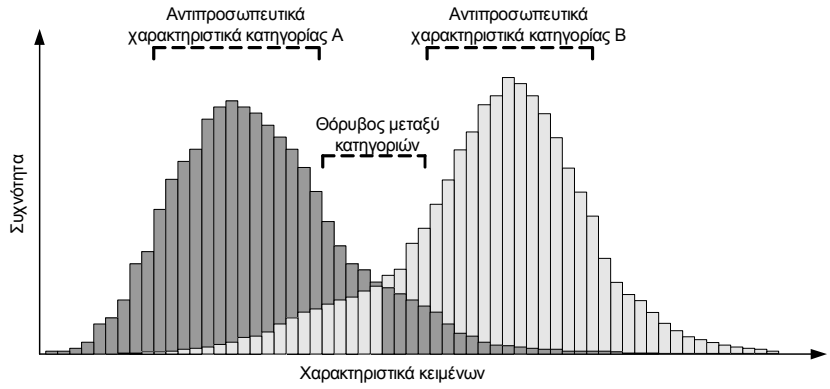
- Η συνεργατική σχέση αλγορίθμου ομαδοποίησης και δυναμικού ορισμού των κέντρων είναι μία σχέση αλληλοϋποστήριξης: ο αλγόριθμος βελτιώνει τη διαμέριση των δεδομένων παρέχοντας ποιοτικότερη πληροφορία στη διαδικασία επιλογής συνθετικών κέντρων, και η τελευταία δημιουργεί αντιπροσώπους που μπορούν να βοηθήσουν τον αλγόριθμο σε περαιτέρω βελτίωση της λύσης κ.ο.κ.
- Από τη δυναμική επιλογή χαρακτηριστικών για τον ορισμό των συνθετικών κέντρων, δεν αλλοιώνεται σε καμία περίπτωση το περιεχόμενο των κειμένων για τα επόμενα βήματα. Κάποια χαρακτηριστικά μπορεί να μη συμπεριλαμβάνονται στο συνθετικό κέντρο της τρέχουσας ομάδας τους, όμως σε επόμενα βήματα μπορεί να είναι ανάμεσα στα κυρίαρχα μίας ομάδας και να παίξουν ρόλο στη διαμόρφωση του συνθετικού κέντρου της.
- Μειώνεται η αναγκαιότητα για έντονο φιλτράρισμα αφού κατά τη διάρκεια της εκπαίδευσης ο αλγόριθμος μειώνει την επιρροή των χαρακτηριστικών θορύβου.

Πριν μελετήσουμε εναλλακτικές για την επιλογή των αντιπροσώπων βάσει της προσέγγισης αυτής, θα πρέπει να έχουμε μία εικόνα για το τι είδους προβλήματα θέλουμε να επιλύσουμε. Έχουμε αναφερθεί στα περισσότερα, όμως το κρισιμότερο ίσως φαινόμενο το οποίο θα πρέπει να αντιμετωπίσουμε είναι πως βάσει της κακής αρχικοποίησης και του φαινομένου της αυτό-ομοιότητας (*self similarity problem*) οι ομάδες που διαμορφώνονται είναι σε πολλές περιπτώσεις συγχωνεύσεις κατηγοριών.

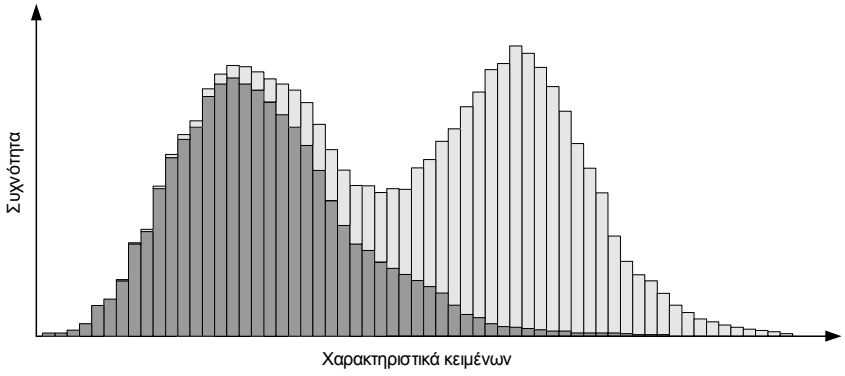
Μπορούμε να παρατηρήσουμε για παράδειγμα τη συγκέντρωση της μεγάλης πλειοψηφίας των κειμένων δύο κατηγοριών σε μία ομάδα οι οποίες δε μπορούν να διασπαστούν από τον αλγόριθμο και τη συνάρτηση ομοιότητας. Αυτό συμβαίνει γιατί παράλληλα σχηματίζονται «άχρηστες» ομάδες που περιέχουν πολύ λίγα κείμενα και πολλές φορές από διαφορετικές κατηγορίες, τα οποία δε μπορούν να αποτελέσουν ανταγωνιστικό κίνητρο για τα εγκλωβισμένα σε ακατάλληλες ομάδες κείμενα.

Το Σχήμα 6.10 θα μας φανεί χρήσιμο στη συνέχεια για να περιγράψουμε ποιοτικά τη συμπεριφορά των τεχνικών ορισμού συνθετικών κέντρων. Εμφανίζει ένα υποθετικό ιστόγραμμα χαρακτηριστικών στο οποίο έχουμε διατάξει εσωτερικά τις κατανομές των χαρακτηριστικών γύρω από τη μέγιστη συχνότητα, ώστε ανάμεσά τους να εμφανίζεται η συχνοτική τομή του κοινού λεξιλογίου. Η τομή αυτή μπορεί να είναι θόρυβος και για τις δύο κατηγορίες, δηλαδή κάποιες συνηθισμένες για τη συλλογή λέξεις, μπορεί όμως να μην αποτελεί και θόρυβο, εκφράζοντας μία πιθανή τομή περιεχομένου.

Στο σημείο αυτό εντοπίζεται το ενδιαφέρον μας. Θα πρέπει δηλαδή να βρούμε έναν τρόπο απόφασης για το ποια κατηγορία θα παραμείνει στην ομάδα και ποια θα πρέπει να αποχωρήσει. Η δυσκολία είναι πως δε μπορούμε να γνωρίζουμε ότι έχουμε συγχώνευση ομάδων για να επέμβουμε ειδικά στις ομάδες που παρουσιάζεται το πρόβλημα. Το πρόβλημα λοιπόν θα πρέπει να εντοπίζεται εμμέσως.



Σχήμα 6.9. Υποθετικές κατανομές (ιστόγραμμα) χαρακτηριστικών δύο κατηγοριών Α, Β οι οποίες έχουν συγχωνευτεί σε μία ομάδα της λύσης.



Σχήμα 6.10. Το αθροιστικό ιστόγραμμα της ομάδας. Οι όροι θορύβου αναδεικνύονται μέσω της συσσώρευσης.

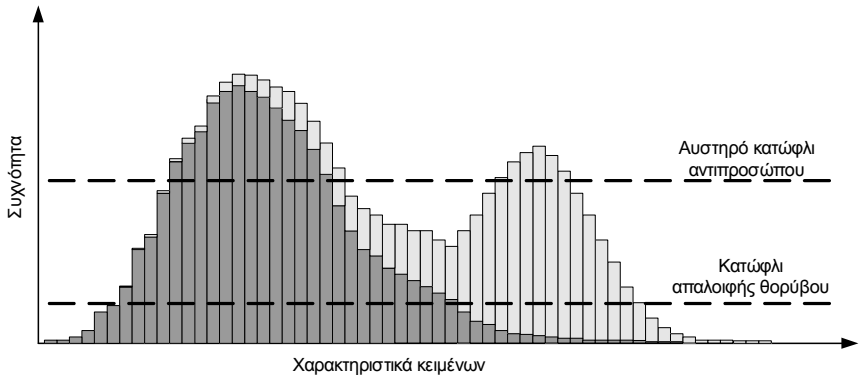
6.6.2.1. Φιλτράρισμα πάνω στον Αλγεβρικό Μέσο

Έχοντας τα κείμενα μίας ομάδας υπολογίζεται ο αλγεβρικός μέσος συλλέγοντας πληροφορία για όλα τα εμφανιζόμενα χαρακτηριστικά της ομάδας, Στη συνέχεια, θέτουμε ένα κατώφλι στους όρους του αντιπροσώπου, διατηρώντας μόνο αυτούς με τα υψηλότερα βάρη στο κανονικοποιημένο γράφημα-διάνυσμα. Αν έχουμε μοντέλο με σχέσεις, διατηρούμε τις σχέσεις που υπάρχουν ανάμεσα στους επιλεγμένους όρους.

Η τεχνική της κατωφλίωσης μοντέλου αναπαράστασης, που παρουσιάστηκε στο Κεφάλαιο 3, αν εφαρμοστεί σε κάθε κείμενο χωριστά μπορεί να προκαλέσει σημαντική απώλεια πληροφορίας. Εδώ όμως λαμβάνεται υπόψη η σχετική σημαντικότητα κάθε χαρακτηριστικού για τη ομάδα. Συνεπώς χαρακτηριστικά που πιθανόν να είχαν

απαλειφθεί από τα κείμενα είναι δυνατόν, βρισκόμενα στην κατάλληλη ομάδα, να συνδιαμορφώσουν τα σημαντικά χαρακτηριστικά της.

Αν θεωρήσουμε ότι υπάρχει μία κυρίαρχη κατηγορία στην ομάδα, τότε η κατωφλίωση μεγέθους θα διατηρήσει στον αντιπρόσωπο όρους που αφορούν κείμενα αυτής της κατηγορίας αναθέτοντας εμμέσως έναν συγκεκριμένο ρόλο στην ομάδα: να εκφράσει την κυρίαρχη κατηγορία.



Σχήμα 6.11. Υποθετική κατανομή χαρακτηριστικών μίας ομάδας και τα αντίστοιχα κατώφλια μοντέλου για τον αντιπρόσωπο.

Επίσης μπορεί να περιορίσει και το θόρυβο ο οποίος μπορεί να μην προέρχεται από το κοινό λεξιλόγιο των κατηγοριών (τα χαρακτηριστικά με μικρές συχνότητες που βρίσκονται στα δύο άκρα του ιστογράμματος. Ο η πληροφορία αυτή μπορεί να είναι θόρυβος μεταξύ διαφορετικών ομάδων κειμένων, υπονοώντας συνηθισμένα χαρακτηριστικά που εμφανίζονται σε πολλές κατηγορίες κειμένων. Μειώνοντας τα επίπεδα του θορύβου αυτού τα κεντροειδή απομακρύνονται, με αποτέλεσμα να δίνουν έμφαση στα κυρίαρχα χαρακτηριστικά που μπορούν να διαχωρίσουν τις κατηγορίες.

Οι κατηγορίες αποκτούν κίνητρο να διαχωριστούν σε διαδοχικά βήματα, όπου όσο περισσότερο επικρατεί η μία από τις δύο τόσο ο αντιπρόσωπος γίνεται πιο ειδικός για τη συγκεκριμένη κατηγορία. Μειώνεται και η ένταση του φαινομένου *self-similarity*. Από την άλλη πλευρά διώχνουμε πολλά χαρακτηριστικά από τον αντιπρόσωπο ευελπιστώντας να διατηρήσουμε χαρακτηριστικά μόνο από την κυρίαρχη κατηγορία της ομάδας, η οποία υποθέτουμε πως έχει υψηλότερο συχνοτικό περιεχόμενο. Οι περιπτώσεις που αδυνατεί να ανιχνεύσει η τεχνική αυτή είναι:

- οι διαφορετικές μορφές που μπορεί να παρουσιάζουν οι κατανομές των χαρακτηριστικών των δύο κατηγοριών. Η κυρίαρχη κατηγορία έχει σίγουρα περισσότερο συχνοτικό περιεχόμενο αλλά δεν έχει απαραίτητα τα χαρακτηριστικά με τις υψηλότερες συχνότητες της ομάδας.
- οι δύο κατηγορίες μπορεί να συμμετέχουν με ίσο ή περίπου ίσο περιεχόμενο στην ομάδα, περίπτωση στην οποία η τεχνική κατωφλίωσης θα διατηρούσε καλύτερη πληροφορία και από τις δύο κατηγορίες.

Θεωρούμε πως η κατωφλίωση του αντιπροσώπου είναι χρήσιμη ως γενική τεχνική η οποία θα πρέπει να χρησιμοποιείται αποκόπτοντας τις χαμηλές συχνότητες της ομάδας (κυρίως το θόρυβο προς άλλες ομάδες). Ουσιαστικά αυτό που επιτυγχάνεται είναι ο διαχωρισμός των δεδομένων μεταφέροντάς τα σε έναν υποχώρο του προβλήματος που αποτελείται από τα χαρακτηριστικά που περιλαμβάνουν οι συνθετικοί αντιπρόσωποι. Ο υποχώρος αυτός είναι αντικείμενο της εκπαίδευσης και εκτιμάται δυναμικά.

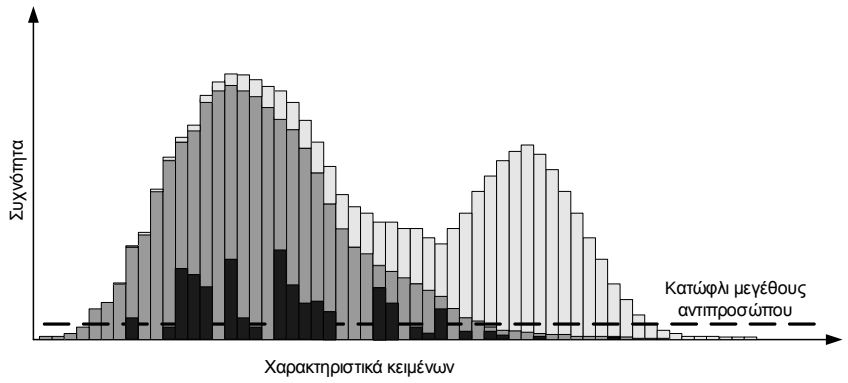
Αν και με λογική ρύθμιση του κατωφλίου μπορεί να επιτύχει καλύτερα αποτελέσματα από τον αλγεβρικό μέσο, δε μπορεί να εντοπίσει το ενδιαφέρον περιεχόμενο παρά μόνο αποκόπτοντας συχνοτικό περιεχόμενο. Τα αρνητικά σημεία της είναι εντονότερα στα αρχικά βήματα της εκπαίδευσης όπου οι ομάδες δεν έχουν ομοιογένεια περιεχομένου. Αν πάλι της δοθεί μία μέτρια αρχικοποίηση αναμένει κανείς να καταφέρει να ξεπεράσει τα μειονεκτήματα της και να δώσει καλά αποτελέσματα.

6.6.2.2. Ο Ενδιάμεσος ως Αντιπρόσωπος Ομάδας

Ο ενδιάμεσος έχει χρησιμοποιηθεί και στο παρελθόν στην ομαδοποίηση κειμένων με το γραφοθεωρητικό μέτρο ομοιότητας χωρίς βάρη, αλλά η επιλογή του είχε να κάνει περισσότερο με την αδυναμία ορισμού του βέλτιστου κέντρου για την ομοιότητα αυτή και την δυσκολία χειρισμού των γραφημάτων ως διανύσματα. Εδώ εξετάζουμε την επιλογή του ενδιάμεσου από μία διαφορετική οπτική γωνία, ως δυναμική επιλογή χαρακτηριστικών από μία ομάδα.

Αναφερθήκαμε στο γεγονός ότι ο ενδιάμεσος, όντας κείμενο μίας κατηγορίας, μπορεί να εξαναγκάσει την ομάδα να δώσει έμφαση στα κείμενα της κατηγορίας του. Από την άλλη πλευρά λόγω των χαρακτηριστικών της κατηγορίας του, για τα οποία δε

μας δίνει πληροφορία, περιορίζει τον αλγόριθμο σε έναν πολύ μικρό υποχώρο των χαρακτηριστικών. Ανάλογα με τα δεδομένα μπορεί να δώσει ακόμα και συγκρίσιμα αποτελέσματα με τις πιο αποτελεσματικές τεχνικές από αυτές που παρουσιάζουμε. Ιδιαίτερα σε εκτελέσεις τυχαίας αρχικοποίησης παρουσιάζει καλύτερη συμπεριφορά από αυτή του αλγεβρικού μέσου λόγω της μεγάλης πόλωσης περιεχομένου.



Σχήμα 6.12. Η κατανομή του ενδιαμέσου (μαύρο χρώμα).

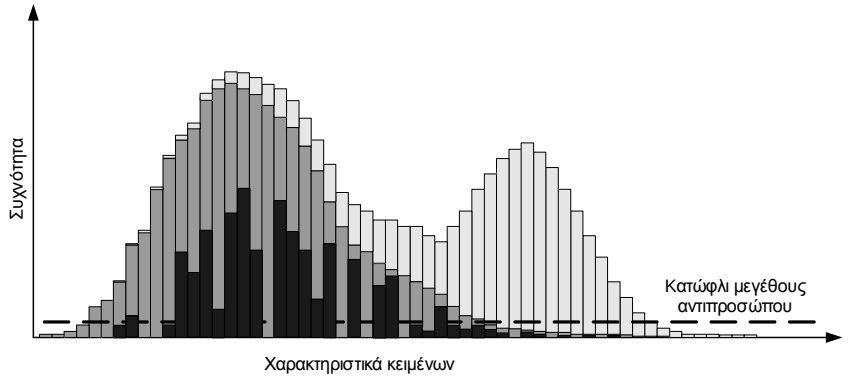
Ο αντιπρόσωπος που δημιουργείται από τον ενδιάμεσο θα πρέπει να φιλτράρεται ελαφρά με την τεχνική κατωφλίωσης μεγέθους αντιπροσώπου διότι θόρυβος υπάρχει και στις αναπαραστάσεις των μεμονωμένων κειμένων.

6.6.2.3. Συνθετικός Αντιπρόσωπος Βασισμένος στους *K-NN* του Ενδιαμέσου

Η ιδέα είναι να εκμεταλλευτούμε ότι ο ενδιάμεσος μπορεί σε κάποιο βαθμό να εντοπίσει την επικρατούσα ομάδα και να αντλήσουμε πληροφορία από τους *K-NN* γείτονες. Λαμβάνοντας τους γείτονες από το εσωτερικό της ομάδας συνδυάζουμε αποτελεσματικά τη βασική οργάνωση και δομή των δεδομένων σε γειτονιές με την ανώτερη περιγραφή των δομών των δεδομένων σε ομάδες. Η γειτονιές των δεδομένων είναι ιδιαίτερα χρήσιμη πηγή άντλησης χαρακτηριστικών ιδιαίτερα σε προβλήματα υψηλής διάστασης.

Με αυτόν τον απλό τρόπο εντοπίζουμε την κατηγορία που επικρατεί σε μία ομάδα, ή επιλέγουμε μία από τις συμμετέχουσες κατηγορίες αν δεν υπάρχει κυρίαρχη, και στη συνέχεια αντλούμε περιεχόμενο από την ομάδα διατηρώντας μία πόλωση προς την

κατηγορία που επιθυμούμε να περιγράφει. Το βασικό φιλτράρισμα κατωφλίου μπορεί να εφαρμοστεί και εδώ για τους λόγους που έχουν αναφερθεί.



Σχήμα 6.13. Η κατανομή του συνθετικού κέντρου βάσει των *K-NN* του ενδιαμέσου (μαύρο χρώμα) σε ομάδα που συγχωνεύει δύο κατηγορίες .

Η τεχνική αυτή αποδεικνύεται πολύ αποτελεσματική. Το θετικό είναι ότι ο ακριβής ορισμός του *K* (μέγεθος γειτονιάς γύρω από τον ενδιάμεσο) δεν παίζει τόσο σημαντικό ρόλο όσο θα υπέθετε κανείς, βάσει εμπειριών από προβλήματα κατηγοριοποίησης. Με λογικές επιλογές των παραμέτρων που εξαρτώνται φυσικά από το μέγεθος της συλλογής κειμένων μπορεί να ξεπεράσει τα προβλήματα που παρουσιάζονται στον αλγεβρικό μέσο. Ιδιαίτερο ενδιαφέρον παρουσιάζουν και οι επιδόσεις του σε τυχαίες αρχικοποιήσεις όπου ξεπερνά τα προβλήματα συσσώρευσης.

6.7. Ο Αλγόριθμος Κ-Συνθετικών Κέντρων

Στη συνέχεια παρουσιάζεται ο αλγόριθμος Κ-Συνθετικών Κέντρων ο οποίος είναι γενίκευση αυτού που παρουσιάστηκε για τον ενδιάμεσο (Σχήμα 6.5). Επίσης, στην ίδια φιλοσοφία μπορεί να βασιστεί και το συνθετικό κέντρο το οποίο δημιουργείται πάνω στον αριθμητικό μέσο με φιλτράρισμα (χωρίς περιορισμό γειτονιάς γύρω από τον ενδιάμεσο). Ο αλγόριθμος κατασκευής του συνθετικού κέντρου που περιγράφεται στο Σχήμα 6.15, τροποποιείται ώστε το σύνολο *Σ* να περιέχει όλα τα κείμενα της ομάδας με αποτέλεσμα να υπολογίζεται ο μέσος στον οποίο εφαρμόζεται φιλτράρισμα.

Επαναλαμβάνουμε πως η βασική διαφορά της τεχνικής αυτής και της παραδοσιακής αλγοριθμικής διαδικασίας του Κ-Μέσων όσων αφορά την

μεγιστοποίηση της αντικειμενικής συνάρτησης, έγκειται στο ότι τα συνθετικά κέντρα δε μπορούν να εγγυηθούν την μονότονη σύγκλιση προς το τοπικό ελάχιστο. Γι' αυτό γίνεται ο έλεγχος για τυχόν μείωση της συνοχής μετά από μία φάση του αλγορίθμου.

K	Ο αριθμός των ομάδων που αναζητούμε
$D=\{d_i\}$, i=1,…,N	Ο αριθμός των κειμένων της συλλογής
DE	Κριτήριο τερματισμού, διαφορά λύσεων δύο διαφορετικών βημάτων
MAX_ITER	Κριτήριο τερματισμού , μέγιστος αριθμός επαναλήψεων
KNN	Το μέγεθος της επιθυμητής γειτονιάς βάσει της οποίας θα δημιουργείται το συνθετικό κέντρο για κάθε ομάδα
FT	Κατώφλι φιλτραρίσματος του συνθετικού κέντρου

Σχήμα 6.14. Είσοδος του αλγόριθμου K-Συνθετικών Κέντρων.

```
    Αλγόριθμος Δημιουργίας Συνθετικού Κέντρού K-Γειτονιάς (Cj)
    {
(1)     Εντόπισε τον ενδιάμεσο dmed της ομάδας Cj
(2)     Δημιούργησε σύνολο Σ = {dmed}

(3)     Αν για το μέγεθος της επιθυμητής γειτονιάς ισχύει KNN > 0
            Εντόπισε τους KNN κοντινότερους γείτονες του ενδιάμεσου κειμένου στην
                ομάδα Cj της τρέχουσας λύσης και εισήγαγέ τους στο σύνολο Σ

(4)     Υπολόγισε τον αριθμητικό μέσο των κειμένων του συνόλου Σ και
            θέσε το αποτέλεσμα ως κέντρο της ομάδας Cj

(5)     Φίλτραρε το κέντρο διατηρώντας μόνο FT όρους
            [διατήρησε και τις ακμές των όρων αυτών στην περίπτωση των γραφημάτων]
    }
```

Σχήμα 6.15. Ψευδοκώδικας αλγορίθμου για τη δημιουργία συνθετικού κέντρου βάσει του ενδιαμέσου μίας ομάδας και των *KNN* κοντινότερων γειτόνων του από τα κείμενα της ίδιας ομάδας.

```
    Αλγόριθμος K-Συνθετικών Κέντρων ομαδικής ενημέρωσης
    {
(1)     Αρχικοποίησε τις K ομάδες δεδομένων αναθέτοντας τα δεδομένα σε αυτές
(2)     Δημιούργησε τα K-Συνθετικά Κέντρα των ομάδων

(3)     Επανέλαβε,
(4)     {   Αποθήκευσε τη τρέχουσα λύση ως Λs-1 του βήματος s-1
(5)         Ανέθεσε κάθε κείμενο di, i=1,…,N στην ομάδα με το κοντινότερο
                συνθετικό κέντρο

(6)         Δημιούργησε τα K-Συνθετικά Κέντρα των ομάδων

(7)         Εκτίμησε την ποιότητα της ομαδοποίησης βάσει της συνοχής
                των ομάδων TCs και των σύνθετων κεντροειδών

(8)         Αν η συνοχή μειώθηκε στο τελευταίο βήμα, TCs < TCs-1
                Τερμάτισε τη διαδικασία και επέστρεψε τη λύση Λs-1
(9)     } μέχρι να μην συμβούν αλλαγές στα κεντροειδή
    }
```

Σχήμα 6.16. Ψευδοκώδικας αλγόριθμου K-Συνθετικών Κέντρων ομαδικής ενημέρωσης.

Παρόλα αυτά να σημειωθεί πως δεν είναι συχνό φαινόμενο να αποκλίνει η διαδικασία. Η συντριπτική πλειοψηφία των λύσεων καταλήγει σε τοπικό ελάχιστο αυξάνοντας μονότονα τη συνοχή των ομάδων. Όσο μάλιστα το συνθετικό κέντρο πλησιάζει τον αριθμητικό μέσο σε περιεχόμενο (δηλ. μεγαλώνουμε τη γειτονιά γύρω από τον ενδιάμεσο, και φιλτράρουμε με μικρό κατώφλι αποκοπής) τόσο το φαινόμενο απόκλισης σπανίζει.

6.8. Ιεραρχική Ομαδοποίηση

Στην ιεραρχική ομαδοποίηση (*hierarchical clustering*) ο αριθμός των ομάδων μεταβάλλεται κατά τις επαναλήψεις των μεθόδων. Εδώ υπάρχουν δύο προσεγγίσεις μεθόδων που διαφοροποιούνται στον τρόπο που μεταβάλλεται ο αριθμός των ομάδων:

- Η πρώτη είναι η ανάλυση από πάνω προς τα κάτω (*top-down analysis*). Αρχικά θεωρείται μία μοναδική ομάδα που περιέχει όλα τα πρότυπα εισόδου, η οποία στη συνέχεια διασπάται σε διαδοχικές επαναλήψεις, ώστε να δημιουργηθούν τελικά οι M επιθυμητές ομάδες. Ο αριθμός των ομάδων αυξάνεται κατά ένα σε κάθε βήμα του αλγορίθμου.
- Εναλλακτικά έχουμε την ανάλυση από κάτω προς τα πάνω (*bottom-up analysis*), όπου ακολουθεί την αντίστροφη διαδικασία. Αρχικά θεωρείται μία ομάδα για κάθε ένα από τα N πρότυπα εισόδου (εν γένει $N > M$). Στη συνέχεια συγχωνεύονται διαδοχικά δύο ομάδες ώστε να δημιουργηθούν τελικά οι M επιθυμητές ομάδες. Ο αριθμός των ομάδων μειώνεται κατά ένα σε κάθε βήμα του αλγόριθμου. Οι μέθοδοι αυτής της κατηγορίας συχνά αναφέρονται ως συσσωρευτικοί (*agglomerative clustering*), υπονοώντας τη συγχώνευση των ομάδων.

Βασικό χαρακτηριστικό των μεθόδων αυτών είναι η απληστία. Για να λάβουν μία απόφαση για το ποια θα είναι η ομάδα η οποία θα διασπαστεί στην πρώτη προσέγγιση και ποιο ζεύγος ομάδων θα συνενωθούν στη δεύτερη, θα πρέπει να υπολογίσουν την βέλτιστη μεταβολή για το βήμα αυτό. Δηλαδή, στην πρώτη περίπτωση θα εντοπίσουν την ομάδα με το μεγαλύτερο εσωτερικό σφάλμα την οποία και θα διασπάσουν, αντιστοίχως στη δεύτερη θα εξεταστούν όλα τα ζεύγη των ομάδων ανά δύο ώστε να βρεθεί το ζεύγος με τη μεγαλύτερη ομοιότητα το οποίο και θα συγχωνευτεί κ.ο.κ.

Οι ιεραρχικές τεχνικές είναι απλές στην εφαρμογή τους, έχουν μελετηθεί αρκετά στο πρόβλημα των κειμένων [9][11] και δίνουν γενικά καλές λύσεις. Τα βασικά μειονεκτήματα είναι η αργή σύγκλιση, το ότι οι αποφάσεις είναι τελικές δηλαδή οι ομάδες που παράγουν είναι εμφωλευμένες και δε δίνουν τη δυνατότητα στα κείμενα να επανακαθορίσουν αυτόνομα τη θέση τους στο σχήμα της λύσης. Αυτό είναι σημαντικό διότι στα πρώτα βήματα του αλγορίθμου συνενώνονται μεμονωμένα κείμενα και αν η πληροφορία των κοντινότερων γειτόνων δεν είναι ποιοτική τότε τα σφάλματα αυτά θα επηρεάσουν ολόκληρη τη διαδικασία ομαδοποίησης.

6.8.1. Αλγόριθμος Συσσωρευτικής Ομαδοποίησης

Όπως φαίνεται και Σχήμα 6.17, σε κάθε επανάληψη του αλγορίθμου οι δυο ομάδες που είναι «περισσότερο όμοιες» σχηματίζουν μία ενιαία ομάδα. Μειώνεται έτσι ο τρέχον αριθμός ομάδων του αλγορίθμου, ο οποίος τερματίζει όταν σχηματιστούν M ομάδες κειμένων. Υπάρχουν τρεις βασικοί τρόποι υπολογισμού της ομοιότητας μεταξύ δύο ομάδων (*inter-cluster similarity*) οι οποίοι και διαχωρίζουν τις μεθόδους:

- Απλού Συνδέσμου (*Single-Link*). Η ομοιότητα υπολογίζεται ως η μέγιστη ομοιότητα μεταξύ ενός ζεύγους κειμένων των ομάδων:

$$cluster\_sim^{(sl)}(C_k, C_l) = \max_{d_i \in C_k, d_j \in C_l} \left\{ sim(d_i, d_j) \right\}.$$

- Πλήρους Συνδέσμου (*Complete-Link*). Η ομοιότητα υπολογίζεται ως η ελάχιστη ομοιότητα μεταξύ ενός ζεύγους κειμένων των ομάδων:

$$cluster\_sim^{(cl)}(C_k, C_l) = \min_{d_i \in C_k, d_j \in C_l} \left\{ sim(d_i, d_j) \right\}.$$

- Μέσου Όρου (*Average-Link, Group Average)*. Η ομοιότητα υπολογίζεται ως η μέση ομοιότητα μεταξύ όλων των ζευγών κειμένων από τις δύο ομάδες:

$$cluster\_sim^{(al)}(C_k, C_l) = \frac{1}{|C_k| \cdot |C_l|} \sum_{d_i \in C_k, d_j \in C_l} sim(d_i, d_j).$$

Στο πρόβλημα των κειμένων, επειδή ο χώρος είναι μεγάλης διάστασης, οι δύο πρώτες προσεγγίσεις είναι ευαίσθητες στον θόρυβο και στα ακραία δεδομένα. Η τρίτη προσέγγιση λαμβάνει υπόψη όλη την ομάδα και δίνει καλύτερα αποτελέσματα [9].

M	Ο αριθμός των ομάδων που αναζητούμε
$D=\{d_i\}$, i=1,…,N	Ο αριθμός των κειμένων της συλλογής

Σχήμα 6.17. Είσοδος αλγόριθμου συσσωρευτικής ομαδοποίησης.

```
    Αλγόριθμος Συσσωρευτικής Ομαδοποίησης
    {
(1)     Θεώρησε N ομάδες οι οποίες περιέχουν ένα κείμενο η κάθε μία
(2)     Υπολόγισε τον πίνακα ομοιότητας μεταξύ ομάδων        /* αρχικά NXN πίνακας */
(3)     Θέσε K = N

(4)     Όσο K > M,
(5)     {   Βρες τις δύο κοντινότερες ομάδες Ci, Cj
(6)         Αν i<=j  Συνένωσε τις Ci, Cj στην Ci,
(7)         Αλλιώς   Συνένωσε τις Ci, Cj στην Cj
(8)         Ενημέρωσε τον πίνακα ομοιότητας για την συγχωνευμένη ομάδα
(9)         Θέσε K = K - 1
        }
    }
```

Σχήμα 6.1. Ψευδοκώδικας αλγόριθμου συσσωρευτικής ομαδοποίησης.

Είναι εύκολο να επεκτείνουμε την εφαρμογή των τεχνικών εύρεσης καλύτερων αντιπροσώπων βάσει του ενδιαμέσου και των *K-NN* γειτόνων του και στον συσσωρευτικό αλγόριθμο. Θεωρώντας τα συνθετικά κέντρα των ομάδων, οι ομοιότητες μεταξύ τους, οι οποίες αποτελούν κριτήριο για την επιλογή της συγχώνευσης, να βασίζονται στην ομοιότητα των αντίστοιχων συνθετικών κέντρων.

ΚΕΦΑΛΑΙΟ 7. ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ

7.1. Συλλογές Εγγράφων

Οι συλλογές εγγράφων είναι σύνολα υπερκειμένων ή απλών κειμένων για τα οποία γνωρίζουμε: α) τον αριθμό και των τύπο των κειμένων, β) τον αριθμό των ομάδων στις οποίες ανήκουν τα κείμενα, γ) την πραγματική κατηγορία κάθε κειμένου η οποία έχει προσδιοριστεί με ανθρώπινα κριτήρια. Αν και πειραματιστήκαμε με συνθετικά δεδομένα που δημιουργήσαμε βάσει του αλγορίθμου που περιγράψαμε στο Κεφάλαιο 2, στόχος μας ήταν να εκτιμήσουμε τις επιδόσεις των τεχνικών που μελετήσαμε σε πραγματικά δεδομένα. Κατά την πειραματική διαδικασία χρησιμοποιήσαμε τις εξής τρεις συλλογές ηλεκτρονικών εγγράφων:

- συλλογή υπερκειμένων *F* (*F-series*), η αρχική συλλογή αποτελείται από 98 HTML κείμενα από 4 κατηγορίες. Αφαιρέσαμε από τη συλλογή 5 κείμενα τα οποία έχουν πολλαπλές ετικέτες. Η τελική μορφή της συλλογής έχει 93 κείμενα: {25, 26, 19, 23}.
- συλλογή υπερκειμένων *J* (*J-series*), η αρχική συλλογή αποτελείται από 185 HTML κείμενα από 10 κατηγορίες. Αφαιρέθηκε ένα διπλότυπο κείμενο, έτσι απέμειναν 184 κείμενα: {19, 19, 19, 19, 20, 16, 19, 19, 18, 16}.

- συλλογή υπερκειμένων *U* (*U-series*), η αρχική συλλογή αποτελείται από 314 HTML κείμενα από 10 κατηγορίες. Αφαιρέσαμε 5 διπλότυπα δημιουργώντας μία συλλογή 309 κειμένων {30, 30, 20, 52, 55, 23, 23, 29, 24, 23}.
- συλλογή φυσικών κειμένων *R* (*R-series*), επιλέχτηκαν 270 κείμενα από 10 κατηγορίες της γνωστής συλλογής *Reuters*-21587 v1.0 [53], {44, 31, 32, 40, 14, 23, 29, 18, 28, 11}. Πρόκειται για σύντομα άρθρα του γνωστού διεθνούς πρακτορείου ειδήσεων. Χαρακτηριστικό της συλλογής αυτής είναι η ύπαρξη ομάδων διαφορετικών μεγεθών και η χαμηλή συνέπεια των γειτονιών. Ο 1-*NN* γείτονας ενός κειμένου ανήκει στην ίδια κατηγορία περίπου κατά 65% έναντι ποσοστών των άλλων συλλογών που ξεπερνούν το 80%.

Τις συλλογές αυτές θα τις χρησιμοποιήσουμε και ως απλά κείμενα αγνοώντας τα προσδιοριστικά της *HTML*. Ο λόγος που επιλέξαμε τα συγκεκριμένα δεδομένα είναι κυρίως ότι είναι μετρίου μεγέθους και ότι έχουν χρησιμοποιηθεί σε σχετικές εργασίες [50][49][47], πέρα από την *R*. Η γνώση της κατηγορίας των κειμένων μέσω των ετικετών που παρέχουν οι συλλογές, θα αποτελέσει την πραγματικότητα (*ground truth*) βάσει της οποίας θα εκτιμήσουμε την απόδοση των μεθόδων.

Στον παρακάτω πίνακα φαίνονται κάποια στατιστικά σχετικά με το μέγεθος κάθε συλλογής. Το μέγεθος ευρετηρίου είναι ο αριθμός των εγγραφών που χρειαζόμαστε στο ευρετήριο για κάθε συλλογή, δηλαδή οι διαφορετικές μη κατευθυνόμενες ακμές και όροι που εμφανίζονται σε αυτή. Το μήκος συλλογής αφορά στο αθροιστικό μέγεθος των μοντέλων, για παράδειγμα μία συλλογή 2 κειμένων, καθένα από τα οποία περιέχει 10 λέξεις και 5 ακμές, έχει μήκος 10+10+5+5 = 30 στοιχεία. Ο αριθμός ακμών και λέξεων ανά κείμενο είναι η μέση τιμή για τα μοντέλα των κειμένων κάθε συλλογής. Να σημειωθεί ότι οι ακμές αναπαριστούν τη σχέση μη κατευθυνόμενης γειτνίασης (r_{undn}) την οποία χρησιμοποιούμε σε όλη την πειραματική μελέτη. Τέλος, αναφέρεται η συνέπεια {1, 3, 5}-*NN* που δείχνει πόσες φορές οι περισσότεροι από τους κοντινότερους γείτονες είναι της ίδιας κατηγορίας με το κείμενο αναφοράς και εκφράζει αρκετά αξιόπιστα τη «δυσκολία» μίας συλλογής.

Συλλογή	Κείμενα	Μέγεθος ευρετηρίου	Μήκος λέξεων συλλογής	Μήκος ακμών συλλογής	Λέξεις ανά κείμενο	Ακμές ανά κείμενο	{1,3,5}-NN
F	93	4925	18774	4883	201.9	52.51	0.97, 0.98, 0.97
J	184	19985	57481	36479	312.4	198.3	0.75, 0.80, 0.82
U	309	12866	51702	20007	167.3	64.8	0.90, 0.94, 0.96
R	270	5195	23811	4811	88.2	17.8	0.63, 0.65, 0.66

Σχήμα 7.1. Στατιστικά για τις τέσσερις συλλογές κειμένων, ύστερα από την προεπεξεργασία (μετ/σμός μορφ/κής ρίζας, αφαίρεση συνηθισμένων λέξεων και λέξεων που εμφανίζονται σε ένα μόνο κείμενο μίας συλλογής).

7.2. Δείκτες Εκτίμησης Ποιότητας Αποτελέσματος

Η αδυναμία γενικού αυστηρού ορισμού της έννοιας «ομάδα» για την ομαδοποίηση δημιουργεί προβλήματα και στην εκτίμηση της ποιότητας του αποτελέσματος. Η τεχνικές εκτίμησης αφορούν ένα παράλληλο με την ομαδοποίηση πρόβλημα και έχουν ιδιαίτερο ερευνητικό ενδιαφέρον. Υπάρχουν πολλοί διαφορετικοί δείκτες οι οποίοι περιγράφουν διαφορετικά χαρακτηριστικά του τελικού διαμερισμού. Ο κυριότερος διαχωρισμός τους είναι σε: α) δείκτες χωρίς επίβλεψη (*unsupervised quality measures*) και β) δείκτες με επίβλεψη (*supervised quality measures*), που αφορά το αν χρησιμοποιούμε πληροφορία από τις πραγματικές κατηγορίες των αντικειμένων. Σε ένα πραγματικό πρόβλημα την πληροφορία αυτή δεν την έχουμε, παρόλα αυτά η συνηθέστερη τακτική είναι να χρησιμοποιούμε γνωστές συλλογές για την πειραματική εκτίμηση της απόδοσης μοντέλων, αλγορίθμων και γενικά συστημάτων *ML*.

7.2.1. Δείκτες Εκτίμησής Ποιότητας Ομαδοποίησης χωρίς Επίβλεψη

7.2.1.1. Μέσο Εσωτερικό Σφάλμα Ομάδων

Υπολογίζουμε την μέση απόσταση του κάθε κειμένου από το κέντρο της ομάδας στην οποία ανήκει χρησιμοποιώντας την συνάρτηση ομοιότητας *sim*(·):

$$MCE_c(c_1,..., c_M) = \frac{1}{N} \cdot \sum_{i=1}^{M} \left[\sum_{d_j \in C_i} \left(1 - sim(d_j, c_i)\right) \right].$$

Ουσιαστικά είναι η συμπληρωματική της αντικειμενικής συνάρτησης (της συνοχής) για τις παραμετρικές μεθόδους της οικογένειας K-Μέσων.

Χρησιμοποιούμε επίσης το μέσο εσωτερικό σφάλμα των ομάδων, *MCE* (*cluster error*), το οποίο υπολογίζεται ως η μέση απόσταση κάθε ζεύγους κειμένων που ανήκουν σε μία ομάδα. Ο λόγος είναι ότι το κέντρο κάθε ομάδας δεν ορίζεται στις συσσωρευτικές μεθόδους, ενώ ακόμα και στον K-Μέσων οι τιμές του MCE_c επηρεάζονται από τον ορισμό του κέντρου (π.χ. αριθμητικός μέσος, συνθετικό κέντρο). Έτσι δε μπορούμε να πάρουμε συγκρίσιμες αριθμητικές εκτιμήσεις για την ποιότητα διαφορετικών εκτελέσεων. Αντίθετα ο *MCE* δίνει συγκρίσιμα αποτελέσματα ανεξάρτητα του κέντρου που επιλέχτηκε για μία εκτέλεση. Ο δείκτης ορίζεται ως εξής:

$$MCE(C_1,..., C_M) = \frac{1}{M} \cdot \sum_{i=1}^{M} \left[\frac{1}{|C_i|} \sum_{d_j \in C_i,\, d_k \in C_i} \left(1 - sim(d_j, d_k)\right) \right].$$

Οι δείκτες αυτοί εξαρτώνται μόνο από τη συνάρτηση ομοιότητας. Οι τιμές τους είναι μεταξύ [0, 1] με τις χαμηλότερες τιμές να υποδηλώνουν καλύτερες λύσεις.

7.2.1.2. Silhouette Coefficient

Ο δείκτης αυτός συνδυάζει τα κριτήρια της εσωτερικής συνοχής και της απόστασης μεταξύ διαφορετικών ομάδων. Ο υπολογισμός βασίζεται σε δύο ποσότητες:

- a_j: η μέση ομοιότητα του αντικειμένου *j* από τα υπόλοιπα αντικείμενα της ομάδας στην οποία ανήκει,
- b_j: η μεγαλύτερη αντίστοιχη ποσότητα για το αντικείμενο *j* και ομάδα διαφορετική της ομάδας στην οποία περιλαμβάνεται το αντικείμενο.

Ο δείκτης για το αντικείμενο *j* ορίζεται ως:

$$Silh_j = \frac{a_j - b_j}{\max\{a_j, b_j\}}.$$

Έτσι, για όλο το αποτέλεσμα η έκφραση είναι:

$$Silh = \frac{1}{N} \cdot \sum_{j=1}^{N} Silh_j.$$

Οι τιμές του δείκτη είναι στο διάστημα [-1, 1] με τις μεγαλύτερες θετικές τιμές να υποδηλώνουν καλύτερο αποτέλεσμα.

7.2.2. Δείκτες Εκτίμησής Ποιότητας Ομαδοποίησης χωρίς Επίβλεψη

7.2.2.1. Rand Index

Ο δείκτης αυτός υπολογίζεται εξετάζοντας κάθε ζεύγος κειμένων της συλλογής, ύστερα από τη διαδικασία ομαδοποίησης. Μετράμε δύο μεγέθη, τις συμφωνίες (*agreements)* και τις ασυμφωνίες (*disagreements)* με τον παρακάτω τρόπο:

- Θεωρούμε μία συμφωνία όταν τα δύο κείμενα ανήκουν:
  - στην ίδια κατηγορία βάσει των ετικετών (*ground truth*), αλλά και στο διαμερισμό που συμπεραίνει η μέθοδος ομαδοποίησης,
  - σε διαφορετικές κατηγορίες βάση των ετικετών και το διαμερισμό που συμπεραίνει η μέθοδος ομαδοποίησης.
- Θεωρούμε μία ασυμφωνία σε κάθε άλλη περίπτωση.

Οι τιμές του δείκτη είναι μεταξύ [0, 1] με βέλτιστο διαχωρισμό όταν $RI = 1$, ενώ ο ακριβής υπολογισμός του δίνεται από τη σχέση:

$$RI = \frac{agreements}{agreements + disagremments}.$$

7.2.2.2. Στατιστικός Δείκτης Ομοιογένειας

Ο στατιστικός δείκτης αυτός (*statistic/purity index*) υπολογίζει το ποσοστό των κειμένων που βρίσκονται σε «σωστή» ομάδα ύστερα από την διαδικασία ομαδοποίησης. Συγκεκριμένα, βρίσκουμε την κυρίαρχη κατηγορία κάθε ομάδας βάσει των ετικετών (*ground truth*). Για παράδειγμα, έστω ομάδα C_i, με στοιχεία από τις κατηγορίες K_A, K_B και κυρίαρχη την K_A, τότε ισχύει:

$$\left|C_i\right| = \left|D_A^{(i)}\right| + \left|D_B^{(i)}\right|, \text{ και βάσει της υπόθεσης } \left|D_A^{(i)}\right| \geq \left|D_B^{(i)}\right|, \text{ έχουμε}$$

$$Purity_i = \frac{\left|D_A^{(i)}\right|}{\left|D_A^{(i)}\right| + \left|D_B^{(i)}\right|},$$

όπου $\left|D_A^{(i)}\right|$ ο αριθμός των κειμένων της κατηγορίας K_A που ανήκουν στην ομάδα του αποτελέσματος C_i (αντίστοιχα για τη K_B). Τελικά, λαμβάνουμε την τιμή του δείκτη για όλο το αποτέλεσμα της ομαδοποίησης από την σχέση:

$$Purity = \frac{1}{M} \cdot \sum_{i=1}^{M} Purity_i \ .$$

Οι τιμές του δείκτη είναι μεταξύ [0, 1] με βέλτιστο διαχωρισμό όταν *SI* = 1. Παρότι γενικά έχουμε *Purity* ≥ 0.5, υπάρχουν περιπτώσεις όπου *Purity* < 0.5. Αυτό συμβαίνει όταν υπάρχουν ομάδες στις οποίες η κυρίαρχη κατηγορία έχει λιγότερα αντικείμενα από το άθροισμα των υπολοίπων αντικειμένων της ομάδας.

7.2.2.3. Εντροπία

Είναι από τους πλέον χρησιμοποιούμενους δείκτες εκτίμησης σε προβλήματα ομαδοποίησης αλλά ειδικότερα και γι' αυτά που αφορούν κείμενα. Συμβολίζοντας με p_{ij} την πιθανότητα ένα αντικείμενο της ομάδας *j* να ανήκει στην κατηγορία *i*, για κάθε ομάδα χωριστά έχουμε:

$$E_j = -\sum_i p_{ij} \cdot \log(p_{ij}),$$

και συνολικά για όλες τις ομάδες του αποτελέσματος:

$$E = \sum_{j=1}^{M} \frac{|C_i|}{N} \cdot E_j.$$

Μια παραδοχή που κάνουμε για να είναι δυνατοί οι υπολογισμοί είναι πως ισχύει: $0 \cdot log(1) = 0$, όταν $p_{ij} = 0$. Οι χαμηλότερες τιμές υποδηλώνουν πιο ομοιογενείς ομάδες, ενώ την ελάχιστη εντροπία λαμβάνουμε όταν έχουμε ένα αντικείμενο ανά ομάδα.

7.2.2.4. F-measure

Το μέτρο αυτό συνδυάζει την ιδέα των *precision* και *recall* από το πεδίο *IR* και χρησιμοποιείται ευρέως για την ποιοτική εκτίμηση των αποτελεσμάτων τεχνικών ομαδοποίησης. Ορίζουμε αρχικά τις ποσότητες για κάθε ζεύγος (κατηγορία *i*, ομάδα *j*):

$$precision\ (\mathrm{i, j}) = n_{ij} / n_j \ ,$$

$$recall\ (\mathrm{i, j}) \quad = n_{ij} / n_i \ ,$$

όπου n_{ij} ο αριθμός των αντικειμένων της ομάδας *j* που ανήκουν στην κατηγορία *i*, n_i ο αριθμός των αντικειμένων της κατηγορίας *i* που και $n_j = |C_j|$. Βάσει αυτών, για την ομάδα *j* και την κατηγορία δεδομένων *i* το *F-measure* ορίζεται ως:

$$F(i,\ j) = \frac{2 \cdot recall(i,\ j) \cdot precision(i,\ j)}{recall(i,\ j) + precision(i,\ j)} \ .$$

Για κάθε κατηγορία *i* υπολογίζεται η μέγιστη τιμή του δείκτη *F*(*i*, *j*), δηλαδή εντοπίζεται η ομάδα του αποτελέσματος η οποία βάσει του δείκτη περιγράφει καλύτερα την κατηγορία *i*. Ο συνολικός δείκτης για το αποτέλεσμα της ομαδοποίησης δίνει τιμές μεταξύ [0, 1] με τη βέλτιστη λύση να περιγράφεται από το $F = 1$. Η ακριβής τιμή του, λαμβάνεται μέσω ζυγισμένου μέσου όρου:

$$F = \sum_i \frac{n_i}{N} \cdot \max\{F(i, j)\}.$$

7.3. Πειραματικά Αποτελέσματα

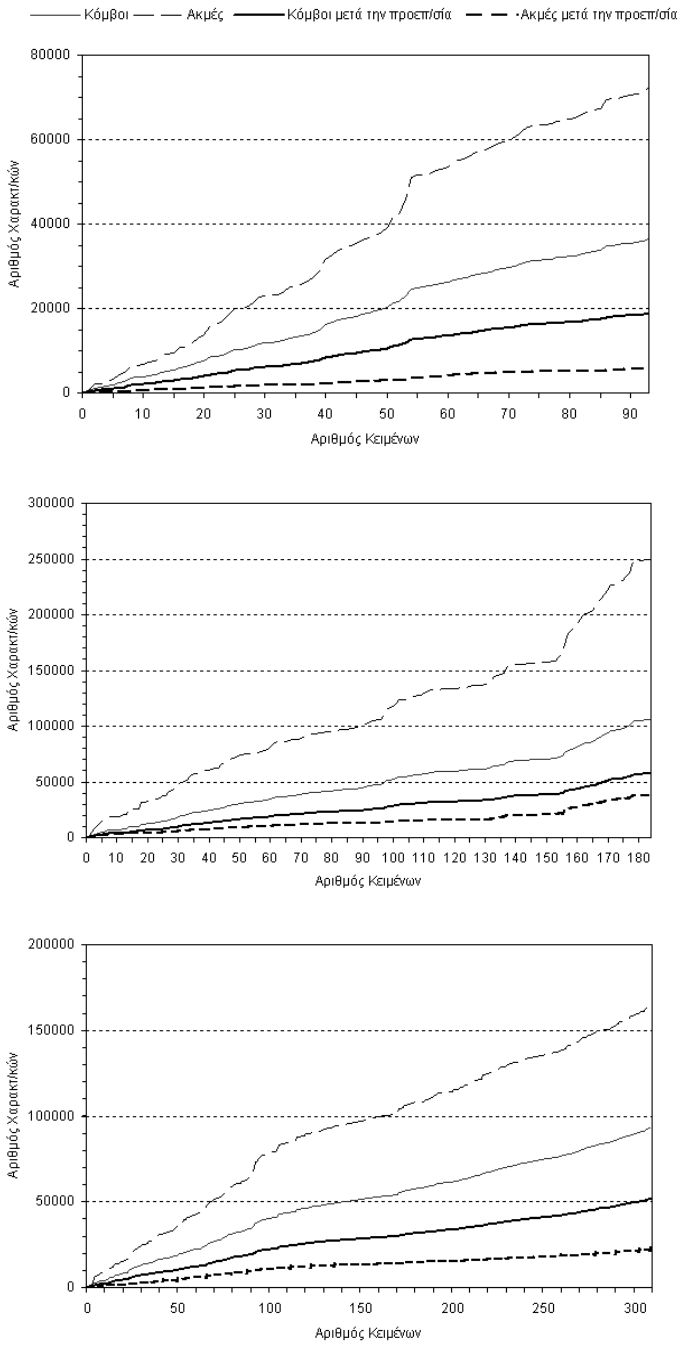
Στην παράγραφο αυτή παρουσιάζονται τα πειραματικά αποτελέσματα της εργασίας. Αρχικά παρουσιάζονται αποτελέσματα σχετικά με τα μοντέλα αναπαράστασης και τις συναρτήσεις ομοιότητας, στη συνέχεια θα αναλύσουμε το φιλτράρισμα βάσει της *K*-γειτονιάς ως τεχνική μείωσης διάστασης και θα κλείσουμε το Κεφάλαιο με συγκριτικά πειραματικά δεδομένα κλασικών αλγόριθμων ομαδοποίησης και του αλγόριθμου K-Συνθετικών Κέντρων.

Rand	Δείκτης Rand.
Purity	Δείκτης ομοιογένειας.
Silh.	Δείκτης ποιότητας Silhouette Coefficient.
F	Δείκτης ποιότητας αποτελέσματος F-measure.
E	Δείκτης ποιότητας αποτελέσματος Entropy.
MCE	Δείκτης ποιότητας αποτελέσματος Mean Cluster Error.
Global-KM	Γενικευμένος K-Μέσων.
HAC	Ιεραρχικός συσσωρευτικός αλγόριθμος.
mean	Αριθμητικός μέσος.
K-NN	Η γειτονιά των *K* κοντινότερων στοιχείων ενός αντικειμένου.
B	Boolean βάρη (αδύνατη η χρήση του IDF)
Freq ή F	Συχνοτικά βάρη χρήσει και του IDF όρου.
Sign ή S	Βάρη σημαντικότητας χρήσει και του IDF όρου.
G	Γραφοθεωρητική συνάρτηση ομοιότητας.
C	Συνημιτονοειδής συνάρτηση ομοιότητας.

Σχήμα 7.2. Συμβολισμοί για την αναφορά αποτελεσμάτων.

7.3.1. Επιβάρυνση Μοντέλων Αναπαράστασης από τις Ακμές των Όρων

Παρουσιάζουμε αρχικά τις παρατηρήσεις σχετικά με το μέγεθος των συλλογών, πριν και αμέσως μετά το στάδιο της βασικής προεπεξεργασίας.



Σχήμα 7.3. Η αύξηση του μήκους των συλλογών ως προς τον αριθμό των κειμένων εισόδου, πριν και μετά το βασικό στάδιο προεπεξεργασίας τους. Από επάνω: συλλογές *F*, *J*, *U*.

Αφού εφαρμόσουμε το μετασχηματισμό μορφολογικής ρίζας (*stemming*), αφαιρέσουμε τις συνηθισμένες λέξεις (*stopwords*) και τις λέξεις που εμφανίζονται σε ένα μόνο κείμενο μίας συλλογής, βλέπουμε πως ο όγκος των δεδομένων και η διάσταση των δεδομένων εισόδου περιορίζονται σημαντικά. Τα παρακάτω σχήματα δείχνουν πόσο αυξάνεται το μήκος κάθε συλλογής (το άθροισμα του μεγέθους του μοντέλου κάθε κειμένου) ως προς τον αριθμό των κειμένων εισόδου.

Οι ακμές, ενώ αρχικά φαίνεται να είναι πολλαπλάσιες των διαφορετικών λέξεων, μετά την προεπεξεργασία καθίστανται κατά πολύ λιγότερες από τους όρους. Από τη μία λοιπόν αντιλαμβάνεται κανείς πως η τάξη πολυπλοκότητας του προβλήματος δεν αλλάζει με την εισαγωγή των γραφικών αναπαραστάσεων και των ακμών μη κατευθυνόμενης γειτνίασης. Από την άλλη, το φαινόμενο αυτό δείχνει και την αδυναμία μηχανικής εξόρυξης σχέσεων ανάμεσα σε όρους των κειμένων. Αναγνωρίζοντας απλά τη διαδοχή των λέξεων ως σχέση τελικά αποκομίζουμε πολλή λιγότερη πληροφορία από αυτή που ενυπάρχει στα κείμενα.

7.3.2. Αναπαράσταση Χρήσει Γραφικών Μοντέλων και Μέτρα Ομοιότητας

Η εισαγωγή των γραφικών αναπαραστάσεων για τα δεδομένα κειμένων εξετάζεται πειραματικά, χρησιμοποιώντας μόνο τις συναρτήσεις ομοιότητας. Η επιλογή αυτή έχει να κάνει με το ότι για να εξάγουμε συμπεράσματα για τα μοντέλα αναπαράστασης είναι προτιμότερο να περιορίσουμε τον αριθμό των παραμέτρων των πειραμάτων. Οι ιδιαιτερότητες που παρουσιάζουν οι τεχνικές ομαδοποίησης δε μας επιτρέπουν να εξετάσουμε εύκολα την επίδραση των ακμών στο επίπεδο της ομοιότητας των δεδομένων, παρά μόνο στο επίπεδο των λύσεων της ομαδοποίησης. Όμως, όπως θα διαπιστώσουμε και στη συνέχεια, οι τεχνικές ομαδοποίησης επηρεάζονται με διαφορετικό τρόπο από την ύπαρξη ακμών ανάλογα με τη δυσκολία του προβλήματος διαμέρισης και των ειδικών ρυθμίσεων των παραμέτρων μίας εκτέλεσης.

Στη βιβλιογραφία η εκτίμηση καταλληλότητας μοντέλων αναπαράστασης, επιλογής χαρακτηριστικών ή ανάθεσης βαρών σε αυτά γίνεται με τεχνικές επίβλεψης. Στην εργασία αυτή, πέραν αυτών της ομαδοποίησης, θα χρησιμοποιήσουμε και τεχνικές όπως η κατηγοριοποίηση *K-NN* για να μελετήσουμε την καταλληλότητα του γραφικού μοντέλου αναπαράστασης και τη χρησιμότητα των ακμών, των μέτρων ομοιότητας, των σχημάτων ανάθεσης βαρών στα χαρακτηριστικά και των τεχνικών

φιλτραρίσματος. Υλοποιήθηκε η τεχνική *unweighted leave-one-out* χωρίς βάρη, όπου θεωρούμε πως γνωρίζουμε την κατηγορία όλων των δεδομένων πλην αυτού που εξετάζεται κάθε φορά, το οποίο ταξινομείται στην πλειοψηφούσα κατηγορία των *K-NN* γειτόνων. Τα συγκεκριμένα πειράματα κατηγοριοποίησης διεξήχθησαν μόνο με την συνάρτηση συνημιτονοειδούς ομοιότητας με συχνοτικά βάρη (*Freq*).

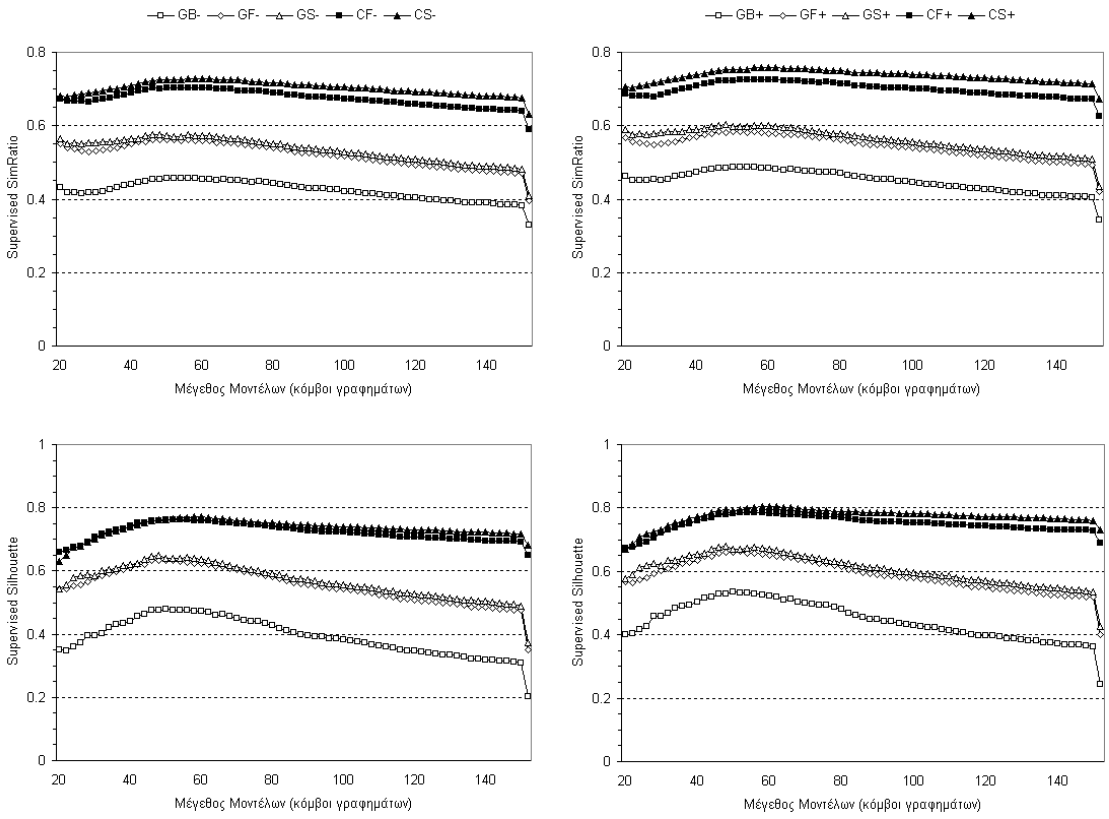
Θεωρώντας ότι έχουμε τον ιδανικό διαχωρισμό μιας συλλογής, κάθε ομάδα περιέχει μία κατηγορία κειμένων, μπορούμε να ορίσουμε τις εξής δύο ποσότητες:

- ως *Supervised SimRatio* (*SSR*) ορίζουμε το μέσο όρο του λόγου της ομοιότητας κάθε κειμένου της συλλογής με τα κείμενα της κατηγορίας του, προς την ομοιότητα που παρουσιάζει με όλα τα κείμενα της συλλογής (συμπεριλαμβάνονται τα κείμενα της δικής του κατηγορίας).

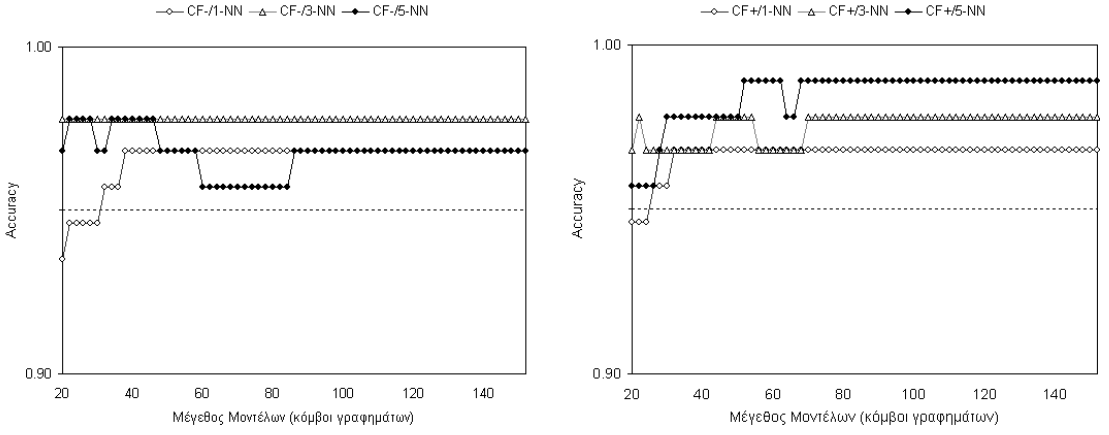
  Το μέγεθος αυτό εκφράζει το πόσο έντονη είναι ή «έλξη» που δέχεται κάθε κείμενο από την πραγματική του κατηγορία σε σχέση με τη θορυβώδη έλξη από τις άλλες κατηγορίες. Οι μεγαλύτερες τιμές του δείκτη υπονοούν ισχυρότερη ομοιότητα με τα κείμενα που επιθυμούμε να ενταχθούν στην ίδια ομάδα από έναν αλγόριθμο ομαδοποίησης.
- ως *Supervised Silhouette Coefficient* (*SSC*) ορίζουμε το *Silhouette* μέτρο που αφορά τον ιδανικό διαμερισμό των δεδομένων. Επειδή εμπλέκει δύο ομάδες δεδομένων, αυτή στην οποία ανήκει ένα κείμενο (η οποία αποτελεί μία ομάδα της ιδανικής διαμέρισης) και τη δεύτερη κοντινότερη ομάδας, εκφράζει κατά κάποιο τρόπο την ευστάθεια της λύσης. Οι μεγάλες τιμές του δείκτη δείχνουν ότι τα κείμενα της συλλογής θα κατέληγαν δυσκολότερα σε κάποια ομάδα με κείμενα που δεν ανήκουν στην ίδια κατηγορία.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα γραφοθεωρητική (*G*) και την συνημιτονοειδή ομοιότητας (*C*). Τα σχήματα βαρών που δοκιμάζονται είναι: το λογικό (*Boolean Weighting – B*), το συχνοτικό (*Frequency Weighting – F*) και τα βάρη βάσει των στοιχείων *HTML* (*Significance Weighting – S*). Στα δύο τελευταία έχει χρησιμοποιηθεί ο όρος *IDF*. Συνεπώς ο συμβολισμός *CF* αφορά τη συνημιτονοειδή ομοιότητα με συχνοτικά *TF·IDF* βάρη, αντίστοιχα *GB* τη γραφοθεωρητική συνάρτηση με λογικά βάρη. Η δεξιά στήλη παρουσιάζει τα αντίστοιχα αποτελέσματα ενσωματώνοντας τις ακμές (‘+’, αντίστοιχα ‘−‘ όταν αυτές αγνοούνται) στα μοντέλα

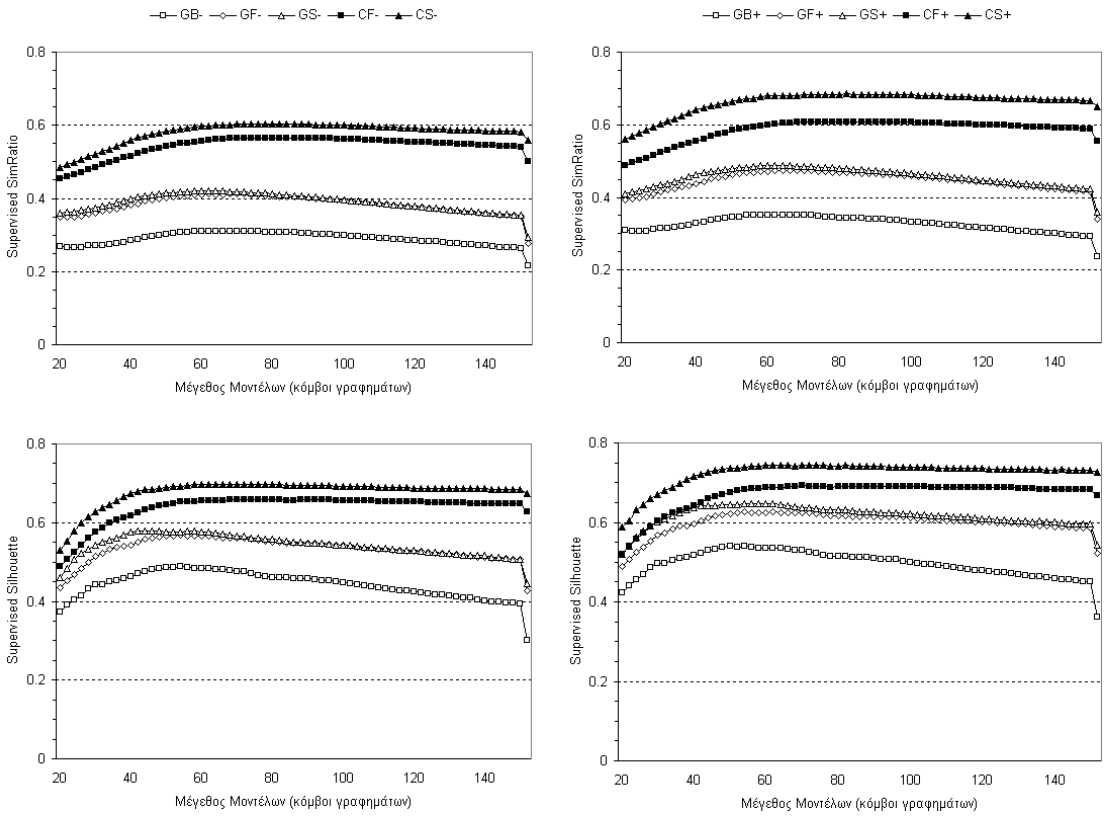
αναπαράστασης. Στον άξονα των *x* έχουμε τον αριθμό των λέξεων του μοντέλου κάθε κειμένου (20 έως 150 όρους). Το τελευταίο πείραμα αφορά τη μη αποκοπή χαρακτηριστικών και εμφανίζεται συνήθως ως απότομη πτώση στις παραστάσεις.



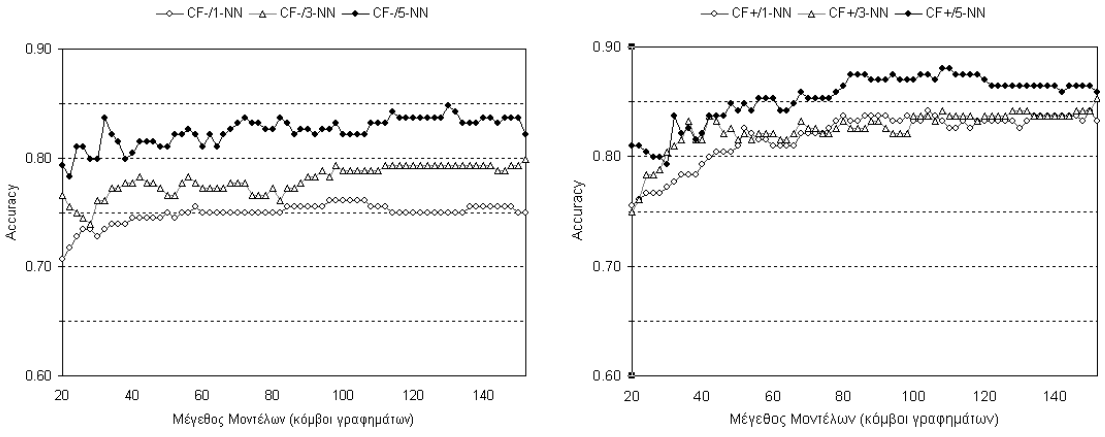
Σχήμα 7.4. Οι δείκτες *SSR* και *SSE*, συλλογή *F*.



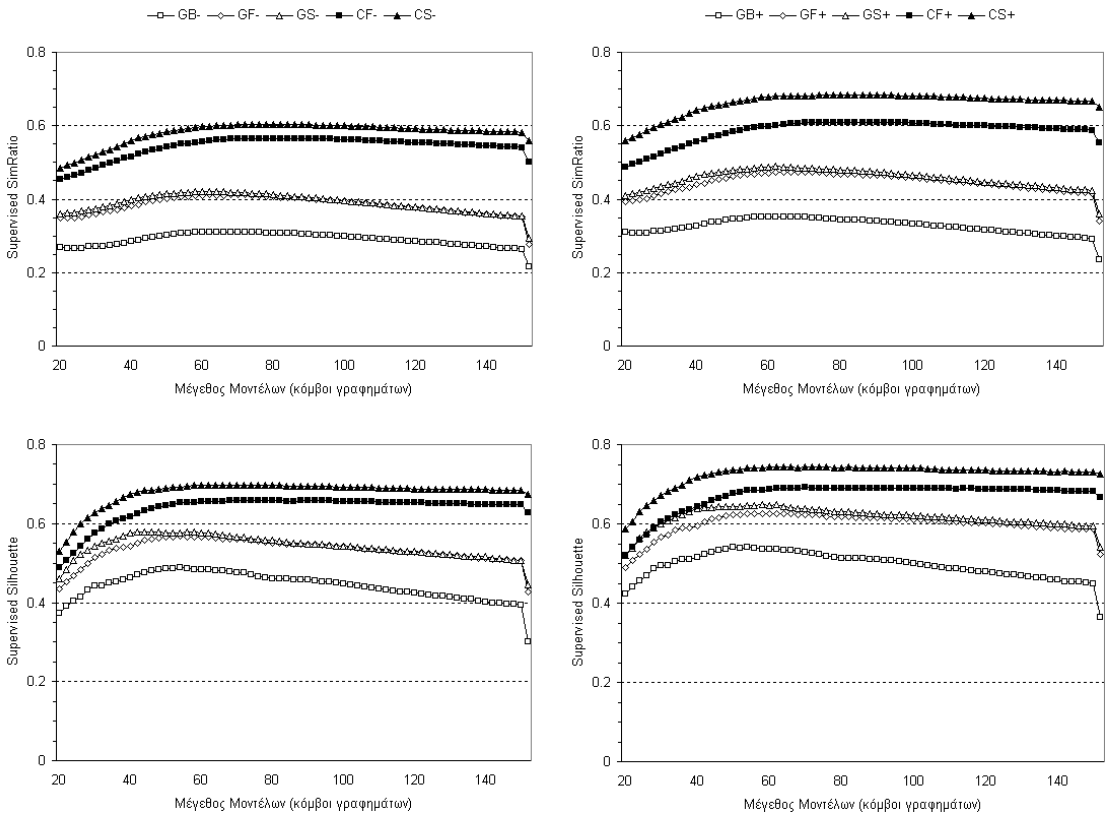
Σχήμα 7.5. Κατηγοριοποίηση *K-NN* με τιμές *K* = {1, 3, 5}, συλλογή *F*.



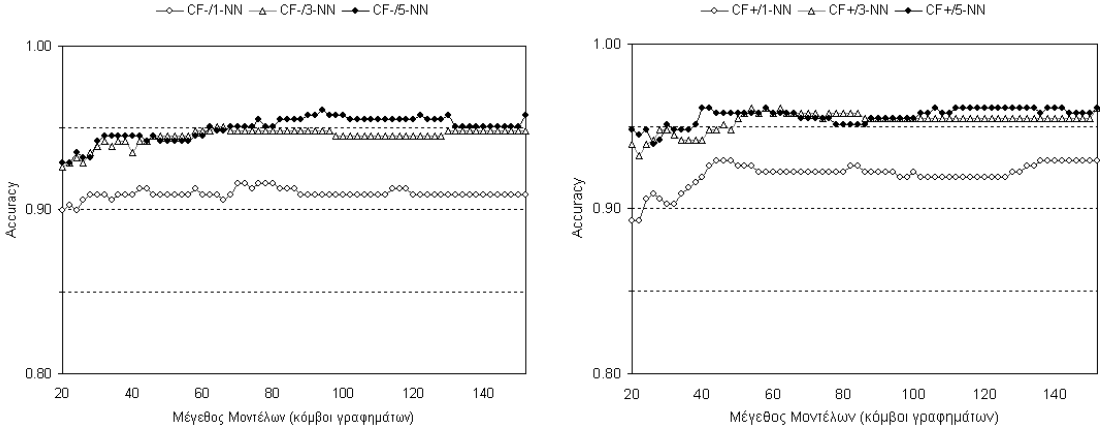
Σχήμα 7.6. Οι δείκτες *SSR* και *SSE*, συλλογή *J*.



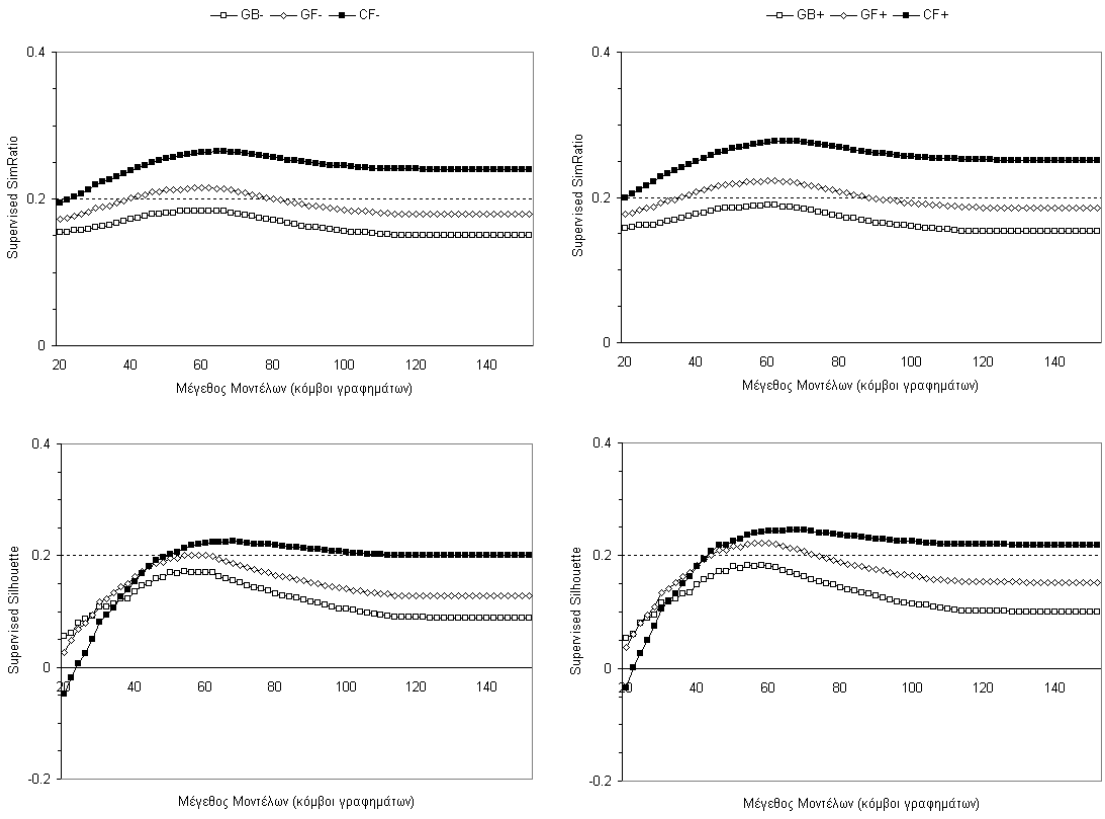
Σχήμα 7.7. Κατηγοριοποίηση *K-NN* με τιμές *K* = {1, 3, 5}, συλλογή *J*.



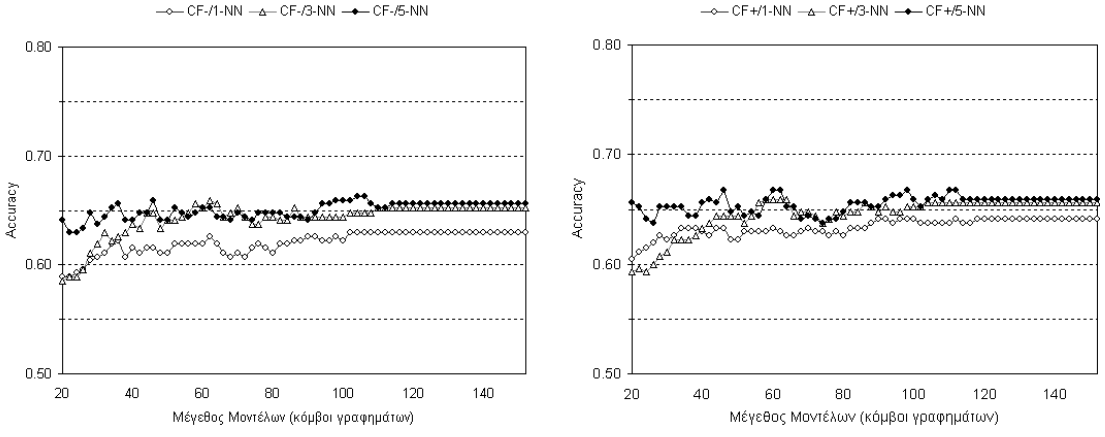
Σχήμα 7.8. Οι δείκτες *SSR* και *SSE*, συλλογή *U*.



Σχήμα 7.9. Κατηγοριοποίηση *K-NN* με τιμές *K* = {1, 3, 5}, συλλογή *U*.



Σχήμα 7.10. Οι δείκτες *SSR* και *SSE*, συλλογή *R*.



Σχήμα 7.11. Κατηγοριοποίηση *K-NN* με τιμές *K* = {1, 3, 5}, συλλογή *R*.

Στο σημείο αυτό μπορούμε να αναφέρουμε μία σειρά παρατηρήσεων. Ως αναφορά τις ακμές, η σύγκριση της εκδοχής κάθε πειράματος ενσωματώνοντας ακμές στα μοντέλα αναπαράστασης (δεξιά στήλη) και χωρίς αυτές (αριστερή στήλη) δείχνει πως ενισχύουν την ομοιότητα των κειμένων με τα κείμενα της ίδιας κατηγορίας. Με άλλα λόγια ενισχύουν τις δομές των δεδομένων σε επίπεδο σχέσεων ομοιότητας, τις οποίες

θα αναζητήσουμε κατά την ομαδοποίηση. Μια βελτίωση της τάξης του 10%, και πλέον που παρατηρείται σε πολλές περιπτώσεις είναι ιδιαίτερα σημαντική για την ποιότητα της πληροφορίας που αντλείται από τη δομή των δεδομένων. Άλλωστε το ότι τα μέτρα εκτίμησης *SSR* και *SSC* εκφράζονται ως μέσες τιμές, σημαίνει πως η βελτίωση αυτή είναι κατά πολύ μεγαλύτερη σε διάφορα υποσύνολα των δεδομένων.

Για τις συναρτήσεις ομοιότητας οι παρατηρήσεις είναι επίσης ξεκάθαρες. Η συνημιτονοειδής ομοιότητα υπερτερεί αρκετά της γραφοθεωρητικής συνάρτησης ομοιότητας, με ακμές ή χωρίς, και με όλα τα σχήματα ανάθεσης βαρών. Η μεταξύ τους διαφορά είναι πολύ μεγάλη στις περισσότερες περιπτώσεις, ενώ το βασικότερο ίσως είναι ότι διαφέρει σε συμπεριφορά όταν αυξάνονται αρκετά τα χαρακτηριστικά των κειμένων. Στη συνέχεια της πειραματικής μελέτης θα παραλείψουμε πειράματα που διεξήχθησαν χρησιμοποιώντας τη γραφοθεωρητική συνάρτηση ομοιότητας, αναφέροντας πως τα αποτελέσματά αυτά ήταν από ελαφρώς έως αρκετά κατώτερα σε ποιότητα αυτών που παρουσιάζονται για τη συνημιτονοειδή ομοιότητα.

Η αποτυχία αυτή, σε μεγάλο βαθμό ερμηνεύεται από την ανάλυση του Κεφαλαίου 4 περί ακαταλληλότητας του προσθετικού υπολογισμού των επιμέρους ομοιοτήτων. Ουσιαστικά εδώ επιβεβαιώνεται και πειραματικά ότι όταν αυξάνονται τα χαρακτηριστικά των μοντέλων η γραφοθεωρητική συνάρτηση δε μπορεί να ξεχωρίσει τα σημαντικά από τα ασήμαντα χαρακτηριστικά. Έτσι, παρατηρείται μία έντονη πτώση των δεικτών ομοιότητας με την αύξηση των μοντέλων. Παρόλα αυτά η εισαγωγή βαρών μπορεί να βοηθήσει αρκετά και τη συνάρτηση αυτή. Το μεγαλύτερο πρόβλημα παρουσιάζει η *GB* που δεν εισάγει βάρη στα χαρακτηριστικά, όπως άλλωστε αναμέναμε.

Μια ακόμα σημαντική παρατήρηση είναι πως με την αύξηση των μοντέλων, αρχικά παρατηρούμε και την αύξηση των δεικτών για όλα τα μέτρα ομοιότητας. Στη συνέχεια οι δείκτες παίρνουν τις μέγιστες τιμές τους, ενώ πλησιάζοντας προς την αναπαράσταση όλων των χαρακτηριστικών κάθε κειμένου οι δείκτες βυθίζονται (τελευταίο πείραμα κάθε γραφικής παράστασης). Αυτό υπονοεί πως, βάσει των συναρτήσεων ομοιότητας, υπάρχει ένα υποσύνολο χαρακτηριστικών στο οποίο οι κατηγορίες γίνονται περισσότερο διαχωρίσιμα. Παρότι χάνεται πληροφορία είναι δυνατό να πετύχουμε καλύτερα αποτελέσματα ομαδοποίησης σε αυτές τις περιπτώσεις. Η διαπίστωση αυτή

είναι χρήσιμη αλλά στην πράξη είναι δύσκολο να εντοπίσουμε κατάλληλα σημεία αποκοπής πληροφορίας, όταν το πρόβλημά μας είναι άνευ επίβλεψης.

Το σχήμα ανάθεσης βαρών σημαντικότητας *SW* (εκτός της *R* συλλογής που περιέχει φυσικά κείμενα) φαίνεται να ενισχύει ακόμα περισσότερο τους δείκτες. Στη συνέχεια θα χρησιμοποιήσουμε κυρίως τα συχνοτικά βάρη ώστε να εξάγουμε συμπεράσματα τα οποία δε θα εξαρτώνται από το συγκεκριμένο σχήμα τιμών.

Τα πειράματα κατηγοριοποίησης με επίβλεψη παρουσιάζουν τα ίδια αποτελέσματα από μία άλλη οπτική, εξετάζοντας κατά πόσο ενισχύεται η ποιότητα των *K*-γειτονιών με την εισαγωγή ακμών. Το συμπέρασμα είναι πως αν υπάρχουν ομάδες στα δεδομένα, οι ακμές μπορούν να βοηθήσουν αρκετά στην ανάδειξη ποιοτικότερων γειτονιών. Ακόμα όμως και στην περίπτωση όπου τα δεδομένα έχουν ασθενείς δομές, όπως συμβαίνει στην συλλογή *R*, δεν προκύπτουν προβλήματα ακατάλληλης αναπαράστασης των δεδομένων.

7.3.3. Σημαντικότητα Ακμών για την Ομαδοποίηση – Συντελεστής Μίξης

Η ακριβής συνεισφορά των βαρών των ακμών στον υπολογισμό της ομοιότητας των κειμένων παίζει σημαντικό ρόλο στην ποιότητα των λύσεων που λαμβάνουμε από έναν αλγόριθμο ομαδοποίησης. Η σειρά πειραμάτων που παρουσιάζεται αφορά τις πειραματικές παρατηρήσεις πάνω στο συντελεστή μίξης (*blend factor*).

Τα αναλυτικά δεδομένα αφορούν τα συχνοτικά βάρη (*Freq*), και περιγράφουν τη συμπεριφορά του αλγορίθμου με τιμές μίξης [0, 1] (0: λαμβάνονται υπόψη μόνο οι λέξεις των κειμένων για τον υπολογισμό ομοιότητας, 1: αντίστοιχα για τις ακμές). Στα δεξιά, τα διαγράμματα στηλών παρουσιάζουν τις τιμές των αντίστοιχων δεικτών εκτίμησης της ποιότητας της ομαδοποίησης του ενοποιημένου μοντέλου χαρακτηριστικών, χρησιμοποιώντας δύο σχήματα βαρών. Σε αυτό το μοντέλο ο συντελεστής μίξης δεν καθορίζεται από το χρήστη, αφήνοντας θα λέγαμε τα δεδομένα να «αυτοκαθορίσουν» τη σχέση περιεχομένου μεταξύ λέξεων και σχέσεων. Τελικά, κατά τον υπολογισμό ομοιότητας χειριζόμαστε ακμές και όρους με τον ίδιο ακριβώς τρόπο. Το ενοποιημένο μοντέλο θα χρησιμοποιηθεί στις επόμενες παραγράφους της πειραματικής μελέτης.

Η ανάλυση αυτή θα περιγράψει καλύτερα τον ακριβή ρόλο που παίζουν οι ακμές των γραφικών μοντέλων. Επίσης, μέσω της διαδικασίας αυτής προκύπτει ως

δυνατότητα η επιλογή μίας κατάλληλης τιμής για τον συντελεστή μίξης. Πριν προχωρήσουμε στα σχετικά αποτελέσματα θα συζητήσουμε τη δυνατότητα αυτή.

Για να επιλέξουμε μία τιμή για τον συντελεστή μίξης (*bf*) σε ένα πλαίσιο επίλυσης από το οποίο απουσιάζει η γνώση τω πραγματικών κατηγοριών των κειμένων, θα πρέπει να επιλέξουμε τα κατάλληλα κριτήρια τα οποία θα μας υποδείξουν μία καλή λύση ομαδοποίησης. Τα κριτήρια χωρίς επίβλεψη που χρησιμοποιήσαμε στην εργασία είναι: ο *Silhouette Coefficient*, και το μέσο σφάλμα *MCE* της λύσης (ή αντίστοιχα η αντικειμενική συνάρτηση), κανένα όμως από τα δύο δεν είναι κατάλληλο να μας υποδείξει μία καλή τιμή για τον συντελεστή μίξης. Η αντικειμενική συνάρτηση MCE_c που αφορά το κεντροειδές έχει ένα επιπλέον πρόβλημα, ότι δε θα μπορούσε να οριστεί στους ιεραρχικούς που δε χρησιμοποιούν κεντροειδές.

Την ακαταλληλότητα της απευθείας χρήσης των μέτρων αυτών μπορούμε να τη δικαιολογήσουμε αναφέροντας ένα απλό θεωρητικό παράδειγμα. Όταν ο $bf = 0$ τότε όλα τα διανύσματα των κειμένων (ας θεωρήσουμε τη διανυσματική αναπαράσταση) έχουν μηδενική μεταβλητότητα στο υποσύνολο των χαρακτηριστικών που ορίζουν οι ακμές της συλλογής. Αυξάνοντας τον *bf* ουσιαστικά αυξάνουμε τη μεταβλητότητα αυτή, με αποτέλεσμα να «αραιώνουν» (αύξηση του χώρου του προβλήματος). Τελικά, όταν $bf = 1$ ο χώρος του προβλήματος ορίζεται βάση μόνο των ακμών αγνοώντας τους όρους. Η μεταβολή αυτή δημιουργεί ένα σημαντικό πρόβλημα σε όλους τους δείκτες χωρίς επίβλεψη, το ότι δε μπορεί να γίνει απευθείας σύγκριση των τιμών που αφορούν δύο πειράματα με διαφορετική ρύθμιση του *bf* διότι ουσιαστικά έχουμε μεταβολές στα μοντέλα των δεδομένων. Αυτό παρατηρείται σε πολλά από τα πειράματα που θα παρουσιάσουμε στη συνέχεια. Είναι σύνηθες αυξάνοντας την επιρροή των ακμών να αυξάνονται οι δείκτες *MCE* και *Silhouette Coefficient* εκφράζοντας μία αντίφαση.

Ο τρόπος με τον οποίο καταφέραμε να ορίσουμε έναν κατάλληλο δείκτη στην εργασία περιγράφεται συνοπτικά. Αναζητώντας ένα αντικειμενικό κριτήριο *Q* μεταξύ εκτελέσεων διαφορετικών τιμών του *bf*, μπορεί να παρατηρήσει κάποιος πως αυτό μπορεί να προκύψει από την εκτίμηση ενός δείκτη χωρίς επίβλεψη αγνοώντας κάποια από χαρακτηριστικά των δεδομένων.

Επιλέγοντας τον δείκτη *MCE* ως κριτήριο *Q*, η διαδικασία περιγράφεται από τα εξής βήματα:

1) Διακριτοποίησε το διάστημα [0, 1] για τις δυνατές τιμές του *bf*.
2) Λύσε το πρόβλημα ομαδοποίησης για *bf* = 0, μόνο για τους όρους των κειμένων και υπολόγισε τον δείκτη Q_0.
3) Για κάθε διακριτή τιμή του *bf* στο [0, 1],
   α. Λύσε το πρόβλημα ομαδοποίησης.
   β. Αφαίρεσε τις ακμές από τα μοντέλα (ουσιαστικά θέσε *bf* = 0).
   γ. Εκτίμησε τον δείκτη $Q_{bf}^{(t)}$, βάσει μόνο των λέξεων.
4) Ανέφερε τη λύση (τιμή *bf*) που παρουσιάζει ελάχιστο $Q_{bf}^{(t)}$.

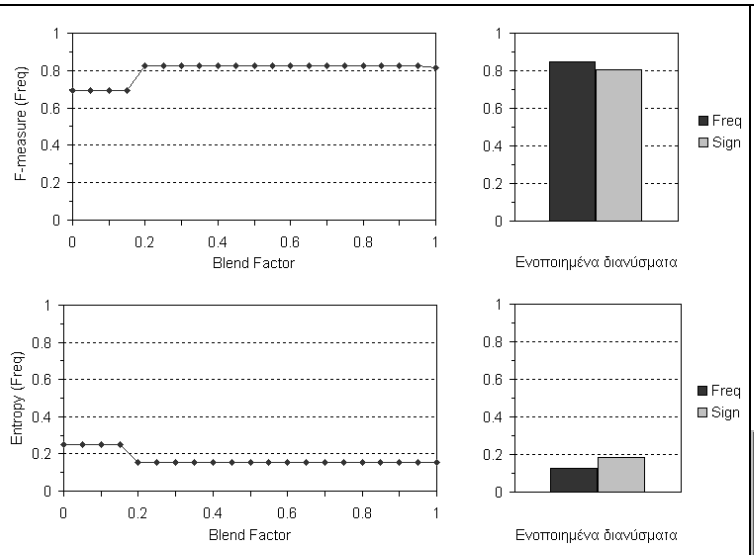
Με άλλα λόγια, όποια και να η τιμή του *bf* κατά την ομαδοποίηση, η εκτίμηση Q_{bf} είναι αντικειμενική γιατί θεωρεί κάθε φορά μόνο τους όρους των κειμένων. Έτσι, ορίζεται μία διαδικασία επικύρωσης πάνω στη λύση (*validation procedure*). Η διαδικασία αυτή θα μπορούσε έχει και την αντίστροφη φορά, ξεκινώντας από την *bf* = 1, απαιτώντας μια λύση που να θεωρείται κατάλληλη από τις ακμές των δεδομένων. Ο συμβιβασμός που κάναμε για να δοκιμάσουμε και τη συνδυασμένη εκδοχή στην πράξη, είναι να συνδυάσουμε και τις δύο αποφάσεις με ίση βαρύτητα. Αν $Q_{\max}^{(t)}$ και $Q_{\max}^{(e)}$ οι δύο μέγιστες τιμές των κριτηρίων λέξεων και ακμών (μέγιστα σφάλματα), ο συντελεστής bf^* εκτιμάται ως:

$$bf^* = \arg\max_{\forall bf} \left\{ 1 - \frac{1}{2} \cdot \left(\frac{Q_{bf}^{(t)}}{Q_{\max}^{(t)}} + \frac{Q_{bf}^{(e)}}{Q_{\max}^{(e)}} \right) \right\}.$$

Με τις ευρετικές τεχνικές αυτές δεν παίρνουμε τη βέλτιστη τιμή για τον συντελεστή μίξης, επιλέγεται μία λύση η οποία τις περισσότερες φορές είναι από τις καλύτερες που προκύπτουν από τις διακριτές τιμές του *bf*. Αν θέλουμε να εμπιστευτούμε τις ακμές τότε εφαρμόζουμε το συνδυαστικό κριτήριο. Η αποτελεσματικότητα του εξαρτάται από την ποιότητα της πληροφορίας που έχουν οι ακμές. Αν για τα δεδομένα είναι παράγοντας που δε βοηθάει τότε ίσως πάρουμε λύσεις χειρότερες και από την Q_0.

Στους παρακάτω αναλυτικούς πίνακες με *Q* συμβολίζεται το συνδυαστικό κριτήριο που μεγιστοποιείται ενώ $Q^{(t)}$ και $Q^{(e)}$ τα δύο επιμέρους κριτήρια που ελαχιστοποιούνται. Σε κατάλληλες στήλες σημειώνονται με '•' οι αποφάσεις τους.

7.3.3.1. Συσσωρευτικός Αλγόριθμος Ιεραρχικής Ομαδοποίησης



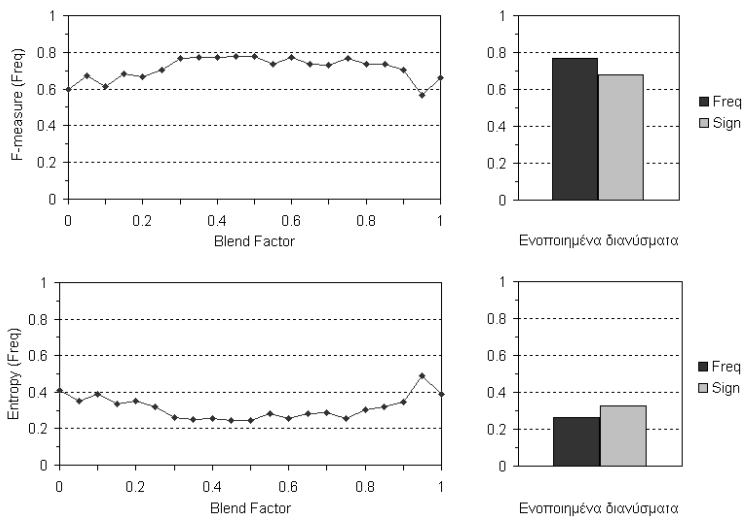
Σχήμα 7.12. Πείραμα μίξης της σημαντικότητας των ακμών στον υπολογισμό των ομοιοτήτων. Ομαδοποίηση με *HAC*, συλλογή *F*.

Βάρη	Blend Factor	Rand	Purity	Silh.	F	E
Freq	-	0.892	0.860	0.671	0.845	0.127
Sign	-	0.846	0.742	0.761	0.804	0.185

Σχήμα 7.13. Ενοποιημένο μοντέλο, ομαδοποίηση με *HAC*, συλλογή *F*.

		Δείκτες ποιότητας ομαδοποίησης						Δείκτες επιλογής *bf*					
Βάρη	Blend Factor	Rand	Purity	Silh.	F	E	MCE	Q	max(Q)	$Q_{bf}^{(t)}$	min($Q_{bf}^{(t)}$)	$Q_{bf}^{(e)}$	min($Q_{bf}^{(t)}$)
Freq	0.00	0.781	0.667	0.622	0.693	0.252	0.867	0.00000		0.86717		0.91681	
	0.05	0.781	0.667	0.627	0.693	0.252	0.8697	0.00000		0.86717		0.91681	
	0.10	0.781	0.667	0.631	0.693	0.252	0.872	0.00000		0.86717		0.91681	
	0.15	0.781	0.667	0.636	0.693	0.252	0.8746	0.00000		0.86717		0.91681	
	0.20	0.866	0.763	0.695	0.826	0.152	0.8699	0.00826	•	0.86006	•	0.90919	•
	0.25	0.866	0.763	0.700	0.826	0.152	0.872	0.00826		0.86006		0.90919	
	0.30	0.866	0.763	0.705	0.826	0.152	0.8748	0.00826		0.86006		0.90919	
	0.35	0.866	0.763	0.710	0.826	0.152	0.8773	0.00826		0.86006		0.90919	
	0.40	0.866	0.763	0.716	0.826	0.152	0.8797	0.00826		0.86006		0.90919	
	0.45	0.866	0.763	0.722	0.826	0.152	0.882	0.00826		0.86006		0.90919	
	0.50	0.866	0.763	0.729	0.826	0.152	0.8846	0.00826		0.86006		0.90919	
	0.55	0.866	0.763	0.735	0.826	0.152	0.8870	0.00826		0.86006		0.90919	
	0.60	0.866	0.763	0.743	0.826	0.152	0.8895	0.00826		0.86006		0.90919	
	0.65	0.866	0.763	0.751	0.826	0.152	0.8920	0.00826		0.86006		0.90919	
	0.70	0.866	0.763	0.760	0.826	0.152	0.8945	0.00826		0.86006		0.90919	
	0.75	0.866	0.763	0.769	0.826	0.152	0.8969	0.00826		0.86006		0.90919	
	0.80	0.866	0.763	0.780	0.826	0.152	0.8990	0.00826		0.86006		0.90919	
	0.85	0.866	0.763	0.792	0.826	0.152	0.9018	0.00826		0.86006		0.90919	
	0.90	0.866	0.763	0.806	0.826	0.152	0.9040	0.00826		0.86006		0.90919	
	0.95	0.866	0.763	0.823	0.826	0.152	0.9070	0.00826		0.86006		0.90919	
	1.00	0.858	0.753	0.859	0.817	0.155	0.9100	0.00689		0.86177		0.90989	

Σχήμα 7.14. Οι δείκτες εκτίμησης του πειράματος. Ομαδοποίηση με *HAC*, συλλογή *F*.



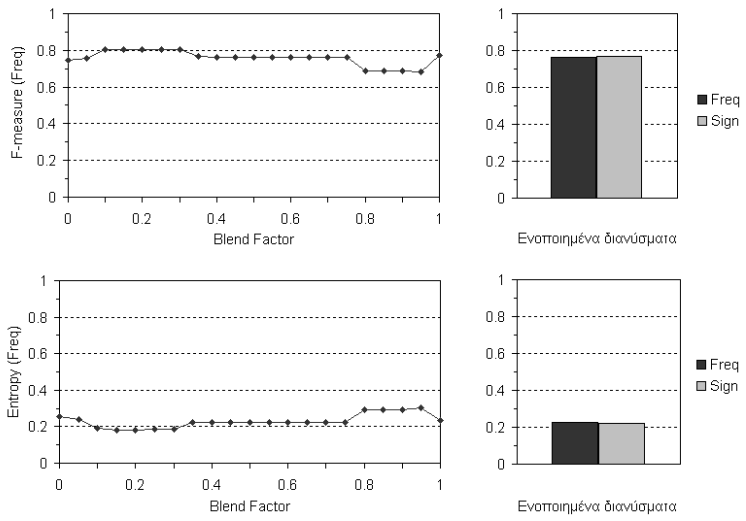
Σχήμα 7.15. Πείραμα μίξης της σημαντικότητας των ακμών στον υπολογισμό των ομοιοτήτων. Ομαδοποίηση με *HAC*, συλλογή *J*.

Βάρη	Blend Factor	Rand	Purity	Silh.	F	E
Freq	-	0.898	0.739	0.508	0.767	0.263
Sign	-	0.872	0.636	0.647	0.678	0.325

Σχήμα 7.16. Ενοποιημένο μοντέλο, ομαδοποίηση με *HAC*, συλλογή *J*.

		Δείκτες ποιότητας ομαδοποίησης						Δείκτες επιλογής *bf*					
Βάρη	Blend Factor	Rand	Purity	Silh.	F	E	MCE	Q	max(Q)	$Q_{bf}^{(i)}$	min($Q_{bf}^{(i)}$)	$Q_{bf}^{(e)}$	min($Q_{bf}^{(i)}$)
Freq	0.00	0.815	0.554	0.378	0.599	0.408	0.8648	0.02360		0.86480		0.86480	
	0.05	0.832	0.641	0.362	0.671	0.352	0.8605	0.03041		0.85897		0.85897	
	0.10	0.817	0.565	0.403	0.616	0.389	0.8674	0.02470		0.86434		0.86434	
	0.15	0.831	0.647	0.405	0.683	0.335	0.8662	0.02840		0.86191		0.86191	
	0.20	0.822	0.636	0.432	0.667	0.350	0.8715	0.02426		0.86587		0.86587	
	0.25	0.848	0.679	0.406	0.706	0.317	0.8595	0.03956		0.85273		0.85273	
	0.30	0.900	0.739	0.426	0.769	0.258	0.8339	0.07111		0.82817		0.82817	
	0.35	0.904	0.745	0.448	0.774	0.251	0.8334	0.07289	•	0.82697	•	0.82697	•
	0.40	0.900	0.745	0.477	0.773	0.254	0.8385	0.06830		0.83135		0.83135	
	0.45	0.902	0.750	0.496	0.780	0.244	0.8387	0.06907		0.83070		0.83070	
	0.50	0.902	0.750	0.512	0.780	0.244	0.8396	0.06907		0.83070		0.83070	
	0.55	0.876	0.701	0.559	0.736	0.281	0.8615	0.04586		0.85035		0.85035	
	0.60	0.899	0.739	0.531	0.771	0.254	0.8457	0.06421		0.83512		0.83512	
	0.65	0.876	0.701	0.589	0.736	0.281	0.8636	0.04586		0.85035		0.85035	
	0.70	0.874	0.696	0.605	0.731	0.287	0.8648	0.04561		0.85068		0.85068	
	0.75	0.895	0.734	0.586	0.769	0.256	0.8505	0.06188		0.83705		0.83705	
	0.80	0.888	0.717	0.575	0.738	0.301	0.8479	0.06500		0.83537		0.83537	
	0.85	0.885	0.707	0.604	0.733	0.317	0.8503	0.06329		0.83678		0.83678	
	0.90	0.883	0.690	0.620	0.703	0.345	0.8446	0.06686		0.83759		0.83759	
	0.95	0.739	0.522	0.685	0.564	0.489	0.9138	0.00000		0.88864		0.88864	
	1.00	0.862	0.647	0.696	0.660	0.388	0.8627	0.04871		0.85302		0.85302	

Σχήμα 7.17. Οι δείκτες εκτίμησης του πειράματος. Ομαδοποίηση με *HAC*, συλλογή *J*.



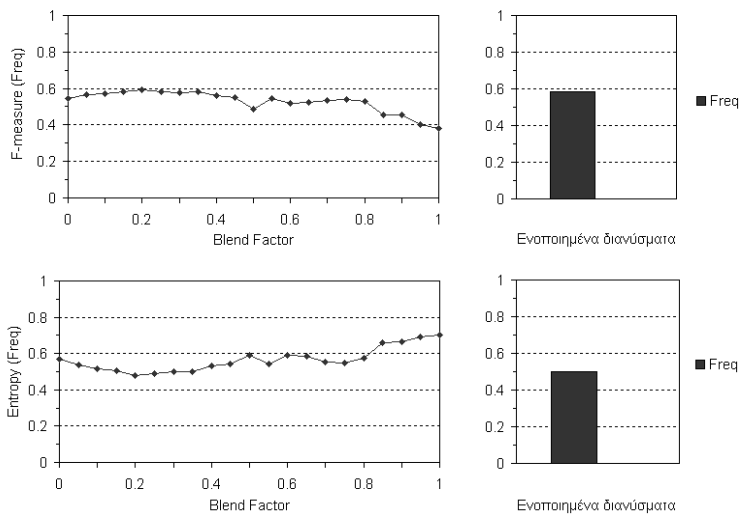
Σχήμα 7.18. Πείραμα μίξης της σημαντικότητας των ακμών στον υπολογισμό των ομοιοτήτων. Ομαδοποίηση με *HAC*, συλλογή *U*.

Βάρη	Blend Factor	Rand	Purity	Silh.	F	E
Freq	-	0.911	0.676	0.749	0.761	0.226
Sign	-	0.913	0.683	0.824	0.767	0.219

Σχήμα 7.19. Ενοποιημένο μοντέλο, ομαδοποίηση με *HAC*, συλλογή *U*.

		Δείκτες ποιότητας ομαδοποίησης						Δείκτες επιλογής *bf*					
Βάρη	Blend Factor	Rand	Purity	Silh.	F	E	MCE	Q	max(Q)	$Q_{bf}^{(f)}$	min($Q_{bf}^{(f)}$)	$Q_{bf}^{(e)}$	min($Q_{bf}^{(f)}$)
Freq	0.00	0.907	0.670	0.673	0.745	0.256	0.8123	0.03291		0.81233		0.87676	
	0.05	0.910	0.676	0.692	0.757	0.237	0.818	0.03060		0.81480		0.87830	
	0.10	0.918	0.738	0.704	0.805	0.189	0.8176	0.03569		0.81146		0.87269	
	0.15	0.921	0.751	0.706	0.806	0.179	0.815	0.04179		0.80560	•	0.86794	
	0.20	0.921	0.751	0.712	0.806	0.179	0.8181	0.04179		0.80560		0.86794	
	0.25	0.920	0.738	0.726	0.806	0.187	0.8241	0.03900		0.80897		0.86939	
	0.30	0.920	0.738	0.734	0.806	0.187	0.8271	0.03900		0.80897		0.86939	
	0.35	0.912	0.676	0.750	0.766	0.221	0.8361	0.03226		0.81520		0.87490	
	0.40	0.911	0.676	0.755	0.764	0.224	0.8378	0.03369		0.81394		0.87366	
	0.45	0.911	0.676	0.763	0.764	0.224	0.8408	0.03369		0.81394		0.87366	
	0.50	0.911	0.676	0.770	0.764	0.224	0.8438	0.03369		0.81394		0.87366	
	0.55	0.911	0.676	0.778	0.764	0.224	0.8468	0.03369		0.81394		0.87366	
	0.60	0.911	0.676	0.787	0.764	0.224	0.8498	0.03369		0.81394		0.87366	
	0.65	0.911	0.676	0.795	0.764	0.224	0.8528	0.03369		0.81394		0.87366	
	0.70	0.909	0.680	0.800	0.763	0.225	0.8583	0.03053		0.81718		0.87590	
	0.75	0.909	0.680	0.810	0.763	0.225	0.8612	0.03053		0.81718		0.87590	
	0.80	0.862	0.615	0.838	0.688	0.294	0.8871	0.00055		0.84838		0.89672	
	0.85	0.862	0.615	0.852	0.688	0.294	0.8895	0.00055		0.84838		0.89672	
	0.90	0.862	0.615	0.866	0.688	0.294	0.8919	0.00055		0.84838		0.89672	
	0.95	0.861	0.602	0.892	0.684	0.302	0.8948	0.00000		0.84881		0.89726	
	1.00	0.906	0.751	0.833	0.771	0.232	0.8557	0.04653	•	0.80919		0.85565	•

Σχήμα 7.20. Οι δείκτες εκτίμησης του πειράματος. Ομαδοποίηση με *HAC*, συλλογή *U*.



Σχήμα 7.21. Πείραμα μίξης της σημαντικότητας των ακμών στον υπολογισμό των ομοιοτήτων. Ομαδοποίηση με *HAC*, συλλογή *R*.

Βάρη	Blend Factor	Rand	Purity	Silh.	F	E
Freq	-	0.831	0.540	0.334	0.586	0.499
Sign	-	-	-	-	-	-

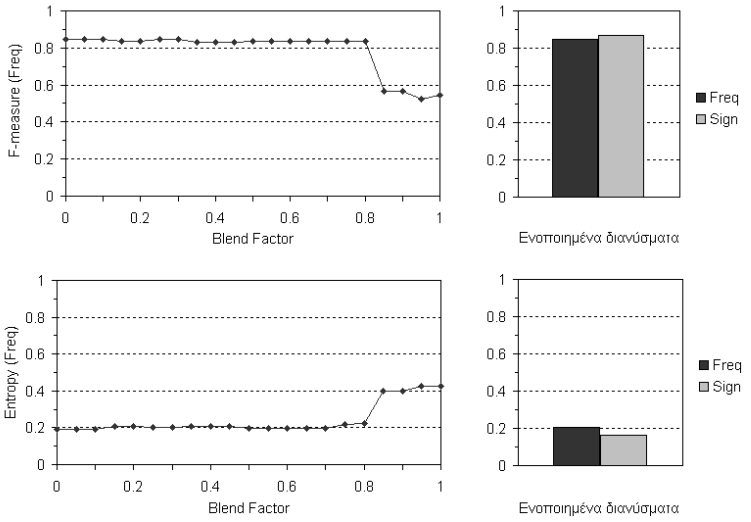
Σχήμα 7.22. Ενοποιημένο μοντέλο, ομαδοποίηση με *HAC*, συλλογή *R*.

		Δείκτες ποιότητας ομαδοποίησης						Δείκτες επιλογής *bf*					
Βάρη	Blend Factor	Rand	Purity	Silh.	F	E	MCE	Q	max(Q)	$Q_{bf}^{(f)}$	min($Q_{bf}^{(f)}$)	$Q_{bf}^{(e)}$	min($Q_{bf}^{(f)}$)
Freq	0.00	0.808	0.507	0.266	0.544	0.566	0.8915	0.02444		0.89150		0.92691	
	0.05	0.823	0.522	0.296	0.566	0.537	0.8922	0.02863		0.89072		0.91984	
	0.10	0.822	0.526	0.320	0.574	0.513	0.8934	0.02872		0.89048		0.91991	
	0.15	0.827	0.541	0.332	0.583	0.503	0.8956	0.02812		0.89123		0.92028	
	0.20	0.834	0.556	0.333	0.593	0.480	0.8926	0.03193		0.88612		0.91831	
	0.25	0.829	0.552	0.321	0.582	0.489	0.892	0.03501		0.88456	•	0.91411	
	0.30	0.823	0.548	0.342	0.578	0.501	0.8963	0.03205		0.88762		0.91656	
	0.35	0.827	0.544	0.354	0.582	0.498	0.8972	0.03262		0.88703		0.91609	
	0.40	0.808	0.533	0.363	0.561	0.534	0.9025	0.02885		0.89228		0.91784	
	0.45	0.800	0.522	0.380	0.553	0.542	0.9051	0.02770		0.89539		0.91686	
	0.50	0.767	0.448	0.383	0.485	0.589	0.9171	0.01588		0.90704		0.92722	
	0.55	0.802	0.522	0.405	0.547	0.545	0.9071	0.02743		0.89872		0.91400	
	0.60	0.786	0.493	0.385	0.520	0.590	0.9104	0.02428		0.90371		0.91485	
	0.65	0.793	0.500	0.401	0.523	0.584	0.9091	0.02639		0.90124		0.91340	
	0.70	0.780	0.500	0.405	0.532	0.554	0.9160	0.01945		0.90836		0.91920	
	0.75	0.815	0.533	0.438	0.541	0.545	0.9006	0.03529	•	0.89589		0.90211	•
	0.80	0.799	0.500	0.464	0.531	0.574	0.908	0.02815		0.90175		0.90958	
	0.85	0.715	0.411	0.518	0.454	0.660	0.9309	0.00603		0.91974		0.93284	
	0.90	0.721	0.411	0.547	0.456	0.663	0.9307	0.00719		0.91832		0.93211	
	0.95	0.698	0.381	0.575	0.403	0.690	0.9344	0.00266		0.92394		0.93492	
	1.00	0.679	0.359	0.644	0.379	0.703	0.9375	0.00000		0.92629		0.93752	

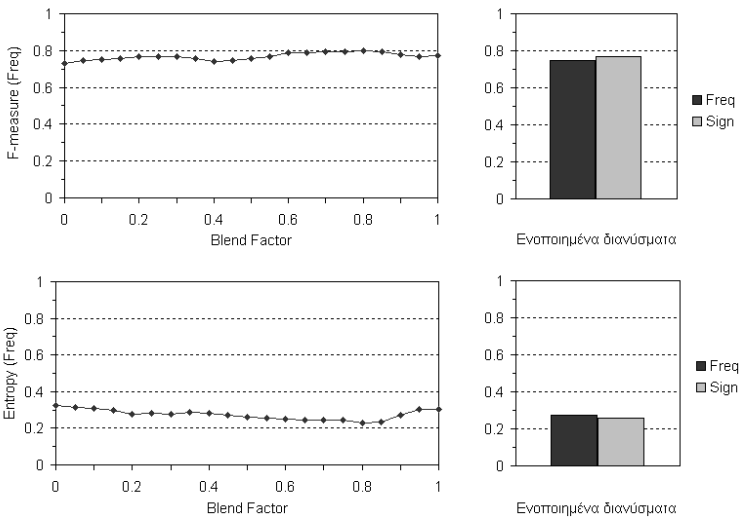
Σχήμα 7.23. Οι δείκτες εκτίμησης του πειράματος. Ομαδοποίηση με *HAC*, συλλογή *R*.

7.3.3.2. Γενικευμένος K-Ενδιαμέσων

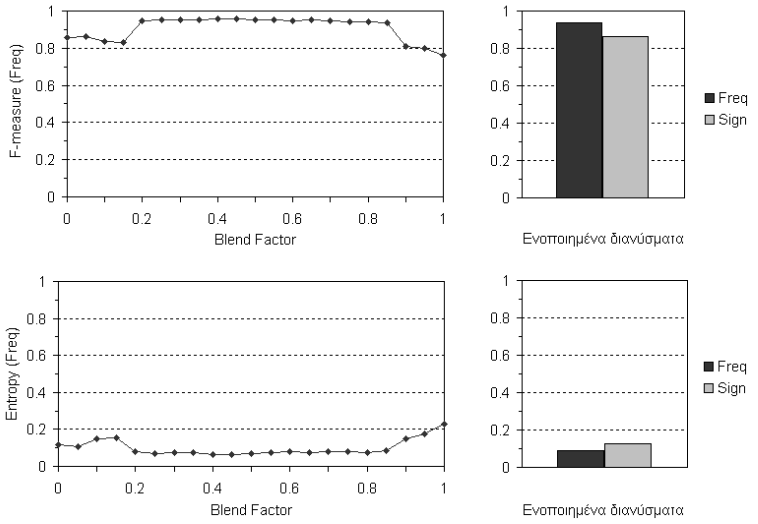
Για τα επόμενα πειράματα παραλείπουμε τους αναλυτικούς πίνακες λύσεων συντομεύοντας την παρουσίαση. Τα αποτελέσματα αφορούν τον ορισμό των ενδιάμεσων κειμένων ως κεντροειδή.



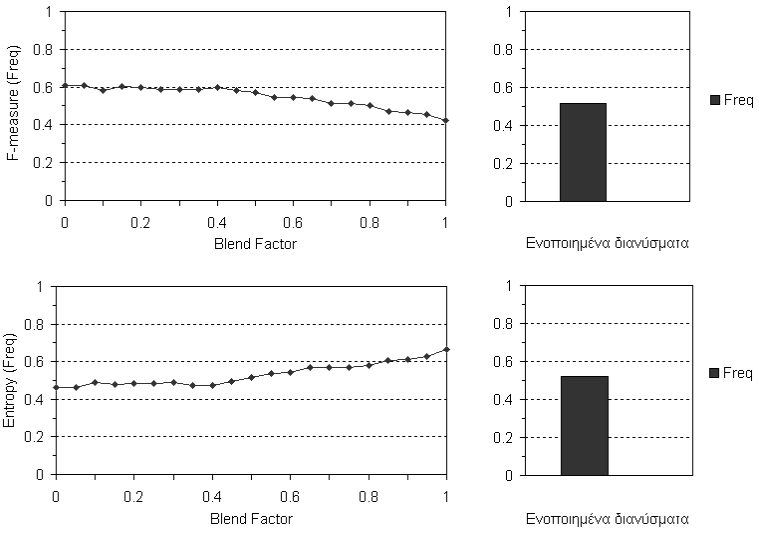
Σχήμα 7.24. Πείραμα μίξης της σημαντικότητας των ακμών στον υπολογισμό των ομοιοτήτων. Ομαδοποίηση με Γενικευμένο K-Ενδιαμέσων, συλλογή *F*.



Σχήμα 7.25. Πείραμα μίξης της σημαντικότητας των ακμών στον υπολογισμό των ομοιοτήτων. Ομαδοποίηση με Γενικευμένο K-Ενδιαμέσων, συλλογή *J*.



Σχήμα 7.26. Πείραμα μίξης της σημαντικότητας των ακμών στον υπολογισμό των ομοιοτήτων. Ομαδοποίηση με Γενικευμένο K-Ενδιαμέσων, συλλογή *U*.



Σχήμα 7.27. Πείραμα μίξης της σημαντικότητας των ακμών στον υπολογισμό των ομοιοτήτων. Ομαδοποίηση με Γενικευμένο K-Ενδιαμέσων, συλλογή *R*.

Ο στόχος της αναφοράς των πειραμάτων μίξης δεν είναι η εκτενής ανάλυση της συμπεριφοράς κάθε αλγορίθμου που υλοποιήσαμε ως προς τις ακμές. Το σημαντικότερο σημείο που θα πρέπει να σταθούμε είναι πως αποτελεί ένα εργαλείο με το οποίο μπορούμε να εκτιμήσουμε το συντελεστή βαρύτητας που θα πρέπει να δώσουμε στις ακμές ώστε να επωφεληθούμε. Αυτό γίνεται με τη διαδικασία επικύρωσης που περιγράφηκε. Επιπλέον, η διαδικασία αυτή θα μπορούσε να εφαρμοστεί για γενικότερα υποσύνολα χαρακτηριστικών (π.χ. τις συχνοτικά ασθενείς

λέξεις) ώστε να εκτιμηθεί σε ένα προεπεξεργαστικό στάδιο η βαρύτητα που θα πρέπει να λάβουν.

Ο συσσωρευτικός αλγόριθμος επωφελείται περισσότερο από την ύπαρξη ακμών, διότι εκμεταλλεύεται άμεσα την ενίσχυση της ομοιότητας των κειμένων ίδιας κατηγορίας που περιγράψαμε σε προηγούμενη παράγραφο.

Αντίθετα, παρατηρήθηκε ότι ο Γενικευμένος αλγόριθμος K-Μέσων επωφελείται ή δεν επηρεάζεται αρνητικά από της ακμές, κυρίως στην περίπτωση των ενδιάμεσων κεντροειδών. Δοκιμάστηκε ο αριθμητικός μέσος (αποτελέσματα που δεν παρατίθενται) αλλά οι λύσεις συνήθως χειροτέρευαν όσο αυξάνεται η επιρροή των ακμών, ακόμα και για συλλογές που ο συσσωρευτικός δείχνει να εκμεταλλεύεται αρκετά την ύπαρξη πληροφορίας ακμών (π.χ. η *J* συλλογή). Μία εξήγηση που μπορεί να δοθεί για το φαινόμενο αυτό είναι πως ο Γενικευμένος K-Μέσων αρχικοποιεί κάθε νέα ομάδα με ένα κείμενο. Αυξάνοντας τη διάσταση του προβλήματος με την εισαγωγή ακμών, και λόγω του προβλήματος της αυτό-ομοιότητας η αρχικοποίηση αυτής της μορφής δεν αποτελεί κίνητρο για να αποσπαστούν κείμενα από τις ήδη διαμορφωμένες ομάδες.

Τέλος, η προσέγγιση του ενοποιημένου μοντέλου που προτείνουμε δεν απαιτεί τον ορισμό συντελεστή μίξης και δείχνει ικανό να ενσωματώσει την πληροφορία των ακμών και να δώσει ικανοποιητικά αποτελέσματα. Τις περισσότερες φορές η ποιότητα της ομαδοποίησης με την προσέγγιση αυτή είναι συγκρίσιμη με την καλύτερη ή τις καλύτερες λύσεις του πειράματος μίξης. Η συμπεριφορά αυτή επιβεβαιώνει και την καταλληλότητα του ορισμού των βαρών των ακμών που θεωρήθηκε στην εργασία.

Για τα βάση με σημαντικότητες παρατηρούμε πως δε δίνουν καλύτερα αποτελέσματα απ τα συχνοτικά βάρη. Μια ερμηνεία για το αποτέλεσμα αυτό είναι πως αποτελεί συχνό φαινόμενο στις ιστοσελίδες να έχουμε κείμενο στο οποίο δίνεται έμφαση από πλευράς μορφής χωρίς όμως να είναι σημαντικό για το νοηματικό περιεχόμενο του κυρίως θέματος. Τέτοια παραδείγματα είναι π.χ ποια εταιρία δημιούργησε τον ιστοχώρο, ή τα μενού πλοήγησης και άλλα. Στη συνέχεια, για την ομαδοποίηση θα αγνοήσουμε τις ετικέτες. Πάντως, η κατεύθυνση βελτίωσης των αποτελεσμάτων θα πρέπει να είναι η ανάθεση βαρών μόνο σε ένα πολύ μικρό σύνολο ετικετών και με βάρη πολύ κοντά στις συχνότητες (π.χ 1.5, 1.2) ώστε να μην παίζουν καταλυτικό ρόλο στην κατανομή του περιεχομένου.

7.4. Μείωση Διάστασης με Φιλτράρισμα Βασισμένο στην Κ-γειτονιά

Η παράγραφος αυτή αναφέρει τα πειραματικά δεδομένα και παρατηρήσεις πάνω στην προσέγγιση για φιλτράρισμα της συλλογής βάσει της τοπικής πληροφορίας των *K-NN* κάθε γειτονιάς (*K-NN* Φιλτράρισμα). Για κάθε συλλογή αναφέρεται ένα σύνολο πειραμάτων στα οποία δίνεται ένα όνομα (στήλη *Πείραμα*). Η στήλη *LFinD* είναι το κατώφλι συχνότητας εμφάνισης σε ένα κείμενο που καθορίζει ποιες λέξεις αμφισβητούνται για την περιεκτικότητα πληροφορίας που έχουν για ένα κείμενο (πριν εφαρμοστεί ο όρος *IDF*) και πρόκειται να εξεταστεί η απαλοιφή τους. Το *k* είναι το μέγεθος της γειτονιάς που συμβουλευόμαστε για το φιλτράρισμα κάθε κειμένου. Το μήκος συλλογής είναι το άθροισμα των όρων που περιέχει κάθε κείμενο (όχι η διάσταση των δεδομένων).

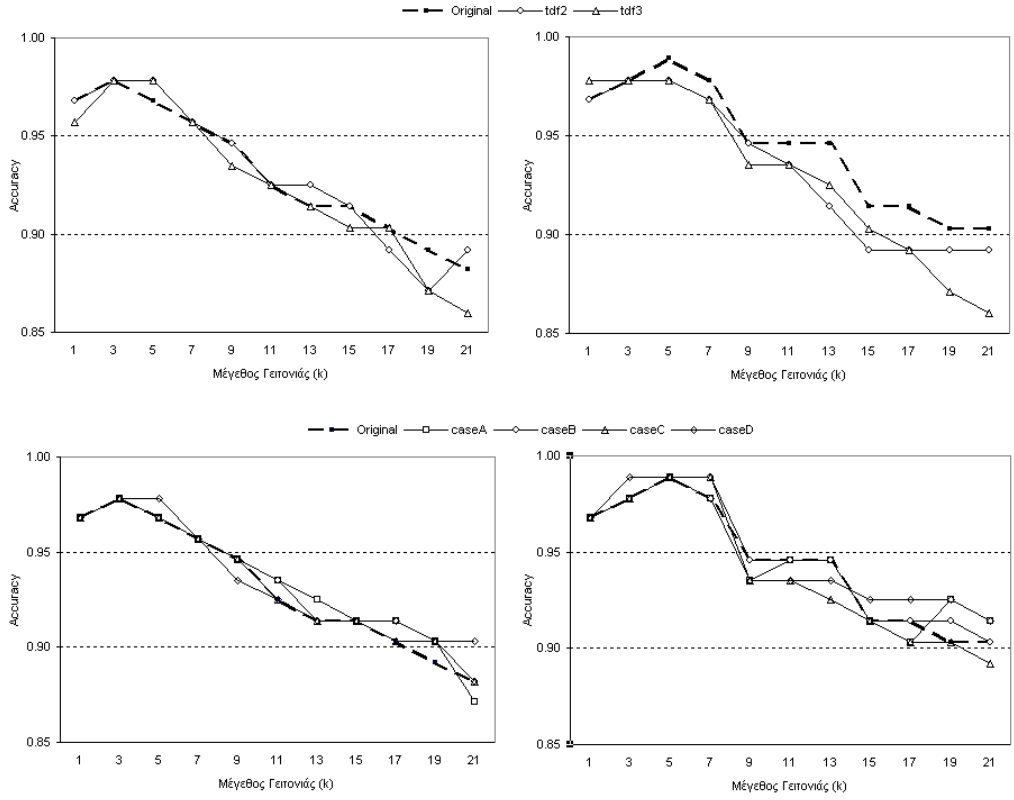
Στη στήλη *Υποψήφιες λέξεις για φιλτράρισμα* αναγράφεται ο αριθμός των λέξεων που αμφισβητείται βάσει του κατωφλίου *LFinD* σε όλη τη συλλογή. Η μείωση μήκους αναφέρει τον αριθμό των λέξεων που αφαιρέθηκαν από όλα τα μοντέλα αναπαράστασης και η συμπίεση εκφράζει τη μείωση των μοντέλων ως προς το μήκος της συλλογής. Στις τελευταίες τρεις στήλες αναγράφονται κατά σειρά: ο μέσος αριθμός λέξεων και ακμών που έχουν τα κείμενα της συλλογής μετά το φιλτράρισμα και τέλος το ποσοστό περιεχομένου που περιέχεται στις ακμές των μοντέλων όταν αυτές χρησιμοποιούνται. Το ποσοστό αυτό αυξάνει διότι το φιλτράρισμα γίνεται πάνω στους όρους όποτε και παραμένουν οι ακμές μικρής συχνότητας στα κείμενα.

Τα τρία πρώτα πειράματα είναι ίδια για όλες τις συλλογές, κατωφλίωση συμμετοχής στη συλλογή (*TDF*) η οποία διώχνει χωρίς κριτήρια τους όρους με συχνότητα χαμηλότερη από *LFinD*. Η μικρή απόκλιση μεταξύ υποψήφιων λέξεων για αποκοπή (στήλη 5) και των λέξεων που τελικά απαλείφονται (στήλη 6) έχει να κάνει με την παράμετρο «χάρης», όπου αν έχουμε πληροφορία από ετικέτες ιδιοτήτων στα κείμενα μπορούμε να ακυρώσουμε την αποκοπή κάποιου όρου (π.χ. αν αυτός παρουσιάζεται στον τίτλο του κειμένου). Η παράμετρος αυτή ορίστηκε με τον ίδιο τρόπο για κάθε πείραμα στην ίδια συλλογή.

Στη συνέχεια παρουσιάζονται δύο σχήματα που αφορούν την κατηγοριοποίηση *K-NN* για γειτονιές μεγέθους 1, …, 21 κειμένων με βήμα αύξησης 2, για μοντέλα που χρησιμοποιούν ακμές (δεξιά) και τα αντίστοιχα χωρίς αυτές (αριστερά). Με τονισμένη διακεκομμένη γραμμή φαίνονται τα αποτελέσματα για τη συλλογή χωρίς φιλτράρισμα.

Πείραμα	LFinD	k	Μήκος Συλλογής	Υποψήφιες λέξεις για φιλτράρισμα	Μείωση μήκους	Συμπίεση	Λέξεις / κείμενο	Ακμές / κείμενο	Περιεχόμενο ακμών %
original	1	-	18774	0	0	0	88.2	18.4	27.0
tdf2	2	-	18774	11129	10968	0.58	27.3	16.5	26.9
tdf3	3	-	18774	14429	14130	0.75	13.2	13.9	26.2
caseA	2	15	18774	11129	2277	0.12	75.5	19.6	28.2
caseB	7	15	18774	17535	2985	0.16	73.7	20.0	28.8
caseC	7	10	18774	17535	4369	0.23	69.5	20.0	29.5
caseD	7	5	18774	17535	6873	0.37	59.7	20.5	30.2

Σχήμα 7.28. Περιγραφή πειραμάτων φιλτραρίσματος, συλλογή *F*.



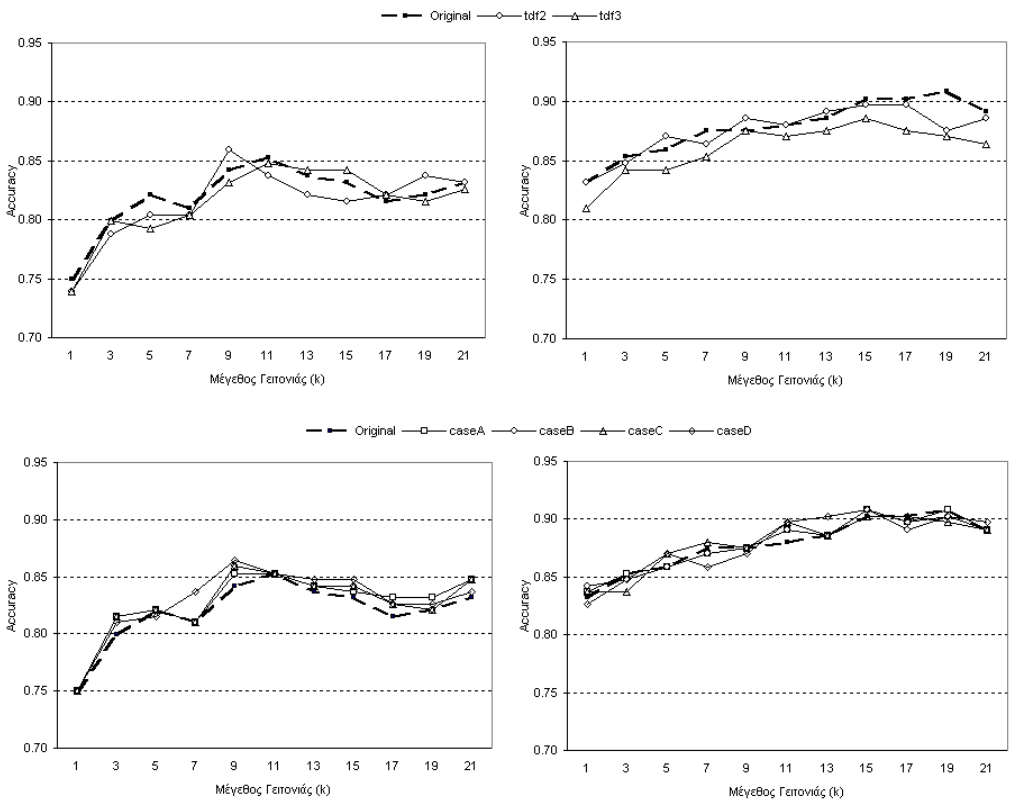
Σχήμα 7.29. Κατηγοριοποίηση *K-NN*, αριστερά: μοντέλα χωρίς ακμές, συλλογή F.

	Μοντέλα χωρίς ακμές					Μοντέλα με ακμές				
Πείραμα	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
Original	0.781	0.667	0.622	0.693	0.252	0.892	0.860	0.671	0.845	0.127
tdf2	0.795	0.710	0.768	0.773	0.229	0.866	0.763	0.796	0.826	0.152
tdf3	0.842	0.796	0.798	0.776	0.195	0.887	0.849	0.815	0.831	0.133
caseA	0.781	0.667	0.643	0.693	0.252	0.892	0.860	0.687	0.845	0.127
caseB	0.892	0.860	0.655	0.845	0.127	0.892	0.860	0.697	0.845	0.127
caseC	0.834	0.806	0.661	0.798	0.198	0.892	0.860	0.716	0.845	0.127
caseD	0.847	0.806	0.732	0.790	0.190	0.892	0.860	0.758	0.845	0.127

Σχήμα 7.30. Ομαδοποίηση των συλλογών με τον *HAC*, συλλογή *F*.

Πείραμα	LFinD	k	Μήκος Συλλογής	Υποψήφιες λέξεις για φιλτράρισμα	Μείωση μήκους	Συμπίεση	Λέξεις / κείμενο	Ακμές / κείμενο	Περιεχόμενο ακμών %
original	1	-	57481	0	0	0.00	312.4	198.3	39.4
tdf2	2	-	57481	30738	30245	0.53	148.0	149.4	39.0
tdf3	3	-	57481	40625	39511	0.69	97.7	124.3	37.7
caseA	2	20	57481	30738	5127	0.09	285.9	195.1	40.6
caseB	2	15	57481	30738	6205	0.11	278.7	193.4	41.0
caseC	5	10	57481	30738	4874	0.08	267.7	191.1	41.7
caseD	5	10	57481	48205	11115	0.19	252.0	190.4	41.2

Σχήμα 7.31. Περιγραφή πειραμάτων φιλτραρίσματος, συλλογή *J*.



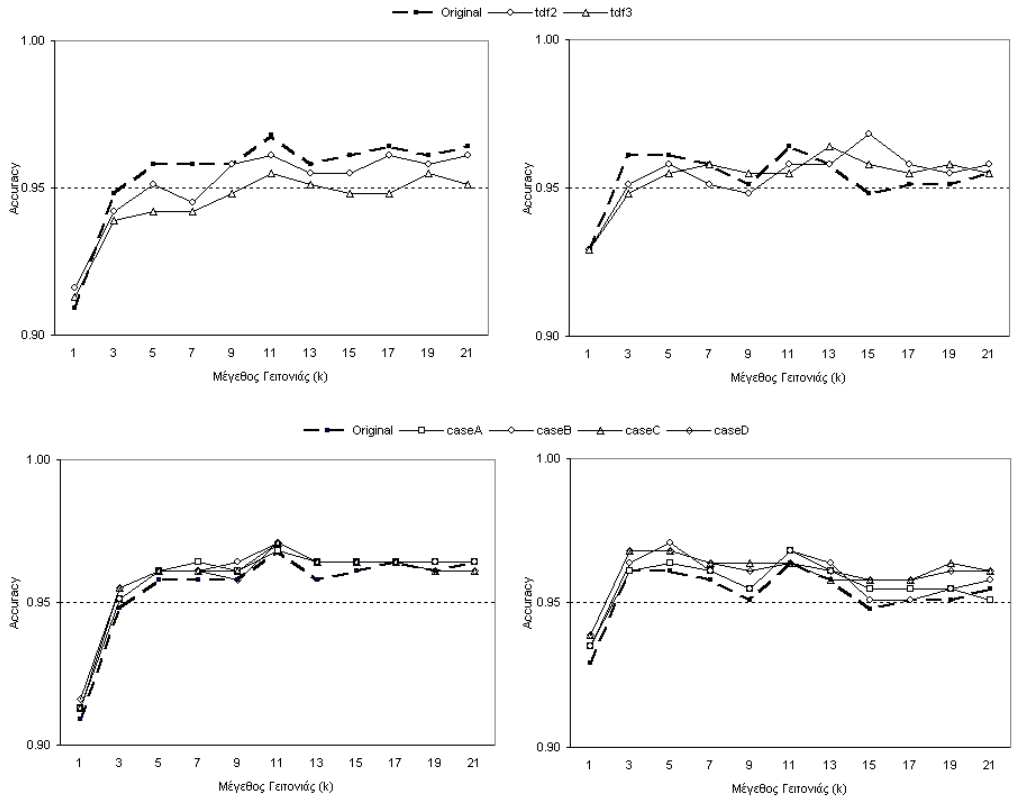
Σχήμα 7.32. Κατηγοριοποίηση *K-NN*, αριστερά: μοντέλα χωρίς ακμές, συλλογή *J*.

	Μοντέλα χωρίς ακμές					Μοντέλα με ακμές				
Πείραμα	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
Original	0.815	0.554	0.378	0.599	0.408	0.898	0.739	0.508	0.767	0.263
tdf2	0.748	0.495	0.543	0.548	0.478	0.891	0.734	0.608	0.766	0.273
tdf3	0.705	0.440	0.680	0.499	0.514	0.891	0.728	0.636	0.762	0.279
caseA	0.815	0.554	0.389	0.599	0.408	0.900	0.745	0.520	0.773	0.257
caseB	0.815	0.554	0.394	0.599	0.408	0.900	0.745	0.526	0.773	0.257
caseC	0.818	0.549	0.389	0.594	0.415	0.895	0.739	0.530	0.769	0.266
caseD	0.804	0.549	0.425	0.600	0.414	0.898	0.739	0.532	0.767	0.263

Σχήμα 7.33. Ομαδοποίηση των συλλογών με τον *HAC*,συλλογή *J*.

Πείραμα	LFinD	k	Μήκος Συλλογής	Υποψήφιες λέξεις για φιλτράρισμα	Μείωση μήκους	Συμπίεση	Λέξεις / κείμενο	Ακμές / κείμενο	Περιεχόμενο ακμών %
original	1	-	51702	0	0	0.00	167.3	64.8	33.6
tdf2	2	-	51702	30738	30819	0.60	67.6	36.3	33.4
tdf3	3	-	51702	40625	37545	0.73	45.8	27.0	33.3
caseA	10	40	51702	50074	4213	0.08	153.7	63.4	34.6
caseB	5	20	51702	46900	7766	0.15	141.6	61.5	35.1
caseC	5	10	51702	46900	12568	0.24	125.6	58.0	35.7
caseD	10	10	51702	50074	12673	0.25	125.3	57.8	35.8

Σχήμα 7.34. Περιγραφή πειραμάτων φιλτραρίσματος, συλλογή *U*.



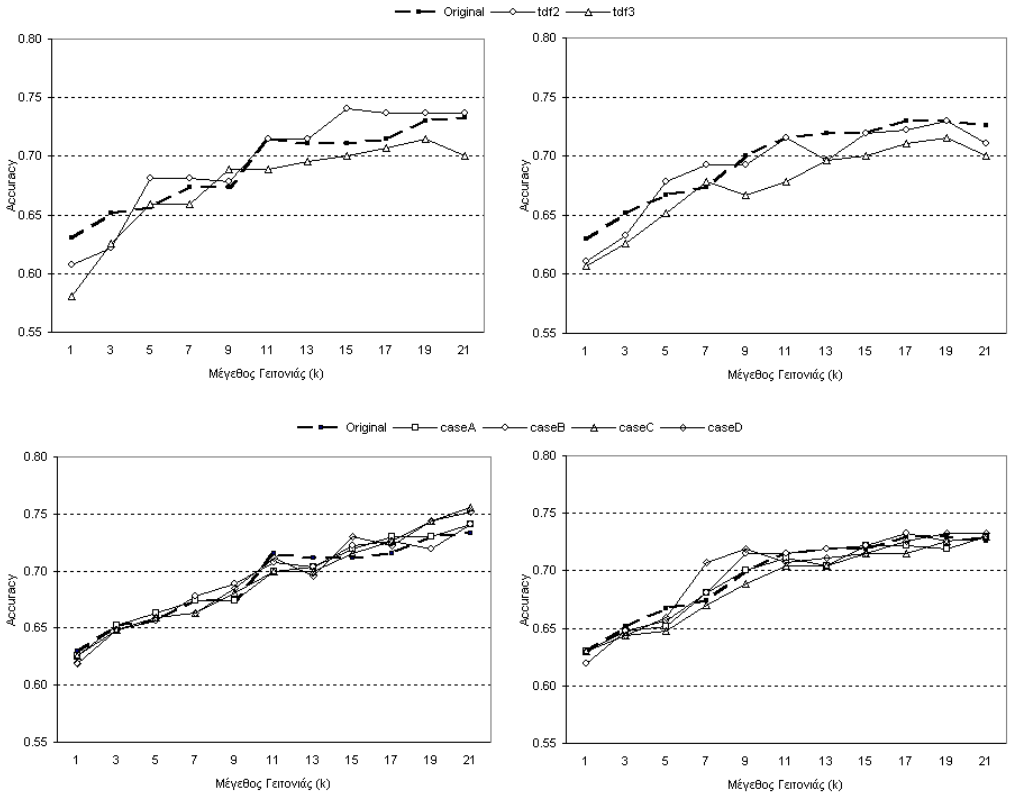
Σχήμα 7.35. Κατηγοριοποίηση *K-NN*, αριστερά: μοντέλα χωρίς ακμές, συλλογή *U*.

	Μοντέλα χωρίς ακμές					Μοντέλα με ακμές				
Πείραμα	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
Original	0.907	0.67	0.673	0.745	0.256	0.911	0.676	0.744	0.761	0.226
tdf2	0.906	0.673	0.74	0.745	0.253	0.912	0.680	0.813	0.770	0.213
tdf3	0.906	0.67	0.766	0.746	0.256	0.912	0.680	0.832	0.770	0.213
caseA	0.916	0.728	0.681	0.794	0.211	0.912	0.680	0.759	0.768	0.216
caseB	0.912	0.722	0.684	0.783	0.234	0.912	0.680	0.768	0.768	0.216
caseC	0.91	0.676	0.717	0.758	0.235	0.912	0.680	0.781	0.768	0.216
caseD	0.918	0.731	0.717	0.799	0.204	0.912	0.680	0.782	0.768	0.216

Σχήμα 7.36. Ομαδοποίηση των συλλογών με τον *HAC*,συλλογή *U*.

Πείραμα	LFinD	k	Μήκος Συλλογής	Υποψήφιες λέξεις για φιλτράρισμα	Μείωση μήκους	Συμπίεση	Λέξεις / κείμενο	Ακμές / κείμενο	Περιεχόμενο ακμών %
original	1	-	23811	0	0	0	88.2	17.8	18.4
tdf2	2	-	23811	16430	16430	0.69	27.3	7.0	16.5
tdf3	3	-	23811	20239	20239	0.85	13.2	3.5	13.9
caseA	2	25	23811	16430	3384	0.14	75.5	16.9	19.6
caseB	4	25	23811	21714	3862	0.16	73.7	16.6	20.0
caseC	2	15	23811	16430	5009	0.21	69.5	16.2	20.0
caseD	4	10	23811	23811	7660	0.32	59.7	14.5	20.5

Σχήμα 7.37. Περιγραφή πειραμάτων φιλτραρίσματος, συλλογή *R*.



Σχήμα 7.38. Κατηγοριοποίηση *K-NN*, αριστερά: μοντέλα χωρίς ακμές, συλλογή R.

	Μοντέλα χωρίς ακμές					Μοντέλα με ακμές				
Πείραμα	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
Original	0.808	0.507	0.266	0.544	0.567	0.831	0.541	0.334	0.586	0.499
tdf2	0.800	0.47	0.473	0.537	0.549	0.826	0.533	0.468	0.574	0.509
tdf3	0.769	0.422	0.475	0.454	0.630	0.801	0.463	0.514	0.504	0.573
caseA	0.813	0.507	0.316	0.553	0.547	0.831	0.541	0.388	0.591	0.489
caseB	0.814	0.504	0.332	0.554	0.543	0.829	0.533	0.384	0.584	0.498
caseC	0.816	0.507	0.326	0.555	0.543	0.829	0.537	0.387	0.585	0.496
caseD	0.817	0.504	0.378	0.556	0.542	0.836	0.559	0.39	0.582	0.492

Σχήμα 7.39. Ομαδοποίηση των συλλογών με τον *HAC*,συλλογή *R*.

Το *K-NN* Φιλτράρισμα αποκόπτει γενικά μικρότερες ποσότητες πληροφορίας σε σχέση με το κατωφλίωση *TDF*. Η δεύτερη φαίνεται να αλλοιώνει την ποιότητα της πληροφορίας των *K-NN* δίνοντας χειρότερα αποτελέσματα ομαδοποίησης με τον *HAC*, αν και δεν αποβαίνει καταστροφικό. Αυτό δικαιολογείται από τα προηγούμενα πειράματα που κάναμε, τα οποία δείχνουν ότι μικρότερα υποσύνολα ισχυρών όρων μπορούν να διατηρήσουν τα βασικά χαρακτηριστικά του χώρου των δεδομένων (διαχωρισιμότητα ομάδων).

Το σημαντικότερο μειονέκτημα είναι η ανυπαρξία κάποιου κριτηρίου συγκράτησης του φιλτραρίσματος *TDF*, καταλήγοντας να μειώνει υπερβολικά τους όρους των κειμένων. Αντίθετα, η προσέγγιση που προτείνεται ενσωματώνει κριτήρια που θέτουν οι ιδιότητες των δεδομένων και συγκεκριμένα η δομή τους σε γειτονιές, διατηρώντας ή ενισχύοντας την ποιότητα τους.

Τελικά ο συσσωρευτικός αλγόριθμος που βασίζεται αρκετά στην πληροφορία αυτή δείχνει να επωφελείται τις περισσότερες των περιπτώσεων. Το ίδιο συμπέρασμα βγαίνει ανεξάρτητα της χρησιμοποίησης ακμών οι οποίες δε, στα αντίστοιχα πειράματα αποτελούν παράγοντα βελτίωσης των λύσεων της ομαδοποίησης και για τις δύο τεχνικές.

7.5. Συνθετικά Κέντρα

7.5.1. Παράδειγμα Εκτέλεσης του K-Μέσων με Διαφορετικά Κεντροειδή

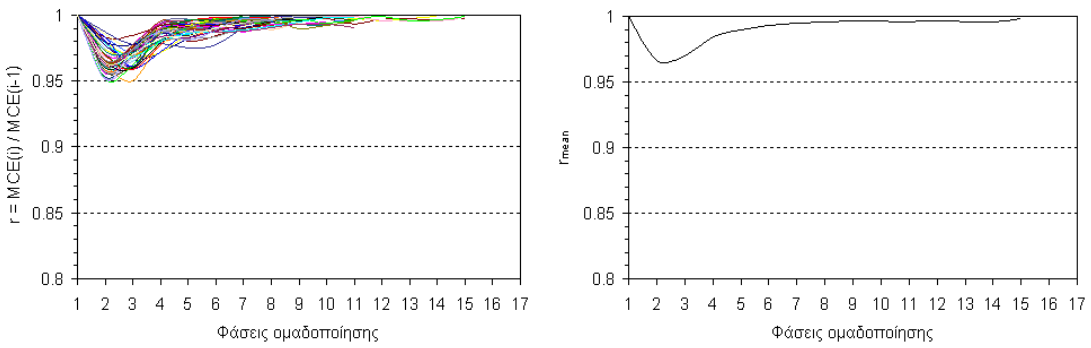
Περνώντας στα συνθετικά κέντρα, θα παρουσιάσουμε ένα παράδειγμα που περιγράφει τη συμπεριφορά τους. Εκτελέσαμε 50 τυχαία πειράματα με τον αλγόριθμο K-Μέσων στη συλλογή *J* λαμβάνοντας τους μέσους όρους. Γενικά στα πειράματα αυτής της σειράς, πέραν του μετασχηματισμού μορφολογικής ρίζας, απαλοιφής των συνηθισμένων λέξεων και αυτών που εμφανίζονται μόνο σε ένα κείμενο της συλλογής, δεν υποβάλλαμε τα δεδομένα σε κανένα άλλο φιλτράρισμα. Η αρχικοποίηση των ομάδων έγινε με ομοιόμορφη ανάθεση όλων των κειμένων στις ομάδες ώστε να προκύψει μία «δύσκολη» αρχική διαμέριση για τον αλγόριθμο. Όπως περιγράψαμε, στην περίπτωση αυτή οι ομάδες ξεκινούν χωρίς κυρίαρχα χαρακτηριστικά. Για τα πειράματα διαφορετικών κεντροειδών η αρχικοποίηση της γεννήτριας τυχαίων αριθμών

έγινε με πανομοιότυπο τρόπο, ώστε συγκριτικά να παρατηρήσουμε τη συμπεριφορά των τεχνικών πάνω στις ίδιες 50 τυχαίες αρχικοποιήσεις.

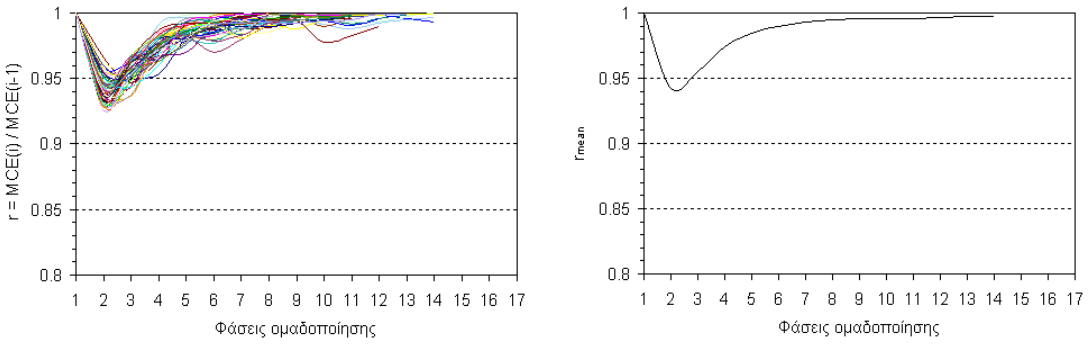
Η ποσότητα που παρακολουθούμε είναι ο λόγος της αντικειμενικής συνάρτησης (απόσταση κειμένων από το κεντροειδές) μεταξύ δύο διαδοχικών βημάτων:

$$r = \frac{MCE_c^{(i)}}{MCE_c^{(i)}}, \quad MCE_c^{(1)} = 1.$$

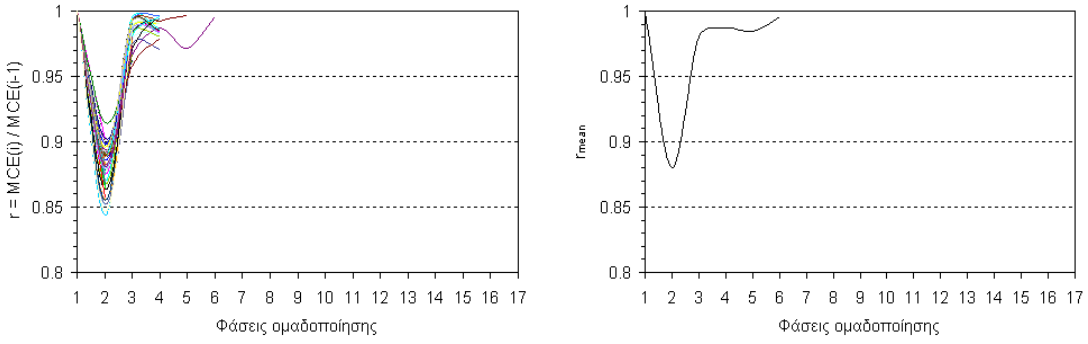
Ο δείκτης αυτός εκφράζει την αναλογία του σφάλματος μεταξύ δύο βημάτων, συνεπώς δεν εξαρτάται από τον τρόπο ορισμού του κέντρου. Μικρότερες τιμές εκφράζουν μεγαλύτερη σχετική μείωση του σφάλματος και ταχύτερη βελτίωση της λύσης. Αριστερά στα σχήματα παραθέτουμε την τιμή του δείκτη *r* στα βήματα κάθε λύσης, ενώ δεξιά φαίνεται η μέση τιμή του λόγου αυτού για όλες τις εκτελέσεις. Όσο μεγαλύτερο είναι το εμβαδόν της γραφικής παράστασης του *r* για μία λύση (με το επάνω όριο των παραστάσεων) τόσο καλύτερο είναι το αποτέλεσμα της ομαδοποίησης, δεδομένης της αρχικοποίησης. Κατ' επέκταση όσο βλέπουμε την καμπύλη της μέσης r_{mean} να καλύπτει μεγαλύτερο χώρο στο γράφημα τόσο βελτιώνονται οι λύσεις της ομαδοποίησης.



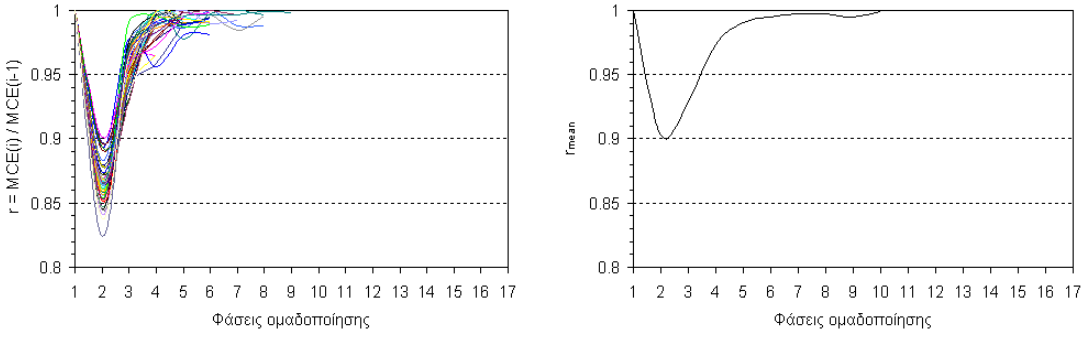
Σχήμα 7.40. Η πρόοδος των 50 εκτελέσεων με τον K-Μέσων και κέντρα τους αριθμητικούς μέσους χωρίς φιλτράρισμα, συλλογή *J*.



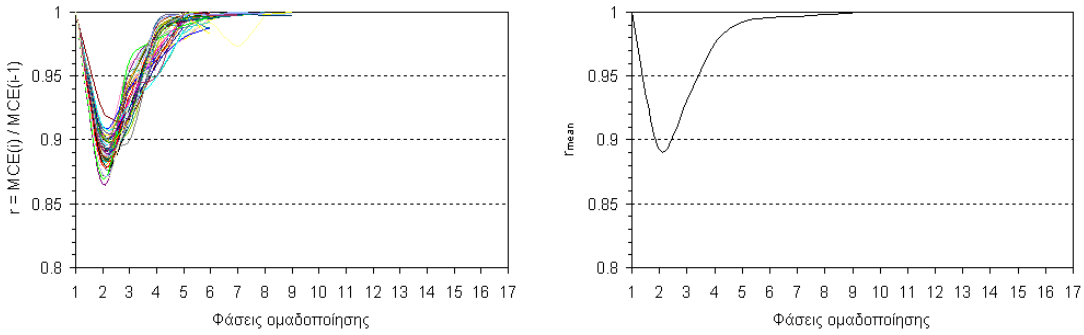
Σχήμα 7.41. Η πρόοδος των 50 εκτελέσεων με τον K-Συνθετικών Κέντρων και κέντρα τους αριθμητικούς μέσους με φιλτράρισμα 160 λέξεων, συλλογή *J*.



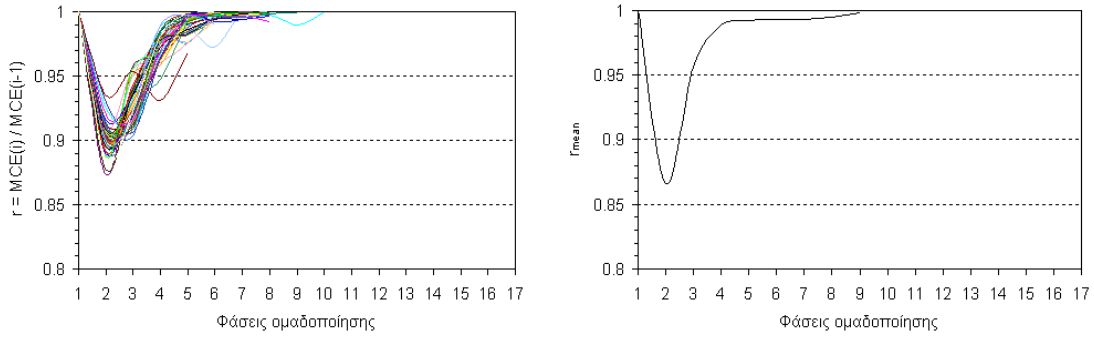
Σχήμα 7.42. Η πρόοδος των 50 εκτελέσεων με τον K-Ενδιαμέσων και κέντρα τα ενδιάμεσα κείμενα χωρίς φιλτράρισμα, συλλογή *J*.



Σχήμα 7.43. Η πρόοδος των 50 εκτελέσεων με τον K-Συνθετικών Κέντρων και κέντρα τις 5-*NN* γειτονιές των ενδιαμέσων με φιλτράρισμα 160 λέξεων, συλλογή *J*.



Σχήμα 7.44. Η πρόοδος των 50 εκτελέσεων με τον K-Συνθετικών Κέντρων και κέντρα τις 10-*NN* γειτονιές των ενδιαμέσων με φιλτράρισμα 160 λέξεων, συλλογή *J*.



Σχήμα 7.45. Η πρόοδος των 50 εκτελέσεων με τον K-Συνθετικών Κέντρων και κέντρα τις 10-*NN* γειτονιές των ενδιαμέσων με φιλτράρισμα 460, συλλογή *J*.

Ο αριθμητικός μέσος εγκλωβίζεται στην άσχημη αρχικοποίηση. Κάνει αρκετά βήματα καθένα από τα οποία βελτιώνει σε μικρό βαθμό τη λύση του προηγούμενου βήματος. Φιλτράροντας τον αριθμητικό μέσο παρατηρούμε μεγαλύτερη δραστηριότητα στα αρχικά βήματα της διαδικασίας, όπου οι λύσεις βελτιώνονται ταχύτερα. Επίσης, ακόμα και στα τελευταία βήματα των ομαδοποιήσεων οι λύσεις βελτιώνονται αισθητά ταχύτερα σε σχέση με τον αριθμητικό μέσο χωρίς φιλτράρισμα.

Στα πειράματα με τον K-Ενδιαμέσων γίνεται ένα δραστικά βελτιωτικό πρώτο βήμα αλλά δεν καταφέρνει να συνεχίσει τη διαδικασία βελτίωσης τους βρίσκοντας καλύτερες ομάδες. Εξηγήσαμε πως το φαινόμενο αυτό είναι αναμενόμενο από τη στιγμή που κάθε κείμενο έχει ένα μικρό υποσύνολο των χαρακτηριστικών της ομάδας που ανήκει. Τελικά η έλλειψη περιεχομένου (αδυναμία μάθησης από τα χαρακτηριστικά της ομάδας) η οποία διαπιστώνεται καθιστά καλή αρχική προσέγγιση

για να ξεφύγει ευκολότερα ο αλγόριθμος από ενδεχόμενη κακή αρχικοποίηση αλλά αδυνατεί να υποστηρίξει όλη τη διαδικασία εκπαίδευσης.

Το ζητούμενο λοιπόν, είναι να εμπλουτιστεί το κεντροειδές με το κατάλληλο περιεχόμενο ώστε που να μπορεί να υποστηρίξει όλη τη διαδικασία εκπαίδευσης παίρνοντας ταυτόχρονα αυστηρότερες αποφάσεις για την κατηγορία που θα περιγράφει. Χρησιμοποιώντας συνθετικά κέντρα βάσει των *K-NN* του ενδιαμέσου παρατηρούμε πως πετυχαίνεται κάτι τέτοιο. Όλες οι λύσεις έχουν ταχύτερη βελτίωση των λύσεων, κοντά στα 2/3 των βημάτων σύγκλισης του αριθμητικού μέσου, διατηρώντας τη σημαντική πληροφορία των ομάδων. Στους πίνακες που ακολουθούν αναγράφονται λεπτομερώς τα χαρακτηριστικά των λύσεων των παραδειγμάτων αυτών στην αρχικοποίηση «τυχαία ανάθεση κειμένων».

7.5.2. Συγκριτικά Αποτελέσματα Τεχνικών Ομαδοποίησης

Παραθέτουμε τα αναλυτικά αποτελέσματα για όλες τις συλλογές δεδομένων που χρησιμοποιήθηκαν. Ο πρώτος πίνακας για κάθε συλλογή περιλαμβάνει τις λύσεις του K-Μέσων που παράγονται με τρεις διαφορετικούς τρόπους αρχικοποίησης. Αναφέρονται οι λύσεις αριθμητικού μέσου με ή χωρίς ακμές (‘+’/‘-’) για τα πλήρη γραφήματα, ώστε να έχουμε μία βάση σύγκρισής για το πόσο καλύτερα είναι τα αποτελέσματα από την παραδοσιακή προσέγγιση. Ο μικρότερος πίνακας που ακολουθεί αναφέρει τις λύσεις του παραδοσιακού ιεραρχικού αλγορίθμου (*HAC – group average*) και του Γενικευμένου K-Μέσων με διάφορες προσεγγίσεις ορισμού κέντρων. Στους πίνακες η στήλη *Φιλτράρισμα Κέντρων* αναφέρει τους όρους που κρατάμε από κάθε κεντροειδές. Όταν χρησιμοποιούνται ακμές τότε διατηρούνται όλες αυτές που συνδέουν όρους που παρέμειναν στο συνθετικό κέντρο μετά το φιλτράρισμα.

				Αρχικοποίηση Κέντρων															
Περιγραφή πειράματος				Μ-μακρινότερα κείμενα					Μ-τυχαία κείμενα					Τυχαία ανάθεση κειμένων				
Αλγ/μος	Ακμές	Κέντρα	Φιλτ/σμα κέντρων	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
K-Μέσων	-	mean	-	0.803	0.737	0.511	0.727	0.257	0.751	0.646	0.436	0.644	0.326	0.727	0.590	0.274	0.589	0.413
	+	mean	-	0.792	0.727	0.504	0.716	0.278	0.742	0.628	0.415	0.626	0.348	0.669	0.473	0.062	0.472	0.509
			460	0.799	0.735	0.524	0.724	0.267	0.750	0.638	0.438	0.636	0.337	0.683	0.502	0.133	0.502	0.485
			260	0.805	0.742	0.538	0.730	0.259	0.758	0.648	0.464	0.647	0.323	0.706	0.543	0.226	0.546	0.446
			160	0.812	0.751	0.552	0.739	0.250	0.763	0.655	0.478	0.655	0.315	0.719	0.570	0.296	0.571	0.422
			80	0.822	0.763	0.571	0.749	0.237	0.786	0.692	0.532	0.688	0.279	0.741	0.605	0.362	0.609	0.388
			40	0.832	0.774	0.590	0.762	0.224	0.808	0.749	0.496	0.737	0.260	0.751	0.623	0.407	0.624	0.367
		medoid	-	0.817	0.770	0.491	0.760	0.265	0.782	0.703	0.440	0.694	0.309	0.783	0.704	0.434	0.693	0.313
		5-NN	-	0.837	0.796	0.522	0.788	0.230	0.808	0.749	0.496	0.737	0.260	0.812	0.751	0.489	0.743	0.262
			460	0.837	0.796	0.531	0.786	0.230	0.800	0.737	0.495	0.725	0.269	0.804	0.737	0.481	0.725	0.273
			260	0.824	0.768	0.524	0.759	0.245	0.802	0.734	0.483	0.721	0.268	0.805	0.735	0.478	0.726	0.273
			160	0.823	0.766	0.510	0.757	0.248	0.800	0.729	0.467	0.719	0.274	0.801	0.728	0.465	0.717	0.279
			80	0.814	0.755	0.495	0.745	0.263	0.793	0.719	0.449	0.708	0.289	0.788	0.706	0.428	0.691	0.302
			40	0.805	0.748	0.502	0.737	0.272	0.780	0.705	0.436	0.690	0.306	0.776	0.699	0.428	0.680	0.317
		10-NN	-	0.864	0.833	0.606	0.823	0.197	0.834	0.780	0.564	0.775	0.228	0.841	0.793	0.567	0.788	0.224
			460	0.864	0.830	0.595	0.820	0.198	0.833	0.777	0.552	0.767	0.234	0.844	0.796	0.557	0.789	0.225
			260	0.849	0.814	0.545	0.799	0.220	0.827	0.770	0.547	0.760	0.241	0.836	0.783	0.555	0.774	0.235
			160	0.855	0.816	0.587	0.804	0.211	0.825	0.768	0.547	0.755	0.243	0.837	0.785	0.573	0.777	0.232
			80	0.867	0.832	0.579	0.821	0.203	0.827	0.767	0.514	0.751	0.250	0.844	0.795	0.535	0.789	0.232
			40	0.858	0.822	0.559	0.805	0.215	0.825	0.765	0.509	0.750	0.255	0.836	0.782	0.520	0.766	0.245
		15-NN	-	0.920	0.897	0.636	0.896	0.124	0.901	0.869	0.613	0.867	0.149	0.908	0.880	0.601	0.879	0.146
			460	0.918	0.895	0.633	0.893	0.128	0.900	0.867	0.611	0.867	0.151	0.910	0.882	0.614	0.881	0.143
			260	0.917	0.896	0.630	0.894	0.131	0.900	0.867	0.611	0.867	0.151	0.904	0.878	0.599	0.876	0.151
			160	0.908	0.878	0.623	0.876	0.142	0.910	0.886	0.614	0.885	0.142	0.888	0.851	0.584	0.849	0.171
			80	0.910	0.886	0.609	0.884	0.146	0.893	0.857	0.593	0.856	0.165	0.889	0.856	0.582	0.852	0.173
			40	0.891	0.852	0.591	0.846	0.163	0.881	0.842	0.574	0.837	0.178	0.886	0.850	0.574	0.849	0.171

Σχήμα 7.46. Αποτελέσματα ομαδοποίησης με τον αλγόριθμο K-Μέσων και K-Συνθετικών Κέντρων τριών διαφορετικών αρχικοποιήσεων, συλλογή *F*.

			Μοντέλα χωρίς ακμές					Μοντέλα με ακμές				
Αλγ/μο	Κέντρα	Φιλτ/σμα κέντρων	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
HAC			0.781	0.667	0.622	0.693	0.252	0.892	0.860	0.671	0.845	0.127
Global-KM	medoid	-	0.871	0.849	0.546	0.848	0.193	0.870	0.849	0.593	0.850	0.204
	mean	-	0.716	0.677	0.668	0.662	0.316	0.756	0.742	0.594	0.731	0.292
	5-NN	160	0.827	0.785	0.552	0.764	0.248	0.869	0.849	0.593	0.842	0.180
	10-NN	160	0.859	0.828	0.588	0.820	0.210	0.860	0.828	0.647	0.817	0.200
	15-NN	160	0.913	0.903	0.628	0.902	0.134	0.914	0.903	0.622	0.903	0.149

Σχήμα 7.47. Αποτελέσματα ομαδοποίησης με τον αλγόριθμο *HAC*, τον Γενικευμένο K-Μέσων και τον Γενικευμένο K-Συνθετικών Κέντρων, συλλογή *F*.

				Αρχικοποίηση Κέντρων															
Περιγραφή πειράματος				Μ-μακρινότερα κείμενα					Μ-τυχαία κείμενα					Τυχαία ανάθεση κειμένων				
Αλγ/μος	Ακμές	Κέντρα	Φιλτ/σμα κέντρων	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
K-Μέσων	-	mean	-	0.887	0.651	0.256	0.648	0.379	0.904	0.698	0.283	0.701	0.316	0.873	0.507	0.037	0.506	0.566
	+	mean	-	0.912	0.722	0.394	0.722	0.313	0.892	0.645	0.324	0.647	0.379	0.864	0.458	0.008	0.456	0.632
			460	0.915	0.729	0.401	0.730	0.303	0.895	0.653	0.334	0.655	0.368	0.870	0.505	0.110	0.510	0.519
			260	0.918	0.737	0.412	0.739	0.290	0.900	0.669	0.351	0.670	0.348	0.886	0.568	0.183	0.571	0.491
			160	0.921	0.747	0.422	0.751	0.274	0.903	0.679	0.366	0.682	0.333	0.896	0.610	0.257	0.615	0.427
			80	0.925	0.759	0.435	0.765	0.258	0.910	0.705	0.397	0.709	0.303	0.907	0.659	0.321	0.667	0.362
			40	0.929	0.772	0.445	0.779	0.240	0.915	0.717	0.413	0.723	0.286	0.910	0.682	0.352	0.695	0.326
		medoid	-	0.910	0.744	0.395	0.744	0.290	0.898	0.698	0.347	0.693	0.337	0.882	0.642	0.277	0.635	0.390
		5-NN	-	0.926	0.787	0.435	0.793	0.238	0.913	0.738	0.392	0.734	0.285	0.907	0.727	0.391	0.725	0.298
			460	0.929	0.791	0.434	0.796	0.235	0.915	0.741	0.395	0.736	0.281	0.911	0.731	0.391	0.725	0.290
			260	0.930	0.794	0.435	0.797	0.233	0.915	0.740	0.399	0.735	0.283	0.914	0.738	0.399	0.731	0.285
			160	0.929	0.793	0.438	0.796	0.234	0.915	0.741	0.400	0.734	0.280	0.916	0.744	0.403	0.740	0.281
			80	0.931	0.795	0.436	0.797	0.229	0.915	0.740	0.391	0.734	0.282	0.916	0.745	0.401	0.735	0.279
			40	0.929	0.791	0.434	0.794	0.232	0.916	0.740	0.395	0.731	0.284	0.918	0.747	0.404	0.738	0.279
		10-NN	-	0.935	0.800	0.453	0.807	0.221	0.922	0.765	0.436	0.767	0.249	0.929	0.776	0.450	0.779	0.236
			460	0.937	0.805	0.457	0.811	0.216	0.926	0.771	0.443	0.773	0.243	0.931	0.784	0.459	0.786	0.228
			260	0.937	0.805	0.458	0.812	0.216	0.927	0.774	0.440	0.777	0.242	0.932	0.783	0.452	0.787	0.228
			160	0.937	0.803	0.455	0.810	0.217	0.928	0.773	0.438	0.775	0.242	0.935	0.797	0.457	0.798	0.217
			80	0.939	0.810	0.458	0.815	0.212	0.929	0.777	0.441	0.778	0.238	0.932	0.785	0.450	0.789	0.225
			40	0.938	0.807	0.451	0.812	0.211	0.930	0.781	0.440	0.783	0.233	0.929	0.780	0.444	0.785	0.227
		15-NN	-	0.937	0.793	0.457	0.799	0.227	0.927	0.761	0.435	0.766	0.252	0.936	0.779	0.437	0.789	0.242
			460	0.937	0.793	0.457	0.799	0.227	0.930	0.770	0.439	0.773	0.245	0.934	0.769	0.434	0.776	0.250
			260	0.937	0.793	0.457	0.799	0.227	0.931	0.774	0.443	0.777	0.241	0.936	0.778	0.437	0.788	0.241
			160	0.937	0.793	0.457	0.799	0.227	0.931	0.775	0.442	0.777	0.239	0.936	0.779	0.442	0.788	0.237
			80	0.937	0.793	0.457	0.799	0.227	0.932	0.779	0.445	0.784	0.233	0.936	0.780	0.432	0.790	0.232
			40	0.937	0.793	0.457	0.799	0.227	0.933	0.779	0.452	0.784	0.232	0.934	0.779	0.440	0.787	0.234

Σχήμα 7.48. Αποτελέσματα ομαδοποίησης με τον αλγόριθμο K-Μέσων και K-Συνθετικών Κέντρων τριών διαφορετικών αρχικοποιήσεων, συλλογή *J*.

			Μοντέλα χωρίς ακμές					Μοντέλα με ακμές				
Αλγ/μο	Κέντρα	Φιλτ/σμα κέντρων	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
HAC			0.815	0.554	0.378	0.599	0.408	0.898	0.739	0.508	0.767	0.263
Global-KM	medoid	-	0.911	0.745	0.246	0.731	0.325	0.919	0.745	0.416	0.747	0.271
	mean	-	0.870	0.652	0.284	0.669	0.356	0.916	0.761	0.480	0.763	0.246
	5-NN	160	0.909	0.717	0.333	0.704	0.297	0.917	0.793	0.459	0.787	0.222
	10-NN	160	0.915	0.728	0.332	0.732	0.278	0.944	0.842	0.442	0.849	0.175
	15-NN	160	0.951	0.832	0.479	0.836	0.175	0.949	0.837	0.383	0.834	0.203

Σχήμα 7.49. Αποτελέσματα ομαδοποίησης με τον αλγόριθμο *HAC*, τον Γενικευμένο K-Μέσων και τον Γενικευμένο K-Συνθετικών Κέντρων, συλλογή *J*.

				Αρχικοποίηση Κέντρων															
Περιγραφή πειράματος				Μ-μακρινότερα κείμενα					Μ-τυχαία κείμενα					Τυχαία ανάθεση κειμένων				
Αλγ/μος	Ακμές	Κέντρα	Φιλτ/σμα κέντρων	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
K-Μέσων	-	mean	-	0.950	0.837	0.573	0.843	0.141	0.936	0.807	0.557	0.806	0.168	0.943	0.825	0.530	0.821	0.166
	+	mean	-	0.951	0.854	0.616	0.854	0.139	0.931	0.796	0.591	0.791	0.181	0.937	0.809	0.542	0.804	0.191
			460	0.953	0.859	0.619	0.859	0.133	0.933	0.800	0.592	0.794	0.176	0.936	0.808	0.544	0.800	0.190
			260	0.955	0.865	0.620	0.864	0.128	0.936	0.810	0.598	0.803	0.168	0.938	0.811	0.552	0.805	0.182
			160	0.955	0.865	0.623	0.866	0.126	0.938	0.814	0.603	0.809	0.163	0.938	0.814	0.555	0.802	0.182
			80	0.956	0.869	0.628	0.867	0.122	0.939	0.820	0.608	0.812	0.158	0.943	0.827	0.574	0.819	0.164
			40	0.960	0.878	0.634	0.878	0.112	0.943	0.830	0.612	0.823	0.147	0.948	0.838	0.588	0.837	0.147
		medoid	-	0.945	0.848	0.598	0.836	0.157	0.924	0.786	0.539	0.768	0.220	0.883	0.686	0.436	0.660	0.315
		5-NN	-	0.945	0.848	0.598	0.836	0.157	0.933	0.813	0.544	0.789	0.193	0.900	0.723	0.461	0.693	0.274
			460	0.952	0.870	0.594	0.851	0.139	0.935	0.819	0.548	0.792	0.189	0.903	0.727	0.464	0.697	0.274
			260	0.952	0.870	0.594	0.852	0.140	0.934	0.817	0.541	0.791	0.192	0.899	0.710	0.435	0.674	0.292
			160	0.953	0.872	0.597	0.854	0.138	0.934	0.818	0.535	0.787	0.194	0.903	0.722	0.434	0.682	0.284
			80	0.950	0.863	0.585	0.845	0.148	0.934	0.815	0.535	0.786	0.199	0.904	0.722	0.431	0.683	0.287
			40	0.950	0.864	0.579	0.845	0.150	0.934	0.816	0.529	0.786	0.201	0.907	0.730	0.422	0.692	0.287
		10-NN	-	0.954	0.881	0.614	0.867	0.127	0.942	0.845	0.578	0.821	0.160	0.929	0.805	0.544	0.774	0.196
			460	0.953	0.880	0.609	0.864	0.130	0.944	0.851	0.575	0.823	0.160	0.927	0.798	0.525	0.762	0.206
			260	0.952	0.878	0.606	0.861	0.131	0.945	0.854	0.578	0.826	0.157	0.928	0.801	0.537	0.767	0.203
			160	0.954	0.882	0.613	0.868	0.129	0.946	0.857	0.582	0.832	0.154	0.930	0.810	0.534	0.772	0.197
			80	0.954	0.882	0.607	0.866	0.130	0.944	0.853	0.576	0.826	0.160	0.929	0.807	0.526	0.767	0.206
			40	0.952	0.877	0.603	0.862	0.135	0.942	0.845	0.567	0.818	0.168	0.931	0.810	0.521	0.773	0.206
		15-NN	-	0.957	0.890	0.625	0.880	0.117	0.947	0.860	0.595	0.838	0.144	0.946	0.858	0.580	0.833	0.145
			460	0.958	0.891	0.627	0.881	0.115	0.948	0.862	0.599	0.841	0.142	0.942	0.848	0.574	0.818	0.155
			260	0.958	0.893	0.628	0.883	0.114	0.948	0.863	0.598	0.840	0.143	0.941	0.844	0.571	0.815	0.158
			160	0.958	0.894	0.626	0.883	0.114	0.948	0.864	0.593	0.840	0.144	0.943	0.850	0.570	0.821	0.154
			80	0.959	0.897	0.619	0.883	0.114	0.950	0.869	0.589	0.844	0.143	0.945	0.854	0.573	0.827	0.147
			40	0.958	0.896	0.617	0.884	0.118	0.947	0.861	0.588	0.838	0.147	0.943	0.850	0.563	0.820	0.163
		20-NN	-	0.957	0.890	0.625	0.880	0.117	0.953	0.874	0.623	0.861	0.126	0.956	0.887	0.613	0.873	0.116
			460	0.962	0.902	0.639	0.897	0.104	0.953	0.874	0.620	0.860	0.126	0.959	0.896	0.623	0.881	0.109
			260	0.962	0.903	0.639	0.898	0.104	0.953	0.874	0.619	0.859	0.127	0.957	0.891	0.618	0.873	0.113
			160	0.960	0.897	0.632	0.889	0.109	0.952	0.872	0.613	0.857	0.129	0.957	0.889	0.612	0.872	0.117
			80	0.964	0.908	0.631	0.897	0.101	0.956	0.884	0.613	0.866	0.124	0.956	0.887	0.603	0.864	0.121
			40	0.964	0.908	0.631	0.897	0.101	0.952	0.872	0.611	0.856	0.133	0.954	0.880	0.587	0.858	0.134

Σχήμα 7.50. Αποτελέσματα ομαδοποίησης με τον αλγόριθμο K-Μέσων και K-Συνθετικών Κέντρων τριών διαφορετικών αρχικοποιήσεων, συλλογή *U*.

			Μοντέλα χωρίς ακμές					Μοντέλα με ακμές				
Αλγ/μο	Κέντρα	Φιλτ/σμα κέντρων	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
HAC			0.907	0.670	0.673	0.745	0.256	0.911	0.676	0.744	0.761	0.226
Global-KM	medoid	-	0.952	0.877	0.594	0.859	0.117	0.970	0.939	0.649	0.938	0.089
	mean	-	0.986	0.964	0.605	0.963	0.054	0.982	0.945	0.651	0.944	0.072
	5-NN	160	0.975	0.948	0.593	0.947	0.071	0.971	0.939	0.647	0.937	0.088
	10-NN	160	0.988	0.968	0.607	0.967	0.052	0.986	0.964	0.605	0.963	0.054

Σχήμα 7.51. Αποτελέσματα ομαδοποίησης με τον αλγόριθμο *HAC*, τον Γενικευμένο K-Μέσων και τον Γενικευμένο K-Συνθετικών Κέντρων, συλλογή *U*.

				Αρχικοποίηση Κέντρων															
Περιγραφή πειράματος				Μ-μακρινότερα κείμενα					Μ-τυχαία κείμενα					Τυχαία ανάθεση κειμένων				
Αλγ/μος	Ακμές	Κέντρα	Φιλτ/σμα κέντρων	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
K-Μέσων	-	mean	-	0.851	0.529	0.170	0.513	0.525	0.855	0.516	0.161	0.503	0.536	0.849	0.452	0.074	0.447	0.630
	+	mean	-	0.852	0.532	0.189	0.519	0.525	0.855	0.509	0.164	0.495	0.543	0.843	0.414	0.035	0.417	0.669
			460	0.855	0.539	0.197	0.525	0.515	0.859	0.523	0.190	0.508	0.522	0.848	0.440	0.071	0.439	0.635
			260	0.858	0.548	0.208	0.531	0.504	0.859	0.523	0.190	0.508	0.522	0.857	0.484	0.131	0.479	0.578
			160	0.863	0.561	0.220	0.546	0.488	0.863	0.537	0.211	0.524	0.502	0.858	0.489	0.151	0.483	0.563
			80	0.870	0.583	0.242	0.566	0.461	0.868	0.554	0.232	0.539	0.483	0.863	0.524	0.183	0.507	0.527
			40	0.873	0.597	0.251	0.578	0.446	0.872	0.573	0.248	0.556	0.464	0.865	0.535	0.204	0.518	0.513
		medoid	-	0.852	0.538	0.191	0.520	0.546	0.853	0.524	0.166	0.497	0.557	0.849	0.494	0.121	0.453	0.601
		5-NN	-	0.864	0.588	0.232	0.564	0.492	0.858	0.558	0.197	0.527	0.518	0.864	0.568	0.214	0.530	0.509
			460	0.864	0.588	0.232	0.564	0.492	0.858	0.558	0.197	0.527	0.518	0.864	0.568	0.214	0.530	0.509
			260	0.865	0.590	0.231	0.566	0.490	0.859	0.561	0.195	0.531	0.515	0.863	0.566	0.208	0.525	0.514
			160	0.863	0.589	0.222	0.565	0.494	0.861	0.567	0.199	0.537	0.511	0.862	0.564	0.204	0.522	0.517
			80	0.862	0.586	0.221	0.560	0.501	0.860	0.559	0.197	0.531	0.521	0.862	0.564	0.203	0.525	0.516
			40	0.861	0.582	0.221	0.555	0.508	0.859	0.558	0.195	0.529	0.528	0.863	0.566	0.203	0.526	0.520
		10-NN	-	0.872	0.605	0.251	0.580	0.461	0.870	0.585	0.241	0.557	0.475	0.872	0.597	0.251	0.558	0.470
			460	0.872	0.608	0.250	0.580	0.456	0.869	0.585	0.241	0.555	0.476	0.873	0.599	0.252	0.559	0.468
			260	0.872	0.608	0.249	0.580	0.456	0.870	0.583	0.239	0.554	0.475	0.873	0.605	0.256	0.564	0.464
			160	0.873	0.611	0.250	0.582	0.454	0.871	0.592	0.240	0.561	0.470	0.871	0.597	0.243	0.552	0.473
			80	0.874	0.618	0.253	0.587	0.455	0.871	0.592	0.236	0.559	0.475	0.872	0.601	0.253	0.559	0.474
			40	0.871	0.609	0.244	0.580	0.466	0.869	0.588	0.234	0.555	0.484	0.870	0.591	0.240	0.547	0.485
		15-NN	-	0.879	0.609	0.266	0.589	0.449	0.877	0.605	0.257	0.575	0.452	0.880	0.617	0.276	0.577	0.445
			460	0.878	0.617	0.261	0.588	0.443	0.878	0.604	0.260	0.576	0.449	0.880	0.617	0.276	0.572	0.444
			260	0.878	0.623	0.261	0.594	0.437	0.878	0.605	0.261	0.575	0.451	0.878	0.613	0.269	0.571	0.448
			160	0.880	0.626	0.266	0.596	0.434	0.879	0.608	0.263	0.578	0.446	0.881	0.624	0.276	0.578	0.440
			80	0.880	0.631	0.267	0.600	0.432	0.879	0.611	0.262	0.579	0.447	0.880	0.621	0.270	0.578	0.446
			40	0.879	0.628	0.265	0.594	0.440	0.878	0.614	0.261	0.579	0.449	0.880	0.621	0.270	0.578	0.446

Σχήμα 7.52. Αποτελέσματα ομαδοποίησης με τον αλγόριθμο K-Μέσων και K-Συνθετικών Κέντρων τριών διαφορετικών αρχικοποιήσεων, συλλογή *R*.

			Μοντέλα χωρίς ακμές					Μοντέλα με ακμές				
Αλγ/μο	Κέντρα	Φιλτ/σμα κέντρων	Rand	Purity	Silh.	F	E	Rand	Purity	Silh.	F	E
HAC			0.808	0.507	0.266	0.544	0.566	0.831	0.541	0.334	0.586	0.499
Global-KM	medoid	-	0.877	0.622	0.216	0.607	0.464	0.860	0.556	0.222	0.516	0.520
	mean	-	0.677	0.370	0.401	0.444	0.646	0.880	0.652	0.301	0.602	0.383
	5-NN	160	0.884	0.678	0.217	0.637	0.444	0.880	0.644	0.273	0.609	0.450
	10-NN	160	0.882	0.652	0.254	0.603	0.436	0.886	0.656	0.289	0.611	0.398
	15-NN	160	0.891	0.667	0.31	0.625	0.379	0.909	0.689	0.302	0.672	0.351

Σχήμα 7.53. Αποτελέσματα ομαδοποίησης με τον αλγόριθμο *HAC*, τον Γενικευμένο K-Μέσων και τον Γενικευμένο K-Συνθετικών Κέντρων, συλλογή *R*.

Τα αποτελέσματα από όλες τις συλλογές επιβεβαιώνουν την ανάλυση του Κεφαλαίου 6, για το φιλτράρισμα των κέντρων και τη δημιουργία συνθετικών κέντρων βάσει των *K*-γειτόνων του ενδιαμέσου.

Σε όλες τις συλλογές το φιλτράρισμα πάνω στον αριθμητικό μέσο αποδείχτηκε καλύτερης ποιότητας κέντρο για την εκπαίδευση. Αυτό επιβεβαιώνεται από τα αποτελέσματα του αλγόριθμου K-Μέσων, ανεξαρτήτου αρχικοποίησης. Η ποιότητα της ομαδοποίησης διαφαίνεται να είναι σε γενικές γραμμές ανάλογη της αυστηρότητας του κατωφλίου φιλτραρίσματος του μέσου. Ένα εκπληκτικό αποτέλεσμα είναι πως από τα πειράματα με τον αριθμητικό μέσο ως κέντρο αναφοράς σε όλες τις περιπτώσεις τα συνθετικά κέντρα που δημιουργούνται από τους αντιπροσώπους των 40 σημαντικότερων λέξεων του (και των ακμών τους) κάθε ομάδας, είναι αυτά που εκπαιδεύουν το σύστημα καλύτερα με τον αλγόριθμο K-Μέσων. Σημειώστε ότι ο αριθμητικός μέσος μπορεί να έχει χαρακτηριστικά της τάξης των αρκετών εκατοντάδων ή και χιλιάδων. Η σημαντικότητα της παρατήρησης αυτής ενισχύεται από την ανεξαρτησία της από την αρχικοποίηση του K-Μέσων.

Ο ενδιάμεσος είναι μία επιλογή που διαχωρίζει τις ομάδες, αρκετές φορές καλύτερα από τον αριθμητικό, όταν ο τελευταίος δε φιλτράρεται. Όμως παρουσιάζει ευαισθησία ως προς την ποιότητα του συνόλου δεδομένων και την αρχικοποίηση των ομάδων. Για παράδειγμα στη συλλογή *U* ο K-Ενδιαμέσων έχει χειρότερη απόδοση από τον μέσο και με τις τρεις αρχικοποιήσεις. Ακόμα, δε συγκρίνεται με τα συνθετικά κέντρα, είτε αυτά βασίζονται στο φιλτράρισμα στον μέσο, είτε στην *K-NN* προσέγγιση.

Τα συνθετικά κέντρα βάσει της *K-NN* προσέγγισης αποδεικνύονται αρκετά ανθεκτικά. Η ρύθμιση του φιλτραρίσματος, με δεδομένο το *K* του μεγέθους της γειτονιάς, δεν παίζει καθοριστικό ρόλο στην απόδοση τους, αν και γενικά το μέγεθος του αντιπροσώπου θα πρέπει είναι περιορισμένο. Αντίθετα το *K* παίζει πιο σημαντικό ρόλο στην απόδοση της τεχνικής, με την έννοια ότι θα πρέπει να είναι αντίστοιχο του μεγέθους της συλλογής. Έχοντας μία σχηματισμένη ομάδα και ένα κατώφλι φιλτραρίσματος που περιορίζει τον αντιπρόσωπο στα σημαντικά χαρακτηριστικά, τότε αυξάνοντας το *K* ουσιαστικά αυξάνουμε την περιοχή δειγματοληψίας των αντιπροσωπευτικών λέξεων. Επειδή η αύξηση του αφορά τη περιοχή του χώρου γύρω από τον ενδιάμεσο, δεν αλλάζει δραματικά την σχετική κατανομή των χαρακτηριστικών στο συνθετικό κέντρο, απλά διαμορφώνει μία καλύτερη εικόνα για τη

σημαντικότητα των όρων της ομάδας. Ακόμα, η ύπαρξη ακμών βοηθά την αναπαράσταση των περιεχομένων μίας ομάδας, η οποία λειτουργεί ως επιπλέον βάρος πάνω στα αντιπροσωπευτικά χαρακτηριστικά της.

Γενικά ο ορισμός του *K* θα πρέπει να λαμβάνει σοβαρά υπόψη το μέγεθος της συλλογής. Η φιλοσοφία που μπορεί να βοηθήσει στον ορισμό του *K* και του κατωφλίου φιλτραρίσματος είναι πως θέτουμε ένα αυστηρό κατώφλι στον κεντροειδές και ένα όχι ιδιαίτερα αυστηρό *K*. Ο ορισμός του ώστε $0.5 \cdot |D| \leq M \cdot K \leq (4/5) \cdot |D|$ είναι μία καλή επιλογή, ώστε να απορρίπτονται κάποια ακραία κείμενα μίας ομάδας.

Οι αλγόριθμοι *HAC* και K-Μέσων χωρίς φιλτράρισμα έχουν παρόμοια συμπεριφορά. Τα αποτελέσματα που αναφέρουμε για τον δεύτερο αφορούν 50 εκτελέσεις, συνεπώς μπορούμε να υποστηρίξουμε ως αναζητώντας στις λύσεις αυτές σίγουρα μπορούμε να εντοπίσουμε λύσεις καλύτερες από αυτή του *HAC*.

Αν και διαφαίνεται θετική στις περισσότερες περιπτώσεις η ύπαρξη ακμών στα μοντέλα, προκύπτει ένα ζήτημα που αφορά την αρχικοποίηση του K-Μέσων. Όταν χρησιμοποιούμε τον μέσο και αρχικοποιούμε με οποιοδήποτε τρόπο εκτός των M-μακρινότερων κειμένων, οι ακμές επιδεινώνουν τις αρχικές συνθήκες. Αυτό όπως εξηγήσαμε έχει να κάνει με την αύξηση της διάστασης του προβλήματος. Παρόλα αυτά τα συνθετικά κέντρα καταφέρνουν να ξεπεράσουν το πρόβλημα αυτό διότι από κάθε ομάδα επιλέγουν έναν υποχώρο αναπαράστασης.

Αξιοσημείωτο είναι επίσης το γεγονός ότι η ομοιόμορφη αρχικοποίηση, αν και γενικά κατώτερης ποιότητας προσέγγιση από τις άλλες εκδοχές, δεν επηρεάζει τα συνθετικά κέντρα ιδιαίτερα. Με κατάλληλα ορισμένο το *K* τα συνθετικά κεντροειδή καταφέρνουν να διαχωρίσουν τις ομάδες καλύτερα και από τους ενδιάμεσους και τις εκδοχές φιλτραρίσματος του αριθμητικού. Τα αποτελέσματα του κάποιες φορές είναι συγκρίσιμα με αυτά της αρχικοποίησης M-τυχαίων κειμένων.

Ο Γενικευμένος αλγόριθμος K-Μέσων χρήσει του αριθμητικού μέσου και πέραν της συλλογή *U*, εμφανίζει τη δυσκολία διαχωρισμού των ομάδων που εξηγήσαμε λόγω της συσσώρευσης. Από την άλλη πλευρά η χρήση του ενδιαμέσου δίνει ικανοποιητικά αποτελέσματα.

Ο αλγόριθμος αυτός γίνεται πολύ αποτελεσματικός όταν χρησιμοποιούμε συνθετικά κέντρα. Στην περίπτωση αυτή παρουσιάζει τα καλύτερα αποτελέσματα σε όλες τις συλλογές. Να σημειώσουμε πως σε δοκιμές που έγιναν στις οποίες

χρησιμοποιήθηκε φιλτράρισμα του αριθμητικού μέσου (τα αποτελέσματα δεν παρουσιάζονται παραπάνω), δεν είχε την ίδια αποτελεσματικότητα με αυτή της *K-NN* προσέγγισης, παρότι έδειχνε ξεκάθαρη βελτίωση σε σχέση με το μη φιλτράρισμα του μέσου.

ΚΕΦΑΛΑΙΟ 8. ΣΥΝΟΨΗ ΣΥΜΠΕΡΑΣΜΑΤΩΝ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Στην εργασία αυτή μελετήθηκε και παρουσιάστηκε το πρόβλημα της ομαδοποίησης κειμένων. Ο πρώτος στόχος από πλευράς παρουσίασης ήταν να αναλυθούν τα βασικά ζητήματα που αφορούν το πρόβλημα μέσα από το πρίσμα της σύγχρονης βιβλιογραφίας και σχετικής ερευνητικής δραστηριότητας. Τα κεφάλαια της εργασίας αφορούν ξεχωριστές διαστάσεις του προβλήματος, η γνώση των οποίων θεωρούμε πως βοηθούν στην καλύτερη προσέγγιση του προβλήματος.

Μέσα από αυτή τη διαδρομή δεύτερος στόχος ήταν να παρουσιάσουμε μια σειρά τεχνικών οι οποίες αναπτύχθηκαν στα πλαίσια της εργασίας και αντιμετωπίζουν διάφορα από τα ζητήματα που αναλύονται. Στο κεφάλαιο αυτό θα κάνουμε μία ανακεφαλαίωση των αποτελεσμάτων και των κύριων συμπερασμάτων που αναφέρουμε στην εργασία, και θα αναφέρουμε τις πιθανές κατευθύνσεις για περαιτέρω διερεύνηση του προβλήματος.

8.1. Σύνοψη Αποτελεσμάτων και Συμπερασμάτων

Παραγωγή Τεχνητών Κειμένων

Αφού έγινε μία εκτενής ανάλυση πάνω στα ιδιαίτερα χαρακτηριστικά των κειμένων και τις στατιστικές ιδιότητές τους, προτάθηκε μία αλγοριθμική διαδικασία για την παραγωγή συνθετικών δεδομένων. Η διαδικασία αυτή συνδυάζει ένα πρότυπο μικτό

μοντέλο επιλογής όρων από ένα σύνολο θεματικών λεξιλογίων, και μία στοχαστική διαδικασία *Zanette-Montemurro* που προσομοιώνει την παραγωγή κειμένου το οποίο ασυμπτωτικά συμφωνεί με τον εμπειρικό νόμο του *Zipf*. Η βασική πρόκληση ήταν να παράγουμε δεδομένα τα οποία πέρα από ρεαλιστικές στατιστικές ιδιότητες να παρουσιάζουν και δομή σε επίπεδο ομάδων. Η διαδικασία παραγωγής που παρουσιάστηκε δίνει τη δυνατότητα καθορισμού μη-ομοιόμορφης «δυσκολίας» στη συλλογή κειμένων μέσα από ένα σύνολο παραμέτρων ρύθμισης.

Μοντέλο Αναπαράστασης Γραφήματος

Το μοντέλο αναπαράστασης μας απασχόλησε ιδιαίτερα, αν και στη βιβλιογραφία επικρατεί η αναπαράσταση των όρων των κειμένων σε διανύσματα. Βάσει της λογικής ότι τα κείμενα περιέχουν πληροφορία που δεν περιορίζεται στην κατανομή των λέξεων την οποία και θα πρέπει να εκμεταλλευτούμε, ορίσαμε ένα γενικευμένο μοντέλο γραφημάτων το οποίο αναπαριστά ως κόμβους τους όρους του κειμένου και ακμές τις σχέσεις ανάμεσα σε αυτούς. Η βασική υπόθεση είναι πως το γράφημα αυτό μπορεί να θεωρείται σύνολο ανεξάρτητων χαρακτηριστικών, τα επιμέρους στοιχεία του γραφήματος. Η γενικότητα του μοντέλου έγκειται στο ότι οι σχέσεις μπορεί να ορίζονται από διαδικασίες εξόρυξης δεδομένων από τα κείμενα και προσδιορίζονται από μοναδικές ετικέτες ακμών. Για παράδειγμα, μπορούν να εισαχθούν σε ένα γράφημα, επιπρόσθετες σχέσεις που εξάγονται από κάποιο εργαλείο (λεξικό, γραμματική τύπου *WordNet*) οι οποίες, ορίζοντας μία κατάλληλη συνάρτηση ανάθεσης βάρους, να αντιμετωπίζονται ως «ένα ακόμα στοιχείο γραφήματος».

Καθορίσαμε μία ειδική κλάση τέτοιων γραφημάτων όπου ως σχέση αναγνωρίζεται η μη-κατευθυνόμενη διαδοχή δύο όρων στο κείμενο. Ο ισχυρισμός μας είναι πως επειδή ο γραπτός λόγος μπορεί να παράγει πολλές νοηματικά ισοδύναμες διατυπώσεις με τους ίδιους βασικούς λεκτικούς όρους, αλλά με διαφορετική αλληλουχία (ή παρεμβολή δευτερεύουσας σημασίας όρων), η ακριβής καταγραφή πλήρων μονοπατιών δεν είναι ουσιώδης πληροφορία. Βάσει της υπόθεσης για ανεξαρτησία των χαρακτηριστικών δίνεται η δυνατότητα για διανυσματοποίηση των κειμένων η οποία βοηθάει στην καλύτερη και αυστηρότερη ανάλυση του προβλήματος χρήσει διανυσματικής άλγεβρας. Ακόμα, αναλύθηκε το σχήμα δομών δεδομένων για την αναπαράσταση των γραφημάτων που χρησιμοποιούμε, το οποίο παρέχει τη δυνατότητα αναζήτησης σχέσεων και όρων σε $O(1)$ χρόνο και τη

γραμμικού χρόνου συσχέτιση δύο γραφημάτων O(|T| + |E|) ως προς το μέγεθος του μικρότερου από τα συσχετιζόμενα γραφήματα.

Η σχέση μη-κατευθυνόμενης γειτνίασης που ορίσαμε δε δημιουργεί ζήτημα αύξησης της τάξης πολυπλοκότητας του προβλήματος. Οι ακμές που αναγνωρίζονται σε ένα κείμενο αρχικά είναι τουλάχιστον διπλάσιες από τους όρους, όμως ύστερα από την προεπεξεργαστική φάση καθίστανται κατά πολύ λιγότερες από αυτούς.

Η Πληροφορία των Σχέσεων Γειτνίασης και οι Συναρτήσεις Ομοιότητας

Σε συνδυασμό με τα βάρη στις ακμές βάσει του τύπου: $w_e = f_e \cdot (1 + ln(min(w_1, w_2))$, που εξαρτάται από την συχνότητα της ακμής f_e και το λογάριθμο του μικρότερου από τα βάρη w_{12}, w_2 των σχετιζόμενων όρων, οι ακμές έδειξαν πειραματικά να ενισχύουν την ομοιότητα μεταξύ των αντικειμένων κάθε κατηγορίας και να δημιουργούν πιο ποιοτική πληροφορία σε επίπεδο κοντινότερων γειτόνων (ενοποιημένο μοντέλο χαρακτηριστικών). Επίσης, η παρατήρηση αυτή επιβεβαιώνεται και για υποσύνολα χαρακτηριστικών των κειμένων. Σε όλα τα σύνολα δεδομένων οι κατηγορίες γίνονται περισσότερο διαχωρίσιμες όταν λαμβάνουμε υποσύνολα ισχυρών χαρακτηριστικών από τα κείμενα. Ο ορισμός του κατάλληλου υποσυνόλου χαρακτηριστικών για κάθε κείμενο είναι δύσκολο να καθοριστεί χωρίς επίβλεψη, παρόλα αυτά υποδεικνύει την ανάγκη λύσης του προβλήματος σε κάποιο υποχώρο χαρακτηριστικών.

Στο Κεφάλαιο 4 η βασική συμβολή της εργασίας ήταν να αναλύσουμε τη συμπεριφορά της δημοφιλούς συνημιτονοειδούς ομοιότητας και του μέτρου της γραφοθεωρητικής ομοιότητας, το οποίο προτάθηκε πιο πρόσφατα στη βιβλιογραφία. Για το δεύτερο προτείναμε την εισαγωγή των βαρών των χαρακτηριστικών στον υπολογισμό της ομοιότητας, και έτσι κατά το ταίριασμα των γράφων δύο κειμένων η ομοιότητα εκφράζεται ως το ποσοστό του κοινού βαρυντικού περιεχομένου. Συγκριτικά με τη συνημιτονοειδή ομοιότητα συμπεράναμε την ακαταλληλότητα των προσθετικών υπολογισμών των στοιχειωδών ομοιοτήτων για κάθε κοινό χαρακτηριστικό των κειμένων. Η βάση του συλλογισμού είναι η ισχύς του εμπειρικού νόμου του *Zipf* και περιέχονται σε ένα κείμενο πολλά χαρακτηριστικά χαμηλής συχνότητας, τα οποία αθροιστικά μπορούν να υπερκεράσουν τη σημαντικότητα των πραγματικά «σημαντικών» χαρακτηριστικών.

Το μειονέκτημα αυτό επιβεβαιώθηκε στην ίδια σειρά πειραμάτων. Η γραφοθεωρητική συνάρτηση χάνει την «αίσθηση» της σημαντικότητας των

χαρακτηριστικών όσο λαμβάνουμε υπόψη μεγαλύτερα τμήματα κειμένων στους υπολογισμούς της ομοιότητας. Η τροποποίηση μας, η οποία εισάγει βάρη, έχει κατά πολύ καλύτερη συμπεριφορά όμως αποδεικνύεται πως απλά μετατοπίζει το πρόβλημα. Συμπεραίνουμε πως η παθογένεια της γραφοθεωρητικής συνάρτησης έγκειται στους προσθετικούς υπολογισμούς που αναφέραμε. Το συμπέρασμα αυτό μπορεί να επεκταθεί και σε άλλες συναρτήσεις που έχουν την ίδια φιλοσοφία για τον υπολογισμό της τομής περιεχομένου. Η συνημιτονοειδής συνάρτηση φαίνεται ανθεκτική στην θορυβώδη πληροφορία των κειμένων και έτσι αποτέλεσε το μέτρο ομοιότητας για την υπόλοιπη πειραματική μελέτη.

Παρουσιάστηκε το ζήτημα της μίξης της σημαντικότητας των ακμών στον υπολογισμό της ομοιότητας και με κατάλληλα πειράματα δείξαμε τη συμπεριφορά ενός παραδοσιακού αλγόριθμου ομαδοποίησης κειμένων (*HAC*) με διαφορετικές τιμές του συντελεστή μίξης (*blending factor - bf*). Για το ζήτημα αυτό προτάθηκε μία ευρετική τεχνική αναγνώρισης μίας κατάλληλης (αλλά όχι βέλτιστης) τιμής μίξης. Διακριτοποιώντας τις τιμές του *bf* εφαρμόζουμε μία διαδικασία επικύρωσης των λύσεων μίξης. Θεωρούμε ως λύσεις αναφοράς τις $bf = 1$ ή/και $bf = 0$, έτσι για να εκτιμηθεί η ποιότητά μίας άλλης λύσης με $0 < bf < 1$ αφαιρούνται οι ακμές από τα μοντέλα (αντίστοιχα οι όροι) και συγκρίνουμε κάποιο δείκτη εκτίμησης χωρίς επίβλεψη. Επιλέγοντας ανάμεσα σε τρία κριτήρια: ένα που αφορά τους όρους, ένα τις ακμές και ένα συνδυαστικό, μπορούμε να επιλέξουμε μία καλή παράμετρο μίξης.

Η παραπάνω διαδικασία είναι ιδιαίτερα ακριβή, γι’ αυτό θεωρήσαμε το ενοποιημένο μοντέλο ανεξάρτητων χαρακτηριστικών που αφήνει τα δεδομένα να καθορίσουν τη σχέση περιεχομένου ανάμεσα στους όρους και τις ακμές. Το μοντέλο αυτό παρουσιάζει καλή συμπεριφορά, όταν οι ακμές βοηθούν τον αλγόριθμο τότε οι λύσεις χρήσει του μοντέλου αυτού δεν είναι χειρότερες από τη λύση που αγνοεί τις ακμές. Σε αντίθετη περίπτωση υπάρχει η πιθανότητα να πάρουμε συγκρίσιμες ή και υποδεέστερες λύσεις. Παρόλα αυτά η απλότητά του, η καλή συμπεριφορά του και η αυξημένη πολυπλοκότητα εφαρμογής μεθόδων εκτίμησης του συντελεστή μίξης, το καθιστούν μία καλή επιλογή. Το ενοποιημένο μοντέλο χρησιμοποιήθηκε σε όλα τα πειράματα της εργασίας , πλην αυτών της μίξης.

Φιλτράρισμα και Συνθετικά κέντρα: Τεχνικές Ενσωμάτωσης Τοπικής Πληροφορίας από τις Γειτονιές Δεδομένων σε Διάφορα Στάδια του Προβλήματος

Ένας από τους βασικούς άξονες της εργασίας ήταν να μελετηθούν τεχνικές οι οποίες θα εκμεταλλεύονταν την τοπική πληροφορία των γειτονιών. Η πληροφορία αυτή είναι ένα στοιχειώδες επίπεδο οργάνωσης των δεδομένων το οποίο είναι προϋπόθεση για την ύπαρξη ομάδων σε αυτά. Έτσι, αναπτύχθηκαν δύο τεχνικές με κοινή αφετηρία.

Η πρώτη τεχνική αφορά τη μείωση των χαρακτηριστικών των μοντέλων η οποία είναι μία έμμεση μείωση διάστασης. Η πρόκληση ήταν η εφαρμογή μίας τέτοιας διαδικασίας χωρίς την ύπαρξη επίβλεψης, όπου οι αντίστοιχες τεχνικές μελετούν τα δεδομένα χρήσει εργαλείων στατιστικής ή της θεωρίας της πληροφορίας. Η φιλοσοφία της είναι να συμβουλεύεται τη γειτονιά ενός κειμένου για να λαμβάνονται αποφάσεις για την απαλοιφή ασθενών συχνοτικά όρων. Το αποτέλεσμα είναι η εξασθένιση της ομοιότητας ενός κειμένου με τα κείμενα που δε μοιάζουν με τη γειτονιά του. Οι παράμετροι, μεγέθους γειτονιάς και κατωφλίου, το οποίο διαχωρίζει του ασθενείς από του ισχυρούς όρους ενός κειμένου, καθορίζουν την ένταση του φιλτραρίσματος.

Η τεχνική αυτή συγκρίθηκε με την «τυφλή» αποκοπή των ασθενών όρων της κατωφλίωσης *TDF* εφαρμόζοντας τον *HAC*. Τα συμπεράσματα είναι πως η κατωφλίωση *TDF* πετυχαίνει μείωση των χαρακτηριστικών των μοντέλων κατά 50-85%, χωρίς κάτι τέτοιο να αποβαίνει πάντα καταστροφικό για τα δεδομένα. Παρόλα αυτά τις περισσότερες φορές υποβαθμίζει την ποιότητα τους επηρεάζοντας κατ' επέκταση και τη διαδικασία εκπαίδευσης.

Από την άλλη πλευρά το *K-NN* Φιλτράρισμα δίνει τη δυνατότητα να αφαιρέσουμε ικανοποιητικές ποσότητες χαρακτηριστικών (10 έως 35%) διατηρώντας την ποιότητα της τοπικής πληροφορίας, με αποτέλεσμα ο *HAC* να επωφελείται, ή να μη επηρεάζεται αρνητικά. Το βασικό χαρακτηριστικό της είναι ότι «σέβεται» τα δεδομένα. Η μείωση των χαρακτηριστικών εξαρτάται από τις ιδιότητες της συλλογής και τη συνέπεια των γειτονιών. Όπως φάνηκε πειραματικά ο ακριβής ορισμός των παραμέτρων δεν παίζει ιδιαίτερα καθοριστικό ρόλο, ενώ ακόμα και σε πειράματα που έγιναν σε μία συλλογή κειμένων με μέτρια συνέπεια γειτονιών (συλλογή *R*) παρατηρήθηκε πως διατηρεί τη δυνατότητα να ενισχύει τη δομή των ομάδων.

Η δεύτερη τεχνική, και από τα σημαντικότερα αποτελέσματα της εργασίας αυτής, είναι η πρότυπη προσέγγιση η οποία ενσωματώνει μία διαδικασία δυναμικής επιλογής χαρακτηριστικών στην οικογένεια αλγορίθμων Κ-Μέσων. Η διαδικασία αυτή τροποποιεί εισάγει έναν πρότυπο ορισμό για το κέντρο μίας ομάδας. Παρότι από μαθηματικής πλευράς ο αριθμητικός μέσος είναι το βέλτιστο κέντρο για μία ομάδα, απέχοντας την ελάχιστη απόσταση από τα δεδομένα της, στην εργασία αυτή αμφισβητήθηκε η καταλληλότητά του ως αντιπρόσωπο της. Οι λόγοι που παραθέσαμε αφορούν τη φύση του προβλήματος και των δεδομένων τα οποία περιέχουν πολλά θορυβώδη και πλεονάζοντα χαρακτηριστικά.

Για κέντρα ορίσαμε συνθετικούς αντιπροσώπους που προκύπτουν βάση δύο επιλογών: α) το αρχικό κέντρο αναφοράς και β) το φιλτράρισμα πάνω σε αυτό. Το κέντρο αναφοράς μπορεί να είναι είτε ο αριθμητικός μέσος, είτε ο ενδιάμεσος μαζί με τους *K-NN* γείτονές του. Ποιοτικά θα λέγαμε πως με τη διαδικασία που προτάθηκε κατά την εκπαίδευση των κέντρων των ομάδων εντοπίζεται δυναμικά ένας υποχώρος χαρακτηριστικών στον οποίο τελικά λύνεται το πρόβλημα, αγνοώντας σημαντικό αριθμό χαρακτηριστικών των ομάδων. Η τεχνική αυτή αποφασίζει τη χρησιμότητα των χαρακτηριστικών σε ένα περιβάλλον συμφραζομένων, τις ομάδες κειμένων, περιορίζοντας σημαντικά την επίδραση του θορύβου στη λύση. Επίσης, ο ενδιάμεσος με τους γείτονές του δημιουργεί πιο «ειδικό» κέντρο για τα χαρακτηριστικά μίας κατηγορίας, χωρίς να επηρεάζεται ιδιαίτερα από το βαθμό ομοιογένειας της ομάδας.

Πειραματικά οι τεχνικές αυτές παρουσίασαν πολύ ενθαρρυντικά αποτελέσματα. Ο αριθμητικός μέσος σε περιπτώσεις κακής αρχικοποίησης, ή περιπτώσεις δύσκολα διαχωρίσιμων ομάδων, παρουσιάζει προβλήματα συσσώρευσης που πηγάζουν από το φαινόμενο της αυτό-ομοιότητας. Αντίθετα οι προσεγγίσεις των συνθετικών κέντρων αποδεικνύονται πιο ανθεκτικές τεχνικές που ταιριάζουν καλύτερα στο πρόβλημα. Η επιλογή των γειτόνων από το εσωτερικό της ομάδας ενισχύει την πιθανότητα να επιλεγούν τελικά περίπου *K* κείμενα της ίδιας κατηγορίας με τον ενδιάμεσο, συνεπώς η αύξηση του *K* σε μία ομοιογενή γενικά ομάδα δεν αναμένεται να αλλοιώνει την κατανομή των σημαντικών χαρακτηριστικών του κέντρου. Για το λόγο αυτό ο ορισμός του *K* δεν είναι από τα ευαίσθητα σημεία της τεχνικής, αρκεί να λαμβάνεται υπόψη το μέγεθος της συλλογής.

Η τεχνική των συνθετικών κέντρων ενσωματώθηκε στον Γενικευμένο Κ-Συνθετικών Κέντρων δίνοντας τα καλύτερα αποτελέσματα σε όλες τις συλλογές

κειμένων. Στην παραδοσιακή του διατύπωση του αριθμητικού μέσου κατά τη διάσπαση των ομάδων μπορεί να θεωρηθεί πως, ο αλγόριθμος αυτός ορίζει ένα πρόβλημα ομαδοποίησης με κακή αρχικοποίηση, διότι η νέα ομάδα έχει ένα μόνο δεδομένο. Αυτός είναι και ο λόγος που σε κάποια πειράματα ο αλγόριθμος αυτός είναι συγκριτικά κατώτερος από τις πολύ απλούστερες προσεγγίσεις μεθόδων.

Αμέσως μετά σε απόδοση έρχεται ο Κ-Συνθετικών Κέντρων με την αρχικοποίηση Μ-μακρινότερων κειμένων που δοκιμάσαμε και συστήνουμε για το πρόβλημα. Και εδώ οι καλύτερες λύσεις παρατηρούνται σε συνδυασμό με τα συνθετικά κέντρα όπου φαίνεται ξεκάθαρα η βελτίωση των λύσεων. Δοκιμάστηκαν τρεις διαφορετικοί τρόποι αρχικοποίησης οι οποίοι επιβεβαίωσαν, ο καθένας με τον δικό του τρόπο, τα συμπεράσματα που αναφέραμε. Ο ιεραρχικός συσσωρευτικός δίνει σχετικά καλά αποτελέσματα συγκρινόμενος όμως μόνο με τον αριθμητικό μέσο.

Στα αρνητικά της τεχνικής καταγράφεται η ύπαρξη παραμέτρων που ορίζονται από τον χρήστη, ο ελαφρώς πολυπλοκότερος υπολογισμός κέντρου (αν και υπάρχει αντισταθμιστικό κέρδος από τη μέτρηση της ομοιότητας ενός μικρότερου αντιπροσώπου με τα *N* κείμενα κατά την ανάθεση των δεδομένων στις ομάδες), και η μη-μονότονη σύγκλισή του. Ιδιαίτερα για το τελευταίο θα πρέπει να πούμε πως δεν είναι έντονο ως φαινόμενο και πως η συχνότητα του είναι αντιστρόφως ανάλογη της ποσότητας των χαρακτηριστικών των ομάδων που αγνοούμε κατά τον ορισμό των κέντρων. Με άλλα λόγια, όσο απομακρυνόμαστε από τον αριθμητικό μέσο τόσο πιθανότερος είναι ο πρόωρος τερματισμός λόγω απόκλισης.

Μια πιο αναλυτική συζήτηση πάνω στα αποτελέσματα υπάρχει στα αντίστοιχα κεφάλαια των ζητημάτων και στο Κεφάλαιο 7, ύστερα από κάθε σειρά πειραμάτων. Στο Σχήμα 8.1 παρουσιάζουμε τα καλύτερα αποτελέσματα των αλγορίθμων: *HAC*, Κ-Μέσων, Κ-Συνθετικών Κέντρων με αρχικοποίηση Μ-μακρινότερων κειμένων, Γενικευμένου Κ-Μέσων και Γενικευμένου Κ-Συνθετικών Κέντρων. Η επιλογή έγινε βάσει των δεικτών με επίβλεψη. Αναγράφονται οι παράμετροι των πειραμάτων, ενώ η ένδειξη '[160]' δηλώνει ότι διεξήχθησαν πειράματα μόνο με τη ρύθμιση αυτή.

Περιγραφή πειράματος									
Αλγ/μος	Συλλογή	Ακμές	Κέντρα	Φιλτ/σμα κέντρων	Rand	Purity	Silh.	F	E
HAC	F	+			0.892	0.860	0.671	0.845	0.127
K-Μέσων		-	mean	-	0.803	0.737	0.511	0.727	0.257
		+	15-NN	-	0.920	0.897	0.636	0.896	0.124
Global-KM		+	mean	-	0.756	0.742	0.594	0.731	0.292
Global-KM		+	15-NN	[160]	0.914	0.903	0.622	0.903	0.149
HAC	J	+			0.898	0.739	0.508	0.767	0.263
K-Μέσων		+	mean	-	0.912	0.722	0.394	0.722	0.313
		+	10-NN	80	0.939	0.810	0.458	0.815	0.212
Global-KM		+	mean	-	0.916	0.761	0.480	0.763	0.246
Global-KM		+	15-NN	[160]	0.914	0.903	0.622	0.903	0.149
HAC	U	+			0.911	0.676	0.744	0.761	0.226
K-Μέσων		+	mean	-	0.951	0.854	0.616	0.854	0.139
		+	10-NN	80	0.964	0.908	0.631	0.897	0.101
Global-KM		+	mean	-	0.982	0.945	0.651	0.944	0.072
Global-KM		-	10-NN	[160]	0.988	0.968	0.607	0.967	0.052
HAC	R	+			0.831	0.541	0.334	0.586	0.499
K-Μέσων		+	mean	-	0.852	0.532	0.189	0.519	0.525
		+	15-NN	80	0.880	0.631	0.267	0.600	0.432
Global-KM		+	mean	-	0.880	0.652	0.301	0.602	0.383
Global-KM		+	20-NN	[160]	0.909	0.689	0.302	0.672	0.351

Σχήμα 8.1. Τα καλύτερα αποτελέσματα από τις βασικές τεχνικές ομαδοποίησης που εξετάστηκαν στην εργασία.

8.2. Επιλογή Κατάλληλων Μεθόδων για Προβλήματα Ομαδοποίησης Κειμένων

Ως αναφορά την επίλυση συγκεκριμένων προβλημάτων μπορούμε να αναφέρουμε τι θα πρέπει να έχουμε κατά νου για να αποφασίσουμε ποια είναι η καταλληλότερη μέθοδος ή επιμέρους επιλογές ρύθμισης.

Καταρχήν οι ακμές θα πρέπει να χρησιμοποιούνται όταν ομαδοποιούμε με τον *HAC*, υπό την έννοια ότι ενισχύουν τις γειτονιές στις οποίες βασίζεται ιδιαίτερα ο αλγόριθμος αυτός στα αρχικά βήματα συνένωσης ομάδων. Η συνέπεια της πληροφορίας των γειτονιών καθορίζει σε μεγάλο βαθμό την ποιότητα των αποτελεσμάτων του. Αν και θεωρείται από τις κλασικότερες προσεγγίσεις για το πρόβλημα, το μεγάλο μειονέκτημά του είναι πως οι αποφάσεις του είναι τελικές και δε δίνει τη δυνατότητα αυτόνομης μετακίνησης κειμένων προς τις κοντινότερές του ομάδες. Είναι χρήσιμη μέθοδος για να παράγουμε μία λύση αναφοράς, την οποία μπορούμε να συγκρίνουμε στη συνέχεια με λύσεις άλλων αλγορίθμων (βάσει του

κριτηρίου συνοχής). Μία επίσης καλή επιλογή είναι να χρησιμοποιείται για αρχικοποίηση του αλγορίθμου Κ-Μέσων.

Ο Κ-Μέσων έχει καλή συμπεριφορά με την προϋπόθεση ότι αρχικοποιείται κατάλληλα. Οι πολλαπλές εκτελέσεις συνήθως δε μπορούν να αποφευχθούν για την αναζήτηση μίας καλής λύσης. Η αρχικοποίηση με τα Μ-Μακρινότερα κείμενα είναι ίσως από τις καλύτερες επιλογές αρχικοποίησης, η οποία με τον κατάλληλο ορισμό του δείγματος αναφοράς μπορεί να γίνει υπολογιστικά προσιτή ακόμα και για μεγάλα σύνολα δεδομένων. Επίσης, όσο καλύτερη είναι μία αρχικοποίηση τόσο λιγότερα βήματα αναμένεται να κάνει ο αλγόριθμος αυτός μέχρι τον τερματισμό του. Συστήνουμε επίσης τη χρήση των συνθετικών κέντρων με αυστηρό κατώφλι φιλτραρίσματος και επαρκές *K*, στους αντιπροσώπους των ομάδων ώστε να χρησιμοποιείται τουλάχιστον το 50% των κειμένων. Αν πάλι θέλουμε μία πιο επιφανειακή και γρήγορη λύση μπορούμε να χρησιμοποιήσουμε τους ενδιάμεσους, ενδεχομένως με ελαφρύ φιλτράρισμα.

Ο Γενικευμένος Κ-Μέσων μπορεί επίσης να εφαρμοστεί με συνθετικά κέντρα είτε με τον ενδιάμεσο είτε μαζί με την *K-NN* γειτονιά του. Ο ίδιος αλγόριθμος δεν έχει αξιόπιστη συμπεριφορά στο συγκεκριμένο πρόβλημα όταν χρησιμοποιείται με τον αριθμητικό μέσο. Η βελτίωση που παίρνουμε με τον αριθμητικό μέσο σε κάποιες περιπτώσεις οι οποίες μπορεί να οφείλονται στα πολλά κείμενα της συλλογής ή στην ευκολία διαχωρισμού των δεδομένων, δε δικαιολογεί το επιπλέον χρονικό κόστος σε σχέση με τις φθηνότερες επιλογές του Κ-Μέσων ή του *HAC*.

Από όσα διαπιστώθηκαν κατά την πειραματική μελέτη, η καλύτερη εκπαίδευση ενός συστήματος μπορεί να επιτευχθεί με τον Γενικευμένο Κ-Συνθετικών Κέντρων. Αν πάλι θέλουμε μία αξιόπιστη λύση σε λιγότερο χρόνο τότε η καλύτερη επιλογή είναι η εκτέλεση ενός μικρού αριθμού πειραμάτων με τον Κ-Συνθετικών Κέντρων, και η επιλογή του καλύτερου από αυτά βάσει της συνοχής των ομάδων.

Αν διαθέτουμε χρόνο για καλύτερη προεπεξεργασία τότε μπορούμε να εφαρμόσουμε κάποια τεχνική μείωσης των χαρακτηριστικών των μοντέλων. Σε περίπτωση που τα κείμενα είναι αρκετά και σε πλήθος, αλλά διαθέτει και αρκετό αριθμό όρων καθένα από αυτά, μπορούμε να δοκιμάσουμε αποκοπή με κατωφλίωση *TDF* με κατώφλι συχνοτήτων μικρότερο του 4. Είναι δυνατό να ελέγχεται διαισθητικά η καταλληλότητα του κατωφλίου βάσει του αριθμού των χαρακτηριστικών που απαλείφουμε και τον μέσο αριθμό χαρακτηριστικών ανά

κείμενο. Αν παρατηρηθεί μεγάλη μείωση χαρακτηριστικών τότε απλά θα πρέπει να θέσουμε υψηλότερο κατώφλι.

Όταν τα δεδομένα δεν είναι πάρα πολλά ή επιθυμούμε να εφαρμόσουμε μία πιο αποτελεσματική προσέγγιση αφαίρεσης χαρακτηριστικών, μπορούμε να εφαρμόσουμε το *K-NN* Φιλτράρισμα με προσεκτική ρύθμιση του *K* ώστε να δίνεται η δυνατότητα να συμβουλευόμαστε αρκετά κείμενα από την ζητούμενη κατηγορία.

8.3. Κατευθύνσεις Μελλοντικής Εργασίας

Μία πρώτη συζήτηση πάνω σε βελτιώσεις έγινε στα αντίστοιχα κεφάλαια κάθε τεχνικής που μελετήσαμε. Για τα συνθετικά δεδομένα είναι δυνατόν να γίνει μία πιο προσεκτική ανάλυση όλων των παραμέτρων παραγωγής των κειμένων. Το φιλτράρισμα με την *K-NN* προσέγγιση θα μπορούσε να μελετηθεί χρήσει της συνάρτησης ομοιότητας κοινών κοντινότερων γειτόνων όπως αναφέραμε. Επίσης, η εφαρμογή της τεχνικής πάνω στο αποτέλεσμα της ομαδοποίησης είναι μία ενδιαφέρουσα εκδοχή. Η ομαδοποίηση ενισχύει την ποιότητα των K-γειτόνων όταν αυτοί επιλέγονται από το εσωτερικό μίας ομάδας.

Για τα μοντέλα αναπαράστασης το επόμενο βήμα διερεύνησης είναι να πειραματιστούμε πάνω σε ένα *Ακτινικό Μοντέλο*, το οποίο θα εξάγει την τοπική πληροφορία συσχετίσεων των όρων ανεξαρτήτου της θέσης και σειράς τους μέσα σε προτάσεις ή τμήματα κειμένου. Η πληροφορία αυτή θα ενσωματώνεται στο ίδιο γραφικό μοντέλο αναπαράστασης και θα έχει ως αποτέλεσμα να έχουμε καλύτερη εικόνα για το ποιες λέξεις έχουν «κοινές εμφανίσεις».

Η τεχνική συνθετικών κέντρων θα πρέπει να βελτιωθεί στην κατεύθυνση της μηχανικής εύρεσης κατάλληλων παραμέτρων για τους αντιπροσώπους. Μία παράλληλη διαδικασία της ομαδοποίησης θα υπολογίζει την παράμετρο *K* και το κατώφλι φιλτραρίσματος ώστε να εκμεταλλευόμαστε πλήρως την ποιότητα που έχουν οι ομάδες (ομοιογένεια) και να προσεγγίζεται καλύτερα οι κατανομές των αντιπροσωπευτικών όρων τους.

Μία εφαρμογή της είναι να δοκιμαστεί σε συνδυασμό με τον *HAC* ορίζοντας αντίστοιχα συνθετικά κέντρα κατά τη συσσώρευση. Επίσης, ίσως φανεί χρήσιμη και σε προβλήματα αυξητικής ομαδοποίησης όπου τα δεδομένα έρχονται ένα-ένα ή κατά μικρές ομάδες. Στην περίπτωση αυτή, η σειρά εισαγωγής των δεδομένων παίζει

σημαντικό ρόλο και οι αρχικές ομάδες που σχηματίζονται μπορούν να απομακρύνουν την εκπαίδευση από την εύρεση μίας καλής συνολικής λύσης. Τα συνθετικά κέντρα που προτείνουμε έχουν τη δυνατότητα να ορίζονται βάσει της τοπικής πληροφορίας και ενδεχομένως να αποδειχτούν πιο ευέλικτα και στο ειδικό αυτό πρόβλημα.

ΑΝΑΦΟΡΕΣ

[1] T.M. Mitchell. "*Machine Learning*", McGraw-Hill International Editions, 1996.

[2] R. Kosala and H. Blockeel. "*Web mining research: A Survey*", ACM SIGKDD Explorations Newsletter, 2(1), pp 1-15, 2000.

[3] Y. Yang and X. Liu. "*A Re-examination of Text Categorization Methods*", In Proceedings of SIGIR'99, 1999.

[4] F. Sebastiani. "*Machine Learning in Automated Text Categorization*", ACM Computing Surveys, Vol. 34(1), pp.1-47, 2002.

[5] I.S. Dhillon. "*Co-clustering documents and words using a Bipartite Spectral Graph Partitioning*", In Proceedings of SIGKDD'01, 2001.

[6] C.M. Bishop. "*Pattern Recognition and Machine Learning*", Springer, 2006.

[7] D.L. Boley. "*Principal Direction Divisive Partitioning*", Data Mining and Knowledge Discovery, Vol. 2(4), pp 325-344, 1998.

[8] L. Etrőz, M. Steinbach and V. Kumar. "*Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach*", In Proceedings of SIAM'01, 2001

[9] Y. Zhao, G. Karypis, and V. Kumar. "*A Comparison of Document Clustering Techniques*", KDD Workshop on Text Mining, 1999.

[10] P. Berkin. "*Survey of Clustering Data Mining Techniques*", Research paper, Accure Software, 2002.

[11] Y. Zhao and G. Karypis. "*Hierarchical Clustering Algorithms for Document Datasets*", Data Mining and Knowledge Discovery, 10, 141-168, 2005.

[12] J.Ghosh and A. Stehl. "*Similarity-Based Clustering: A Comparative Study*", Grooуping Multidimensional Data, Springer, pp 73-97, 2006.

[13] I. Dhillon and Y. Guan. "*Iterative Clustering of High Dimensional Text Data Augmented by Local Search*",.In Proceedings of ICDM'02, 2002.

[14] Y. Zhao and G. Karypis. "*Criterion Functions for Document Clustering*",

Technical Report #01-04, November 2001.

[15] W.Cohen and Y. Singer. "*Context-Sensitive Learning Methods for Text Categorization*", ACM Transactions on Information Systems, Vol. 17(2), pp 141-173, April 1999.

[16] S. Dumais, J. Platt and D. Heckerman. "*Inductive Learning Algorithms and Representations for Text Categorization*", In Proceedings of the 7th International Conference on Information and Knowledge Management, 1998.

[17] G. Karypis and E.H. Han. "*Concept Indexing: A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization*", Technical Report #00-016, 2000.

[18] J. Moore, E.H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar and B. Mobasher. "*Web Page Categorization and Feature Selection Using Association Rule and Principal Component Clustering*", Army High Performance Computing Research Center, U.S.A Publications, February 1998.

[19] D. Cai, X. He and J. Han. "*Document Clustering Using Locality Preserving Indexing*", IEEE Transactions on Knowledge and Data Engineering, Vol. 117(12), December 2005.

[20] Y. Wang and J. Hodges. "*Document Clustering with Semantic Analysis*", In Proceedings of the 39th Hawaii International Conference on System Sciences, IEEE, 2006.

[21] H. Kargupta, I. Hamzaoglu and B. Stafford. "*Distributed data mining using an agent based architecture*", In Proceedings of Knowledge Discovery and Data Mining, pp 211–214. AAAI Press, 1997.

[22] D. Freitag and A. McCallum. "*Information Extraction with HMMs and Shrinkage*", In Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction, 1999.

[23] Σ. Μαστρογιαννάκης. "*Ταξινόμηση κειμένων με χρήση στατιστικών μεθόδων*", ΜΕ, Ιωάννινα 2003.

[24] D. Boley, M. Gini, R. Gross, S. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher and J. Moore. "*Partitioning-Based Clustering for Web Document Categorization*", Decision Support Systems, Vol. 27, pp 329-341, 1999.

[25] A. Purandare and T. Pedersen. "*Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces*", In Proceedings of the Conference on Computational Natural Language Learning, May 2004.

[26] S. Scott and S. Matwin. "*Feature Engineering for Text Classification*", In Proceedings of the 16th International Conference on Machine Learning ICML-99, pp 379–388, 1999.

[27] S. Soderland. "*Learning Information Extraction Rules for Semi-structured and Free Text*", Machine Learning, Vol. 34(1-3), pp 233-272, 1999.

[28] W.W. Cohen. "*Learning to Classify English text with ILP Methods*", In Proceedings of the 5th International Workshop on Inductive Logic Programming, pp 3-24, 1995.

[29] M. Junker, M. Sintek and M. Rinck. "*Learning for Text Categorization and Information Extraction with ILP*", In James Cussens, editor, Proceedings of the 1st Workshop on Learning Language in Logic, pp 84-93, 1999.

[30] C. Apte, F. Damerau and S. Weiss. "*Text Mining with Decision Trees*", In Proceedings of the Conference on Automated Learning and Discovery, 1998.

[31] C. Ding, X. He: "*K-Nearest-Neighbor Consistency in Data Clustering*: *Incorporating Local Information into Global Optimization*", SAC'04, March 2004.

[32] I.S. Dhilon, James Fan and Y. Guan. "*Efficient Clustering of Very Large Document Collections*", In R. Grossman, G. Kamath, and R. Naburu, editors, Data Mining for Scientific and Engineering Applications, Kluwer Academic Publications, 2001.

[33] M.F. Porter. "*An algorithm for suffix stripping. Program*", Vol. 14(3), pp 130-137, July 1980.

[34] D. Hull and Grefenstette. "*Stemming algorithms: A case study for detailed evaluation*", J. Am. Soc. Inf. Sci., Vol. 47(1), pp 70-84, 1996.

[35] W. Francis. "*Frequency Analysis of English Usage: Lexicon and Grammar*", Houghton Mifflin, 1982.

[36] R.T. Lo, B. He and I. Ounis. "*Automatically Building a Stopword List for an Information Retrieval System*", 5th Dutch-Belgium Retrieval Workshop (DIR'05), 2005.

[37] C. Fox. "*Lexical Analysis and Stoplists*", In Information Retrieval – Data Structures & Algorithms, Prentice Hall, pp 102-130, 1992.

[38] I.H. Witten, A. Moffat and T.C. Bel. "*Managing Gigabytes*", 2nd edn Morgan-Kaufman, 1999.

[39] W.J. Wilbur. "*Global Term Weights for Document Retrieval Learned from TREC Data*", Journal of Information Science, Vol. 27(5), pp. 303-310, 2001.

[40] J.H. Williams Jr. and M.P. Perriens. "*Automatic Full Text Indexing and Searching System*", presented at IBM Information Systems Symposium, 1968.

[41] M. Lan and C.L. Tan. "*A Comprehensive Study on Term Weighting Schemes*

for Text Categorization with Support Vector Machines", In Proceedings of WWW'05, May 2005.

[42] Y. Yang and J.P. Pedersen. "*A Comparative Study on Feature Selection in Text Categorization*", The Fourteenth International Conference on Machine Learning, pp. 412-420, 1997.

[43] G. Salton. "*Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*", Addison-Wesley, MA, 1989.

[44] K. Aas and L. Eikvil. "*Text Categorization: A survey*", Technical Report 941, Norwegian Computing Center, June 1999.

[45] G. Salton, A. Wong, and C. Yang. "*A Vector Space Model for Automatic Indexing*", Communications of the ACM, Vol. 18(11), pp 613-620, November 1975.

[46] G. Salton and M.J. McGill. "*Introduction to Modern Information Retrieval*", McGraw-Hill, 1983.

[47] K.M. Hammouda and M.S. Kamel. "*Efficient Phrase-Based Document Indexing for Web-Document Clustering*", IEEE, 2003.

[48] J. Fuerkranz, T. Mitchell, and E. Rilloff. "*A Case-study in Using Linguistic Phrases for Text Categorization on the WWW*", M. Sahami, editor, In learning for Text Categorization Papers from the 1998 AAAI Workshop (Technical Report WS-98-05), 1998.

[49] A. Schenker, M. Last, H. Bunke and A. Kandel. "*Clustering of Web Documents Using a Graph Model*", Web Document Analysis: Chalenges and Opportunities, eds. A. Antonacopoulos and J. Hu, to appear.

[50] A.Schenker, M.Last, H. Bunke and A.Kandel. "*A Comparison of Two Novel Algorithms for Clustering Web Documents*", N00039-01-1-2248.

[51] H. Bunke and K. Shearer. "*A Graph Distance Metric Based on the Maximal Common Subgraph*", Pattern Recognition Letters, Vol. 19, pp 255-259, 1998.

[52] A. Likas, N. Vlassis and J. J. Verbeek. "*The Global K-Means Clustering Algorithm*", Pattern Recognition, Vol. 36, pp 451-461, 2003.

[53] D.D Lewis. "*Reuters-21578 Document Corpus v1.0*", http://kdd.ics.ucs.edu/databases/reuters21578/reuters21578.html.

[54] A. Kalogeratos and A. Lykas. "*A Significance-based Graph Model for Clustering Web Documents*", In Proceedings of SETN '06, 2006.

[55] O. Zamir and O. Etzioni. "*Web Document Clustering: A feasibility Demonstration*", In Proceedings of 21st International ACM SIGIR Conference,

pp 46-54, 1998.

[56] O. Zamir and O. Etzioni. "*Grouper: A Dynamic Clustering Interface to Web Search Results*", Computer Networks, Vol. 31(11-16), pp 1361-1374, 1999.

[57] J.W. Wilbur and K. Sirotkin. "*The Automatic Iidentification of Stop Words*", J. Inf. Sci., Vol. 18, pp 45-55, 1992.

[58] Y. Yang: "*Noise Reduction in a Statistical Approach to Text Categorization*", In Proceedings of SIGIR'95, pp 256-263, 1995.

[59] Y. Yang and W.J. Wilbur: "*Using Corpus Statistics to Remove Redundant Words in Text Categorization*", In J. Amer. Soc. Inf. Sci, 1996.

[60] D. Haussler. "*Convolution Kernels on Discrete Structures*", In Technical Report UCS-CRL-99-10. UC Santa Cruz, 1999.

[61] M. Collins and N. Duffy. "*Parsing with a Single Neuron: Convolution Kernel for Natural Language Problems*", In Technical Report UCS-CRL-01010, UC Santa Cruz, 2001.

[62] N. Cristianini and J. Shawe-Taylor. "*An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*", Cambridge University Press, 2000.

[63] D. Lin. "*An Information Theoretic Definition of Similarity*", In Proceedings of 15th International Conference on Machine Learning, pp 296-304, 1998.

[64] B. Messemer and H. Bunke. "*A New Algorithm for Error-tolerant Subgraph Isomorphism Detection*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, pp 493-504, May 1998.

[65] A. Sanfeliu and K. Fu. "*A Distance Measure Between Attributed Relational Graphs for Pattern Recognition*", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 13, pp 353-362, May-June 1983.

[66] K. Tsuda and T. Kudo. "*Clustering Graphs by weighted Substructure Mining*", In Proceedings of the 23rd International Conference on Machine Learning, pp 953-960, 2006.

[67] A. Schenker, M. Last, H. Bunke and A. Kandel. "*Classification of Web documents Using a Graph Model*", In Proceedings of ICDAR'03, August 2003.

[68] P.N. Tan, M. Steinbach and V. Kumar. "*Introduction to Data Mining*", Addison-Wesley Longman, 2005.

[69] A. Hotho. "*Using Ontologies to Improve Text Clustering and Classification Task*", In Proceedings of GfKl'05, February 2005.

[70] M. Teboule, P. Berkhin and I. Dhillon, Y. Guan, and J. Kogan: "*Clustering with Entropy-Like k-Means Algorithms*", Grouping Multidimentional Data, Springer, pp 127-160, February 2006.

[71] G.W. Milligan and M.C. Cooper. "*An Examination of Procedures for Determining the Number of Clusters in a Data Set*", Psychometrica, Vol. 50(2), pp 159-179, June 1985.

[72] E. Dimitriadou, S.Dolničar and A. Weingessel. "*An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets*", Psychometrica, Vol. 67(3), pp 137-160, September 2002.

[73] C.A. Sugar and G.M. James. "*Finding the Number of Clusters in a Data Set: An Information Theoretic Approach*", Journal of the American Statistical Association, Vol. 98, 2003.

[74] R. Tibshirani, G. Walther and T. Hastie. "*Estimating the Number of Clusters in a Data Set via the Gap Statistic*", Journal of the Royal Statistical Society. Series B (Statistical and Methodology), Vol. 63(2), pp 411-423, 2001.

[75] G.B. Mufti, P. Bertrand and L.E. Moubarki. "*Determining the Number of Groups from Measures of Cluster Stability*", ASMDA 2005, May 2005

[76] T. Honkela, S. Kaski, K. Lagus and T. Kohonen. "*WEBSOM—Self-organizing Maps of Document Collections*", Neurocomputing, Vol. 21, pp 101-117, May 1998.

[77] E. Chlebus and O. Rahul. "*Estimating Parameters of the Pareto Distribution by Means of Zipf's Law: Application to Internet Research*", In Proceedings of IEEE GLOBECOM'05, Vol. 2, December 2005.

[78] G.K. Zipf. "*The Psycho-biology of Language, an Introduction to Dynamic Philology*", Cambridge, Mass.: MIT Press, 1936.

[79] H. Simon. "*On a Class of Skew Distributions*", Biometrika, Vol. 42, pp 435-440, 1955.

[80] B.B Mandelbrot. "*The Fractal Geometry of Nature*", New York, Freeman, 1983.

[81] G.A. Miller. "*Some Effects of Intermittent Silence*", *The* American Journal of Psychology, Vol. 70(2), pp 311-314, June 1957.

[82] H.S. Heaps. "*Information Retrieval: Computational and Theoretical Aspects*", Academic Press, October 1978.

[83] W. Ebeling and T. Pöschell. "*Entropy and Long-range Correlations in Literary English*", Europhysics Letters, Vol. 26, p. 241-246, 1994.

[84] I. Kanter and D.A Kessler. "*Markov Processes: Linguistics, Zipf's Law and Long-range Correlations*", Physical Review Letters, Vol. 74, pp 4559-4562, May 1995.

[85] D.H Zanette and A.Montemurro. "*Dynamics of Text Generation with Realistic Zipf's Distribution*", Journal of Quantitative Linguistics, Vol. 12, pp 29-40, December 2005.

[86] P. Barford, A. Bestavros, A. Bradley and M. Crovella. "*Changes in Web Client Access Patterns: Characteristics and Caching Implications*", World Wide Web, Vol. 2(1-2), pp 15-28, June 1999.

[87] L.A. Adamic and B.A Huberman. "*Zipf's Law and the Internet*", Glottometrics, Vol. 3, pp 143-150, 2002.

[88] M.W Berry, S.T Dumais and G.W O'Brien. "*Using Linear Algebra for Intelligent Information Retrieval*", SIAM Review, Vol. 37(4), pp 573-595, December 1995.

ΠΑΡΑΡΤΗΜΑ

Εκτίμηση παραμέτρων της *Zipf* κατανομής

Η κατανομή αυτή παρουσιάζει την εξής ενδιαφέρουσα ιδιότητα. Αν προβάλουμε τα δεδομένα σε λογαριθμική κλίμακα, τοποθετώντας στον άξονα *x* τα στοιχεία σε φθίνουσα σειρά βάσει των λογαριθμικών συχνοτήτων και στον άξονα *y* τις λογαριθμικές συχνότητες εμφάνισης, τότε παρατηρούμε να σχηματίζουν ένα ευθύγραμμο τμήμα το οποίο κλίνει προς τα δεξιά. Το γεγονός αυτό δείχνει πως σε λογαριθμική κλίμακα υπάρχει γραμμική σχέση ανάμεσα στη συχνότητα εμφάνισης ενός όρου και τη θέση του στην φθίνουσα ακολουθία δειγμάτων. Πιο συγκεκριμένα αυτό μπορεί να γίνει σαφές:

$$x_r = \frac{c}{r^\beta} \Leftrightarrow \log x_r = \log \frac{c}{r^\beta} \Leftrightarrow \log x_r = -\beta \log r + \log c .$$

Η παραπάνω γραμμικότητα ισχύει μόνο στην περίπτωση που ισχύει ιδανικά ο νόμος του *Zipf*, για σταθερές παραμέτρους, γεγονός το οποίο ήδη αναφέρθηκε πως είναι ιδιαίτερα απίθανο σε πραγματικά δεδομένα. Συνεπώς, η εύρεση των παραμέτρων είναι ένα πρόβλημα εκτίμησης των παραμέτρων της ευθείας η οποία ταιριάζει καλύτερα στα μετασχηματισμένα δεδομένα (στη λογαριθμική κλίμακα).

Στη βιβλιογραφίας έχει παρουσιαστεί η εκτίμηση των παραμέτρων χρήσει της μεθόδου ελαχίστων τετραγώνων. Αν θεωρήσουμε κάθετη διαταραχή στα δεδομένα, δηλαδή στον άξονα των συχνοτήτων, η οποία είναι ουσιαστικά και η απόκλιση του μοντέλου από τα πραγματικά χαρακτηριστικά των δεδομένων, τότε η μπορούμε να εκφράσουμε την αναζητούμενη ευθεία σε λογαριθμική πάντα κλίμακα:

$$\log \hat{x}_r = -\beta \log r + \log c$$

με $\hat{x}_r$ την εκτιμώμενη τιμή – συχνότητα για το *r*-οστό στοιχείο της διάταξης. Μπορούμε στο σημείο αυτό να διατυπώσουμε το πρόβλημα εύρεσης της κατάλληλης ευθείας ως πρόβλημα ελαχιστοποίησης των τετραγωνικών σφαλμάτων:

$$(\beta, c) = \arg\min_{\beta,c} \{E(\beta, c)\}$$

Αρκεί συνεπώς να ορίσουμε τη συνάρτηση $E(\beta,c)$ την οποία επιθυμούμε να ελαχιστοποιήσουμε:

$$E(\beta, c) = \left(\log x_r - \log \hat{x}_r\right)^2 = \left(\log x_r - \beta \log r + \log c\right)^2.$$

Παίρνοντας τις μερικές παραγώγους και θέτοντας $\alpha = \log c$ προκύπτει το παρακάτω σύστημα δύο εξισώσεων με δύο αγνώστους:

$$(\Sigma 1) \qquad \frac{\partial E}{a} = -2\sum_{r=1}^{N}\left[\log x_r + \beta \log r - a\right] = 0$$

$$(\Sigma 2) \qquad \frac{\partial E}{\beta} = 2\sum_{r=1}^{N}\left[\log r\left(\log x_r + \beta \log r - a\right)\right] = 0$$

Από την ($\Sigma 1$) παίρνω:

$$\sum_{r=1}^{N}\log x_r + \beta\sum_{r=1}^{N}\log r - \sum_{r=1}^{N} a = 0 \Leftrightarrow a = \frac{1}{n}\left[\sum_{r=1}^{N}\log x_r + \beta\sum_{r=1}^{N}\log r\right]$$

Και αντικαθιστώντας την παράμετρο α της ($\Sigma 1$) στην ($\Sigma 2$):

$$\sum_{r=1}^{N}\log r \log x_r + \beta\sum_{r=1}^{N}\left(\log r\right)^2 - a\sum_{r=1}^{N}\log r = 0 \Leftrightarrow$$

$$\beta\left[\sum_{r=1}^{N}\left(\log r\right)^2 - \frac{1}{n}\left(\sum_{r=1}^{N}\log r\right)^2\right] = \frac{1}{n}\left(\sum_{r=1}^{N}\log x_r\right)\left(\sum_{r=1}^{N}\log r\right) - \sum_{r=1}^{N}\log r \log x_r \Leftrightarrow$$

Τέλος πολλαπλασιάζοντας με n/n:

$$\beta = \frac{\left(\sum_{r=1}^{N}\log x_r\right)\left(\sum_{r=1}^{N}\log r\right) - n\sum_{r=1}^{N}\log r \log x_r}{n\sum_{r=1}^{N}\log r^2 - \left(\sum_{r=1}^{N}\log r\right)^2}.$$

Επίσης: $a = \log c \Leftrightarrow c = e^a$.

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Ο Αργύρης Καλογεράτος γεννήθηκε το 1982 στην Πάτρα. Αποφοίτησε από το 13^{ο} Ενιαίο Λύκειο της ίδιας πόλης και εισήχθη το 2001 στο προπτυχιακό πρόγραμμα σπουδών του Τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων. Το πτυχίο του έλαβε το 2005 και έπειτα παρακολούθησε το μεταπτυχιακό πρόγραμμα σπουδών από το οποίο αποφοίτησε τον Νοέμβριο του 2007.