Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων
Και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης &
Υπολογιστων

Ευφυείς Τεχνικές Σημασιολογικής Ανάλυσης και Αναζήτησης Κειμένου

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΓΕΡΑΣΙΜΟΥ ΣΠΑΝΑΚΗ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2006)

Αθήνα, Ιούλιος 2012

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων & Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης &
Υπολογιστων

Ευφυείς Τεχνικές Σημασιολογικής Ανάλυσης και Αναζήτησης Κειμένου

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΓΕΡΑΣΙΜΟΥ ΣΠΑΝΑΚΗ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2006)

Συμβουλευτική Επιτροπή: Ανδρέας - Γεώργιος Σταφυλοπάτης
Παναγιώτης Τσανάκας
Στέφανος Κόλλιας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 19η Ιουλίου 2012.

...	...	...
Α.-Γ. Σταφυλοπάτης	Π. Τσανάκας	Στ. Κόλλιας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.
...	...	...
Γ. Στάμου	Γ. Καραγιάννης	Γ. Μαΐστρος
Λέκτορας Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επ. καθηγητής Ε.Μ.Π.

...
Ελένη Ευθυμίου
Ερευνήτρια Α΄
Ινστιτούτο ΑΘΗΝΑ/ΙΕΛ

Αθήνα, Ιούλιος 2012

...

ΓΕΡΑΣΙΜΟΣ ΣΠΑΝΑΚΗΣ

Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

ΠΡΟΛΟΓΟΣ

Αντικείμενο αυτής της διατριβής αποτελεί η εξερεύνηση της χρήσης ευφυών τεχνικών σε μεθοδολογίες ανάλυσης κειμένου. Η υπολογιστική νοημοσύνη και ο ρόλος που μπορεί να έχει στην ανάλυση της κειμενικής πληροφορίας (άφθονης στις μέρες μας), αξιοποιώντας πλούσιες πηγές γνώσης (ελεύθερα διαθέσιμες μέσω του Παγκόσμιου Ιστού) αποτέλεσαν το κίνητρο για την εκπόνηση αυτής της διατριβής. Είναι αποτέλεσμα μιας ερευνητικής πορείας που ξεκίνησε το 2007 στο Εργαστήριο Ευφυών Συστημάτων και επηρέασε καθοριστικά τη διαμόρφωση της προσωπικότητάς μου και κυρίως τον τρόπο σκέψης μου. Ο πλούτος των γνώσεων που απέκτησα, οι προκλήσεις που αντιμετώπισα και κυρίως η όλη ερευνητική διαδικασία που οδήγησε στα αποτελέσματα αυτής της διατριβής περιγράφουν στο ελάχιστο μερικά από τα χαρακτηριστικά αυτής της μακρόχρονης πορείας.

Σε όλη τη διάρκεια της εκπόνησης της διατριβής, παρών ήταν ο επιβλέπων καθηγητής κ. Ανδρέας-Γεώργιος Σταφυλοπάτης, τον οποίο θα ήθελα να ευχαριστήσω θερμά για την εμπιστοσύνη που έδειξε στο πρόσωπό μου, τη στήριξη και την πολύτιμη βοήθειά του τόσο σε επιστημονικό όσο και προσωπικό επίπεδο. Οι γνώσεις, η εμπειρία και το διαρκές ενδιαφέρον του με συνοδεύουν σε όλη τη μακρόχρονη συνεργασία μας από τις προπτυχιακές μου σπουδές μου έως και σήμερα, και έχουν συμβάλει καθοριστικά στη διαμόρφωση της παρούσας διατριβής αλλά και της προσωπικότητάς μου. Ευχαριστίες οφείλω και στους καθηγητές ΕΜΠ κ. Στέφανο Κόλλια και κ. Παναγιώτη Τσανάκα, μέλη της συμβουλευτικής επιτροπής, για το ενδιαφέρον και τη στήριξή τους αυτά τα χρόνια. Θα ήθελα επίσης να ευχαριστήσω τους κ. Γιώργο Καραγιάννη, Καθηγητή ΕΜΠ, κ. Γιάννη Μαΐστρο, Επίκουρο Καθηγητή ΕΜΠ, κ. Γιώργο Στάμου, Λέκτορα

ΕΜΠ, καθώς και την κ. Ελένη Ευθυμίου, Διδάκτορα Γλωσσολογίας, Ερευνήτρια Α΄ στο Ινστιτούτο ΑΘΗΝΑ/ΙΕΛ για την τιμή που μου έκαναν να είναι μέλη της επιτροπής αξιολόγησης της διατριβής.

Ξεχωριστό ρόλο στην ολοκλήρωση της διατριβής διαδραμάτισε ο διδάκτωρ και μεταδιδακτορικός ερευνητής του Εργαστηρίου Ευφυών Συστημάτων κ. Γιώργος Σιόλας, ο οποίος είχε καθημερινή ενασχόληση και εμπλοκή με την ερευνητική μου εργασία και με παρότρυνε και συμβούλευε για τις κατευθύνσεις της έρευνάς μου. Επίσης, θα ήθελα να ευχαριστήσω όλα τα (παλαιά και νέα) μέλη του Εργαστηρίου Ευφυών Συστημάτων και ιδιαίτερα τους Απόστολο Μαρακάκη, Χρήστο Πατερίτσα, Μηνά Περτσελάκη, Δημήτρη Φροσυνιώτη, Αλέξανδρο Χορταρά διδάκτορες ΕΜΠ, καθώς και τους Γιώργο Αλεξανδρίδη, Άρη Λαναρίδη, Γιώργο Στρατογιάννη, Χρήστο Φερλέ, Χριστίνα Χριστάκου υποψήφιους διδάκτορες ΕΜΠ για τη συνεργασία που είχαμε, την προθυμία να βοηθήσουν σε οποιοδήποτε ζήτημα παρουσιαζόταν στη διάρκεια εκπόνησης της διατριβής αυτής.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, τους φίλους μου και τα αγαπημένα μου πρόσωπα για τη συμπαράστασή τους σε όλα τη διάρκεια εκπόνησης της διατριβής.

Γεράσιμος Σπανάκης

Αθήνα, Ιούλιος 2012

ΠΕΡΙΛΗΨΗ

Η ραγδαία αύξηση του όγκου των διαθέσιμων ψηφιακών εγγράφων τα τελευταία χρόνια δημιουργεί την ανάγκη δημιουργίας συστημάτων οργάνωσης και διαχείρισής τους. Η κειμενική πληροφορία με τη μορφή ψηφιακών εγγράφων αποτελεί μία τεράστια πηγή πληροφοριών που αναπτύσσεται μέρα με τη μέρα λόγω και της εξάπλωσης του Παγκόσμιου Ιστού, ο οποίος σήμερα διαθέτει μεγάλες ποσότητες ελεύθερου κειμένου. Ο όγκος των διαθέσιμων εγγράφων απαιτεί αποδοτικές τεχνικές αποθήκευσης και αναπαράστασής τους στον υπολογιστή καθώς και αποτελεσματικές μεθόδους οργάνωσης, διαχείρισης, αναζήτησης και επεξεργασίας. Η συνεισφορά της διατριβής εντάσσεται στην περιοχή της αποδοτικής και πιο πλήρους αναπαράστασης εγγράφων και της αποτελεσματικής ανάλυσης προβλημάτων που σχετίζονται με τα έγγραφα (οργάνωση, ομαδοποίηση κτλ).

Η χρήση της γλώσσας είναι εξαιρετικά πολύπλοκη, κάτι το οποίο δημιουργεί διάφορα προβλήματα στην προσπάθεια αναπαράστασης εγγράφων στον υπολογιστή: Τα νοήματα που κρύβονται σε ένα κείμενο λόγω των σχέσεων που υπάρχουν ανάμεσα στις λέξεις, η πληροφορία που υπονοείται λόγω πρότερης ή εκ φύσεως γνώσης, παρομοιώσεις, μεταφορές κτλ είναι μερικά από τα προβλήματα που ανακύπτουν και σε συνδυασμό με την υπολογιστική πολυπλοκότητα που εισάγεται λόγω του μεγάλου όγκου των εγγράφων δεν έχουν επιτρέψει μέχρι σήμερα να βρεθεί ένα σταθερό και αποδοτικό μοντέλο αναπαράστασης.

Στα πλαίσια της διατριβής εξετάζονται οι βασικές μονάδες αναπαράστασης των εγγράφων (συλλαβές, λέξεις, προτάσεις/φράσεις) και πιο συγκεκριμένα το μοντέλο του χώρου διανυσμάτων (Vector Space Model, VSM), το οποίο χρησιμοποιείται ευρέως για

την αναπαράσταση εγγράφων. Οι λέξεις αποτελούν την κυριότερη μονάδα αναπαράστασης εγγράφων και παρά τα μειονεκτήματα που παρουσιάζουν ως μονάδα αναπαράστασης (μεγάλος χώρος αναζήτησης, διάσπαση ομάδων λέξεων κτλ) παραμένουν έως και σήμερα στο επίκεντρο των περισσοτέρων μοντέλων. Εξάλλου, δεν είναι τυχαίο πως οι μηχανές αναζήτησης στον Παγκόσμιο Ιστό λειτουργούν βάσει λέξεων-κλειδιών. Βάσει της ιδέας του ότι οποιαδήποτε ομοιότητα ή σχέση μεταξύ εγγράφων μπορεί να αναχθεί στον καθορισμό της σχέσης των λέξεων που τα αποτελούν και βάσει της παρατήρησης πως η αναζήτηση με λέξεις-κλειδιά παραμένει ο κυριότερος τρόπος αναζήτησης, η διατριβή προτείνει μία μέθοδο προσδιορισμού της σημασιολογικής σχέσης λέξεων. Στόχος είναι να βρεθεί ένα βαθμωτό μέτρο που θα ποσοτικοποιεί την οποιαδήποτε σχέση (συνωνυμία, υπερωνυμία, αντωνυμία κτλ) υπάρχει μεταξύ δύο οποιωνδήποτε λέξεων και γιαυτό το σκοπό αξιοποιεί την πληροφορία που παρέχεται από το ιεραρχικό λεξικό WordNet καθώς και τα λεξικο-συντακτικά πρότυπα που εξάγονται από τα αποτελέσματα αναζήτησης για τις εν λόγω λέξεις που επιστρέφονται από κάποια μηχανή αναζήτησης.

Αναγνωρίζοντας τις αδυναμίες ενός μοντέλου αναπαράστασης με λέξεις αλλά και των περιορισμών που θέτει ένα βαθμωτό μέτρο συσχέτισης λέξεων, το επόμενο βήμα της διατριβής είναι η εισαγωγή ενός νέου μοντέλου αναπαράστασης που δε θα βασίζεται στις λέξεις του εγγράφου, αλλά θα εισάγει σημασιολογικό περιεχόμενο στην αναπαράσταση βάσει των εννοιών (concepts) (οι οποίες μπορεί να αποτελούνται από παραπάνω της μιας λέξης). Για το σκοπό αυτό αξιοποιείται η Wikipedia που λόγω του αυξανόμενου όγκου της και της δομής της (ιεραρχική δόμηση, πλήρεις καλογραμμένες προτάσεις, κατατοπιστικοί τίτλοι άρθρων κτλ) παρέχει πολλές δυνατότητες ενίσχυσης της σημασιολογίας των εγγράφων, μέσω χαρακτηριστικών που κατασκευάζονται από γνωρίσματα τα οποία εξάγονται από τη Wikipedia. Το μοντέλο που παρουσιάζεται οδηγεί σε αναπαραστάσεις τόσο πιο πλούσιες (σημασιολογικά) όσο και πιο συμπιεσμένες (από άποψη απαιτήσεων χώρου) σε σχέση με το κλασσικό μοντέλο VSM.

Αφού πλέον υπάρχει διαθέσιμο ένα καλύτερο μοντέλο αναπαράστασης εγγράφων, η διατριβή πηγαίνει στο επόμενο επίπεδο και δεν εξετάζει πλέον τις σχέσεις ανάμεσα στις λέξεις του εγγράφου, αλλά τα θέματα με τα οποία ασχολούνται μεγάλες συλλογές

εγγράφων, προτείνοντας δύο μεθοδολογίες ομαδοποίησης εγγράφων βάσει του περιεχομένου τους. Και στις δύο μεθοδολογίες κυρίαρχο ρόλο διαδραματίζει το μοντέλο αναπαράστασης εγγράφων βάσει της Wikipedia χρησιμοποιώντας τα χαρακτηριστικά που έχουν κατασκευαστεί. Η πρώτη μεθοδολογία βασίζεται στις πιο σημαντικές έννοιες της συλλογής των εγγράφων που εξετάζεται και δημιουργεί γρήγορα και αποδοτικά μία ιεραρχική δενδρική δομή ομάδων στις οποίες κατανέμονται τα έγγραφα βάσει του περιεχομένου τους. Η δομή είναι κατευθυνόμενη από το χρήστη ως προς το βάθος και το πλάτος του δέντρου (και συνακόλουθα ελέγχονται και οι θεματικές περιοχές στις οποίες χωρίζονται τα έγγραφα). Η δεύτερη μεθοδολογία αξιοποιεί τους Αυτο-Οργανούμενους Χάρτες (Self Organizing Maps, SOM) ως εργαλείο για την ομαδοποίηση εγγράφων. Μέσα από τρία βήματα υλοποιείται η εκπαίδευση ενός Αυτο-Οργανούμενου Χάρτη, τροποποιημένου τόσο ως προς την αρχικοποίηση και τη δημιουργία του πλέγματος των νευρώνων (που γίνεται βάσει μιας διαδικασίας που βασίζεται στο μοντέλο αναπαράστασης με χρήση της Wikipedia), όσο και ως προς τη διαδικασία της εκπαίδευσης, επιταχύνοντάς τη καταλυτικά. Στο τέλος της εκπαίδευσης παρέχεται η δυνατότητα ιεραρχικής ομαδοποίησης των παρόμοιων νευρώνων του Χάρτη σε ομάδες, ενώ η οπτικοποίηση δίνει με ακρίβεια την τοπολογική σχέση των ομάδων (θεματικών περιοχών).

Η ουσιαστική συμβολή της διατριβής συνοψίζεται στη δυνατότητα χρήσης ευφυών τεχνικών με αξιοποίηση διαφόρων πηγών γνώσης, ώστε να βελτιωθούν ζητήματα που έχουν να κάνουν με την αποδοτική αναπαράσταση και αντιμετώπιση ζητημάτων ανάλυσης των ολοένα και μεγαλύτερων σε όγκο εγγράφων. Κάθε μεθοδολογία που αναπτύχθηκε αξιολογήθηκε πειραματικά με χρήση συνόλων δεδομένων, τα οποία χρησιμοποιούνται ευρέως από την επιστημονική κοινότητα ενώ έγιναν και συγκρίσεις με τις σημαντικότερες μεθόδους στο κάθε πεδίο έρευνας.

ABSTRACT

The rapid proliferation of digital text documents during the last years raises the need to create efficient organization and management systems. The textual content of digital documents is a huge source of information that grows every day, assisted by the global growth of the Internet, which contains large quantities of plain text. The size of available documents demands efficient ways to store and represent information using computers, as well as efficient methods for its organization, management, search and editing. The contribution of this PhD thesis rests in the fields of efficient and more thorough representation of documents and the efficient analysis of problems related with them (organization, classification etc.)

The use of language is particularly complicated and creates various problems in the attempt to represent documents using computers. The meanings and senses that exist latently in a document because of the context, the relations between the words, the implied information (derived from natural or earlier knowledge), metaphors etc. are some of the problems that emerge, which, combined with the computational complexity that is introduced by the large size of documents, have not allowed for a stable and efficient way of representation to be found until today.

In this PhD thesis we examine the basic units of document representation (syllables, words, phrases/sentences) and specifically the Vector Space Model (VSM) which is widely used for document representation based on words. Words are the main unit of document representation and, despite the drawbacks (large search space, breaking of multi-words etc.) introduced, they still are the base of most models. It is no random fact that World Wide Web search engines utilize keywords.

Based on the idea that documents consist of words and therefore, every similarity or relation between documents can be reduced to the determination of the relation between the words that consist them and given the observation that searching with keywords is the main way of searching, this PhD thesis proposes a method of determining the semantic relation between words. The main goal is to find a scalar measure that quantifies any relation (synonym, antonym, hypernym etc.) between any two words. For this reason, the method utilizes the information provided by the hierarchical dictionary WordNet as well as the lexico-syntactic patterns extracted from the search results returned by a search engine using those words.

Recognizing the weaknesses of a representation model based on words as well as the limitations that are bound by the scalar measurement of word relatedness, the next step in this PhD thesis is the introduction of a new representation model, which is not based on document words, but rather includes semantics in the representation, based on its named entities-concepts (which can contain more than one word). Wikipedia is utilized for this purpose, since its increasing size and rich structure (hierarchical organization, full well-written sentences, informative article headers etc.) provide many capabilities to enhance the document semantics with characteristics created by attributes derived from Wikipedia. The presented model leads to representations not only richer, but also more compressed ones, compared to the VSM model.

The introduction of a better representation mode allows the PhD thesis to examine not only the relations between the words-senses of a document, but also the topics appearing in large document collections through the proposal of two document clustering techniques (according to their content). In both techniques the Wikipedia based representation model plays an important role utilizing the constructed features. The first technique is based on the most important concepts of the document collection examined and creates fast and efficiently a hierarchical tree structure of the groups that documents are clustered to, according to their content. The depth and width of the tree structure is fully controlled by the user. The second

technique utilizes Self Organizing Maps (SOM) as a tool for document clustering. Self Organizing Map's initialization and initial neuron structure (grid) is modified to the original SOM algorithm (utilizing the Wikipedia based representation model) and training is carried out in three steps in accelerated time compared to the original process. At the end of training, user has the possibility to hierarchically organize similar neurons of the map.

The actual contribution of this PhD thesis is summarized in the possibility of using intelligent techniques and utilizing various source of knowledge in order to improve aspects or solve problems related to the efficient representation and analysis of the growing in size documents. In order to derive useful conclusions, at every stage of the research and for each proposed method the results of appropriately designed and performed experiments and comparisons are provided, which do not focus only on the overall evaluation of the methodologies, but, in parallel, intend to justify the particular choices and to prove their merits.

Κατάλογος Συντμήσεων

20-NG	:	20-NewsGroup
BOW	:	Bag-of-Words
CHC	:	Conceptual Hierarchical Clustering
DoSO	:	Document Self Organizer
IDF	:	Inverse Document Frequency
NP	:	Noun Phrase
POS	:	Part-of-Speech
rSVM	:	Regression Support Vector Machine
SOM	:	Self-Organizing Map
SVM	:	Support Vector Machine
TF	:	Term Frequency
URL	:	Uniform Resource Locator
VSM	:	Vector Space Model
WWW	:	World Wide Web

Κεφάλαιο 1

Εισαγωγή

Στη σύγχρονη εποχή με τον ολοένα αυξανόμενο όγκο πληροφοριών, καθίσταται αναγκαία η σωστή διαχείριση και οργάνωση της διαθέσιμης πληροφορίας. Η άναρχη ανάπτυξη του Παγκόσμιου Ιστού (WWW) μας επιτρέπει να τον θεωρήσουμε ως μια τεράστια αποθήκη δεδομένων κειμένου, τα οποία όμως τις περισσότερες φορές δεν είναι οργανωμένα όπως θα ήθελε ένας χρήστης προκειμένου να εντοπίσει την πληροφορία που τον ενδιαφέρει. Η ανάγκη για αποδοτικές και υψηλής ποιότητας μεθόδους ανάλυσης και αναζήτησης κειμένου καθίσταται επιτακτική.

1.1 Προβλήματα - Προκλήσεις

Ξεκινώντας από τον προσωπικό υπολογιστή του καθενός και επεκτείνοντας την αναζήτηση στον Παγκόσμιο Ιστό, η διαθεσιμότητα εγγράφων είναι πολύ μεγάλη και αυτό έχει σαν συνέπεια να δυσχεραίνεται η αποτελεσματική διαχείρισή τους. Ο κυριότερος τρόπος πρόσβασης στα έγγραφα που υπάρχουν στο Παγκόσμιο Ιστό είναι μέσω της διαδικασίας εξόρυξης κειμένου (text mining) δηλαδή της ανάκτησης πληροφορίας υψηλής ποιότητας από αυτά τα έγγραφα με διάφορες τεχνικές (όπως στατιστικές). Οι μηχανές αναζήτησης (όπως έχουν αναπτυχθεί στον Παγκόσμιο Ιστό) και άλλοι κατάλογοι οργάνωσης χρησιμοποιούνται για τη διαχείριση των ψηφιακών εγγράφων και την αποτελεσματική ανάκτησή τους. Παρόλα αυτά, τα αποτελέσματα από μία αναζήτηση σε μία αντίστοιχη μηχανή στο WWW δεν είναι πολλές φορές τα αναμενόμενα για το χρήστη (ειδικά όσο το ερώτημα γίνεται πιο σύνθετο). Αν δοκιμάσει κάποιος να αναζητήσει έγγραφα σχετικά με τη λέξη apple, εύκολα διαπιστώνει τη σύγχυση που υπάρχει ανάμεσα στο γνωστό φρούτο και στη γνωστή εταιρεία. Το πρόβλημα που παρατηρείται εδώ, είναι πως παρ΄ όλη την εξέλιξη των αλγορίθμων εξόρυξης κειμένου (που χρησιμοποιούνται από μηχανές αναζήτησης), η αναζήτηση σήμερα εξακολουθεί να γίνεται με βάση λέξεις-κλειδιά (keywords). Ξεκάθαρα απαιτείται ενίσχυση προς δύο κυρίως κατευθύνσεις:

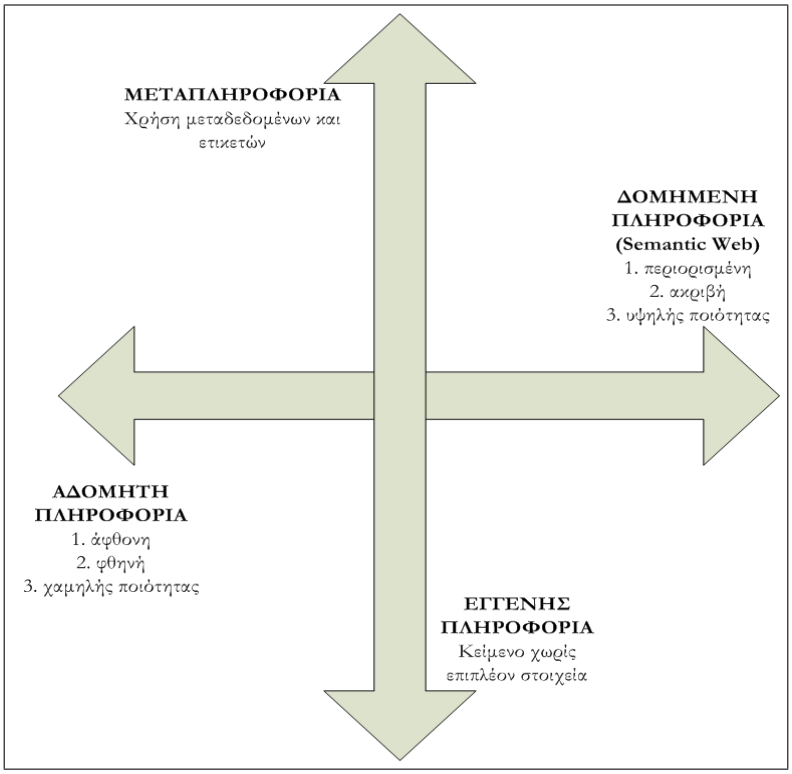
- τη δυνατότητα ερωτήσεων στο WWW με φυσική γλώσσα και όχι μόνο με τη χρήση απλών λέξεων ή φράσεων,
- τη βελτίωση των αναπαραστάσεων και συσχετίσεων της γραπτής πληροφορίας στο WWW.

Η ενίσχυση αυτή όμως δεν είναι τόσο εύκολη. Το προηγούμενο παράδειγμα με τη λέξη apple είναι ένα μικρό δείγμα της ιδιαιτερότητας της γλώσσας και των προβλημάτων που ανακύπτουν και τα οποία καλείται σήμερα να αντιμετωπίσει η επιστημονική

κοινότητα (ενίοτε με συνεργασίες με πολλούς κλάδους της επιστήμης: τεχνητή νοημοσύνη, γλωσσολογία κλπ). Το μείζον ερώτημα που δημιουργείται είναι αν μπορούν να αναπτυχθούν ευφυείς τεχνικές που θα δίνουν τη δυνατότητα στον υπολογιστή να αντιμετωπίσει τα έγγραφα όπως περίπου ο ανθρώπινος νους. Η απάντηση στο ερώτημα αυτό, αυτόματα επιλύει και όλα τα σημαντικότερα ζητήματα που έχουν να κάνουν με διαχείριση εγγράφων (ανάκτηση, ομαδοποίηση κτλ).

1.1.1 Η οργάνωση της πληροφορίας στη σημερινή εποχή

Ο Παγκόσμιος Ιστός σήμερα κατακλύζεται από δισεκατομμύρια έγγραφα τα οποία περιέχουν τις περισσότερες φορές ελεύθερο κείμενο χωρίς κάποια οργάνωση. Ο Σημασιολογικός Ιστός (Semantic Web) [Shadbolt et al., 2006] αποτελεί μια προσπάθεια καλύτερης οργάνωσης των περιεχομένων του (κλασσικού) Ιστού που θα οδηγήσει σε περισσότερες δυνατότητες εξαγωγής πληροφορίας από τα έγγραφα και θα βοηθήσει τους υπολογιστές να κατανοούν πιο εύκολα την κειμενική (και άλλη) πληροφορία. Η οργάνωση της πληροφορίας στα ψηφιακά έγγραφα μπορεί να φανεί καλύτερα στο Σχήμα 1.1.



Σχήμα 1.1: *Άξονες οργάνωσης της πληροφορίας στα κειμενικά δεδομένα*

Η ουσία του Σημασιολογικού Ιστού έγκειται στο ότι το νόημα της διαθέσιμης πληροφορίας πρέπει να αποδίδεται πολύ καλά σε μορφή κατανοητή τόσο από ανθρώπους, όσο και από ηλεκτρονικούς υπολογιστές, διευκολύνοντας τη μεταξύ τους συνεργασία. Ο τρόπος με τον αναπτύσσεται ο Σημασιολογικός Ιστός είναι ανάλογος με αυτόν του σύγχρονου Παγκόσμιου Ιστού. Το πρώτο βήμα στη διαμόρφωση του Παγκόσμιου Ιστού του μέλλοντος είναι η δημιουργία νησίδων πληροφορίας οργανωμένης σημασιολογικά. Οι νησίδες αυτές σιγά-σιγά θα διασυνδεθούν μεταξύ τους προσφέροντας δυνατότητες για την ανάπτυξη προηγμένων και ευφυών εφαρμογών. Το επόμενο βήμα είναι η αξιοποίηση του πλέγματος πληροφοριών που δημιουργείται, σύμφωνα με τις προοπτικές

που παρέχει η σημασιολογική (πια) οργάνωσή τους. Η δυνατότητα επέκτασης αυτής της δομής του Σημασιολογικού Ιστού σε οποιαδήποτε κλίμακα (από έναν προσωπικό υπολογιστή και από μία μικρή συλλογή κειμένων σε μεγάλα δίκτυα υπολογιστών και τεράστιους όγκους κειμένων) παρέχει σημαντικά πλεονεκτήματα στο πεδίο οργάνωσης κειμενικής πληροφορίας: λογικότερη οργάνωση εγγράφων, αποτελεσματικότερη αναζήτηση κειμένων, αυτοματοποίηση και ολοκλήρωση εφαρμογών επεξεργασίας και ανάλυσης κειμένου κ.τ.λ.

Αυτή η νέα μορφή του Ιστού απαιτεί συστήματα τα οποία θα οργανώνουν τις πληροφορίες με συγκεκριμένες απαιτήσεις (απαιτείται σημασιολογικό περιεχόμενο, οντολογίες, πράκτορες κλπ). Τα ζητήματα που πρέπει να αντιμετωπιστούν από τα συστήματα τεχνολογιών Σημασιολογικού Ιστού (αλλά και γενικότερα) που ασχολούνται με κειμενική πληροφορία είναι τα ακόλουθα:

- Συλλογή κειμένων: Το πρώτο και απαραίτητο βήμα είναι η απαραίτητη συλλογή των αντίστοιχων πληροφοριών που αποτελούν τη βάση δημιουργίας οποιουδήποτε συστήματος.

- Αναπαράσταση/Επισημείωση κειμένων: Το κάθε κείμενο που συλλέγεται πρέπει να περιγραφεί με όμοιο τρόπο με τα υπόλοιπα και όσο το δυνατόν πιο αποδοτικά (τόσο για το υπολογιστικό σύστημα όσο και για τον άνθρωπο που το διαχειρίζεται).

- Αποσαφήνιση εννοιών: Η αποσαφήνιση του σωστού νοήματος των λέξεων που περιέχονται στα κείμενα (και κατ΄ επέκταση του θέματος των κειμένων) αποτελεί πάντα μία ενεργή πρόκληση, καθώς οι λέξεις έχουν διαφορετικά νοήματα, αλλάζουν και εξελίσσονται διαχρονικά.

- Οργάνωση κειμένων: Στο πλαίσιο αυτό εντάσσονται μεθοδολογίες και τεχνικές που θα οργανώσουν/ομαδοποιήσουν/κατατάξουν τα κείμενα σε κατηγορίες βάσει των απαιτήσεων του χρήστη και βάσει των μεταξύ τους σημασιολογικών (ή άλλων) ομοιοτήτων/διαφορών.

- Ενημέρωση συλλογής: Ένα σημαντικό ζήτημα είναι πως θα ενημερωθεί μια συλλογή κειμένων με τις τελευταίες πληροφορίες ή πως θα γίνει μία αλλαγή στα κείμενα της συλλογής αυτής κ.ο.κ.

- Διαχείριση ερωταποκρίσεων: Περιλαμβάνει ακριβώς τον τρόπο λειτουργίας μιας μηχανής αναζήτησης στο WWW. Ο χρήστης υποβάλει ερωτήματα σχετικά με αυτό που τον ενδιαφέρει και το σύστημα πρέπει να μπορεί να αποκριθεί ανάλογα.

- Παρουσίαση αποτελεσμάτων: Υπάρχει ανάγκη για όσο το δυνατόν καλύτερη οπτική απεικόνιση του αποτελέσματος και ουσιαστική ευχρηστία στην ανάκτηση των αποτελεσμάτων.

Ο Σημασιολογικός Ιστός βασίζεται εν πολλοίς στη χρήση οντολογιών που προσφέρουν δυνατότητες για περιγραφή και αναπαράσταση διαφόρων θεματικών περιοχών ενισχύοντας έτσι το σημασιολογικό περιεχόμενο των κειμένων. Παρόλα αυτά, η κατασκευή οντολογιών είναι μία διαδικασία ακριβή, ενίοτε και χρονοβόρα, επομένως καταβάλλονται προσπάθειες να αξιοποιηθούν ευφυείς τεχνικές για την οργάνωση και αξιοποίηση της αδόμητης (και φθηνής) πληροφορίας που υπάρχει άφθονη στο WWW.

Η παρούσα διατριβή προσπαθεί να απαντήσει αποτελεσματικά σε αρκετές από τις παραπάνω προκλήσεις εντάσσοντάς έτσι τις μεθοδολογίες που αναπτύσσονται στη γενικότερη σκοπιά που υπάρχει περί πλήρους αξιοποίησης των μεγάλων ποσοτήτων κειμενικής πληροφορίας που υπάρχουν στον Παγκόσμιο Ιστό.

1.1.2 Αναπαράσταση κειμένου και προβλήματα

Το κυριότερα προβλήματα που αντιμετωπίζονται κατά την απόπειρα αναπαράστασης ενός κειμένου/εγγράφου έχουν να κάνουν με το πως θα δοθεί η δυνατότητα αντιστοίχησης μιας λέξης σε μία έννοια (μια διαδικασία δηλαδή, παρόμοια με αυτήν που κάνει ο ανθρώπινος νους όταν διαβάζει ένα κείμενο). Υπάρχουν διάφορες κατηγορίες προβλημάτων που μελετώνται αναλυτικά στο Κεφάλαιο 2. Πιο συγκεκριμένα, παρατηρούνται οι εξής δύο βασικοί τύποι προβλημάτων που έχουν να κάνουν με:

- εννοιακές/κατηγορικές σχέσεις και ιδιότητες: Οφείλονται στις ιδιότητες των εννοιών και είναι ανεξάρτητες της γλώσσας που είναι γραμμένο ένα κείμενο. Για παράδειγμα η λέξη *"σώμα"* σχετίζεται τόσο με τη λέξη *"χέρι"* όσο και με τη λέξη *"άνθρωπος"* λόγω των αντίστοιχων συσχετίσεων που υπάρχουν μεταξύ αυτών των λέξεων.

  Παραδείγματα εννοιακών σχέσεων είναι οι μεταβατικές σχέσεις υπερωνυμίας-υπωνυμίας (γνωστές και ως σχέσεις IS-A), οι σχέσεις μερωνυμίας (γνωστές και ως σχέσεις PART-OF) και η σχέση αντωνυμίας. Είναι φανερό πως ο υπολογιστής δεν έχει δυνατότητα να εντοπίσει (χωρίς κάποια πρότερη γνώση) τέτοιες σχέσεις.

- λεκτικές σχέσεις και ιδιότητες: Αφορούν συγκεκριμένες λεκτικές οντότητες (πέραν της έννοιας που έχουν) και εξαρτώνται από την εκάστοτε γλώσσα. Για παράδειγμα, η αναφορά στη φράση *"λευκό κρασί"* οδηγεί στο συμπέρασμα πως πρόκειται για τα κρασιά που το χρώμα τους είναι ανοιχτόχρωμο, κίτρινο, ξανθό (αλλά όχι ακριβώς λευκό) και προφανώς έχει νόημα μόνο εφόσον χρησιμοποιηθεί ως φράση (δηλαδή και με τις δύο λέξεις μαζί).

  Γενικότερα, μια λέξη μπορεί να έχει διάφορες έννοιες και η διάκριση μεταξύ τους επιτυγχάνεται στο εκάστοτε περιβάλλον χρήσης της λέξης (περικείμενη πληροφορία-context). Το συγκεκριμένο πρόβλημα είναι η λεγόμενη αμφισημία (word ambiguity) και περιλαμβάνει τις περιπτώσεις της ομωνυμίας/πολυσημίας (όταν δηλαδή οι λέξεις γράφονται με τον ίδιο ακριβώς τρόπο αλλά έχουν διαφορετική σημασία, για παράδειγμα η λέξη *"άτομο"* όταν αναφέρεται σε πρόσωπα και στο σωματίδιο της φυσικής ή λέξη *"γράμμα"* που αναφέρεται μπορεί να αναφέρεται στο χαρακτήρα του αλφαβήτου και στην επιστολή).

  Επίσης, στην κατηγορία αυτή εντάσσεται και το πρόβλημα της συνωνυμίας των λέξεων (όταν κάποιες λέξεις μοιράζονται μία τουλάχιστον κοινή έννοια), όπως για παράδειγμα οι λέξεις *"λιθάρι"* και *"πέτρα"*. Γενικά ισχύει πως για μία έννοια είναι πιθανό να υπάρχουν περισσότερες της μιας λεκτικές οντότητες που την εκφράζουν και συνιστούν το Σύνολο Συνωνύμων για την έννοια αυτή, όπως επίσης υπάρχουν και διάφοροι βαθμοί συνωνυμίας.

  Τέλος, υπάρχουν και κάποιες ομάδες λέξεων (multi-words) που έχουν νόημα μόνο εφόσον χρησιμοποιούνται μαζί στο κείμενο και φυσικά χρειάζονται επιπλέον γνώσεις (τόσο από τον άνθρωπο και πολύ περισσότερο από τον υπολογιστή)

πέραν των γραμματικών και εννοιακών για την κατανόηση του πραγματικού περιεχομένου τους. Για παράδειγμα, υπάρχουν συνδυασμοί λέξεων που συνθέτουν εκφράσεις όπως *"διαστημικό λεωφορείο"* ή *"γεωγραφικό μήκος"* που αν χρησιμοποιηθούν μόνες τους έχουν τελείως διαφορετικό νόημα.

Με την εισαγωγή των παραπάνω προβλημάτων, εισάγεται αντίστοιχα και το πρόβλημα της ομοιότητας/σχετικότητας μεταξύ λέξεων (ή άλλων μονάδων που απαρτίζουν ένα κείμενο). Ο πρώτος βαθμός συσχέτισης έχει να κάνει με ομοιότητα λόγω κάποιας από τις σχέσεις που περιγράφηκαν παραπάνω (συνωνυμία/αντωνυμία, υπερωνυμία/υπωνυμία). Σε δεύτερο βαθμό εξετάζεται η συγγένεια λέξεων βάσει της χρήσης τους στα κείμενα, αν δηλαδή κάποιες λέξεις τείνουν να εμφανίζονται σε παρόμοια λεκτικά περιβάλλοντα. Σε αυτή την περίπτωση δεν είναι απαραίτητο να υπάρχει σχέση (σημασιολογική ή άλλη) μεταξύ των λέξεων.

Αντιμετωπίζοντας εννοιολογικά το έγγραφο εισάγεται το ερώτημα-πρόβλημα του ποια είναι η βασική μονάδα έννοιας (και ποια η σχέση της με τη βασική μονάδα δομής του εγγράφου): το μόρφημα; η λέξη; η πρόταση; Η απάντηση στο ερώτημα αυτό δεν είναι μονοσήμαντη καθώς το πραγματικό νόημα ενός εγγράφου καθορίζεται από τη συνισταμένη-σημασία που δίνει η ταυτόχρονη παρουσία όλων των εννοιολογικών μονάδων εντός της εξεταζόμενης κειμενικής ποσότητας, δηλαδή χρειάζεται να υπάρχει γνώση για ολόκληρο το περιεχόμενο ενός εγγράφου καθώς η παρουσία επιπλέον μονάδων (π.χ. επιπλέον λέξεων) μπορεί να τροποποιεί το νόημα.

Από τα παραπάνω γίνεται φανερό πως προκύπτουν άπειροι συνδυασμοί μεταξύ των μονάδων της γλώσσας δημιουργώντας έτσι ένα μεγάλο χώρο για τη λύση του προβλήματος (που είναι η κατανόηση του πραγματικού νοήματος των εγγράφων). Αναγκαία είναι η εισαγωγή αφενός αριθμητικών αναπαραστάσεων των εγγράφων (δυαδικές, διανύσματα, πίνακες κτλ) που θα εξυπηρετήσουν την όποια επεξεργασία και ανάλυση αλλά και η αξιοποίηση ευφυών στατιστικών τεχνικών και τεχνικών μηχανικής μάθησης για την αποδοτική ανάλυση.

1.1.3 Η σπουδαιότητα της σημασιολογικής γνώσης στην επεξεργασία και ανάλυση λόγου

Από τα προβλήματα που περιγράφηκαν παραπάνω, γίνεται σαφές πως η σημασιολογική γνώση και πληροφορία που προέρχεται από συγγένειες / έννοιες / κατηγοριοποιήσεις λέξεων είναι μεγάλης σημασίας για την κατανόηση από τον υπολογιστή της πραγματικής σημασίας των κειμένων, ώστε να μπορούν να αναπτυχθούν εφαρμογές επεξεργασίας και ανάλυσης αυτών. Παραδείγματα τέτοιων εφαρμογών είναι:

- Μοντελοποίηση και Παραγωγή Γλώσσας: Πρόκειται για την κατανόηση της λειτουργίας της γλώσσας με πρόβλεψη των κατάλληλων λεκτικών ακολουθιών ή αντικατάσταση με παρεμφερείς. Σπουδαίες εφαρμογές υπάρχουν στην αυτόματη αναγνώριση ομιλίας ή χειρόγραφου κειμένου, στην ορθογραφική διόρθωση καθώς και σε διαλογικά συστήματα ή στη μηχανική μετάφραση.

- Εξαγωγή Πληροφορίας (Information Extraction): Πρόκειται για εφαρμογές που ο χρήστης εισάγει σε ένα σύστημα μια ερώτηση (ή πληροφορία εν γένει) και αναμένει είτε μια απάντηση είτε μία λίστα κειμένων που περιέχουν τη ζητούμενη πληροφορία. Σε τέτοιες εφαρμογές είναι αυτονόητη η ανάγκη χρήσης σημασιολογικής γνώσης ώστε να βρεθούν παρόμοιες λέξεις ή να προσδιοριστεί το θέμα

του κειμένου. Πολύ γνωστές εφαρμογές στην κατηγορία αυτή είναι οι μηχανές αναζήτησης στο διαδίκτυο.

- Εξόρυξη Κειμένου (Text Mining): Πρόκειται για την εξαγωγή πληροφορίας υψηλού επιπέδου (συνήθως μη-πρότερα γνωστής) από διάφορες γραπτές πηγές με στόχο το συνδυασμό και την αξιοποίησή τους τους ώστε να προκύψουν νέα γεγονότα ή υποθέσεις. Η διαφορά από την (πιο γενική) Εξαγωγή Πληροφορίας είναι πως στην τελευταία ο χρήστης συνήθως γνωρίζει την απάντηση (έχει δοθεί από κάποιον άλλο) και το πρόβλημα στην ουσία είναι η ανάδειξη μιας ήδη γνωστής πληροφορίας. Αντίθετα, στην Εξόρυξη Κειμένου στόχος είναι η ανακάλυψη πληροφορίας άγνωστης, που δεν έχει καταγραφεί έως τώρα.

- Αποσαφήνιση Σημασίας Λέξεων (Word Sense Disambiguation): Είναι η διαδικασία προσδιορισμού της σωστής σημασίας μιας αμφίσημης/πολύσημης λέξης.

Γενικά, θα μπορούσε κανείς να πει πως η συγκέντρωση σημασιολογικής γνώσης από λεξικά, θησαυρούς, δίκτυα, οντολογίες και άλλες πηγές γνώσης με σκοπό τη βελτίωση εφαρμογών που έχουν να κάνουν με επεξεργασία και ανάλυση κειμένων είναι ένα βήμα παραπέρα στη δημιουργία (ακόμα πιο) ευφυών συστημάτων που θα λειτουργούν ολοένα και περισσότερο με τη λογική του ανθρώπινου νου.

1.1.4 Κρίσιμα προβλήματα

Κρίσιμα θέματα τα οποία αποτελούν πάντα ανοιχτά και ενδιαφέροντα ερευνητικά ζητήματα στον κλάδο της εξόρυξης και ανάκτησης κειμένου είναι τα ακόλουθα :

1. Αναπαράσταση κειμένου: Για τον ευκολότερο χειρισμό τους τα κείμενα είναι απαραίτητο να μετατρέπονται εύκολα και αποδοτικά από τη μορφή που είναι διαθέσιμα σε κάποια άλλη που να περιγράφει επαρκώς το περιεχόμενό τους. Είναι σαφές πως το συγκεκριμένο ζήτημα αποτελεί από μόνο του ξεχωριστή ερευνητική κατεύθυνση (να υπάρχουν δηλαδή τεχνικές πετυχημένης αναπαράστασης εγγράφων) αλλά είναι και άρρηκτα συνδεδεμένο με τα επόμενα τρία προβλήματα (πως θα αναπαρασταθούν τα έγγραφα προκειμένου να γίνει επεξεργασία τους).

2. Σημασιολογική Συσχέτιση Λέξεων (Word Semantic Relatedness): Τα μέτρα σημασιολογικής συσχέτισης λέξεων χρησιμοποιούνται σε πολλές εφαρμογές όπως αποσαφήνιση σημασίας λέξεων (word sense disambiguation), επέκταση ερωτήσεων (query expansion), επισήμανση σελίδων στο WWW (web page annotation) κ.τ.λ.

3. Κατηγοριοποίηση/Ομαδοποίηση Εγγράφων (Text Categorization/Clustering): Σε τεράστιες συλλογές εγγράφων είναι ιδιαίτερα χρήσιμη η αποδοτική ομαδοποίησή τους ώστε να επιτυγχάνεται καλύτερη οργάνωση, περίληψη, πλοήγηση και ανάκτησή τους. Ειδικά η ομαδοποίηση εγγράφων (δεν περιλαμβάνει καμία εκ των προτέρων γνώση για τις κατηγορίες των εγγράφων) είναι ένα δύσκολο έργο από τη φύση του που δυσχεραίνεται λόγω των φαινομένων της "αραιότητας" (sparsity) και της μεγάλης διάστασης των δεδομένων κειμένου αλλά και της πολύπλοκης σημασιολογίας της φυσικής γλώσσας.

4. Εξαγωγή Θέματος από Κείμενο (Topic Extraction): Οι περισσότερες σύγχρονες εργασίες θεωρούν πως στα κείμενα υπάρχουν διάφορα θέματα, δηλαδή οι

λεκτικές μονάδες που συγκροτούν ένα κείμενο ανήκουν σε διάφορα θέματα και όταν συνθέτονται μαζί δημιουργούν το έγγραφο που εξετάζεται. Η ανακάλυψη αυτών ακριβώς των θεμάτων που υπάρχουν ως "κρυμμένη" πληροφορία στα κείμενα αποτελεί ένα χρήσιμο εργαλείο για την επισήμανση αυτών των εγγράφων και κατόπιν τη χρήση τους για την οργάνωση, περίληψη και αναζήτησή τους.

1.2 Συνεισφορά της διατριβής

Η συνεισφορά της διατριβής συνοψίζεται στα παρακάτω σημεία :

1. Μελέτη των δυνατοτήτων χρήσης εξωτερικών πηγών γνώσης στις υπάρχουσες μεθοδολογίες αναπαράστασης κειμένου και πιο συγκεκριμένα :

   - Χρήση του λεξικού WordNet για την εξαγωγή περισσοτέρων σχέσεων συσχέτισης μεταξύ λέξεων,
   - Χρήση των αποτελεσμάτων αναζήτησης στο WWW για εξαγωγή πληροφορίας για τη σημασιολογική συσχέτιση λέξεων όταν αναζητούνται πληροφορίες για αυτές,
   - Ολοκληρωμένη χρήση της Wikipedia για τη δημιουργία ενός νέου μοντέλου αναπαράστασης εγγράφων: Ενσωμάτωση άρθρων της Wikipedia σε κάθε έγγραφο μέσα από την αντιστοίχηση ενός ή περισσοτέρων λέξεων με το αντίστοιχο άρθρο της Wikipedia, δημιουργώντας έτσι τις *έννοιες* (concepts) του εγγράφου και χρησιμοποιώντας αυτές αντί του μοντέλου BOW. Χρησιμοποιείται όλη η πληροφορία που παρέχεται από τη Wikipedia δηλαδή περιεχόμενο άρθρων, εισερχόμενοι/εξερχόμενοι σύνδεσμοι κτλ ενώ παράλληλα αναπτύσσεται και μέθοδος αποσαφήνισης των εννοιών.

2. Ανάπτυξη ευφυών τεχνικών για την αποτελεσματικότερη ανάλυση κειμένου σε διάφορες εφαρμογές όπως:

   - Σημασιολογική συσχέτιση λέξεων: Ανάπτυξη υβριδικής τεχνικής που εκμεταλλεύεται τον αριθμό εμφανίσεων στα αποτελέσματα αναζήτησης, τους τίτλους, τα snippets και τα urls που επιστρέφονται από μία μηχανή αναζήτησης στο WWW. Η μέθοδος ενισχύει το μοντέλο BOW με εξαγωγή λεξικο-συντακτικών προτύπων από τα αποτελέσματα αναζήτησης για όλες τις σημασιολογικές σχέσεις που υπαγορεύονται από το λεξικό WordNet (συνώνυμα, υπερώνυμα, αντώνυμα, μερώνυμα). Χρήση μηχανών διανυσμάτων υποστήριξης (Support Vector Machines, SVM) για την επίλυση του προβλήματος παλινδρόμησης ("regression") που προκύπτει.
   - Ιεραρχική Ομαδοποίηση Εγγράφων: Αναπτύσσονται δύο τεχνικές ομαδοποίησης εγγράφων (ειδικά αναπτυχθείσες για έγγραφα). Και οι δύο μέθοδοι λειτουργούν ιεραρχικά, δημιουργώντας δενδρικές δομές κάνοντας εύκολη την πλοήγηση στα έγγραφα αλλά και την αναζήτησή τους.
   - Οπτικοποίηση Μεγάλων Συλλογών Εγγράφων: Με αξιοποίηση του μοντέλου των Αυτο-Οργανούμενων Χαρτών (SOM) λαμβάνονται δισδιάστατες απεικονίσεις των νευρώνων στους οποίους αντιστοιχίζονται τα έγγραφα, αποτυπώνοντας πλήρως τις τοπολογικές σχέσεις μεταξύ των εγγράφων αλλά και των θεματικών περιοχών στις οποίες εντάσσονται, πραγματοποιώντας

έτσι και εξαγωγή των θεμάτων με τα οποία ασχολείται η συλλογή εγγράφων.

1.3 Δομή της διατριβής

Η διατριβή αυτή αποτελείται από ακόμη 5 Κεφάλαια.

Το Κεφάλαιο 2 εισάγει τον αναγνώστη στο πρόβλημα της σημασιολογίας που δυσχεραίνει την αναπαράσταση κειμένων στον υπολογιστή. Παρουσιάζονται τα προβλήματα της σημασιολογίας λέξεων, φράσεων, προτάσεων και καθορίζονται οι στόχοι ενός μοντέλου αναπαράστασης εγγράφων στον υπολογιστή (ενσωμάτωση όσο το δυνατόν περισσότερης πληροφορίας και αποδοτική χρήση). Γίνεται μία πρώτη εισαγωγή στο κύριο μοντέλο αναπαράστασης (μοντέλο χώρου διανυσμάτων, VSM) και εντοπίζονται τα προβλήματα που ανακύπτουν και καθιστούν αναγκαία την ενίσχυση της σημασιολογίας του υπολογιστή.

Στο Κεφάλαιο 3 εξετάζεται το ζήτημα της στατιστικής σημασιολογίας και αναπτύσσεται μία μεθοδολογία για την ποσοτικοποίηση της σημασιολογικής συσχέτισης λέξεων. Αξιολογώντας το ρόλο των λέξεων στις σημερινές μεθόδους αναπαράστασης και αναζήτησης εγγράφων (π.χ. στις μηχανές αναζήτησης), η μεθοδολογία προτείνει ένα βαθμωτό μέτρο που εξετάζει όλες τις πιθανές σχέσεις δύο λέξεων και αποφασίζει μέσω μιας τιμής για τη σχετικότητα δύο οποιωνδήποτε λέξεων. Η μεθοδολογία συγκρίνεται με τις υπάρχουσες μεθόδους της βιβλιογραφίας.

Αναγνωρίζοντας τις αδυναμίες της αναπαράστασης εγγράφων με λέξεις (αλλά και των περιορισμών που παρουσιάζονται μέσω της χρήσης ενός βαθμωτού μέτρου σημασιολογικής συσχέτισης λέξεων), στο Κεφάλαιο 4 γίνεται μία επισκόπηση των μεθόδων αναπαράστασης εγγράφων. Αναπτύσσονται τα κυριότερα μοντέλα που έχουν προταθεί και διερευνηθεί και κατόπιν παρουσιάζεται αναλυτικά το μοντέλο αναπαράστασης βάσει των εννοιών που αναπτύχθηκε στα πλαίσια της διατριβής με χρήση πληροφορίας από τη Wikipedia. Το μοντέλο αυτό αφενός εισάγει περισσότερη σημασιολογία, αφετέρου συμπιέζει κατά πολύ το χώρο αναπαράστασης (σε σχέση με το κλασσικό μοντέλο χώρου διανυσμάτων).

Στο Κεφάλαιο 5 εισάγεται ο χώρος των εννοιών (αντί του χώρου των λέξεων) για την αντιμετώπιση του ζητήματος της ομαδοποίησης εγγράφων και παρουσιάζονται αναλυτικά οι μεθοδολογίες που αναπτύχθηκαν (συγκριτικά με τις μεθόδους της βιβλιογραφίας). Το αποτέλεσμα είναι η κατασκευή ιεραρχικών ομάδων στις οποίες χωρίζονται τα έγγραφα μιας συλλογής ενώ μέσω της δεύτερης μεθοδολογίας είναι δυνατή και η τοπολογική αναπαράσταση των σχέσεων μεταξύ των εγγράφων (και συνακόλουθα των θεμάτων τα οποία πραγματεύονται) σε ένα δισδιάστατο χώρο.

Τέλος, στο Κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα και πορίσματα της διατριβής καθώς και καταγράφονται οι μελλοντικές κατευθύνσεις στη συγκεκριμένη περιοχή έρευνας.

□

Κεφάλαιο 2

Το πρόβλημα της σημασιολογίας

"He was white and shaken, like a dry martini."
P.G. Wodehouse, Cocktail Time (1958)

2.1 Γενικά για τη σημασιολογία

Η σημασιολογία της ανθρώπινης γλώσσας, ο τρόπος δηλαδή μεταβίβασης νοημάτων μέσω της γλώσσας, είναι ένα αντικείμενο που απασχολεί επιστήμονες από πολλούς χώρους (γλωσσολόγους, φιλοσόφους κτλ) και πρόσφατα και επιστήμονες από το χώρο της τεχνητής νοημοσύνης. Η βασικότερη θεώρηση που μπορεί να κάνει κανείς (και είναι και η πιο λογική στον ανθρώπινο νου) είναι πως η η σημασία μιας πρότασης, συντίθεται από τις σημασίες των λέξεων που την αποτελούν, οι οποίες δίνονται από ορισμούς, όπως αυτούς που βρίσκει κανείς στα λεξικά. Αυτή η αντιμετώπιση σύμφωνα με τον [Saeed, 1997] οδηγεί σε 3 προβλήματα:

- Το αναδρομικό πρόβλημα του ορισμού των λέξεων: Εφόσον για τη σημασία των λέξεων χρειάζεται η αναφορά σε λεξικά, τότε για τον καθορισμό της σημασίας κάθε λέξης, απαιτείται η εύρεση της σημασίας του ορισμού της. Για να επιτευχθεί αυτό θα πρέπει για κάθε λέξη του ορισμού, να γίνει αναφορά ξανά στο λεξικό κ.ο.κ. οδηγώντας αρκετά γρήγορα σε φαύλους κύκλους.

- Η ακρίβεια των ορισμών: Το νόημα των λέξεων βρίσκεται στο μυαλό των ανθρώπων που μιλούν τη γλώσσα, σαν ένα είδος γνώσης. Η γνώση αυτή συχνά συγχέεται με τη γενική γνώση των ανθρώπων για τον κόσμο (ενώ δε θα πρέπει να υποτιμηθεί το γεγονός πως κάθε άνθρωπος έχει διαφορετικές γνώσεις απ' όλους τους άλλους). Για παράδειγμα, αν για τη λέξη "*φάλαινα*" χρησιμοποιηθεί ο ορισμός "*μεγάλο θηλαστικό που ζει στη θάλασσα*" τότε είναι έγκυρο να θεωρηθεί άραγε πως όποιος γνωρίζει τη σημασία της λέξης "*θηλαστικό*" δε γνωρίζει τη σημασία της λέξης "*φάλαινα*"; Ένα άλλο πρόβλημα είναι ότι οι χρήστες δε χρησιμοποιούν κάθε λέξη με την ίδια σημασία. Για παράδειγμα μπορεί ένας φυσικός να καταλαβαίνει τη λέξη "*ὁρμή*" διαφορετικά από κάποιον που δεν ξέρει φυσική.

- Η ιδιαιτερότητα της περίστασης: Πολλές φορές σε μία πρόταση, η πραγματική σημασία της χρήσης μιας λέξης είναι κατανοητή μόνο με χρήση της κοινής λογικής. Για παράδειγμα, η πρόταση "*ωραίος καιρός!*" έχει πολύ διαφορετικό νόημα όταν λέγεται μια ηλιόλουστη μέρα και όταν λέγεται μια μέρα που βρέχει καταρρακτωδώς (ειρωνικά). Επίσης, ίσως κάποιες φορές είναι αναγκαία η κοινή λογική

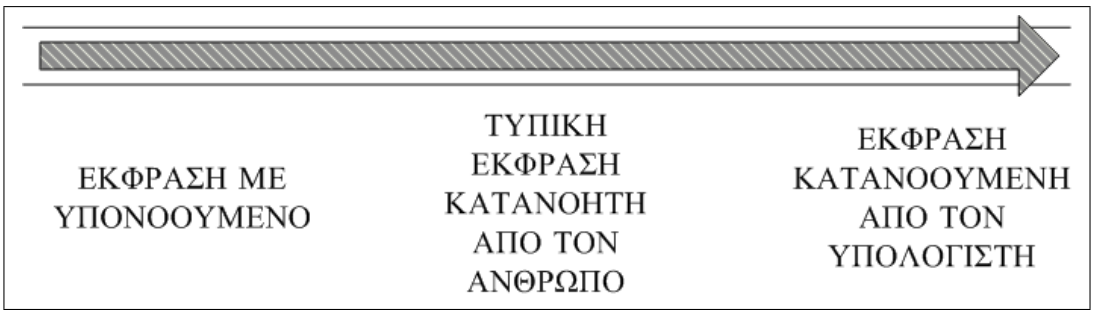
κάποιου για να βγει κάποιο συμπέρασμα. Π.χ. κάποιος που λέει σε ένα φίλο του σε ένα μπαρ *"πήγε τρεις η ώρα"* και στην ουσία εννοεί *"Θέλω να φύγουμε"*.

Έχουν γίνει διάφορες προσπάθειες να αντιμετωπιστούν τα παραπάνω προβλήματα (όπως με χρήση μεταγλώσσας ή διαχωρισμό της σημασίας σε δύο συνιστώσες (μία για κάθε χρήση και μία εξαρτώμενη από την περίσταση)) αλλά γενικά παραμένουν ανοιχτά προβλήματα της επιστημονικής κοινότητας.

Από τα παραπάνω γίνεται σαφής η πολυπλοκότητα της σημασιολογίας της γλώσσας και επομένως οι δυσκολίες που ανακύπτουν στην οριοθέτηση των διαφόρων προβλημάτων. Τα προβλήματα αυτά γίνονται πιο έντονα, όταν χρειάζεται η γλώσσα (με τη μορφή ενός κειμένου) να αναπαρασταθεί στον υπολογιστή (ο οποίος ως γνωστόν δεν έχει γενική γνώση αλλά μόνον ότι του παρέχει ο χρήστης). Ανακύπτει λοιπόν το μεγάλο στοίχημα της δημιουργίας κειμενικού περιεχομένου κατανοητού (και επομένως επεξεργάσιμου) από τον υπολογιστή. Θεωρώντας το φάσμα της σημασιολογίας των δεδομένων του Σχήματος 2.1 μπορεί να πει κανείς πως όσο πιο "δεξιά" βρίσκεται το περιεχόμενο, τόσο πιο πολύ υπερισχύουν τα εξής χαρακτηριστικά:

- προσβασιμότητα και δυνατότητα αξιοποίησης από υπολογιστές,
- λιγότερη αμφισημία ή αμφισβήτηση,
- επαναχρησιμοποίηση και επεκτασιμότητα,
- σθεναρότητα σε αλλαγές,
- δυσκολία υλοποίησης

Πόσες όμως είναι οι περιπτώσεις που οι εκφράσεις που εμφανίζονται σε ένα κείμενο βρίσκονται "δεξιά" στον άξονα αυτό;



Σχήμα 2.1: *Φάσμα σημασιολογίας*

2.2 Προβλήματα στην προσπάθεια κατανόησης της σημασιολογίας

Από πολύ παλιά υπάρχει η άποψη [W.N. & D.A., 1955] πως από τη μελέτη μεγάλων συλλογών κειμένου, είναι δυνατό να κατανοηθεί η ανθρώπινη σημασιολογία και η πραγματική χρήση των λέξεων. Η στατιστική μελέτη της συνεμφάνισης λεκτικών μονάδων σε μεγάλες συλλογές κειμένων έγινε ακόμα πιο έντονη λόγω της διαθεσιμότητας εγγράφων (παλαιότερα μέσω της ψηφιοποίησής τους, τα πρόσφατα χρόνια μέσω της μεγάλης ανάπτυξης του WWW). Η μελέτη και ανάλυση αυτών των συλλογών κειμένου δεν έρχεται όμως χωρίς προβλήματα. Τέσσερα είναι τα πιο βασικά:

- Δε λαμβάνεται υπόψιν η εγγενής παραμόρφωση των διαφόρων πτυχών του σημασιολογικού χώρου λόγω της αλληλεξάρτησης των εννοιών,
- Δεν ανιχνεύονται συνεμφανίσεις αφηρημένων δομών (ομάδες λέξεων που εμφανίζονται σε συγκεκριμένη μορφή π.χ.), ειδικά όταν δεν έχουν σχέση μεταξύ τους,
- Υπάρχει έλλειψη της εξωτερικής γνώσης που ένα άτομο λαμβάνει μέσω της μάθησης και της εμπειρίας,
- Γίνεται η υπόθεση πως οι λέξεις αποτελούν ατομικές οντότητες.

2.2.1 Η εγγενής παραμόρφωση του σημασιολογικού χώρου

Μία βασική αρχή που υπάρχει στις περισσότερες μεθοδολογίες ανάλυσης κειμένων ([Lund & Burgess, 1996], [Fletcher & Linzie, 1998]) είναι πως "η σημασία μιας λέξης μπορεί να θεωρηθεί ως μία θέση στο σημασιολογικό χώρο και η διάσταση του χώρου αυτού καθώς και η θέση της κάθε λέξης σε αυτόν μπορούν να βρεθούν από τις αποστάσεις μεταξύ των διαφόρων λέξεων". Η άποψη αυτή υπονοεί πως βρίσκοντας τις αποστάσεις (δηλαδή πόσο διαφέρουν σημασιολογικά) ανάμεσα σε όλες τις λέξεις είναι δυνατόν να γίνει μία τοποθέτηση σε ένα ν-διάστατο χώρο όπου το ν θα καθορίζει τις σημασιολογικές συνιστώσες. Άμεση συνέπεια της άποψης αυτής είναι πως οι λέξεις θα έχουν συγκεκριμένες σταθερές τοποθεσίες στο σημασιολογικό χώρο, κάτι το οποίο δεν είναι εντελώς λανθασμένο αλλά παραβλέπει το γεγονός πως οι θέσεις στο σημασιολογικό χώρο είναι (σημασιολογικά) αλληλεξαρτώμενες. Η ορθή θεώρηση είναι πως πρέπει κάθε λέξη να μπορεί να μετακινηθεί στο σημασιολογικό χώρο ανάλογα με τη σημασία (που εξαρτάται και από το λεξιλογικό περιβάλλον) στο οποίο θα χρησιμοποιηθεί.

Για παράδειγμα, η λέξη *στυλό* βρίσκεται (στο σημασιολογικό χώρο) πιο κοντά σε λέξεις όπως *μολύβι, γράφω, χαρτί* κλπ αλλά αν κάποιος εκεί που π.χ. γράφει χρησιμοποιήσει το στυλό για να ξύσει τα μαλλιά του, τότε αυτομάτως το *στυλό* μετακινείται σημασιολογικά προς άλλες λέξεις όπως *ξύνω, φαγούρα* κλπ. Επομένως, πέραν του εντοπισμού της κύριας σημασίας μιας λέξης είναι απαραίτητη και η εξέταση της λέξης (και της σημασίας της) στο περιβάλλον που χρησιμοποιείται κάθε φορά.

2.2.2 Εντοπισμός αφηρημένων δομών συνεμφάνισης

Η ανάκτηση γνώσεων για τη σημασιολογία μιας συγκεκριμένης λέξης επιτρέπει την αξιολόγηση της ποιότητας των σχέσεων ανάμεσα σε αυτή τη λέξη και σε άλλες. Υπάρχουν τουλάχιστον δύο βάσεις για αυτές τις σχέσεις [Chalmers et al., 1992]. Για παράδειγμα, η πρόταση *Ο Γιάννης είναι πραγματικό κυπαρίσσι* αναφέρεται σε κάποια χαρακτηριστικά του Γιάννη και του κυπαρισσιού (στη συγκεκριμένη περίπτωση είναι το *ψηλός* και *λεπτός*).

Όμως, στην πρόταση *Ο Γιάννης είναι πραγματικός κυνηγός με τις γυναίκες* δεν υπονοείται πως ο Γιάννης κυνηγάει τις γυναίκες και τις σκοτώνει, αλλά θεωρείται πως η σχέση του με τις γυναίκες είναι καλή, αναλογικά με τη σχέση του κυνηγού με το θήραμά του. Η αναλογία του πρώτου παραδείγματος είναι καθαρά προσδιοριστική (αναφέρεται στα κοινά "επιφανειακά" χαρακτηριστικά) ενώ η αναλογία του δεύτερου παραδείγματος είναι κυρίως σχετική και βασίζεται σε συμπεριφοριακά χαρακτηριστικά.

Το πρώτο είδος αναλογιών μπορεί να εντοπιστεί από τεχνικές συνεμφάνισης ενώ το δεύτερο (που αποτελεί και τη βάση όλων των αναλογιών [Gentner, 1983]) όχι.

2.2.3 Χρήση πληροφοριών εξωτερικής γνώσης

Ένα εξίσου σημαντικό πρόβλημα που παρατηρείται είναι πως υπάρχει σημαντικός όγκος πληροφορίας που θεωρείται γνωστός από τον άνθρωπο και είναι σχετικός με την πρόταση που αναφέρεται κάποιο κείμενο. Για παράδειγμα, έστω η φράση *οι πατέρες είναι πάντα αρσενικού γένους* όπως και η φράση *Υπάρχει ένας ακήρυχτος πόλεμος ανάμεσα στο Ισραήλ και την Παλαιστίνη*. Και στις δύο αυτές προτάσεις-δηλώσεις υπάρχει γνώση η οποία δεν είναι γνωστή (εκ των προτέρων) σε κανένα πρόγραμμα ανάλυσης κειμένου, αλλά οι άνθρωποι έχουν τη δυνατότητα να κάνουν χρήση (πολλές φορές υποσυνείδητα) των σχετικών και σημασιολογικών πληροφοριών για τη σημασία μιας λέξης ή μιας πρότασης, π.χ. γιατί ένας πατέρας είναι πάντα αρσενικού γένους ή ποια ακριβώς είναι η κατάσταση που επικρατεί ανάμεσα σε Παλαιστίνη και Ισραήλ.

2.2.4 Οι λέξεις δεν είναι ατομικές οντότητες

Έστω ένα παράδειγμα για μία ερώτηση βασισμένη στο υποσυνείδητο από την εργασία [French, 1988]:
"Με βαθμό από το 1 (άσχημο) έως το 10 (άριστο) αξιολογείστε το όνομα Flugly ως:

- όνομα μιας διάσημης ηθοποιού του Hollywood,
- όνομα ενός λογιστή σε μια ταινία του W.C. Fields"

Οι άνθρωποι μπορούν να κάνουν αυτή την αξιολόγηση πολύ εύκολα. Το όνομα Flugly θεωρείται κακής ποιότητας για όνομα μιας ηθοποιού, ενώ μπορεί να θεωρηθεί αξιοπρεπές για ένα λογιστή. Πως προκύπτει όμως αυτή η γνώση, εφόσον θεωρείται πως κανείς δεν έχει ξαναδεί τη λέξη Flugly;

Πιο συγκεκριμένα, η λέξη Flugly περιέχει το όχι ευχάριστο για το αυτί λαρυγγικό g, επίσης έχει το φώνημα ug και περιέχει και τη λέξη ugly.

Το ζήτημα που καταδεικνύει το παραπάνω παράδειγμα είναι πως οι λέξεις περιέχουν κρίσιμες πληροφορίες (που προκύπτουν από τα μορφήματα και τα φωνήματα που εμπεριέχουν) που συνεισφέρουν στη σημασία της λέξης και στον τρόπο της αντίληψής της από τους ανθρώπους. Η πιθανή επέκταση της ανάλυσης σε επίπεδο γραμμάτων ή συλλαβών δεν είναι σίγουρο πως θα έχει τα επιθυμητά αποτελέσματα, χώρια του ότι θα οδηγήσει το πρόβλημα σε έναν υπολογιστικά μεγαλύτερο χώρο.

2.3 Σημασιολογία στο επίπεδο των λέξεων

Η λεκτική σημασιολογία (lexical semantics) παραδοσιακά έχει τους εξής στόχους:

1. να αναπαραστήσει τη σημασία κάθε λέξης,
2. να καταδείξει πως οι σημασίες των λέξεων μιας γλώσσας σχετίζονται.

Το βασικότερο πρόβλημα είναι πως μία λέξη μπορεί να σχετίζεται τόσο με άλλες λέξεις που βρίσκονται στην ίδια πρόταση αλλά και με λέξεις που δεν είναι παρούσες. Για παράδειγμα, αν κάποιος διαβάσει την πρόταση *Μόλις είδα τη μητέρα σου να βγαίνει από το σπίτι*, καταλαβαίνει πως ο γράφων είδε μια γυναίκα. Αυτό το οποίο υπονοείται είναι πως είτε υπάρχει κάποια σχέση μεταξύ των λέξεων *μητέρα* και *γυναίκα*, είτε πως η λέξη *μητέρα* περιέχει ένα σημασιολογικό στοιχείο *γυναίκα* ως μέρος της σημασίας της.

Σε κάθε περίπτωση αυτό που καθίσταται σαφές είναι πως οι σχέσεις μεταξύ λέξεων είναι σημείο αναφοράς στον τρόπο με τον οποίο δομείται η γλώσσα.

Υπάρχουν διάφοροι τρόποι με τους οποίους μπορεί να σχετίζονται δύο λέξεις και πιο συγκεκριμένα μία λέξη μπορεί να ανήκει ταυτόχρονα σε περισσότερες της μιας σχέσεις, δημιουργώντας ένα δίκτυο συνδέσεων μεταξύ των λέξεων, το λεγόμενο *λεξικό*. Παρακάτω παρατίθενται οι συνηθέστερες μορφές σχέσεων μεταξύ λέξεων.

2.3.1 Ομωνυμία

Θεωρείται η πιο απλή και σημασιολογικά η λιγότερο ενδιαφέρουσα σχέση ανάμεσα σε λέξεις. Παραδοσιακά, η ομωνυμία θεωρείται ως μία σχέση μεταξύ λέξεων που έχουν την ίδια προφορά αλλά διαφορετική ορθογραφία και σημασία, π.χ. σήκω-σύκο, μίλα-μήλα, λίπη-λύπη κτλ.

2.3.2 Πολυσημία

Υπάρχει μια ειδοποιός διαφορά ανάμεσα στην ομωνυμία και στην πολυσημία. Και οι δύο ασχολούνται με πολλαπλές έννοιες της ίδιας γραφής μιας λέξης αλλά η ομωνυμία προϋποθέτει πως δεν υπάρχει καμία σχέση ανάμεσα στις διαφορετικές έννοιες της ίδιας λέξης. Αυτό είναι και η διαφορά της σε σχέση με την πολυσημία, στην οποία οι έννοιες θεωρούνται σχετιζόμενες μεταξύ τους (και γιαυτό στα διάφορα λεξικά παρουσιάζονται κάτω από το ίδιο λήμμα). Επίσης, τα πολύσημα έχουν κοινή ετυμολογία, σε αντίθεση με τα ομώνυμα, τα οποία εξελίχτηκαν από διαφορετικές πηγές και συμπτωματικά κατέληξαν να έχουν την ίδια μορφή, π.χ. *γράμμα* (στοιχείο αλφαβήτου, επιστολή) κτλ.

2.3.3 Συνωνυμία

Η συνωνυμία αφορά διαφορετικές λέξεις που έχουν όμως την ίδια σημασία, π.χ. ήρεμος-ήσυχος, μάζα-ύλη, πλάνη-απάτη, συνοχή-ενότητα, δικηγόρος-συνήγορος κτλ. Από τα προηγούμενα παραδείγματα γίνεται φανερό πως η συνθήκη της απόλυτης συνωνυμίας δεν ισχύει (σχεδόν ποτέ) για τους εξής λόγους:

- δύο λέξεις σπάνια εμφανίζουν πλήρη ταυτότητα περιγραφικής, εκφραστικής και κοινωνικής σημασίας γιατί συνήθως διαφοροποιούνται σε κάποια από τα τρία είδη σημασίας: π.χ. οι λέξεις αστυνομικός-μπάτσος έχουν την ίδια περιγραφική σημασία αλλά διαφοροποιούνται στην εκφραστική καθώς το δεύτερο πολλές φορές είναι φορτισμένο με τη συναισθηματική στάση του γράφοντα. Αντίστοιχα, οι λέξεις μπακάλικο-παντοπωλείο έχουν ίδια περιγραφική σημασία αλλά διαφέρουν ως προς το επίπεδο ύφους.

- ακόμα και αν δύο λέξεις εμφανίζουν ταυτότητα στο είδος της σημασίας υπάρχουν περιπτώσεις που δε μπορεί να χρησιμοποιηθεί η ίδια λέξη σε όλα τα περιβάλλοντα. Για παράδειγμα στην πρόταση *Η θάλασσα είναι ήρεμη/ήσυχη* υπάρχει ταυτότητα μεταξύ των συνωνύμων ήρεμος/ήσυχος, δεν ισχύει όμως το ίδιο για την πρόταση *Άσε με ήσυχο/ήρεμο*.

2.3.4 Αντωνυμία

Η αντωνυμία αφορά λέξεις που έχουν εντελώς αντίθετη σημασία. Η έννοια βέβαια του "αντιθέτου" είναι ιδιαίτερα γενική, με παρόμοιο τρόπο που περιγράφηκε και η έννοια

της συνωνυμίας παραπάνω. Υπάρχουν διάφορες σχέσεις και βαθμοί αντωνυμίας που περιγράφονται παρακάτω:

1. απλά αντώνυμα: Πρόκειται για ζεύγη λέξεων συμπληρωματικά, δηλαδή το "θετικό" του ενός υποδηλώνει το "αρνητικό" του άλλου και αντίστροφα, π.χ. ζωντανός-νεκρός, επιτυχία-αποτυχία (ενός διαγωνίσματος). Η χρήση δηλαδή σε μια πρόταση της λέξης *ζωντανός* υπονοεί το *όχι νεκρός*,

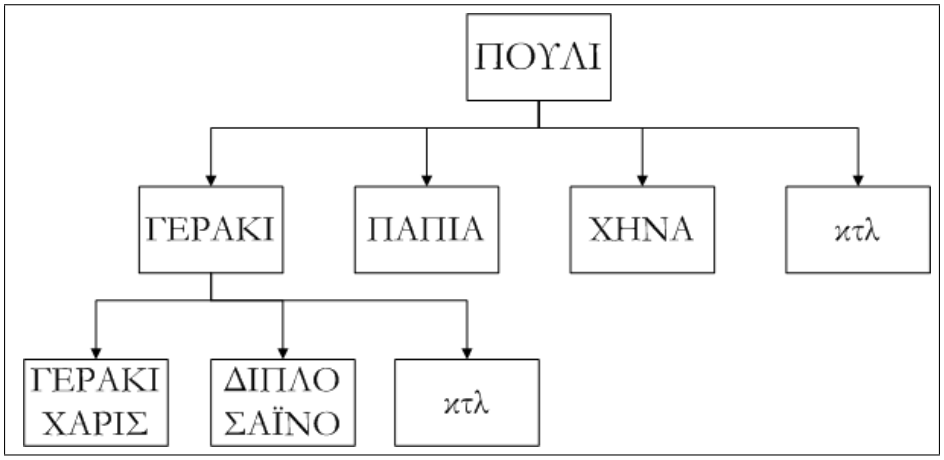
2. διαβαθμισμένα αντώνυμα: Πρόκειται για ζεύγη λέξεων που το "θετικό" του ενός δεν υποδηλώνει απαραίτητα το "αρνητικό" του άλλου, π.χ. πλούσιος-φτωχός, γρήγορος-αργός, όμορφος-άσχημος. Η σχέση αυτή έχει να κάνει κυρίως με επίθετα και έχει δύο χαρακτηριστικά: συνήθως υπάρχουν και ενδιάμεσες λέξεις (π.χ. ανάμεσα στις λέξεις *ζεστός-κρύος* υπάρχει η λέξη *χλιαρός*), οι όροι είναι συνήθως σχετικοί (π.χ. ένα *χοντρό μολύβι* είναι πιο λεπτό από ένα *λεπτό κορίτσι*) και τέλος συνήθως ο ένας όρος είναι πιο κοινός (π.χ. η κοινή ερώτηση είναι *πόσο ψηλός είσαι;* και όχι *πόσο κοντός είσαι;*),

3. "ανάποδα" αντώνυμα: Η ανάποδη σχέση υπάρχει συνήθως μεταξύ λέξεων που περιγράφουν κίνηση και η μία δείχνει μία κατεύθυνση και η άλλη την αντίθετη, π.χ. πάνω-κάτω, μέσα-έξω, άνοδος-κάθοδος,

4. αντίστροφα αντώνυμα: Αφορά λέξεις που έχουν μεταξύ τους μία σχέση και αναλόγως τη σκοπιά χρησιμοποιείται η μία ή η άλλη, π.χ. εργοδότης-εργαζόμενος, κατέχω-ανήκω.

2.3.5 Υπωνυμία/Υπερωνυμία

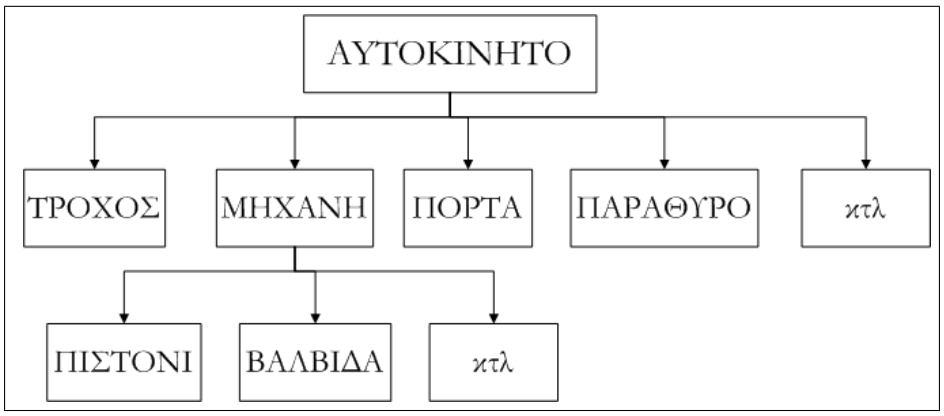
Η υπωνυμία είναι μία σχέση συμπερίληψης. Ένα υπώνυμο περιλαμβάνει τη σημασία μιας πιο γενικής λέξης, π.χ. οι λέξεις *γάτα* και *σκύλος* είναι υπώνυμα της λέξης *ζώο*, οι λέξεις *αδελφή* και *μητέρα* είναι υπώνυμα της λέξης *γυναίκα*. Η πιο γενική λέξη καλείται υπερώνυμο. Πολλές λέξεις ενός λεξικού συνδέονται μεταξύ τους με σχέσεις υπωνυμίας/υπερωνυμίας και αν παρασταθούν όλες, τότε το προκύπτον σημασιολογικό δίκτυο σχηματίζει μια *ταξινομία*. Παράδειγμα μιας τέτοιας ταξινομίας φαίνεται στο Σχήμα 2.2. Στη συγκεκριμένη περίπτωση το *διπλοσάινο* είναι υπώνυμο του *γερακιού* και αυτό είναι υπώνυμο του *πουλιού*.



Σχήμα 2.2: *Παράδειγμα ταξινομίας*

2.3.6 Μερωνυμία

Η μερωνυμία είναι ένας όρος που χρησιμοποιείται για να περιγράψει μία σχέση PART-OF ανάμεσα σε λέξεις, π.χ. οι λέξεις *εξώφυλλο* και *σελίδα* είναι μερώνυμα του *βιβλίου*. Η μερωνυμία μπορεί να αναπαρασταθεί επίσης με ταξινομικές σχέσεις (ιεραρχικά) όπως ακριβώς η σχέση υπωνυμίας. Ένα παράδειγμα φαίνεται στο Σχήμα 2.3.



Σχήμα 2.3: *Παράδειγμα μερωνυμίας*

Τα χαρακτηριστικά της μερωνυμίας είναι λιγότερο σαφή από αυτά της υπωνυμίας. Τα μερώνυμα μπορεί να διαφέρουν ανάλογα με το πόσο απαραίτητο είναι το "μέρος" για το "όλον" (π.χ. η *μύτη* θεωρείται απαραίτητη για το *πρόσωπο* όχι όμως η *σοφίτα* για το *σπίτι*). Επίσης, η μερωνυμία διαφέρει από την υπωνυμία όσον αφορά τη μεταβατικότητα. Σε μία σχέση υπωνυμίας ισχύει πάντα η μεταβατική ιδιότητα, π.χ. το *νύχι* είναι μερώνυμο του *δαχτύλου* και το *δάχτυλο* είναι μερώνυμο του *χεριού*, έτσι μπορεί κανείς να πει πως το *χέρι* έχει *νύχια*. Δεν ισχύει όμως το ίδιο για τη μερωνυμία *τζάμι-παράθυρο-δωμάτιο*, καθώς δε μπορεί κανείς να πει πως το *δωμάτιο* έχει *τζάμι*.

2.3.7 Ο ρόλος των λέξεων στη σημασιολογία

Πολλοί θεωρούν πως οι λέξεις και οι σημασίες τους είναι χρήσιμες εφόσον παρέχουν τα απαραίτητα εφόδια για την κατανόηση του νοήματος μιας πρότασης (και επαγωγικά ολόκληρου του κειμένου). Από την περιγραφή των παραπάνω σχέσεων γίνεται φανερό πως οι αλληλεπιδράσεις ανάμεσα στις λέξεις είναι πολλές και σε αρκετές περιπτώσεις αλληλεπικαλυπτόμενες (δύο λέξεις μπορεί να συνδέονται με παραπάνω της μιας σχέσεις). Δημιουργείται έτσι ένα πλούσιο σημασιολογικό δίκτυο ανάμεσα σε λέξεις που καθορίζει τι σημαίνουν οι λέξεις αλλά και πως χρησιμοποιούνται.

Γνωστικά, η σημασία της λεξικής σημασιολογίας έγκειται στο γεγονός πως οι λέξεις αποτελούν ονόματα για συγκεκριμένες έννοιες/ιδέες (concepts). Για παράδειγμα, στις περισσότερες γλώσσες δεν υπάρχει μία και μόνη λέξη που να ονοματίζει τη μυρωδιά ενός ροδάκινου ή τη μαλακή περιοχή δέρματος στο κάτω μέρος του πήχη του χεριού (παρότι θα μπορούσε). Επιπρόσθετα, είναι αρκετά συνηθισμένο για μία γλώσσα να επιλέγει να λεξικογραφήσει διαφορετικές λέξεις, ανάλογα με τις συνήθειες, την παράδοση και την εξέλιξή της. Ένα παράδειγμα (δανεισμένο από τη φιλοσοφία) αναφέρει πως δεν υπάρχει καμία γλώσσα που να έχει μία λέξη που να περιγράφει αντικείμενα τα οποία ήταν πράσινα πριν την 1η Ιανουαρίου 2000 και γαλάζια μετά. Ο καθένας θα μπορούσε να επινοήσει μια τέτοια λέξη (π.χ. *πρασάζια*) αλλά προκύπτουν αυτόματα πολλά ερωτήματα: Είναι εύλογη η επιλογή μιας τέτοιας λέξης; Μπορεί να υπάρξει μια τέτοια λέξη; Ποια είναι

η σχέση της με τις υπόλοιπες λέξεις και πως εντάσσεται στο όλο οικοδόμημα της γλώσσας;

Ο εντοπισμός συστηματικών προτύπων που καθορίζουν τις σημασίες και τη χρήση των λέξεων μπορεί να δώσει απαντήσεις σε ερωτήματα όπως τα παραπάνω. Γενικά υπάρχουν δύο παράλληλες κατευθύνσεις: από τη μία πλευρά οι φυσικές κλάσεις των λέξεων που μοιάζουν συντακτικά μεταξύ τους (π.χ. η κλάση των ουσιαστικών) και έχουν παρόμοιο τρόπο χρήσης και από την άλλη οι σημασιολογικές κλάσεις των λέξεων δηλαδή οι λέξεις που έχουν την ίδια έννοια.

Καταλήγοντας, μπορεί κανείς να πει πως η μελέτη του πως η θεμελιώδης δομή της σημασίας μιας λέξης αλληλεπιδρά με τις διάφορες συντακτικές δομές (όπως χρησιμοποιούνται στο γραπτό και προφορικό λόγο), βοηθά στην εξερεύνηση της φύσης της γλώσσας, της αντίληψης και της γνώσης. Αυτή τη δυνατότητα σε συνδυασμό με τη δυναμική της στατιστικής σημασιολογίας, αξιοποιεί η διατριβή στο Κεφάλαιο 3.

2.4 Σημασιολογία σε επίπεδο φράσεων, προτάσεων και δομής

Προχωρώντας παραπέρα από το επίπεδο των λέξεων, η σημασιολογία γίνεται πιο σύνθετη. Ποια είναι η επόμενη δομική μονάδα που πρέπει να μελετηθεί; Η φράση; Η πρόταση; Η παράγραφος; Στις επόμενες παραγράφους γίνεται μία προσπάθεια συστηματικοποίησης της προσέγγισης.

2.4.1 Ομάδες λέξεων (Multi-words)

Ο όρος *ομάδες λέξεων* ή *πολυλέξεις* (Multi-words) αναφέρεται σε δύο ή περισσότερες λέξεις που εμφανίζουν μεταξύ τους μία συνοχή (συντακτική και σημασιολογική). Εμφανίζονται σε διάφορες συντακτικές και γραμματικές μορφές ανάλογα με τη γλώσσα που μελετάται κάθε φορά, π.χ. σαν δύο ουσιαστικά σε ονομαστική (*ατμοσφαιρική πίεση*) ή το ένα να βρίσκεται σε γενική (*άρμα μάχης*), σαν συνδυασμός ουσιαστικών/επιθέτων με άρθρα ή προθέσεις (*η ώρα η καλή*, *κεραυνός εν αιθρία*), σαν ρηματικές εκφράσεις (*είμαι έξω φρενών*) κ.ο.κ.

Οι ομάδες λέξεων έχουν το κύριο χαρακτηριστικό [Baldwin et al., 2003] πως δεν έχουν καμία ομοιότητα μεταξύ τους σε επίπεδο λεξιλογικό, μορφολογικό, συντακτικό, σημασιολογικό, στατιστικό, επομένως δεν υπάρχει δυνατότητα σαφούς εντοπισμού τους. Επιπροσθέτως, εμφανίζονται να είναι "μη-αποσυνθέσιμες", δηλαδή δε γίνεται να αντικαταστήσεις μία ή και περισσότερες από τις λέξεις τους με άλλη, χωρίς να χαθεί η σημασία (και πολλές φορές το νόημα). Στις περισσότερες γλώσσες, οι λεξιλογικές πηγές (όπως λεξικά, θησαυροί, οντολογίες) παρουσιάζουν έλλειψη κάλυψης και αναφοράς των πολυλεκτικών ομάδων, κάτι το οποίο κάνει τον εντοπισμό τους ακόμα πιο δύσκολο.

Τα προβλήματα που παρουσιάζουν οι πολυλεκτικές ομάδες συνοψίζονται στα ακόλουθα:

- δε μπορούν να προσδιοριστούν σαφή όρια για το θεματικό πεδίο των πολυλεκτικών ομάδων,
- από υπολογιστικής απόψεως, η αναγνώριση των πολυλεκτικών ομάδων είναι ένα ανοιχτό ερευνητικά (και χρήσιμο) ζήτημα σε όλες τις εφαρμογές που εμπλέκεται

Πίνακας 2.1: *Κύρια χαρακτηριστικά σημασιολογίας σε επίπεδο προτάσεων*

Χαρακτηριστικό	**Παράδειγμα**
Συνωνυμία	Ο Γιάννης είναι ελεύθερος Ο Γιάννης δεν έχει παντρευτεί ποτέ
Συνεπαγωγή	Ο Κώστας σκότωσε τη Μαρία Η Μαρία είναι νεκρή
Αντίφαση	Ο Γιάννης μόλις γύρισε από τη Θεσσαλονίκη Ο Γιάννης δεν έχει πάει ποτέ στη Θεσσαλονίκη
Προϋπόθεση	Η Μαδρίτη έχει δήμαρχο γυναίκα Υπάρχει δήμαρχος στη Μαδρίτη
Ταυτολογία	Οι πλούσιοι άνθρωποι είναι πλούσιοι

κείμενο αλλά απουσιάζουν από τα περισσότερα λεξικά λογικού μεγέθους,

- δεν υπάρχει αντιστοιχία ανάμεσα στις διάφορες γλώσσες καθώς για μία πολυλεκτική ομάδα σε μία γλώσσα δεν υπάρχει ακριβής αντιστοιχία στις υπόλοιπες.

Υπάρχουν αρκετές μεθοδολογίες εντοπισμού των πολυλεκτικών ομάδων ([Lin, 1999], [Schone & Jurafsky, 2001], [Bannard et al., 2003], [Mccarthy et al., 2007], [Katz, 2006], [Piao et al., 2006], [Aline et al., 2007], [Korkontzelos & Manandhar, 2010]) και αξιοποιούν κυρίως στατιστικά μέτρα συνεμφάνισης ή άλλες μη-επιβλεπόμενες γραφοειδείς μεθόδους. Το κυριότερο μειονέκτημά τους είναι το γεγονός πως δεν υπάρχει ένα σαφές πλαίσιο επιβεβαίωσης της ορθότητας του εντοπισμού των πολυλεκτικών μονάδων. Οι περισσότερες μεθοδολογίες αξιολογούνται ξεχωριστά ή σε μικρές ομάδες σε διαφορετικές συλλογές κειμένων και με διαφορετικές παραμέτρους, επομένως είναι δύσκολος ο έλεγχός τους. Παρόλα αυτά, εξακολουθεί να παραμένει η ανάγκη εντοπισμού των πολυλεκτικών εκφράσεων σε ένα κείμενο, καθώς περιέχουν σημαντικό σημασιολογικό περιεχόμενο.

Ειδική περίπτωση ομάδων λέξεων αποτελούν οι ονοματικές φράσεις. Πρόκειται για ομάδες λέξεων που περιλαμβάνουν συνηθέστερα κάποιο ουσιαστικό σε συνδυασμό με κάποιο επίθετο, κάποιον προσδιορισμό κτλ (π.χ. *Πρόεδρος των Ηνωμένων Πολιτειών*). Η σημασία τους περιγράφεται αναλυτικά στην Παράγραφο 4.2.1. Σε σχέση με τις ομάδες λέξεων έχουν το πλεονέκτημα πως με απλά εργαλεία επεξεργασίας φυσικής γλώσσας (π.χ. κάποιον αναλυτή που επισημαίνει τις λέξεις με τα αντίστοιχα μέρη λόγου) είναι δυνατό να εντοπιστούν συνδυασμοί που οδηγούν σε ονοματικές φράσεις. Επίσης, θεωρείται πως στις ονοματικές φράσεις περιέχεται το σημαντικότερο νόημα των προτάσεων (σε σχέση με λέξεις που ανήκουν σε άλλα μέρη του λόγου όπως π.χ. τα ρήματα).

2.4.2 Σημασιολογία στα υπόλοιπα δομικά στοιχεία των κειμένων

Εξετάζοντας τη σημασιολογία ενός κειμένου σε επίπεδο προτάσεων (ή/και φράσεων) το ζήτημα γίνεται αρκετά σύνθετο, παρόλα αυτά υπάρχουν διάφορα χαρακτηριστικά που μπορούν να συναντηθούν σε προτάσεις και συγκεντρώνονται στον Πίνακα 2.1.

Τα παραπάνω χαρακτηριστικά αποτελούν μερικά από τα προβλήματα που παρατηρούνται σε επίπεδο προτάσεων και (γενικεύοντας) σε επίπεδο κειμένων μεταξύ τους.

Πως μπορούν δύο προτάσεις συνώνυμες να αναγνωριστούν και να θεωρηθούν ισοδύναμες; Πως μπορεί η γνώση (που συγκεντρώνεται από μία προϋπόθεση) να συγκεντρωθεί; Η απάντηση τέτοιων ερωτημάτων δεν είναι εύκολη χωρίς επαρκείς αναπαραστάσεις κειμένου και κυρίως χωρίς την εισαγωγή εξωτερικής γνώσης (ο αναγνώστης εισάγεται σχετικά στην Παράγραφο 4.3).

2.5 Αναζητώντας μοντέλα αναπαράστασης κειμένου στον υπολογιστή

Τα περισσότερα κείμενα που υπάρχουν σήμερα (και υπάρχουν σε διάφορες διαθέσιμες ηλεκτρονικές μορφές) δε μπορούν να διαβαστούν και να κατανοηθούν από τον υπολογιστή όπως συμβαίνει με τον άνθρωπο. Ο άνθρωπος βλέποντας ένα κείμενο μπορεί αυτομάτως να αναγνωρίσει τα συστατικά που το αποτελούν (προτάσεις, λέξεις αλλά και σημεία στίξης, ειδικά σύμβολα κλπ) και πολύ περισσότερο μπορεί να κατανοήσει τη σημασία τόσο μιας λέξης μεμονωμένα αλλά και σε σχέση με τις περικείμενες λέξεις (context). Για να μπορέσει ο υπολογιστής (αρχικά) να αναγνωρίσει τους όρους ενός κειμένου (συστατικά), πρέπει να προηγηθεί μία διαδικασία μετατροπής του σε άλλη μορφή (συχνά καλείται και δεικτοδότηση) που στην ουσία αποτελεί την αναπαράσταση του κειμένου στον υπολογιστή. Πάνω σε αυτή τη διαδικασία βασίζεται και το μοντέλο χώρου διανυσμάτων (Vector Space Model) που παρουσιάζεται αναλυτικά στην Παράγραφο 2.5.2.

Ο στόχος της αναπαράστασης κειμένων είναι να βρεθεί ένα μοντέλο το οποίο θα ικανοποιεί δύο, συχνά αλληλοσυγκρουόμενες, προϋποθέσεις:

- την ενσωμάτωση όσο το δυνατόν περισσότερης από την πληροφορία που περικλείουν τα κείμενα
- την αποδοτική χρήση του ώστε να είναι απλή και κυρίως γρήγορη.

Η κατασκευή ενός τέτοιου μοντέλου αποτελεί ακόμη και σήμερα μεγάλη πρόκληση για το ερευνητικό πεδίο της αναπαράστασης κειμένων, κυρίως λόγω του μεγάλου όγκου των κειμένων αλλά και της τεράστιας ποσότητας πληροφορίας που περικλείεται σε αυτά (φανερής ή υπονοούμενης). Το πρώτο πρόβλημα που εντοπίζεται είναι πως η υπολογιστική πολυπλοκότητα που προκύπτει αν αναπαρασταθεί ένα κείμενο όσο το δυνατόν πληρέστερα είναι απαγορευτική για οποιαδήποτε εφαρμογή. Ας σκεφτεί κανείς πως η γλώσσα διαθέτει εκατοντάδες χιλιάδες λέξεις, ενώ ο αριθμός των τρόπων με τους οποίους μπορούν να συνδυαστούν και να σχηματίσουν εκφράσεις, προτάσεις και κείμενα είναι πολύ μακριά από τα όρια δυνατοτήτων οποιουδήποτε υπολογιστικού συστήματος σήμερα και πιθανότατα για αρκετά ακόμη χρόνια. Το δεύτερο πρόβλημα έχει να κάνει με τη φύση του γραπτού λόγου: Σχεδόν πάντοτε υπάρχει κάποιου είδους πληροφορία "κρυμμένη" πίσω από τις λέξεις. Αυτή μπορεί να προκύπτει για παράδειγμα από το στυλ ή ύφος γραφής ή από τον τρόπο χειρισμού των σημείων στίξης. Νοήματα που προκύπτουν από μεταφορές και παρομοιώσεις είναι εύκολα αντιληπτά από τον άνθρωπο άλλα όχι από μια μηχανή. Τα ορθογραφικά λάθη σηματοδοτούν για το σύστημα μία καινούρια, εντελώς άγνωστη, λέξη ενώ ζητήματα όπως η αισθητική, επιστημονική ή ψυχαγωγική αξία ενός κειμένου είναι αμφιλεγόμενα ακόμα και για τους ανθρώπους.

Στις επόμενες Παραγράφους γίνεται μια πρώτη εισαγωγή στους τρόπους αναπαράστασης εγγράφων με χρήση του υπολογιστή, έχοντας κατά νου τις δύο προϋποθέσεις που παρουσιάστηκαν παραπάνω.

2.5.1 Συμπιεσμένες μορφές αναπαράστασης κειμένων

Σχετικά με την πρώτη προϋπόθεση που αναφέρθηκε στην προηγούμενη Παράγραφο, το κυριότερο ερώτημα που πρέπει να απαντηθεί είναι τι ακριβώς θα αποτελέσει τη βασική μονάδα της αναπαράστασης (ποιοι θα είναι δηλαδή οι όροι του). Οι επιλογές είναι αρκετές και συνοψίζονται στον Πίνακα 2.2.

Πίνακας 2.2: *Επίπεδα αναπαράστασης κειμένου στον υπολογιστή*

Επίπεδο	**Όρος**
Πρωταρχικό	Χαρακτήρας
Υπο-συλλαβικό	Μόρφημα
Υπο-λεκτικό	Συλλαβή
Λεκτικό	Λέξη (token)
Πολυ-λεκτικό	Προτάσεις/φράσεις κ.α.
Συντακτικό / Σημασιολογικό	Σημασίες/έννοιες κ.α.

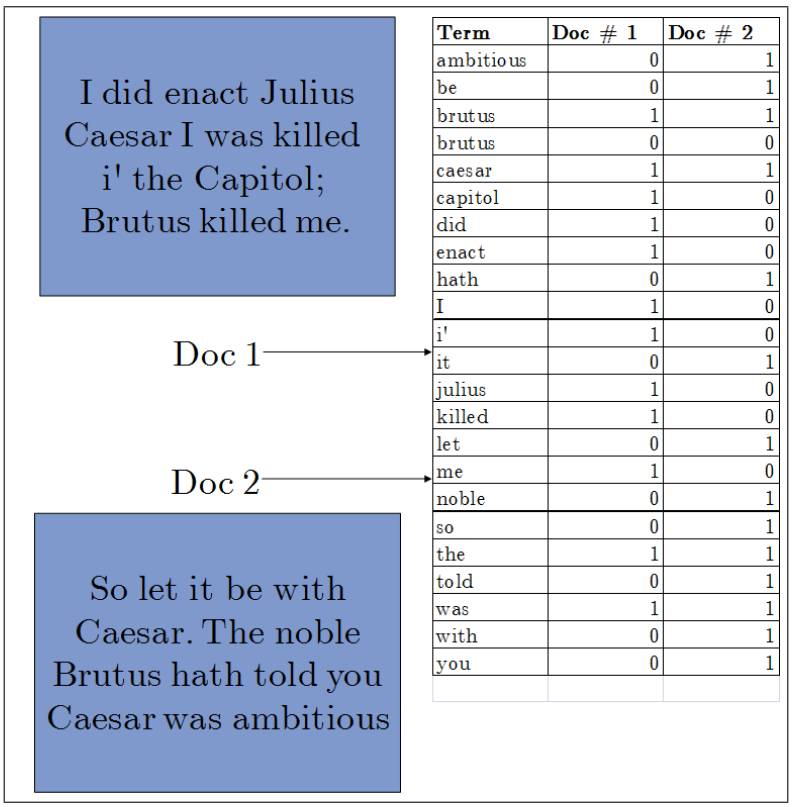
Οι μέχρι τώρα αλγόριθμοι θεωρούν ως πιο αποδοτική την αναπαράσταση με λέξεις, εξ ου και στις μηχανές αναζήτησης επικρατούν ακόμη οι ερωτήσεις με λέξεις-κλειδιά. Αφού καθοριστεί η βασική μονάδα αναπαράστασης(όρος), το επόμενο βήμα είναι να καθοριστεί το βάρος κάθε τέτοιας μονάδας στο κείμενο. Και εδώ έχουν προταθεί διάφορα σχήματα που καθορίζουν τη βαρύτητα ενός όρου όπως τα ακόλουθα:

- λογικό μοντέλο (boolean): ανάλογα με την παρουσία ή όχι του όρου στο έγγραφο,
- συχνότητα όρου στο έγγραφο: πόσες φορές εμφανίζεται κάθε όρος σε κάθε έγγραφο,
- συχνότητα όρου στην ομάδα εγγράφων: αν υπάρχουν πολλά έγγραφα διαθέσιμα, σε πόσα από αυτά εμφανίζεται ένας όρος.

Επιλέγοντας το είδος των όρων και ένα σχήμα στάθμισης του βάρους τους στο κείμενο, η (συμπιεσμένη) αναπαράσταση του κειμένου στον υπολογιστή είναι πλέον έτοιμη. Ένα παράδειγμα της διαδικασίας για δύο έγγραφα, χρησιμοποιώντας το λογικό μοντέλο (boolean) φαίνεται στο Σχήμα 2.4. Από το σχήμα αυτό γίνεται φανερό πως μπορούν να συγκριθούν δύο (ή περισσότερα) έγγραφα βάσει των κοινών όρων.

2.5.2 Το μοντέλο χώρου διανυσμάτων (Vector Space Model, VSM)

Από το παράδειγμα του Σχήματος 2.4, γίνεται εμφανής ο τρόπος λειτουργίας του βασικού μοντέλου αναπαράστασης, του μοντέλου χώρου διανυσμάτων (Vector Space Model, VSM, ή Bag-of-Words, BOW). Στο μοντέλο αυτό, θεωρείται καταρχάς μία συλλογή εγγράφων (στο παραπάνω παράδειγμα είναι δύο, αλλά συνήθως είναι πολύ περισσότερα), η οποία είναι επιθυμητό να αναπαρασταθεί. Για το σκοπό αυτό χρησιμοποιούνται διανύσματα (για κάθε έγγραφο της συλλογής), μεγέθους όσο το λεξιλόγιο της συλλογής εγγράφων που χρησιμοποιείται σαν βάση). Αξίζει να σημειωθεί πως το μοντέλο BOW αγνοεί τη σειρά των λέξεων και ασχολείται μόνο με το πόσες φορές εμφανίζονται. Έτσι, σε μία συλλογή, κάθε έγγραφο j αναπαρίσταται δηλαδή από ένα διάνυσμα $\mathbf{d}_j = (w_{1,j}, w_{2,j}, ..., w_{M,j})$ όπου m είναι το σύνολο των διαφορετικών λέξεων



Term	Doc # 1	Doc # 2
ambitious	0	1
be	0	1
brutus	1	1
brutus	0	0
caesar	1	1
capitol	1	0
did	1	0
enact	1	0
hath	0	1
I	1	0
i'	1	0
it	0	1
julius	1	0
killed	1	0
let	0	1
me	1	0
noble	0	1
so	0	1
the	1	1
told	0	1
was	1	1
with	0	1
you	0	1

Σχήμα 2.4: *Παράδειγμα αναπαράστασης 2 εγγράφων με χρήση του λογικού μοντέλου*

και $w_{i,j}$ είναι το βάρος της λέξης i στο εν λόγω έγγραφο j. Το βάρος αυτό (η τιμή δηλαδή της συνιστώσας κάθε διανύσματος) δείχνει τη συμβολή της συγκεκριμένης λέξης στη σημασιολογία του εγγράφου. Στο παράδειγμα του Σχήματος 2.4 τα βάρη είναι δυαδικά (απουσία/παρουσία όρου) αλλά συνήθως επιλέγεται κάποια σταθμισμένη τιμή που υπολογίζεται από τις ακόλουθες εξισώσεις:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.1}$$

όπου $n_{i,j}$ είναι ο αριθμός των εμφανίσεων του εξεταζόμενου όρου t_i στο έγγραφο d_j και ο παρανομαστής αντιστοιχεί στο άθροισμα όλων των εμφανίσεων όλων των όρων στο έγγραφο d_j.

$$idf_i = \log \frac{N}{df_i} \tag{2.2}$$

όπου N είναι ο συνολικός αριθμός των εγγράφων και df_i είναι ο αριθμός των εγγράφων που περιέχουν τον όρο t_i.

Βάσει αυτών των εξισώσεων, το βάρος ενός όρου i σε ένα έγγραφο j υπολογίζεται από την παρακάτω εξίσωση :

$$w_{i,j} = (tf - idf)_{i,j} = tf_{i,j} \cdot idf_i \tag{2.3}$$

όπου μία υψηλή τιμή επιτυγχάνεται από όρους υψηλής συχνότητας (στο συγκεκριμένο έγγραφο) και χαμηλής συχνότητας σε όλη τη συλλογή. Επομένως, τα βάρη τείνουν να αποκλείουν πολύ κοινούς όρους.

Κατ' αυτόν τον τρόπο εφόσον υπάρχουν N διαφορετικά έγγραφα και M διαφορετι-

κοί όροι, το αποτέλεσμα είναι η δημιουργία ενός πίνακα όρων-εγγράφων που οι γραμμές αναπαριστούν τα έγγραφα και οι στήλες τους όρους και οι συνιστώσες του πίνακα είναι τα βάρη $w_{i,j}$ (όπως περιγράφηκαν παραπάνω). Δημιουργείται δηλαδή ένας πίνακας βαρών W μεγέθους $N \times M$ όπως φαίνεται παρακάτω.

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & ... & w_{1,M} \\ w_{2,1} & w_{2,2} & ... & w_{2,M} \\ ... & ... & ... & ... \\ w_{N-1,1} & w_{N-1,2} & ... & w_{N-1,M} \\ w_{N,1} & w_{N,2} & ... & w_{N,M} \end{bmatrix} \tag{2.4}$$

2.5.3 Η ανάγκη για ενίσχυση της σημασιολογίας του υπολογιστή

Με τη διαδικασία που περιγράφηκε στην προηγούμενη Παράγραφο, περιορίζεται ο όγκος της πληροφορίας των κειμένων στις δομικές μονάδες τους και στα βάρη τους. Ένα από τα σημαντικότερα ζητήματα που δυσχεραίνει την αναζήτηση με λέξεις-κλειδιά είναι η αγνόηση τυχόν σημασιολογικών σχέσεων μεταξύ λέξεων (και το οποίο είναι ανέφικτο εφόσον ο υπολογιστής δεν έχει την ανθρώπινη ευφυΐα). Τα κυριότερα προβλήματα της σημασιολογίας αναλύθηκαν στην Παράγραφο 2.2 και στις επόμενες δύο (2.3 και 2.4) έγινε προσπάθεια αποσύνθεσης των προβλημάτων αυτών σε επίπεδο λεκτικό, προτασιακό και ευρύτερο.

Από την παραπάνω μελέτη, γίνεται λοιπόν επιτακτική η ανάγκη αξιοποίησης διαθέσιμων πόρων ώστε να παρασχεθεί στον υπολογιστή η απαραίτητη γνώση για να προσομοιάσει τη δράση του ανθρώπου. Για παράδειγμα στο Σχήμα 2.4 που επιλέγεται η δυαδική αναπαράσταση με βάση τις λέξεις, ο υπολογιστής δεν έχει τρόπο να γνωρίζει πως οι όροι *Julius Caesar* πρέπει να μείνουν ενωμένοι και όχι να σπάσουν σε 2 διαφορετικούς (διότι καταστρέφεται έτσι το νόημά τους). Ο μοναδικός τρόπος να κατανοήσει ο υπολογιστής τέτοια προβλήματα είναι μέσω της εισαγωγής γνώσης για την ενίσχυση της σημασιολογίας του υπολογιστή.

Προσπάθειες για την εισαγωγή σημασιολογίας έχουν γίνει αξιοποιώντας την τυχόν μορφοποίηση της πληροφορίας (formatting) στο κείμενο, παρόλα αυτά αυτό είναι κάτι που στη γενική περίπτωση δεν ισχύει αφού τα περισσότερα κείμενα (στο WWW) είναι διαθέσιμα σε ελεύθερη μορφή (plain text). Μοιραία, καθίσταται αναγκαία η σύνδεση των πόρων του κειμένου με πηγές γνώσης, ώστε να δοθεί η δυνατότητα στον υπολογιστή να κατανοήσει μέρος αυτών που κατανοεί ο άνθρωπος κοιτώντας ένα κείμενο.

Τέτοιες περιπτώσεις πληροφορίας (και επομένως γνώσης για τον υπολογιστή) που θα αντιμετωπίσουν τα προβλήματα που αναφέρθηκαν είναι π.χ. οι ακόλουθες:

- πηγές συνωνύμων λέξεων (π.χ. garbage συνώνυμο με rubbish),
- ιεραρχίες λέξεων (π.χ. η λέξη animals είναι υπερσύνολο της λέξης cat)
- λίστες πολυλεκτικών ομάδων (π.χ. Julius Caesar, President of the USA)

Ανάλογα με το περιεχόμενο των κειμένων, μπορούν να χρησιμοποιηθούν οντολογίες, λεξικά (όπως το WordNet) και ταξινομίες είτε με γενικότερο θέμα (π.χ. οικονομία, βιολογία) ή πιο συγκεκριμένο (π.χ. μακροοικονομία, εισόδημα, ιολογία κτλ).

Τα τελευταία χρόνια η εξέλιξη του Παγκόσμιου Ιστού σε ένα καθολικό, αποκεντρωμένο, πλουραλιστικό πληροφοριακό σύστημα, συνεχώς αναπτυσσόμενο με τεράστιο όγκο πληροφορίας (ελεύθερα διαθέσιμης), έχει οδηγήσει σε δομές πληροφορίας όπως η Wikipedia που το κύριο χαρακτηριστικό τους είναι πως παραμένουν πάντα ενημερωμένες on-line (σε αντιδιαστολή με τις οντολογίες ή τα λεξικά τα οποία απαιτείται να ενημερώνονται χειρωνακτικά και off-line).

Η χρήση τέτοιων πηγών γνώσης εισάγει διάφορα πλεονεκτήματα στην ανάλυση κειμένου:

1. Περιορίζεται η ποσότητα των διαφορετικών προτύπων και συνδυασμών των γλωσσικών φαινομένων: Η γνώση που παρέχεται από κάποια εξωτερική πηγή δίνει τη δυνατότητα δημιουργίας περιορισμών, ώστε οι αναπαραστάσεις κειμένου να γίνονται πιο πλούσιες σημασιολογικά αλλά και να αποκτούν καλύτερο νόημα,

2. Αντιμετωπίζονται ζητήματα συνωνυμίας, πολυσημίας, αναφορών και σχέσεων,

3. Η εξωτερική γνώση μπορεί να αξιοποιηθεί σε ζητήματα προεπεξεργασίας κειμένου (πριν την τελική του αναπαράσταση στον υπολογιστή) ώστε να δημιουργηθεί τόσο ένας συνεπής λεξιλογικός χώρος αναφοράς, όσο μία συνεπής ιεραρχία εννοιών.

Φυσικά, η εισαγωγή τέτοιων πηγών γνώσης δεν έρχεται χωρίς διλήμματα και πιθανά μειονεκτήματα. Η εξαντλητική χρήση όλων των δυνατών πηγών γνώσης και η εξονυχιστική τους αναζήτηση ώστε να εντοπιστούν όλες οι δυνατές μορφές αξιοποίησής στους στα έγγραφα αυξάνει υπέρογκα τον υπολογιστικό φόρτο, σε σημείο απαγορευτικό για οποιαδήποτε επεξεργασία. Στο Κεφάλαιο 4 θα γίνει μία αναλυτικότερη παρουσίαση των μοντέλων αναπαράστασης κειμένων (και των απλών αλλά και αυτών με εξωτερική πηγή γνώσης) και θα εισαχθεί ένα μοντέλο αναπαράστασης εγγράφων που αξιοποιεί (χωρίς όμως να επιβαρύνει απαγορευτικά) εξωτερική πηγή.

□

Κεφάλαιο 3

Αξιοποίηση στατιστικής σημασιολογίας για την ποσοτικοποίηση της συσχέτισης λέξεων

"You shall know a word, by the company it keeps"
J. R. Firth (1957)

3.1 Γενικά περί στατιστικής σημασιολογίας και σημασιολογικής συσχέτισης λέξεων

Η ολοένα και μεγαλύτερη διαθεσιμότητα κειμένων (με τη συνακόλουθη πλούσια συλλογή λέξεων, φράσεων κτλ) οδηγεί στο εύλογο ερώτημα: Είναι δυνατόν μέσα από την παρατήρηση των προτύπων που εμφανίζονται οι διάφορες λέξεις να εξαχθεί πληροφορία για τη σημασία τους; Και αν ναι, ποια είναι τα όρια σε αυτή την προσέγγιση;

Η στατιστική σημασιολογία [Furnas et al., 1984] είναι ένας γενικός όρος που χρησιμοποιείται για να απαντήσει σε αυτή την ερώτηση, περιλαμβάνει δηλαδή όλες εκείνες τις μεθόδους που μελετούν το πως μπορούν να εφαρμοστούν τα στατιστικά πρότυπα της χρήσης λέξεων, ώστε να κατανοηθεί η σημασία τους. Πιο αναλυτικά, εξετάζεται η σημασιολογική ομοιότητα μεταξύ μονάδων κειμένου (λέξεις, φράσεις, προτάσεις κτλ) βάσει της στατιστικής ανάλυσης προτύπων συνεμφάνισης των μονάδων αυτών. Η στατιστική σημασιολογία εστιάζει στις σημασίες των λέξεων και στις μεταξύ τους σχέσεις σε αντίθεση με τις παραδοσιακές τεχνικές εξόρυξης κειμένου που ασχολούνται με ολόκληρα έγγραφα, συλλογές εγγράφων, εξαγωγή ονομάτων κτλ.

Η λέξη αποτελεί τη βασικότερη (και πιο φυσική) μονάδα που απαρτίζει τα κείμενα, επομένως είναι λογικό σε πρώτη φάση να εξεταστεί η δυνατότητα εκτίμησης μεθόδων ανάλυσης κειμένου σε αυτό το πρώτο επίπεδο. Όπως έχει αναφερθεί ήδη στα προηγούμενα Κεφάλαια, ο υπολογιστής δεν έχει τη δυνατότητα να γνωρίζει τη σχέση δύο λέξεων με τον ίδιο τρόπο που έχει ανθρώπινος νους, δυνατότητα που θα ήταν ιδιαίτερα χρήσιμη σε όλες τις εφαρμογές που επεξεργάζονται κείμενα.

Είναι φανερό πως ο σημασιολογικός συσχετισμός δύο λέξεων είναι πολύπλοκος και δεν είναι μονοσήμαντος. Υπάρχουν πολλές πλευρές και προβλήματα της σημασιολογίας (όπως περιγράφηκαν στο Κεφάλαιο 2) τα οποία εισάγουν δυσκολίες στην αξιολόγηση

της συσχέτισης λέξεων. Στο Κεφάλαιο αυτό εξετάζεται η δυνατότητα " ευθυγράμμισης " όλων των πλευρών σημασιολογίας για τη δημιουργία ενός (κατά το δυνατόν) αντικειμενικού ποσοτικού βαθμωτού μέτρου εκτίμησης της σημασιολογικής συσχέτισης δύο (ή και περισσότερων λέξεων), ενός μέτρου δηλαδή της απόστασης στο σημασιολογικό χώρο.

Στο σημείο αυτό, πρέπει να γίνει διάκριση μεταξύ του όρου της *σημασιολογικής ομοιότητας* και της *σημασιολογικής συσχέτισης* λέξεων [Budanitsky & Hirst, 2006]. Η σημασιολογική ομοιότητα έχει να κάνει κυρίως με το χαρακτηριστικό της συνωνυμίας ή/και της υπερωνυμίας λέξεων. Για παράδειγμα οι λέξεις car και automobile είναι συνώνυμες ενώ η λέξη vehicle είναι υπερώνυμο της λέξης car. Η σημασιολογική συσχέτιση ποσοτικοποιεί το βαθμό στον οποίο δύο λέξεις σχετίζονται, θεωρώντας όχι μόνο τη συνωνυμία αλλά και κάθε άλλη πιθανή σχέση. Για παράδειγμα υπάρχουν λέξεις που συνδέονται με τη σχέση της μερωνυμίας (η λέξη finger και η λέξη hand), αντωνυμίας (λέξεις με αντίθετη σημασία όπως hot, cold) και γενικά κάθε άλλης σχέσης που δεν είναι απαραίτητο να περιγράφεται όπως οι παραπάνω (όπως οι λέξεις penguin και Antarctica).

Στο υπόλοιπο μέρος του Κεφαλαίου παρουσιάζονται οι υπάρχουσες τεχνικές σημασιολογικής συσχέτισης λέξεων και κατόπιν παρουσιάζεται αναλυτικά μία μεθοδολογία που αναπτύχθηκε στη βάση της στατιστικής σημασιολογίας και υπολογίζει τη συσχέτιση δύο λέξεων. Ο στόχος της προσέγγισης (βάσει και όσων αναφέρθηκαν παραπάνω) είναι διττός:

- η εξέταση της ανάλυσης κειμένου στο επίπεδο της λέξης (πιο συνηθισμένο επίπεδο έως τώρα),
- η εξαγωγή ενός βαθμωτού μέτρου που θα ποσοτικοποιεί με ένα νούμερο τη σημασιολογική συσχέτιση δύο λέξεων και θα ενσωματώνει όλες τις σχέσεις (π.χ. συνώνυμες, μερώνυμες) αξιολογώντας το πόσο σχετικές είναι δύο λέξεις.

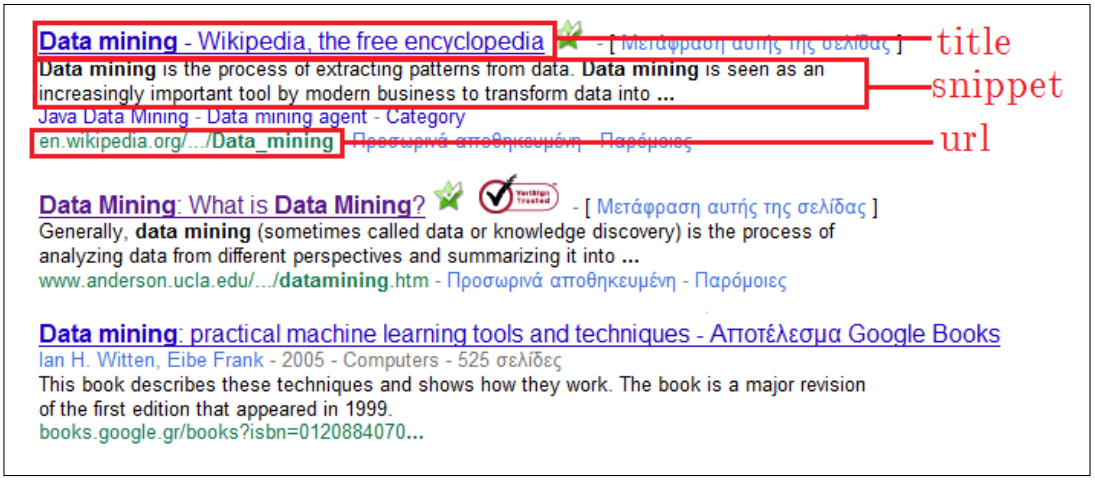
3.2 Υπάρχουσες τεχνικές σημασιολογικής συσχέτισης λέξεων

Τα μέτρα σημασιολογικής συσχέτισης λέξεων χρησιμοποιούνται σε πολλές εφαρμογές όπως αποσαφήνιση έννοιας λέξης [Resnik, 1999], επέκταση ερωτήσεων [Vélez et al., 1997], επισήμανση σελίδων στο WWW [Cimiano et al., 2004], επομένως είναι λογικό κατά καιρούς να έχουν προταθεί πολλά μέτρα.

Δημιουργούμενες με το χέρι βάσεις όπως το WordNet [Fellbaum, 1998] κωδικοποιούν τις σχέσεις μεταξύ των λέξεων. Μέχρι τώρα έχουν προταθεί διάφορα μέτρα που ορίζουν τη συσχέτιση μεταξύ λέξεων, βασιζόμενα στη δομή του WordNet ([Resnik, 1995], [Lin, 1998], [Jarmasz, 2003], [Grefenstette, 1992], [Budanitsky & Hirst, 2006]). Ειδικότερα, στη μέθοδο των [Jiang & Conrath, 1997] αξιοποιείται η συσχέτιση λέξεων βάσει της θέσης τους στην ιεραρχική οργάνωση του WordNet σε συνδυασμό με τις κλασσικές μεθόδους ανάκτησης πληροφορίας από έγγραφα. Όμως είναι κοινά αποδεκτό, πως οι μέθοδοι που βασίζονται σε λεξιλογικές πηγές που συντηρούνται με το χέρι, έχουν διάφορα μειονεκτήματα όπως: χρόνος και προσπάθεια για την διατήρηση των δεδομένων, απουσία διάφορων ονομάτων, νεολογισμών, τεχνικών όρων κ.τ.λ..

Από την άλλη μεριά, οι σελίδες του WWW έχουν εξελιχτεί σε μία πολύ χρήσιμη πηγή πληροφοριών για διάφορα ζητήματα υπολογισμού συσχέτισης. Ένα απόσπασμα

αποτελεσμάτων μηχανών αναζήτησης φαίνεται στο Σχήμα 3.1, από το οποίο φαίνονται οι χρήσιμες πληροφορίες που επιστρέφονται οργανωμένες σε μια δομή που περιέχει τον τίτλο της σελίδες, ένα απόσπασμα από το έγγραφο που περιέχει τη λέξη (snippet) και το URL της σελίδας. Τα δομημένα αποτελέσματα μπορούν να αποτελέσουν μία πηγή γνώσης για τη χρήση της λέξης (ή των λέξεων που αναζητούνται) και μάλιστα δυναμικά προσαρμοζόμενης ανάλογα με την οργάνωση των αποτελεσμάτων από τη μηχανή αναζήτησης.



Σχήμα 3.1: *Παράδειγμα πληροφοριών από αποτελέσματα μηχανής αναζήτησης για τον όρο data mining*

Στην εργασία του [Turney, 2001] προτάθηκε η χρήση του μέτρου Point-Wise Mutual Information σε συνδυασμό με την ανάκτηση πληροφοριών από τα αποτελέσματα αναζήτησης στο WWW για την αναγνώριση συνωνύμων. Στην εργασία των [Sahami & Heilman, 2006] υπολογίστηκε η σημασιολογική ομοιότητα μεταξύ δύο ερωτήσεων χρησιμοποιώντας τα snippets που επιστρέφονται από μία μηχανή αναζήτησης για τις συγκεκριμένες ερωτήσεις. Για κάθε ερώτηση συλλέγονται τα snippets από μία μηχανή αναζήτησης τα οποία σχηματίζουν ένα διάνυσμα βαρών $TF - IDF$ (όπως αναφέρθηκε στην Παράγραφο 2.5.2). Η σημασιολογική ομοιότητα μεταξύ των δύο ερωτήσεων ορίζεται εύκολα ως το εσωτερικό γινόμενο μεταξύ των δύο κεντροειδών των δύο διανυσμάτων.

Στην εργασία των [Chen et al., 2006], προτείνεται ένα μοντέλο "διπλού" ελέγχου του κειμένου που επιστρέφεται από τα snippets για τον υπολογισμό της σημασιολογικής ομοιότητας. Για δύο λέξεις P και Q, συλλέγονται τα snippets για κάθε λέξη που επιστρέφονται από μία μηχανή αναζήτησης. Κατόπιν, μετριούνται οι εμφανίσεις της λέξης P στα snippets της λέξης Q και οι εμφανίσεις της λέξης Q στα snippets της λέξης P (έτσι εξηγείται ο διπλός έλεγχος). Αυτές οι τιμές συνδυάζονται γραμμικά και υπολογίζεται ένα μέτρο ομοιότητας μεταξύ P και Q.

Στην εργασία των [Strube & Ponzetto, 2006], γίνεται χρήση της Wikipedia ως εξής: Δοθέντος ενός ζεύγους λέξεων w_1, w_2, το σύστημα (Wikirelate!) βρίσκει τα άρθρα της Wikipedia που περιέχουν τις δύο αυτές λέξεις στους τίτλους τους. Κατόπιν, η σημασιολογική συσχέτιση υπολογίζεται βάσει του περιεχομένου των άρθρων αυτών ή βάσει της απόστασης υπολογισμένης στην ιεραρχία των κατηγοριών της Wikipedia.

Με χρήση της Wikipedia υπολογίζουν και οι [Gabrilovich & Markovitch, 2007] τη σημασιολογική συσχέτιση λέξεων (ή φράσεων). Η κύρια διαφορά τους από τη μέθοδο του Wikirelate! είναι πως δίνουν τη δυνατότητα συσχέτισης των λέξεων όχι μόνο με

άρθρα της Wikipedia που περιέχουν τις λέξεις αλλά με άρθρα που έχουν σχέση εν γένει με τις λέξεις. Η μέθοδος που ακολουθείται κατασκευάζει ένα σημασιολογικό διερμηνέα που αντιστοιχίζει τις λέξεις (ή φράσεις) της φυσικής γλώσσας σε μία ακολουθία (διάνυσμα) άρθρων της Wikipedia σταθμισμένη βάσει της σχετικότητας με την είσοδο (λέξεις).

Στην εργασία [Iosif & Potamianos, 2007] εισάγεται ένα μη-επιβλεπόμενο μοντέλο, στο οποίο για να υπολογιστεί η ομοιότητα μεταξύ δύο λέξεων P και Q, είναι απαραίτητη η μεταφόρτωση ενός αριθμού από τα έγγραφα που επιστρέφει η μηχανή αναζήτησης στις υψηλότερες θέσεις για την ερώτηση P AND Q και κατόπιν εφαρμόζονται δύο μέτρα "ευρέως περιεχομένου" (wide-context) και "περιορισμένου περιεχομένου" (narrow context).

Στην εργασία των [Bollegala et al., 2007] προτάθηκε μία μέθοδος που εκμεταλλεύεται πλήρως, τα συνολικά αποτελέσματα της αναζήτησης (page counts) και το κείμενο των αποσπασμάτων (snippets) που επιστρέφονται από μια μηχανή αναζήτησης. Ορίζεται ένα μέτρο ομοιότητας ανάμεσα στις λέξεις P και Q χρησιμοποιώντας τον αριθμό των επιστρεφόμενων σελίδων για τις ερωτήσεις για τις λέξεις P, Q και εξάγοντας λεξικο-συντακτικά πρότυπα από τα αποσπάσματα κειμένου που επιστρέφονται από τη μηχανή αναζήτησης.

Συνοπτικά, τα χαρακτηριστικά των διαφόρων εργασιών της βιβλιογραφίας με τις καλύτερες επιδόσεις φαίνονται στον Πίνακα 3.1.

Πίνακας 3.1: *Χαρακτηριστικά των διαφόρων μεθόδων σημασιολογικής ομοιότητας/συσχέτισης λέξεων*

	Αποτελέσματα αναζήτησης	page counts	snippets	Wikipedia	Μεταφόρτωση εγγράφων	Πρότυπα	WordNet	Παραδοσιακές Τεχνικές IR
Παραδοσιακά μετρικά	X	X						
Sahami	X		X					X
CODC	X		X					
SemSim	X	X	X			X	X	
Jiang							X	X
CS(W/WS)					X			X
Wikirelate!				X				X
Gabrilovich				X				X

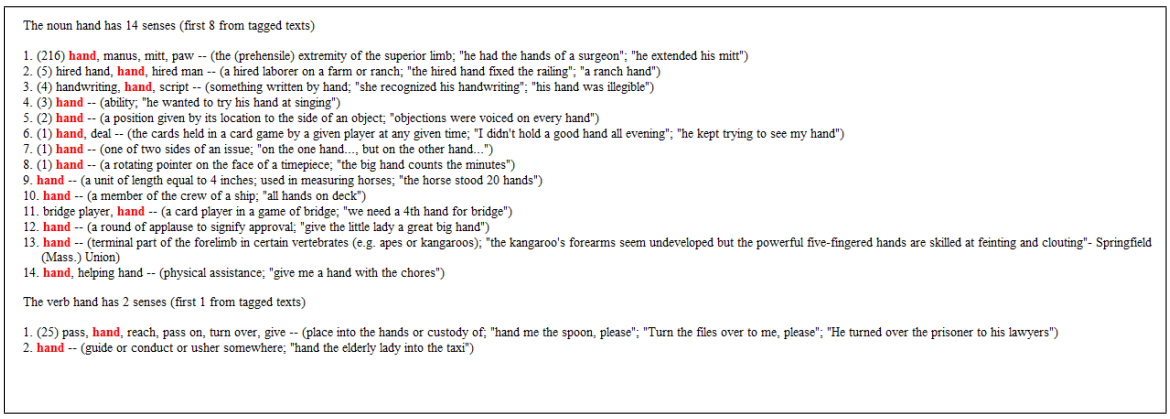
Η μεθοδολογία που αναπτύσσεται στο υπάρχον Κεφάλαιο βασίζεται στο WordNet το οποίο παρουσιάζεται αναλυτικά αμέσως μετά.

3.2.1 Η λεξιλογική βάση WordNet

Το WordNet αποτελεί μία λεξιλογική βάση που οργανώνεται γύρω από λογικές ομάδες που καλούνται synsets (έννοιες). Κάθε synset περιλαμβάνει μία λίστα συνώνυμων λέξεων ή ομάδων λέξεων (collocations) όπως "take in", συνοδευόμενη από το μέρος του λόγου (Part of Speech, POS) που ανήκει και δείκτες που περιγράφουν τις σχέσεις ανάμεσα στα διάφορα synsets. Μία λέξη ή ομάδα λέξεων μπορεί να εμφανίζεται σε περισσότερα του ενός synset και σε περισσότερα του ενός μέρη του λόγου.

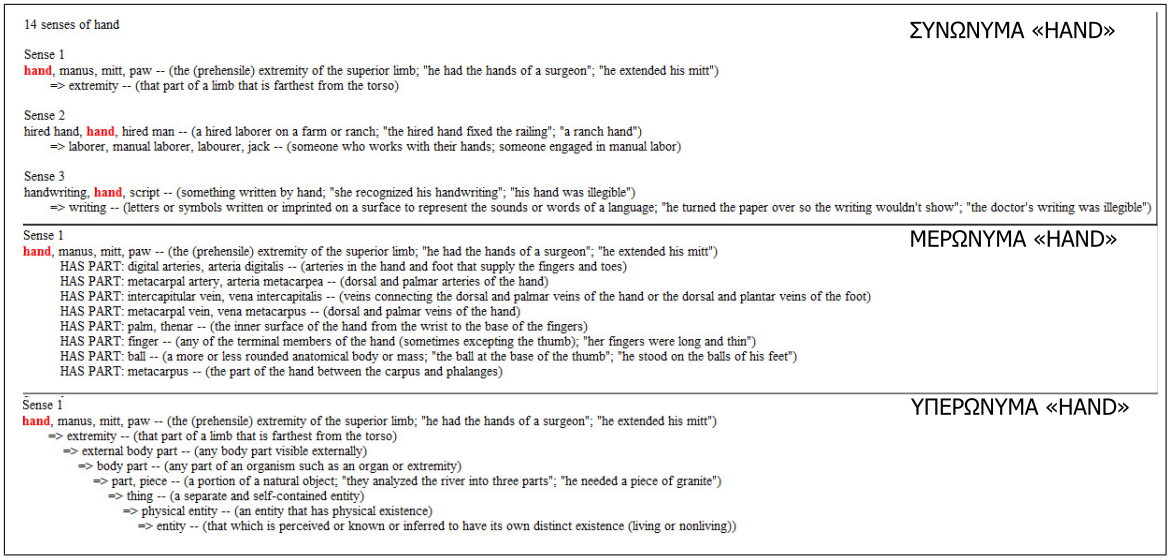
Ένα παράδειγμα για τη λέξη hand φαίνεται στο Σχήμα 3.2.

Οι δείκτες καθορίζουν δύο είδη σχέσεων: λεξιλογικές και σημασιολογικές. Οι λεξιλογικές σχέσεις έχουν να κάνουν με τις λέξεις που μοιάζουν μεταξύ τους μορφολογικά. Οι σημασιολογικές σχέσεις (που έχουν λιγότερο τετριμμένο περιεχόμενο και προσφέρουν δυνατότητες οργάνωσης των λέξεων) περιγράφουν τις σχέσεις μεταξύ



Σχήμα 3.2: *Synsets από το WordNet για τη λέξη hand (ουσιαστικό και ρήμα)*

των λέξεων βάσει της σημασίας τους και περιλαμβάνουν: υπερωνυμίες/υπωνυμίες, αντωνυμίες, συνεπαγωγές και μερωνυμίες. Στο Σχήμα 3.3 φαίνονται μερικά παραδείγματα τέτοιων σχέσεων για τη λέξη hand.



Σχήμα 3.3: *Συνώνυμα, μερώνυμα και υπερώνυμα για κάποιες έννοιες της λέξης hand από το WordNet*

Τα ουσιαστικά και ρήματα οργανώνονται ιεραρχικά βάσει των σχέσεων υπερωνυμίας/υπωνυμίας μεταξύ των synsets ενώ τα επίθετα οργανώνονται σε ομάδες βάσει ενός κεντρικού synset. Θα μπορούσε κανείς να πει πως το WordNet "ξέρει" πως το *μαύρο* είναι το αντίθετο του *λευκού* και πως τα *πουλιά* είναι *ζώα*. Εξαιτίας αυτών των ιδιοτήτων, αρκετές εργασίες χρησιμοποίησαν το WordNet ώστε να εμπλουτίσουν την αναπαράσταση του κειμένου [Breaux & Reed, 2005].

3.3 Μεθοδολογία εξαγωγής σημασιολογικής ομοιότητας λέξεων

Η μέθοδος που αναπτύσσεται με στόχο τη δημιουργία ενός βαθμωτού μέτρου εκτίμησης της σημασιολογικής συσχέτισης λέξεων βασίζεται στην αξιοποίηση των πληροφοριών

που εξάγονται από τα αποτελέσματα των μηχανών αναζήτησης για τις λέξεις αυτές. Αναλυτικά τα βήματα της μεθόδου είναι τα ακόλουθα:

1. Χρήση κειμένου (τίτλος και snippets) για τα αποτελέσματα αναζήτησης κάθε λέξης ώστε να αναπαρασταθεί διανυσματικά με χρήση του μοντέλου BOW και εκτίμηση ενός πρώτου μέτρου συσχέτισης (Rel_{BOW}, Παράγραφος 3.3.1),

2. Κατασκευή ενός δεύτερου μέτρου εκτίμησης της συσχέτισης (Rel_{SVM}) που θα χρησιμοποιεί διανύσματα για κάθε ζεύγος λέξεων, των οποίων οι συνιστώσες θα αποτελούνται από βάρη που έχουν να κάνουν με μέτρα συνεμφάνισης των λέξεων και με τα λεξικο-συντακτικά πρότυπα που εξάγεται από τα αποτελέσματα αναζήτησης. Πιο αναλυτικά, η διαδικασία για την κατασκευή των διανυσμάτων αυτών περιλαμβάνει τα εξής:

   (α′) Υπολογισμός μέτρων συνεμφάνισης και εξαγωγή των λεξικο-συντακτικών προτύπων που εμφανίζονται στον τίτλο, στο snippet και το url από τα αποτελέσματα αναζήτησης για το ζεύγος λέξεων που εξετάζεται (Παράγραφος 3.3.2),

   (β′) Επιλογή των σημαντικότερων από αυτά τα χαρακτηριστικά ώστε να μορφώσουν το τελικό διάνυσμα χαρακτηριστικών (Παράγραφος 3.3.3),

   (γ′) Εκπαίδευση μιας μηχανής διανυσμάτων υποστήριξης (rSVM, regression support vector machine) βάσει ενός συνόλου λέξεων (των οποίων είναι γνωστή η σχέση) και το αποτέλεσμα της οποίας θα χρησιμοποιηθεί ως ένα δεύτερο μέτρο για τον υπολογισμό συσχέτισης των λέξεων (Παράγραφος 3.3.4)

Ένας γραμμικός συνδυασμός των δύο μέτρων (του Rel_{BOW} και του Rel_{SVM}) επιτρέπει την ποσοτικοποίηση της συνολικής σημασιολογικής συσχέτισης μεταξύ δύο λέξεων (Rel_{total}):

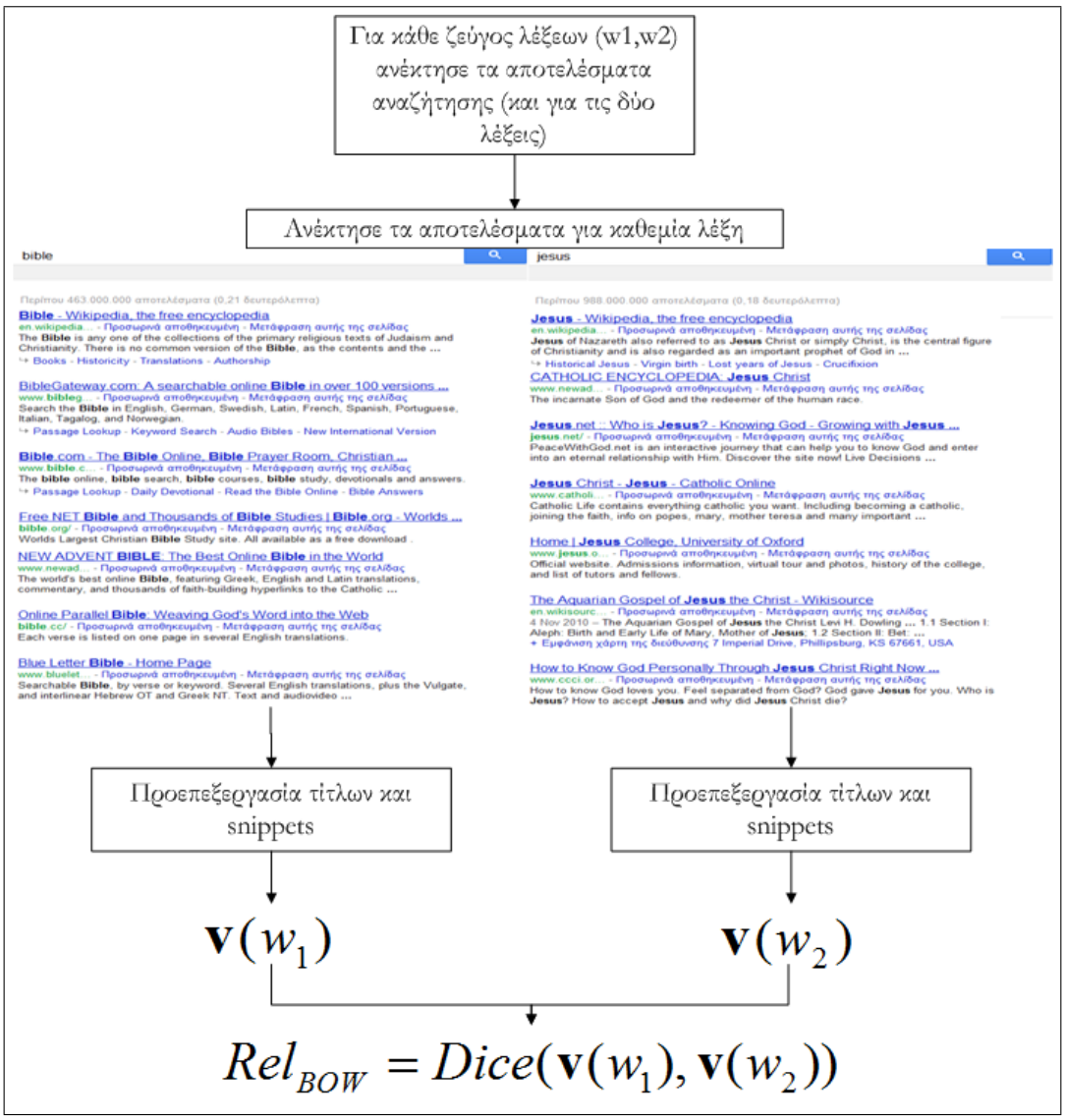
$$Rel_{total} = \lambda Rel_{SVM} + (1 - \lambda) Rel_{BOW} \tag{3.1}$$

όπου το $\lambda \in (0, 1)$ είναι μια παράμετρος βάρους κάθε μεθόδου στο τελικό υβριδικό μέτρο.

3.3.1 Περιγραφή του μέτρου Rel_{BOW}

Η παραδοσιακή μέθοδος αναπαράστασης εγγράφων (Bag of Words, BOW) περιγράφηκε στο Κεφάλαιο 2.5.2 (εξισώσεις 2.1-2.3) και υιοθετείται και εδώ. Στην προσέγγισή θεωρείται ένα σύνολο δεδομένων (S) αποτελούμενο από λέξεις (w). Έχει ήδη αναφερθεί πως τα αποτελέσματα αναζήτησης στο WWW οργανώνονται σε 3 τμήματα: τον τίτλο της σελίδας, το απόσπασμα κειμένου (snippet) από τη σελίδα (που συνήθως περιέχει τη λέξη που αναζητείται) και το URL της σελίδας. Τα τρία αυτά μέρη περιέχουν πολύτιμες πληροφορίες για το νοηματικό περιεχόμενο μιας λέξης και η υπόθεση που γίνεται (και επιβεβαιώνεται από τα πειράματα που γίνονται) είναι πως κάθε τμήμα έχει διαφορετική συνεισφορά και γιαυτό το λόγο θα εξεταστούν χωριστά. Στόχος είναι η αναπαράσταση κάθε λέξης w με ένα έγγραφο d, το οποίο δημιουργείται από τα αποτελέσματα αναζήτησης στο WWW για την εν λόγω λέξη w. Εποπτικά, η διαδικασία φαίνεται στο Σχήμα 3.4.

Πιο συγκεκριμένα, ανακτώνται τα 1000 πρώτα αποτελέσματα για τη λέξη w και κατόπιν ενώνονται όλοι οι τίτλοι και τα αποσπάσματα (snippets) σε ένα έγγραφο (d).



Σχήμα 3.4: *Εποπτική διαδικασία υπολογισμού του μέτρου* Rel_{BOW}

Έτσι, κάθε λέξη στο σύνολο δεδομένων αναπαρίσταται από ένα τέτοιο έγγραφο, στο οποίο εφαρμόζονται οι συνήθεις τεχνικές κειμενικής προεπεξεργασίας (απαλοιφή λέξεων χωρίς νόημα (stop-words), περιστολή (stemming). Επομένως, προκύπτει ένα σύνολο εγγράφων όπου κάθε έγγραφο d αντιστοιχεί σε μία λέξη w. Μετά από αυτή τη διαδικασία, αν επιλεγούν όλοι οι διαφορετικοί όροι (t_i) που εμφανίζονται σε όλα τα έγγραφα d, δημιουργείται το λεξικό ($dict$) του συνόλου δεδομένων. Παραλείπονται οι όροι των οποίων η ποσότητα $\sum_{j=0}^{N}(tf - idf)_{i,j}$ είναι μικρότερη από ένα κατώφλι (κάποια πολύ μικρή τιμή). Χρησιμοποιείται η εξίσωση 2.3, όπου το i αναφέρεται στους όρους t_i, το j αναφέρεται στα έγγραφα d και N είναι ο συνολικός αριθμός των εγγράφων (δηλαδή των λέξεων που υπάρχουν στο σύνολο δεδομένων). Εναλλακτικά, αν δεν τεθεί κατώφλι μπορούν να συμπεριληφθούν όλοι οι όροι t_i.

Για κάθε έγγραφο d (επομένως για κάθε λέξη w) δημιουργείται ένα διάνυσμα $\mathbf{v}(w)$, στο οποίο κάθε συνιστώσα i τίθεται στην τιμή $tf - idf$ (όπως προκύπτει από την εξίσωση 2.3) του όρου t_i για το συγκεκριμένο έγγραφο. Επομένως, προκύπτουν $|S|$ διανύσματα (όπου $|S|$ είναι ο αριθμός των λέξεων του συνόλου δεδομένων). Η σχέση μεταξύ δύο λέξεων w_1, w_2 υπολογίζεται από τον συντελεστή Dice μεταξύ των αντιστοί-

χων διανυσμάτων, χρησιμοποιώντας την παρακάτω εξίσωση :

$$Rel_{BOW}(w_1, w_2) = \frac{2 \cdot \mathbf{v}(w_1) \cdot \mathbf{v}(w_2)}{||\mathbf{v}(w_1)||^2 + ||\mathbf{v}(w_2)||^2} \tag{3.2}$$

όπου : $\mathbf{v}(w_1) \cdot \mathbf{v}(w_2)$ είναι το εσωτερικό γινόμενο των δύο διανυσμάτων και $||\mathbf{v}||$ είναι η ευκλείδια νόρμα του διανύσματος $\mathbf{v}$. Αναλυτικά ο αλγόριθμος φαίνεται στο Σχήμα 3.5.

Υπολογισμός Rel_{BOW}

1: $ListOfTerms_{tot} \leftarrow NULL$
2: **for all** λέξεις w από το S **do**
3: - ένωσε τίτλους και snippets από τα 1000 πρώτα αποτελέσματα του $query(w)$ σε ένα έγγραφο d
4: - απάλοιψε λέξεις χωρίς νόημα, εφάρμοσε περιστολή και κανονικοποίηση
5: - πάρε τη λίστα με τους όρους $ListOfTerms_w$ που αντιπροσωπεύουν το έγγραφο d (δηλαδή τη λέξη w) και υπολόγισε το tf με χρήση της εξίσωσης 2.1
6: **end for**
7:
8: **for all** διαφορεικούς όρους t_i που εμφανίζονται στη λίστα $ListOfTerms_w$ **do**
9: - υπολόγισε idf με χρήση της εξίσωσης 2.2
10: - υπολόγισε το $tf - idf_i$ με χρήση της εξίσωσης 2.3
11: **if** $\sum_{j=0}^{N} tf - idf_{i,j} \geq threshold$ **then**
12: - βάλε το t_i στη λίστα $ListOfTerms_{tot}$
13: **end if**
14: **end for**
15:
16: - Η λίστα $ListOfTerms_{tot}$ χρησιμοποιείται για το σχηματισμό του διανύσματος v
17: **for all** words w from dataset S **do**
18: - Σχημάτισε και γέμισε με τιμές το (w) που αντιπροσωπεύει τη λέξη w
19: **end for**
20:
21: **for** λέξεις w_1 και w_2 **do**
22: - Υπολόγισε το μέτρο Rel_{BOW} βάσει της εξίσωσης 3.2
23: **end for**

Σχήμα 3.5: *Περιγραφή του μέτρου* Rel_{BOW}

Η μέθοδος που περιγράφηκε έχει 3 μειονεκτήματα : (1) σπάει τις εκφράσεις που αποτελούνται από δύο ή περισσότερες λέξεις σε εντελώς ανεξάρτητους όρους, (2) θεωρεί τις λέξεις με πολλές σημασίες σαν μία και (3) αντιστοιχίζει συνώνυμες λέξεις σε διαφορετικούς όρους, οπότε όταν εφαρμόζεται μόνη της έχει μικρή αποτελεσματικότητα. Επομένως, είναι απαραίτητη η ενσωμάτωση σημασιολογικών πληροφοριών και εννοιολογικών προτύπων για την ενίσχυση της δυνατότητας πρόβλεψης της σχετικότητας λέξεων.

3.3.2 Εξαγωγή χαρακτηριστικών από τα αποτελέσματα αναζήτησης στο WWW

3.3.2.1 Μέτρα βασισμένα σε μετρητές σελίδων (page counts)

Η βασική ιδέα που υποστηρίζει αυτά τα μέτρα είναι το γεγονός πως η συνεμφάνιση λέξεων στα αποτελέσματα αναζήτησης είναι ένδειξη σημασιολογικής σχετικότητάς τους. Αυτή η συνεμφάνιση εκφράζεται από τις σελίδες που επιστρέφονται από την αναζήτηση για P *AND* Q (έστω πως P και Q είναι ένα ζεύγος λέξεων και το *AND* είναι λογικό). Παρόλα αυτά, είναι λογικό να μη ληφθεί υπόψιν μόνο το αποτέλεσμα της αναζήτησης P *AND* Q, αλλά και ο αριθμός των εγγράφων που περιέχουν μόνο τη μία λέξη (P ή Q) ξεχωριστά.

Θεωρούνται τέσσερα δημοφιλή μέτρα συνεμφάνισης που εκφράζουν τη σημασιολογική σχετικότητα με χρήση των μετρητών σελίδων. Ο συμβολισμός που χρησιμοποιείται για τους μετρητές σελίδων(page counts) από μία μηχανή αναζήτησης είναι ο ακόλουθος: Το $H(P)$ συμβολίζει τον αριθμό σελίδων για την αναζήτηση P, $H(Q)$ συμβολίζει τον αριθμό σελίδων για την αναζήτηση Q, και $H(P \bigcap Q)$ συμβολίζει τον αριθμό σελίδων για την αναζήτηση P AND Q.

Ο συντελεστής Jaccard ορίζεται (για δύο λέξεις P και Q) ως εξής :

$$Jaccard(P,Q) = \frac{H(P \bigcap Q)}{H(P) + H(Q) - H(P \bigcap Q)} \tag{3.3}$$

Ο συντελστής Overlap ορίζεται ως εξής :

$$Overlap(P,Q) = \frac{H(P \bigcap Q)}{\min(H(P), H(Q))} \tag{3.4}$$

Ο συντελεστής Dice ορίζεται ως εξής :

$$Dice(P,Q) = \frac{2 \cdot H(P \bigcap Q)}{H(P) + H(Q)} \tag{3.5}$$

Τέλος, ο συντελεστής PMI, Point-Wise Mutual Information) ορίζεται ως εξής :

$$PMI(P,Q) = \log \frac{\frac{H(P \bigcap Q)}{N}}{\frac{H(P)}{N} \cdot \frac{H(Q)}{N}} \tag{3.6}$$

όπου το N αναφέρεται στον αριθμό των σελίδων που εξετάζονται από τη μηχανή αναζήτησης. Ο αριθμός αυτός μπορεί να θεωρηθεί $N = 10^{10}$, σύμφωνα και με τις σελίδες που επιστρέφονται από τη μηχανή αναζήτησης Google (γενικά κάποιος πολύ μεγάλος αριθμός θεωρείται κατάλληλος).

Μερικά παραδείγματα υπολογισμού των παραπάνω μέτρων για κάποια τυχαία ζεύγη λέξεων φαίνονται στον Πίνακα 3.2. Ήδη από τα πολύ απλά μέτρα αυτά φαίνεται πως λέξεις ανόμοιες όπως τα ζεύγη fruit/car, monk/implement παρουσιάζουν μικρές τιμές, αντίθετα με λέξεις πιο όμοιες (coast/shore, cat/dog). Και πάλι τονίζεται πως ενδιαφέρει η σχετική ομοιότητα (δηλαδή η σχετικότητα) των λέξεων (π.χ. οι λέξεις cat-dog θεωρούνται εξίσου συναφείς με τις συνώνυμες coast-shore).

Πίνακας 3.2: *Παράδειγμα υπολογισμού των μέτρων συνεμφάνισης στα αποτελέσματα αναζήτησης*

	coast/shore	cat/dog	journey/voyage	bird/bridge	fruit/car	monk/implement
$H(P)$	962.000.000	3.050.000.000	680.000.000	712.000.000	1.150.000.000	108.000.000
$H(Q)$	382.000.000	2.590.000.000	380.000.000	135.000.000	4.770.000.000	245.000.000
$H(P \cup Q)$	22.3000.000	1.910.000.000	206.000.000	96.000.000	654.000.000	6.370.000
$Jaccard(P,Q)$	0.1989	0.5121	0.2412	0.1278	0.1242	0.0184
$Overlap(P,Q)$	0.5838	0.7375	0.5421	0.7111	0.5687	0.0590
$Dice(P,Q)$	0.3318	0.6773	0.3887	0.2267	0.2209	0.0361
$PMI(P,Q)$	0.7831	0.3834	0.9016	0.9995	0.0764	0.3815

3.3.2.2 Εξαγωγή λεξικο-συντακτικών προτύπων από τον τίτλο, το snippet και το url

Έστω το ακόλουθο παράδειγμα αποτελέσματος αναζήτησης για την ερώτηση *"cats AND dogs"* (Το AND είναι λογικό):

Title: Friends of Cats & Dogs - Home
Snippet: The Friends of Cats and Dogs Foundation (Los Amigos de Gatos y Perros Foundacion) seeks to help stray animals, pets, and their owners through various ...
URL: www.fcdf.org/cats-and-dogs.html

Φαίνεται πως και οι δύο λέξεις (cats, dogs) εμφανίζονται και στον τίτλο, και στο snippet και στο url με τα ακόλουθα πρότυπα :

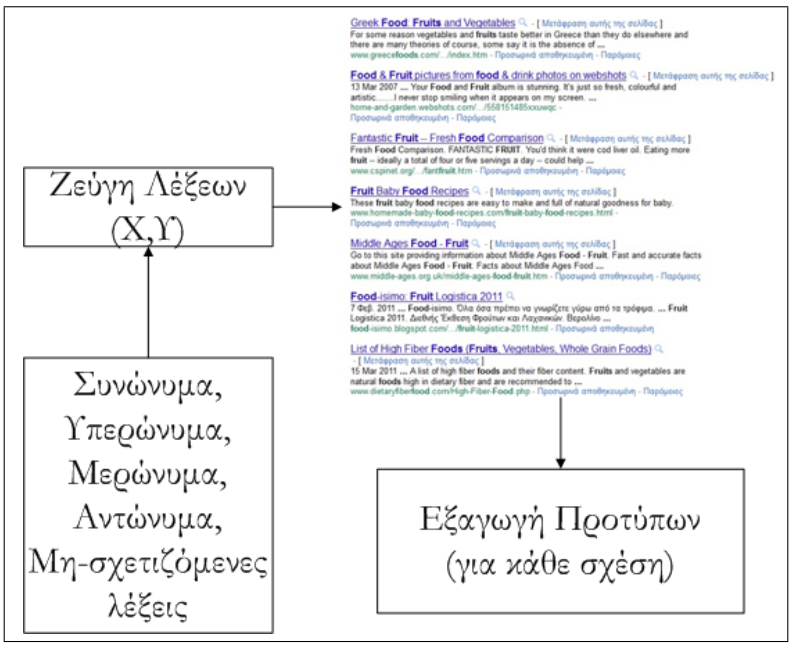
- *"Cats & Dogs"* στον τίτλο,
- *"Cats and Dogs"* στο snippet,
- *"cats-and-dogs"* στο url.

Εάν οι λέξεις αναζήτησης αντικατασταθούν από τους χαρακτήρες X και Y αντίστοιχα, τότε εξάγεται ένα πρότυπο "X & Y" για τον τίτλο, ένα πρότυπο "X and Y" για το snippet και ένα πρότυπο "X-and-Y" για το URL.

Έτσι, θεωρώντας όλες τις σχέσεις σημασιολογικής ομοιότητας που υπάρχουν στο WordNet (συνωνυμία, υπερωνυμία, μερωνυμία, αντωνυμία), υπάρχει η δυνατότητα να εξαχθούν πάρα πολλά τέτοια πρότυπα για ζεύγη λέξεων ανάλογα με τη μεταξύ τους σχέση, με μια διαδικασία που περιγράφεται στο Σχήμα 3.6.

Όσον αφορά στην παρούσα μέθοδο, υλοποιήθηκε ο αλγόριθμος που φαίνεται στο Σχήμα 3.7 για την εξαγωγή λεξικο-συντακτικών προτύπων, επεκτείνοντας των αλγόριθμο του [Bollegala et al., 2007], λαμβάνοντας υπόψιν όχι μόνο τα snippets αλλά και τους τίτλους και τα URLs.

Δεδομένου ενός συνόλου T από ζεύγη λέξεων, ανακτώνται τα αποτελέσματα του Google για αυτό το ζεύγος λέξεων και γίνεται χωριστή επεξεργασία στον τίτλο, στο snippet και στο url. Για κάθε ένα από αυτά τα τμήματα και για όλα τα αποτελέσματα διατηρούνται μόνο εκείνα που περιέχουν και τις δύο λέξεις. Εξετάζεται το περιεχόμενο μεταξύ των δύο αυτών λέξεων και παραλείπονται όσες περιπτώσεις έχουν παραπάνω από 4 λέξεις ή/και λέξεις διαφορετικές από μία ευρεία λίστα κοινών λέξεων (Διαδικασία *GetPatterns* στο Σχήμα 3.7). (Οι υπόλοιπες λέξεις είναι πολύ συγκεκριμένες για το ζευγάρι των λέξεων και παύουν να είναι γενικά πρότυπα). Υπολογίζεται η συχνότητα κάθε προτύπου και κατασκευάζεται μία λίστα των πιο κοινών προτύπων για τον τίτλο, το snippet και το url των αποτελεσμάτων.



Σχήμα 3.6: *Διαδικασία εξαγωγής προτύπων βάσει των ήδη γνωστών σχέσεων μεταξύ λέξεων*

Είναι προφανές, πως για τον εντοπισμό των πιο σημαντικών προτύπων, τα οποία και καθορίζουν τις σχέσεις μεταξύ των λέξεων, πρέπει να χρησιμοποιηθεί ένα μεγάλο σύνολο δεδομένων εκπαίδευσης. Για την εξαγωγή προτύπων για όλες τις σχέσεις λέξεων που υποστηρίζονται από το WordNet, επιλέγονται 5 διαφορετικές κλάσεις, που κάθε μία περιέχει 1800 ζεύγη συνωνύμων, 1800 ζεύγη υπερωνύμων, 1800 ζεύγη μερωνύμων, 1800 ζεύγη αντωνύμων και 1800 ζεύγη μη-σχετιζόμενων λέξεων, όλες προερχόμενες από το WordNet. Με την επεξεργασία των αποτελεσμάτων για όλα τα ζεύγη λέξεων, προκύπτουν πολλά διαφορετικά πρότυπα που περιγράφουν τις 4 σχέσεις που αναφέρθηκαν παραπάνω αλλά και πρότυπα που εμφανίζονται σε μη-σχετιζόμενες λέξεις.

3.3.3 Επιλογή χαρακτηριστικών για την εξαγωγή των σημαντικότερων προτύπων

Προφανώς, δεν είναι εφικτό να εκπαιδευτεί οποιοδήποτε μοντέλο (στην περίπτωσή μας ένα rSVM) με έναν πάρα πολύ μεγάλο αριθμό προτύπων, οπότε καθίσταται αναγκαία η χρήση κάποιας τεχνικής επιλογής χαρακτηριστικών ώστε να επιλεγούν τα πρότυπα που έχουν τη μεγαλύτερη διακριτική ικανότητα μεταξύ των κλάσεων.

Χρησιμοποιείται η ιδέα της επιλογής χαρακτηριστικών μεταξύ πολλών κλάσεων των [Forman, 2004] για να βρεθούν τα πιο σημαντικά πρότυπα για κάθε μία από τις 4 σημασιολογικές κατηγορίες. Ο αλγόριθμος φαίνεται αναλυτικά στο Σχήμα 3.8.

Ο στόχος είναι η κατάταξη των προτύπων βάσει της ικανότητάς τους να διακρίνουν κάθε αρχική κλάση (συνώνυμα, υπερώνυμα, μερώνυμα, αντώνυμα), που από δω και στο εξής θα καλείται θετική κλάση, απέναντι σε μία άλλη (που θα καλείται αρνητική κλάση), αποτελούμενη από όλες τις υπόλοιπες κλάσεις μαζί (δηλαδή αν τα συνώνυμα είναι η θετική κλάση, τότε η αρνητική θα περιέχει έναν ανάλογο αριθμό από υπερώνυμα, αντώνυμα και μερώνυμα). Για την επίτευξη του στόχου αυτού ορίζεται ένα κριτήριο M. Για κάθε πρότυπο που έχει εξαχθεί, κατασκευάζεται ένας πίνακας συνάφειας όπως

Εξαγωγή προτύπων από τα αποτελέσματα αναζήτησης

1: **for all** ζεύγη λέξεων (A, B) από όλες τις κλάσεις (συνώνυμα, υπερώνυμα, μερώνυμα, αντώνυμα, μη-σχετιζόμενες) λέξεων **do**
2: $T \leftarrow$ GetAllTitles($query(A, B)$)
3: $S \leftarrow$ GetAllSnippets($query(A, B)$)
4: $U \leftarrow$ GetAllURLS($query(A, B)$)
5: $NT \leftarrow 0$
6: $NS \leftarrow 0$
7: $NU \leftarrow 0$
8: **end for**
9:
10: **for all** τίτλους t από το σύνολο T **do**
11: $NT \leftarrow NT +$ GetPatterns(t, A, B)
12: $TitlePatterns \leftarrow$ CountFrequency NT
13: **end for**
14: **return** $TitlePatterns$
15:
16: **for all** snippets s από το σύνολο S **do**
17: $NS \leftarrow NS +$ GetPatterns(s, A, B)
18: $SnippetPatterns \leftarrow$ CountFrequency NS
19: **end for**
20: **return** $SnippetPatterns$
21:
22: **for all** urls u από το σύνολο U **do**
23: $NU \leftarrow NU +$ GetPatterns(u, A, B)
24: $URLPatterns \leftarrow$ CountFrequency NU
25: **end for**
26: **return** $URLPatterns$

Σχήμα 3.7: *Αλγόριθμος εξαγωγής προτύπων από τα αποτελέσματα αναζήτησης*

φαίνεται στον Πίνακα 3.3. Σε αυτόν τον πίνακα, το A δηλώνει τη συχνότητα του προτύπου στη θετική κλάση, το B δηλώνει τη συχνότητα του προτύπου στην αρνητική κλάση, το U δηλώνει τη συνολική συχνότητα όλων των προτύπων για τη θετική κλάση και το V δηλώνει τη συνολική συχνότητα όλων των προτύπων για την αρνητική κλάση.

Πίνακας 3.3: *Επιλογή χαρακτηριστικών rSVM: Πίνακας συνάφειας για το πρότυπο x*

	Πρότυπο x	Πρότυπο εκτός x
Συχνότητα για την εκάστοτε θετική κλάση	A	$C = U - A$
Συχνότητα για την εκάστοτε αρνητική κλάση	B	$D = V - B$

Για τη μέθοδο που αναπτύσσεται ορίζονται τα μέτρα Mutual Information (MI) και Information Gain (IG) [Manning & Schútze, 1999] ως ακολούθως :

$$MI = \log \frac{\frac{A}{N}}{\frac{(A+B)}{N} \cdot \frac{(A+C)}{N}} \tag{3.7}$$

Επιλογή προτύπων για την εξεύρεση των πιο σημαντικών

1: **for all** κλάσεις c συνωνύμων, υπερωνύμων, αντωνύμων, μερωνύμων **do**
2: - κατάταξε όλα τα γνωρίσματα σύμφωνα με το κριτήριο M για το υποζήτημα του διαχωρισμού της κλάσης c σε σχέση με όλες τις άλλες κλάσεις μαζί
3: - αποθήκευσε την κατάταξη χαρακτηριστικών για την κλάση c
4: **end for**
5:
6: **while** λίστα χαρακτηριστικών (*list*) δεν έχει συμπληρωθεί **do**
7: - κάλεσε το δρομολογητή Round-Robin για την επιλογή της επόμενης κλάσης c
8: - επέλεξε το επόμενο γνώρισμα f από την M-κατάταξη για την κλάση c
9: - συγχώνευσε το f με τη λίστα *list* αν δεν υπάρχει ήδη
10: **end while**

Σχήμα 3.8: *Αλγόριθμος επιλογής των σημαντικότερων λεξικο-συντακτικών προτύπων*

$$IG = e(A + C, B + D) - [\frac{A}{N} \cdot e(A, B) + \frac{C}{N} \cdot e(C, D)] \tag{3.8}$$

όπου

$$e(x, y) = -\frac{x}{x+y} \cdot \log \frac{x}{x+y} - \frac{y}{x+y} \cdot \log \frac{y}{x+y} \tag{3.9}$$

και βάσει αυτών ορίζεται ένα καινούριο μέτρο σημαντικότητας των προτύπων (pattern importance, PI):

$$PI = \begin{cases} IG, \text{ εάν} & MI > 0 \\ 0, \text{ εάν} & MI \leq 0 \end{cases} \tag{3.10}$$

Το μέτρο αυτό όπως ορίζεται παραλείπει πρότυπα που έχουν τιμή για το MI αρνητική, επειδή αρνητική τιμή για το MI υπονοεί (μέσω της εξίσωσης 3.7) λογαριθμικό όρισμα με τιμή μικρότερη του 1, δηλαδή ότι ο παρανομαστής είναι μεγαλύτερος του αριθμητική. Από σειρά πειραμάτων, φαίνεται πως αυτό συμβαίνει διότι είτε το B ή/και το C είναι ένας μεγάλος αριθμός, ένδειξη ότι το πρότυπο είναι (αρκετά) πιο συχνό στις υπόλοιπες κλάσεις παρά σε αυτήν που εξετάζουμε.

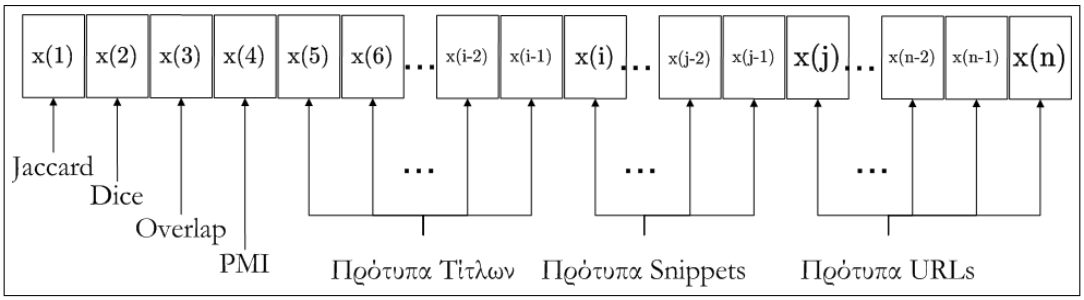
Με χρήση της εξίσωσης 3.10, δημιουργούνται κατατάξεις χαρακτηριστικών για κάθε κλάση και έπειτα μέσα από μία διαδικασία Round Robin επιλέγονται τα πιο σημαντικά (βάσει της διακριτικής τους ικανότητας) γνωρίσματα για κάθε κλάση σε σχέση με τις άλλες. Μπορεί επίσης να χρησιμοποιηθεί ένα κατώφλι στην επιλογή των χαρακτηριστικών, ανάλογο με το πόσο μεγάλο διάνυσμα χαρακτηριστικών είναι επιθυμητό. Στον Πίνακα 3.4 παρουσιάζονται τα 5 πιο διακριτικά πρότυπα για τον τίτλο, το snippet και το url. Επισημαίνεται πως τα πρότυπα αυτά είναι διαφορετικά για κάθε τμήμα των αποτελεσμάτων, το οποίο υπονοεί πως υπάρχει επιπλέον πληροφορία στη δομή των αποτελεσμάτων αναζήτησης αλλά και πως τα πρότυπα αυτά δεν εξαρτώνται από τη σχέση από την οποία έχουν προέλθει (συνωνυμία κτλ) αλλά καθορίζουν (γενικά) τη συσχέτιση των λέξεων.

Πίνακας 3.4: *5 σημαντικότερα πρότυπα από άποψη ικανότητας καθορισμού της σημασιολογικής συσχέτισης λέξεων (ανεξαρτήτως είδους σχέσης)*

Τίτλοι	**Snippets**	**URLs**
X or Y	X, Y	X-or-Y
X in Y	X for Y	X+Y
X (Y	X in Y	X/Y
X: Y	X & Y	XandY
X vs. Y	X, or Y	X-Y

3.3.4 Δημιουργία του διανύσματος χαρακτηριστικών και εκπαίδευση rSVM

Το διάνυσμα χαρακτηριστικών για κάθε ζεύγος λέξεων (P και Q) κατασκευάζεται σύμφωνα με την ακόλουθη διαδικασία: Συλλέγονται τα αποτελέσματα από το Google για την αναζήτηση *"P AND Q"* (το AND είναι λογικό) και διαχωρίζονται οι τίτλοι, τα snippets και τα url. Σύμφωνα με τη διαδικασία που περιγράφηκε στην Παράγραφο 3.3.2.1 υπολογίζονται τα μέτρα ομοιότητας βασισμένα στα page counts και ενσωματώνονται με τα χαρακτηριστικά που δημιουργούνται από τη διαδικασία του κεφαλαίου 3.3.2.2. Με τη διαδικασία αυτή δημιουργείται ένα διάνυσμα χαρακτηριστικών διάστασης όσων είναι τα γνωρίσματα που έχουμε επιλέξει (ανάλογα με το κατώφλι που τίθεται στη διαδικασία Round-Robin) συν τέσσερα που είναι τα μέτρα των page counts. Το διάνυσμα αυτό φαίνεται εποπτικά στο Σχήμα 3.9.



Σχήμα 3.9: *Περιγραφή διανύσματος για τον υπολογισμό του μέτρου* Rel_{SVM}

Με τη χρήση ενός συνόλου εκπαίδευσης που περιέχει ζεύγη λέξεων των οποίων ξέρουμε το μέτρο της συσχέτισης εκπαιδεύεται ένα rSVM. Μετά την εκπαίδευση έχουμε τη δυνατότητα υπολογισμού της συσχέτισης μεταξύ δύο οποιονδήποτε λέξεων. Στις μηχανές διανυσμάτων υποστήριξης για παλινδρόμηση (regression) (rSVM) [Collobert et al., 2001], ο σκοπός είναι η προσέγγιση της συναρτησιακής εξάρτησης μιας μεταβλητής y από ένα σύνολο ανεξάρτητων μεταβλητών x (που αποτελούν το χώρο εισόδου, συμβολιζόμενου και ως X).

Στα προβλήματα παλινδρόμησης, η αρχική υπόθεση είναι πως η σχέση μεταξύ των ανεξάρτητων και των εξαρτημένων μεταβλητών δίνεται από μία ντετερμινστική συνάρτηση συν κάποιο θόρυβο ϵ:

$$y = f(\mathbf{x}) + \epsilon \tag{3.11}$$

Το ζήτημα είναι να βρεθεί μία συναρτησιακή μορφή για την f που να μπορεί σωστά να προβλέπει νέες περιπτώσεις που δεν έχουν δοθεί για εκπαίδευση στο SVM. Αυτό

γίνεται μέσα από την εκπαίδευση του μοντέλου σε ένα σύνολο δεδομένων, μια διαδικασία που περιλαμβάνει την σταδιακή βελτιστοποίηση μιας συνάρτησης σφάλματος :

$$error = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i + C\sum_{i=1}^{N}\xi_i' \tag{3.12}$$

όπου C είναι η παράμετρος χωρητικότητας που καθορίζει την εξισορρόπηση μεταξύ σφάλματος εκπαίδευσης και του περιθωρίου, $\mathbf{w}$ είναι το διάνυσμα των χαρακτηριστικών ($\mathbf{w} \in X$), i είναι ένας δείκτης που δείχνει τα N παραδείγματα εκπαίδευσης ξ_i, ξ_i' είναι παράμετροι για το χειρισμό μη-διαχωρίσιμων δεδομένων.

Η ελαχιστοποίηση του σφάλματος γίνεται σύμφωνα με τις ακόλουθες εξισώσεις :

$$\mathbf{w}^T \cdot \phi(x_i) + b - y_i \leq \epsilon + \xi_i' \tag{3.13}$$

$$y_i - \mathbf{w}^T \cdot \phi(x_i) - b \leq \epsilon + \xi_i \tag{3.14}$$

$$\xi_i', \xi_i \geq 0 \tag{3.15}$$

όπου b είναι μία σταθερά ($b \in \mathbb{R}$) και $\phi(x_i)$ είναι μια συνάρτηση που χρησιμοποιείται για την περίπτωση που τα δεδομένα στο σύνολο εκπαίδευσης είναι μη γραμμικά (όπως εδώ) και μέσω αυτής της συνάρτησης γίνεται προσπάθεια για γραμμικό διαχωρισμό.

Οι εξισώσεις 3.3.4 έως 3.3.4 που περιγράφουν το πρόβλημα της βελτιστοποίησης επιλύονται με πολλαπλασιαστές Lagrange (λ_i) [Baraff, 1996] και η επιφάνεια απόφασης είναι:

$$L = \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \phi(x_i)\cdot\phi(x_j) \tag{3.16}$$

με τον περιορισμό της εξίσωσης 3.3.4 και τον επιπλέον περιορισμό:

$$\sum_{i=1}^{N}\lambda_i y_i = 0 \tag{3.17}$$

Το εσωτερικό γινόμενο $\phi(x_i)\cdot\phi(x_j)$ ονομάζεται συνάρτηση πυρήνα (kernel function) K και μπορεί να είναι γραμμική, πολυωνυμική, ακτινική συνάρτηση βάσης (Radial Basis Function, (RBF)), σιγμοειδής κτλ.

Η μορφή του πυρήνα RBF (που επιλέγεται στα επόμενα) είναι η ακόλουθη (για δεδομένα x_i, x_j):

$$K(x_i, x_j) = e^{-\gamma\cdot d(x_i, x_j)} \tag{3.18}$$

όπου γ είναι μια παράμετρος (προσδιορίζεται πειραματικά) και d είναι η ευκλείδεια απόσταση των διανυσμάτων.

Το μοντέλο εκπαίδευσης βελτιστοποιείται σε σχέση με τις παραμέτρους C και γ. Ο RBF πυρήνας έχει τα καλύτερα αποτελέσματα από τις άλλες περιπτώσεις (γραμμικός, πολυωνυμικός) και γιαυτό επιλέγεται για τα πειράματα.

3.4 Έλεγχος μεθοδολογίας εξαγωγής σημασιολογικής συσχέτισης λέξεων

3.4.1 Σύνολα δεδομένων (datasets)

Η μεθοδολογία που περιγράφηκε στην Παράγραφο 3.3 ελέγχθηκε σε σύνολα δεδομένων (datasets) της βιβλιογραφίας που ασχολούνται με το ίδιο ζήτημα. Πιο συγκεκριμένα, για τη διαδικασία της εκπαίδευσης του SVM (μέτρο Rel_{SVM}, Παράγραφος 3.3.4, χρησιμοποιήθηκε το σύνολο δεδομένων (ζεύγη λέξεων) Similarity-353 [Finkelstein et al., 2002] από το οποίο εξαιρέθηκαν τα 28 ζεύγη του συνόλου δεδομένων Miller-Charles [Miller & Charles, 1991] τα οποία χρησιμοποιήθηκαν για το μέτρο Rel_{BOW} (Παράγραφος 3.3.1) αλλά και για τον έλεγχο της μεθόδου. Το σύνολο Miller-Charles αποτελείται από 28 ζεύγη λέξεων των οποίων η ομοιότητα έχει αξιολογηθεί από 38 ανθρώπους. Τα ζεύγη έχουν τιμές ομοιότητας που κυμαίνονται από 0 (πλήρως ανόμοιες λέξεις) έως 4 (απόλυτη συνωνυμία). Και τα δύο σύνολα δεδομένων παρουσιάζονται στο Παράρτημα Β′.

Στο σημείο αυτό θα πρέπει να τονιστεί η ανυπαρξία ενός μεγάλου και αξιόπιστου συνόλου δεδομένων με σημασιολογικά σχετικές (και άσχετες) λέξεις. Το WordNet παρέχει πολλά ζεύγη λέξεων και χρησιμοποιώντας ένα μέτρο (όπως αυτά που περιγράφονται στην Παράγραφο 3.2), είναι δυνατό να παραχθούν πολλά ζεύγη λέξεων μαζί με την τιμή ομοιότητάς τους. Γενικά όμως, η αξιολόγηση της συσχέτισης δύο λέξεων ακόμα και από ανθρώπους είναι σχετική και βασίζεται σε πολλές παραμέτρους, επομένως γενικά είναι δύσκολο να υπάρχει ένα σύνολο δεδομένων επαρκές.

3.4.2 Επιλογή παραμέτρων

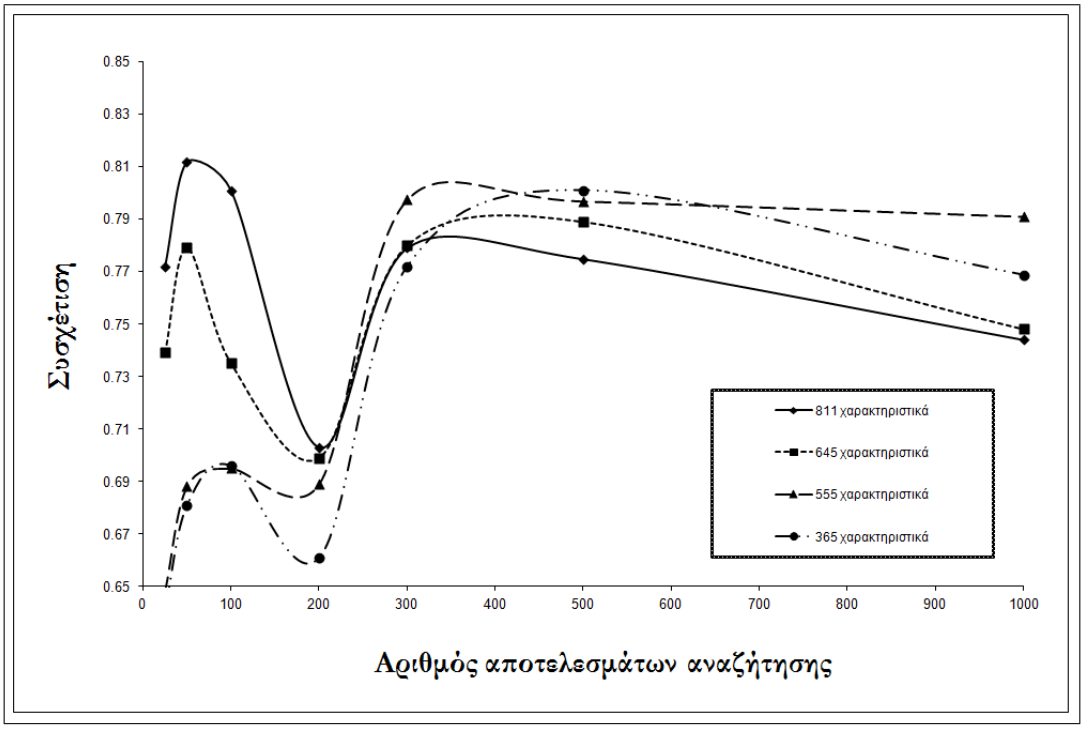
Για το μέτρο Rel_{BOW}, ανακτήθηκαν τα 1000 πρώτα αποτελέσματα αναζήτησης στο WWW για τις διαφορετικές λέξεις του συνόλου Miller-Charles και έπειτα με τη διαδικασία που περιγράφεται στο Σχήμα 3.5, υπολογίζεται η ομοιότητα μεταξύ των ζευγών λέξεων του συνόλου. Κάθε λέξη αναπαρίσταται από ένα έγγραφο που αποτελείται από τους τίτλους και τα snippets των 1000 πρώτων αποτελεσμάτων που επιστρέφονται από τη μηχανή αναζήτησης Google. Προφανώς, το λεξικό στη συγκεκριμένη περίπτωση (κάθε λέξη του λεξικού είναι ένας όρος στο διάνυσμα αναπαράστασης των εγγράφων) περιέχει όλες τις λέξεις από όλα τα έγγραφα. Το μειονέκτημα του μέτρου αυτού, είναι πως αν θέλει κανείς να υπολογίσει τη σημασιολογική συσχέτιση κάποιας λέξης που δεν υπάρχει στο λεξικό (με κάποια άλλη), θα πρέπει να επανασχηματιστούν τα διανύσματα χαρακτηριστικών ώστε να αντανακλούν τις αλλαγές που εισάγονται από την καινούρια λέξη (και άρα το καινούριο έγγραφο).

Για το μέτρο Rel_{SVM}, εκπαιδεύεται ένα rSVM [Joachims, 1999] με πυρήνα RBF. Στην πειραματική διαδικασία δοκιμάστηκαν και οι υπόλοιποι συνήθεις πυρήνες (γραμμικός, πολυωνυμικός κτλ) αλλά οι RBF είναι οι πιο αποτελεσματικοί και γιαυτό επιλέχθηκαν. Οι παράμετροι που χρησιμοποιήθηκαν για το SVM (και θεωρούνται βέλτιστες) φαίνονται στον Πίνακα 3.5.

Η συμπεριφορά του μέτρου Rel_{SVM} σε σχέση με τον αριθμό των προτύπων που χρησιμοποιήθηκαν σαν γνωρίσματα και σε σχέση με τον αριθμό των αποτελεσμάτων αναζήτησης που χρησιμοποιήθηκαν φαίνεται στο Σχήμα 3.10 ως προς την συσχέτιση των αποτελεσμάτων στο σύνολο Miller-Charles. Παρατηρείται πως χρήση ενός μικρού αριθμού αποτελεσμάτων και ενός μικρού αριθμού χαρακτηριστικών δίνει μικρές τιμές

Πίνακας 3.5: *Βέλτιστες παράμετροι για το SVM που χρησιμοποιείται στο μέτρο* Rel_{SVM}

Παράμετρος	**Τιμή**
Πυρήνας	RBF
γ	0.1
C	20



Σχήμα 3.10: *Επιρροή του αριθμού αποτελεσμάτων αναζήτησης και χαρακτηριστικών στο μέτρο* Rel_{SVM}

συσχέτισης. Καθώς αυξάνεται ο αριθμός των χαρακτηριστικών και (ή) ο αριθμός των αποτελεσμάτων, επιτυγχάνεται αύξηση της συσχέτισης. Επιπλέον, φαίνεται πως όσο περισσότερα αποτελέσματα και γνωρίσματα χρησιμοποιούνται, η συσχέτιση αρχίζει και φθίνει, πιθανώς επειδή το διάνυσμα χαρακτηριστικών γίνεται αρκετά αραιό. Καλύτερη συσχέτιση επιτυγχάνεται όταν χρησιμοποιείται ένας σχετικά μικρός αριθμός αποτελεσμάτων και ένας σχετικά μεγάλος αριθμός χαρακτηριστικών. Αυτό εξηγείται κυρίως από το γεγονός, πως μόνο τα πρώτα αποτελέσματα αναζήτησης του Google περιέχουν σημαντικές και ποιοτικές πληροφορίες για το ζεύγος των λέξεων, αλλά και πάλι χρειάζεται προσοχή καθώς ενδεχόμενη χρήση πολύ μικρού αριθμού αποτελεσμάτων οδηγεί σε μείωση της τιμής της συσχέτισης ξανά. Τα πειράματα έδειξαν πως η χρήση 50 αποτελεσμάτων αναζήτησης με ένα πλούσιο διάνυσμα (διάστασης 811) επιλέγοντας ένα χαμηλό κατώφλι στη διαδικασία που περιγράφεται στο Σχήμα 3.8 για την επιλογή των προτύπων.

Έπειτα από τον υπολογισμό των δύο ανεξάρτητων μέτρων, έγινε πειραματισμός στο σύνολο εκπαίδευσης σχετικά με την παράμετρο λ της εξίσωσης με στόχο τη στάθμιση των δύο μέτρων. Τα αποτελέσματα που παρουσιάζονται εδώ, ανακτήθηκαν με $\lambda = 0.7$ και κανονικοποιούνται στο διάστημα $[0, 1]$ για τις ανάγκες της σύγκρισης με προηγούμενες μεθόδους.

3.4.3 Αποτελέσματα

Η μέθοδος που αναπτύχθηκε συγκρίθηκε ως προς την επίδοσή της για τα 28 ζεύγη του συνόλου Miller-Charles σε σχέση με τις κυριότερες μεθόδους της βιβλιογραφίας. Τα αποτελέσματά παρουσιάζονται στον Πίνακα 3.6. Η μέθοδος [Jiang & Conrath, 1997] χρησιμοποιεί τη σημασιολογική ιεραρχική δομή του WordNet και επιτυγχάνει την καλύτερη συσχέτιση ανάμεσα στις " ανταγωνιστικές " μεθόδους που βασίζονται σε οντολογίες. Στις μεθόδους που χρησιμοποιούν αποτελέσματα αναζήτησης, τα μέτρα που βασίζονται στα page counts έχουν τη χαμηλότερη συσχέτιση με το PMI να είναι το πιο ακριβές. Οι μέθοδοι των [Sahami & Heilman, 2006] και [Iosif & Potamianos, 2007] (μη-επιβλεπόμενες) επιτυγχάνουν μία μέτρια τιμή συσχέτισης ενώ η μέθοδος των [Bollegala et al., 2007] έχει τη δεύτερη καλύτερη τιμή χρησιμοποιώντας όμως μόνο συνώνυμες λέξεις.

Πίνακας 3.6: *Σύγκριση μεθόδων σημασιολογικής συσχέτισης λέξεων στο σύνολο Miller-Charles*

Μέτρα	miller-charles	jaccard	dice	overlap	PMI	Sahami	CODC	SemSim	Jiang	Binary CS	Προτεινόμενη
cord-smile	0.13	0.06	0.07	0.08	0.03	0.09	0	0	0.35	0.4	0.23
rooster-voyage	0.08	0.02	0.02	0.03	0.04	0.20	0	0.02	0.08	0	0.29
noon-string	0.08	0.13	0.14	0.11	0.30	0.08	0	0.02	0.18	0.16	0.32
glass-magician	0.11	0.08	0.08	0.27	0.37	0.14	0	0.18	0.68	0.18	0
monk-slave	0.55	0.22	0.23	0.12	0.71	0.10	0	0.38	0.39	0.19	0.32
coast-forest	0.42	0.92	0.92	0.32	0.55	0.25	0	0.41	0.29	0.76	0.45
monk-oracle	1.1	0.07	0.08	0.07	0.32	0.05	0	0.33	0.34	0.47	0.31
lad-wizard	0.42	0.07	0.08	0.05	0.33	0.15	0	0.22	0.32	0.37	0.31
forest-graveyard	0.84	0.10	0.11	0.35	0.59	0	0	0.55	0.19	0.11	0.19
food-rooster	0.89	0.04	0.05	0.39	0.28	0.08	0	0.06	0.4	0.35	0.45
coast-hill	0.87	1	1	0.40	0.53	0.29	0	0.87	0.71	0.18	0.42
car-journey	1.16	0.37	0.39	0.43	0.22	0.19	0.29	0.29	0.33	0.52	0.44
crane-implement	1.68	0.22	0.24	0.13	0.61	0.15	0	0.13	0.59	0.1	0.25
brother-lad	1.66	0.16	0.17	0.33	0.52	0.24	0.38	0.34	0.28	0.58	0.41
bird-crane	2.97	0.24	0.26	0.27	0.57	0.22	0	0.88	0.73	0.59	0.60
bird-cock	3.05	0.33	0.35	0.23	0.53	0.06	0.50	0.59	0.73	0.44	0.53
food-fruit	3.08	0.91	0.91	1	0.57	0.18	0.34	1	0.63	0.79	0.79
brother-monk	2.82	0.18	0.19	0.37	0.55	0.27	0.55	0.38	0.91	0.63	0.47
asylum-madhouse	3.61	0.07	0.08	0.13	1	0.21	0	0.77	0.97	0.51	0.76
furnace-stove	3.11	0.33	0.35	0.14	0.96	0.31	0.93	0.89	0.39	1	0.91
magician-wizard	3.5	0.25	0.27	0.23	0.78	0.23	0.67	1	1	0.59	0.94
journey-voyage	3.84	0.50	0.52	0.21	0.52	0.52	0.42	1	0.88	0.75	0.89
coast-shore	3.7	0.88	0.89	0.53	0.71	0.38	0.52	0.95	0.99	0.5	0.61
implement-tool	2.95	0.50	0.52	0.55	0.53	0.42	0.42	0.68	0.97	0.8	0.72
boy-lad	3.76	0.18	0.19	0.52	0.54	0.47	0	0.97	0.88	0.67	0.75
automobile-car	3.92	0.55	0.57	0.69	0.37	1	0.69	0.98	1	0.76	1
midday-noon	3.42	0.18	0.20	0.25	0.94	0.29	0.86	0.82	1	0.74	0.92
gem-jewel	3.84	0.37	0.39	0.15	0.68	0.21	1	0.69	1	0.53	0.95
Correlation	**1**	**0.27**	**0.28**	**0.38**	**0.62**	**0.58**	**0.69**	**0.83**	**0.83**	**0.71**	**0.88**

Ας σημειωθεί πως μέθοδοι που αξιοποιούν τα snippets από τα αποτελέσματα αναζήτησης (Sahami, CODC, Bollegala) παρουσιάζουν εμφανώς καλύτερα αποτελέσματα δείχνοντας αφενός τη χρησιμότητά τους και αφετέρου γιατί η προτεινόμενη μέθοδος που αξιοποιεί πλήρως τόσο τα αποτελέσματα αναζήτησης όσο και τις σημασιολογικές σχέσεις του WordNet επιτυγχάνει την υψηλότερη συσχέτιση (έως τώρα). Επίσης, το σύνολο Miller-Charles, (MC) περιέχει αρκετές αμφίσημες λέξεις (π.χ. father, oracle, crane) το οποίο δημιουργεί πρόβλημα στα μετρικά που βασίζονται στα page counts και δε θεωρούν το περιβάλλον που χρησιμοποιείται η λέξη.

Στον Πίνακα 3.7 φαίνεται η συνεισφορά κάθε μέτρου (Rel_{BOW}, Rel_{SVM}) αλλά και των επιμέρους (page counts, λεξικο-συντακτικά πρότυπα) στο τελικό μέτρο (Rel_{total}). Για την επιρροή των page counts εκπαιδεύτηκε το rSVM με διανύσματα 4 χαρακτηριστικών (όσα και τα μέτρα που βασίζονται στα page counts) ενώ αντίστοιχα η ίδια διαδικασία έγινε για τα λεξικο-συντακτικά πρότυπα. Από τα νούμερα του Πίνακα γίνεται σαφές πως

η μεγαλύτερη επιρροή στο τελικό αποτέλεσμα είναι από τα λεξικο-συντακτικά πρότυπα (προφανώς λόγω των πολύ κοινών λέξεων του συνόλου MC υπάρχουν πολλές συνυπάρξεις τους σε τίτλους, snippets και urls), παρόλα αυτά η βελτίωση που επιτυγχάνεται, αφενός με τη χρήση των page counts, αφετέρου με τη χρήση του Rel_{BOW} δείχνει πως το κάθε μέτρο και η κάθε μέθοδος έχει τη δική της συνεισφορά. Τα page counts και το Rel_{BOW} είναι σίγουρα απαραίτητα σε περιπτώσεις ζευγών λέξεων που δεν εμφανίζονται συχνά σε πολλά πρότυπα, ενώ ο συνδυασμός τους με τα λεξικο-συντακτικά πρότυπα βελτιώνει αισθητά το τελικό αποτέλεσμα.

Πίνακας 3.7: *Επιρροή των διαφόρων συνιστωσών στο συνολικό μέτρο σημασιολογικής συσχέτισης* Rel_{total}

Μέθοδος	**Συσχέτιση στο σύνολο MC**
Rel_{BOW}	0.64
Rel_{SVM}	0.81
Μόνο page counts	0.57
Μόνο λεξικο-συντακτικά πρότυπα	0.77
Rel_{total}	0.88

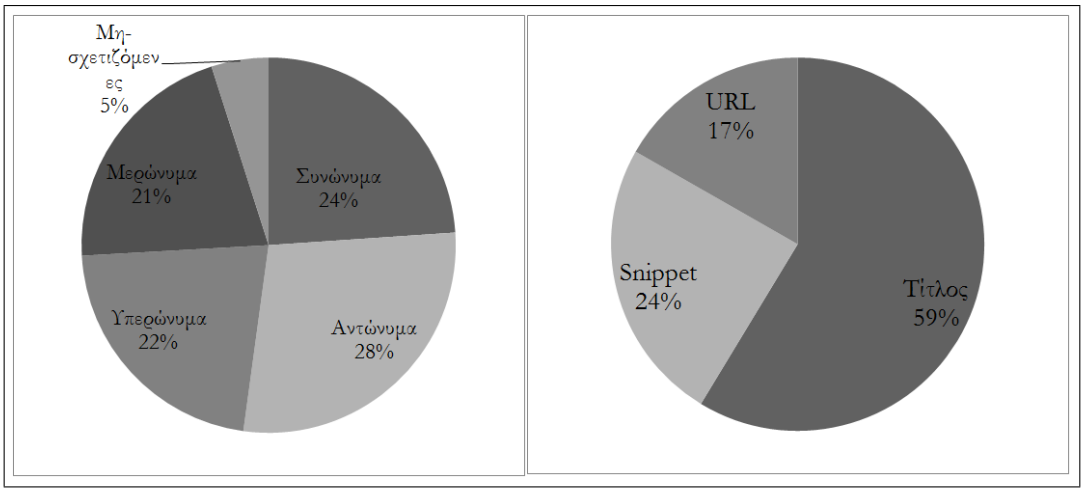
Τέλος, στο Σχήμα 3.11 φαίνεται η κατανομή των προτύπων που απαρτίζουν τις συνιστώσες του διανύσματος του rSVM όσον αφορά:

- ποια είναι η σημασιολογική σχέση (συνωνυμία, αντωνυμία, υπερωνυμία, μερωνυμία) στην οποία συναντάται πιο συχνά το συγκεκριμένο πρότυπο και
- ποιο είναι το μέρος των αποτελεσμάτων αναζήτησης (τίτλος, snippet, url) στο οποίο συναντάται πιο συχνά το συγκεκριμένο πρότυπο (κανονικοποιημένο σε σχέση με το μέγεθος των ανάλογων κειμένων σε λέξεις και χαρακτήρες).

Από το Σχήμα αυτό γίνονται σαφή τα εξής:

- Επιτυγχάνεται η " ευθυγράμμιση" όλων των δυνατών σχέσεων του WordNet, ώστε συνεισφέρουν περίπου ισότιμα στον υπολογισμό του μέτρου συσχέτισης των λέξεων. Επαναλαμβάνεται πως η μεθοδολογία εξετάζει γενικά τη σχέση δύο λέξεων και όχι μόνο αν είναι συνώνυμες,
- Αναλογικά, τα περισσότερα πρότυπα εξάγονται από τους τίτλους των αποτελεσμάτων, κάτι που εξηγείται λογικά εφόσον οι τίτλοι γενικά είναι κατατοπιστικοί, περιγραφικοί και σαφείς για το περιεχόμενο των εγγράφων τους. Η πληροφορία αυτή επιβεβαιώνει και το γεγονός πως υπάρχει ξεχωριστή βαρύτητα ανάλογα με το μέρος των αποτελεσμάτων αναζήτησης που εξετάζεται, κάτι που επιβεβαιώνει την αρχική υπόθεση πως υπάρχει πληροφορία στη δομή των αποτελεσμάτων των μηχανών αναζήτησης.

□



Σχήμα 3.11: *Προέλευση προτύπων στο διάνυσμα του rSVM*

Κεφάλαιο 4

Αναπαράσταση κειμένου

Η πετυχημένη αναπαράσταση εγγράφων αποτελεί βασική προϋπόθεση για την οποιαδήποτε διαδικασία αποδοτικής ανάλυσης κειμενικών δεδομένων. Αυτό συμβαίνει εξαιτίας της ανικανότητας των περισσοτέρων αλγορίθμων ανάλυσης κειμένου να ερμηνεύουν σωστά τα έγγραφα στην πρωτότυπη μορφή τους, οπότε καθίσταται αναγκαία μια διαδικασία δεικτοδότησης (indexing) από ένα κείμενο d_j σε μία πιο συμπιεσμένη αναπαράσταση του περιεχομένου του. Το πρόβλημα της επιλογής του κατάλληλου μοντέλου αναπαράστασης για ένα έγγραφο δεν έχει μοναδική λύση, αφού εξαρτάται από τι θεωρεί ο καθένας σημαντικό σε κάθε κείμενο (το λεγόμενο σημασιολογικό και εννοιολογικό πρόβλημα).

Μία πρώτη εισαγωγή στα μοντέλα αναπαράστασης εγγράφων έγινε στην Παράγραφο 2.5, όπου και παρουσιάστηκε το βασικό μοντέλο αναπαράστασης εγγράφων (VSM). Παρακάτω θα γίνει μία κύρια επισκόπηση των μεθόδων αναπαράστασης κειμένου και θα προταθεί το μοντέλο που αναπτύχθηκε με χρήση εξωτερικής γνώσης στα πλαίσια της διατριβής.

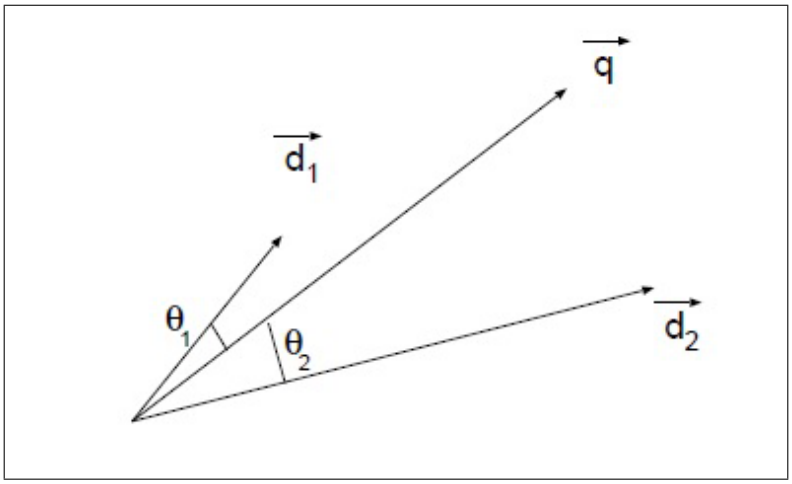
4.1 Επισκόπηση μοντέλου χώρου διανυσμάτων (Vector-Space-Model)

Όπως αναφέρθηκε ήδη στο Κεφάλαιο 2.5.2, το κλασσικό μοντέλο αναπαράστασης εγγράφων είναι το μοντέλο του χώρου διανυσμάτων (Vector Space Model, VSM ή "Bag-of-Words", BOW), το οποίο αναπαριστά κάθε έγγραφο μιας συλλογής (από έστω συνολικά M το πλήθος έγγραφα) με ένα διάνυσμα διάστασης όσες οι διαφορετικές λέξεις της συλλογής (έστω N), οδηγώντας σε έναν πίνακα διάστασης $N \times M$ όπως φαίνεται παρακάτω:

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & ... & w_{1,M} \\ w_{2,1} & w_{2,2} & ... & w_{2,M} \\ ... & ... & ... & ... \\ w_{N-1,1} & w_{N-1,2} & ... & w_{N-1,M} \\ w_{N,1} & w_{N,2} & ... & w_{N,M} \end{bmatrix} \tag{4.1}$$

Κατά καιρούς έχουν προταθεί διάφορες προσεγγίσεις ως προς τον τρόπο υπολογισμού των tf και idf αλλά η γενική ιδέα παραμένει ίδια (όπως διατυπώθηκε στις εξισώσεις 2.1 έως 2.3). Το μοντέλο VSM αποτελεί το σημαντικότερο εργαλείο στα συστήματα ανάκτησης πληροφορίας (Information Retrieval) εδώ και δεκαετίες [Berry

et al., 1999] λόγω του βασικού του πλεονεκτήματος που είναι η υποστήριξη της σχετικής κατάταξης εγγράφων με ετερογενείς μορφές σε σχέση με ερωτήσεις (queries). Οι ερωτήσεις αναπαρίστανται από διανύσματα της ίδιας διάστασης με τα έγγραφα. Η κατάταξη συσχέτισης των εγγράφων σε σχέση με τις ερωτήσεις γίνεται βάσει μιας συνάρτησης ομοιότητας σε σχέση με το διάνυσμα της ερώτησης. Όσον αφορά στη συνάρτηση ομοιότητας, πολλά συστήματα χρησιμοποιούν το συνημίτονο της γωνίας των διανυσμάτων ως εύκολα υλοποιήσιμο και με θετικότερα αποτελέσματα σε σχέση με την Ευκλείδεια Απόσταση: Για παράδειγμα στο Σχήμα 4.1 το έγγραφο d_1 είναι "πιο κοντά" στην ερώτηση q αν χρησιμοποιηθεί η απόσταση συνημιτόνου ενώ το έγγραφο d_2 είναι "πιο κοντά" αν χρησιμοποιηθεί η ευκλείδεια απόσταση. Διάφορα άλλα μέτρα απόστασης έχουν προταθεί για διαφορετικές εφαρμογές και βάσεις δεδομένων. Αντίστοιχα, η ίδια συνάρτηση ομοιότητας μπορεί να χρησιμοποιηθεί και για την απευθείας σύγκριση εγγράφων μεταξύ τους.



Σχήμα 4.1: *Συσχέτιση δύο διανυσμάτων εγγράφων με το διάνυσμα μιας ερώτησης*

4.1.1 Προεπεξεργασία κειμένων

Για καλύτερη απόδοση του μοντέλου VSM λαμβάνουν χώρα δύο τεχνικές προεπεξεργασίας των κειμένων πριν τη μόρφωση του διανύσματος αναπαράστασης: η αφαίρεση των κοινών λέξεων (stop words) και η περιστολή (stemming) των καταλήξεων. Η αφαίρεση των κοινών λέξεων παραλείπει άρθρα, προθέσεις (π.χ. and, the), λέξεις δηλαδή οι οποίες εμφανίζονται πολλές φορές και σε όλα (σχεδόν) τα έγγραφα, άρα η διακριτική τους ικανότητα είναι πολύ μικρή και έτσι μειώνεται επιπλέον και ο υπολογιστικός φόρτος. Η περιστολή των καταλήξεων αποσκοπεί στο να αφήσει τις λέξεις μόνο με τη ρίζα τού λήμματός τους, ώστε να αντιμετωπίζονται ως όμοιες λέξεις που διαφέρουν γραμματικά μόνο (π.χ. learn, learning, learner).

Η προεπεξεργασία αυτή μειώνει κάπως τον υπολογιστικό φόρτο και βελτιώνει τα αποτελέσματα, παρόλα αυτά δεν καταφέρνει να ξεπεράσει τα δύο βασικά προβλήματα του μοντέλου που είναι η μεγάλη διάσταση των διανυσμάτων και η άγνοια για τη σημασιολογική/συντακτική σχέση των λέξεων.

4.1.2 Τεχνικές μείωσης της διάστασης του διανύσματος αναπαράστασης

Πέραν από τη χρήση διαφορετικών μονάδων ως συνιστώσες για την κατασκευή του διανύσματος αναπαράστασης, έχουν αναπτυχθεί και τεχνικές απευθείας μείωσης της διάστασης του διανύσματος (με χρήση είτε των λέξεων, είτε των όρων, είτε άλλης μονάδας). Στις τεχνικές μείωσης διάστασης που αναπτύσσονται παρακάτω γίνεται αναφορά σε λέξεις αλλά εννοείται πως η διαδικασία είναι η ίδια στην περίπτωση άλλων μονάδων. Στόχος της διαδικασίας είναι η κατάταξη όλων των λέξεων των εγγράφων που υπάρχουν διαθέσιμα, βάσει κάποιου κριτηρίου, ώστε να μπορέσει να γίνει επιλογή των σημαντικότερων.

4.1.2.1 Μέθοδος μείωσης διάστασης βάσει της συχνότητας στα έγγραφα

Ως συχνότητα μιας λέξης i στα έγγραφα (Document Frequency) ορίζεται ο όρος df_i όπως υπάρχει στην Εξίσωση 2.2, δηλαδή ως ο αριθμός των εγγράφων (από τα διαθέσιμα της συλλογής) στα οποία εμφανίζεται ο όρος i.

Η βασική ιδέα πίσω από αυτό το μέτρο είναι πως πολύ σπάνιες λέξεις (άρα με μικρή συχνότητα) είτε δεν περιέχουν σημαντική πληροφορία είτε δεν επηρεάζουν τη συνολική επίδοση. Συνήθως η διαδικασία αυτή εφαρμόζεται μετά την αφαίρεση των κοινών λέξεων και την περιστολή των καταλήξεων.

Με τη χρήση της συχνότητας στα έγγραφα (DF) σαν κριτήριο επιλογής των συνιστωσών (άρα των λέξεων), μόνο οι όροι με υψηλές συχνότητες χρησιμοποιούνται. Παρά την απλότητα της μεθόδου, θεωρείται πως μπορεί να έχει ικανοποιητικά αποτελέσματα [Yang & Pedersen, 1997]. Τέλος, είναι αυτονόητο πως αν υπάρχουν m έγγραφα και n λέξεις η πολυπλοκότητα της μεθόδου είναι $\mathcal{O}(mn)$.

4.1.2.2 Λανθάνουσα Σημασιολογική Δεικτοδότηση (Latent Semantic Indexing, LSI)

Βάσει του μέσου τετραγωνικού σφάλματος, η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis, PCA) είναι η καλύτερη γραμμική μέθοδος δευτέρας τάξεως για τη μείωση της διάστασης δεδομένων και βασίζεται στον πίνακα συνδιακύμανσης των μεταβλητών [Fodor, 2002]. Κάποιες φορές στο πεδίο της επεξεργασίας κειμένου είναι γνωστή ως Διάσπαση Ιδιαζουσών Τιμών (Singular Value Decomposition (SVD).

Η SVD δέχεται σαν όρισμα έναν πίνακα $\mathbf{X}$ και τον αναπαριστά με έναν πίνακα $\hat{\mathbf{X}}$ με μικρότερη διάσταση και με τέτοιο τρόπο ώστε η απόσταση (νόρμα-2) ανάμεσα στους δύο πίνακες (στους δύο χώρους) να είναι ελάχιστη :

$$\mathbf{\Delta} = \left\| \mathbf{X} - \hat{\mathbf{X}} \right\| \tag{4.2}$$

Η νόρμα-2 για τους πίνακες είναι το ισοδύναμο της Ευκλείδειας απόστασης για τα διανύσματα.

Πιο αναλυτικά, ο στόχος της PCA (μοιραία και της SVD) είναι η μείωση της διάστασης των δεδομένων με την εύρεση λίγων ορθογώνιων διαστάσεων που είναι γνωστές ως κύριες συνιστώσες και είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών με τη μεγαλύτερη διακύμανση. Η πρώτη κύρια συνιστώσα είναι ο γραμμικός συνδυασμός με

τη μεγαλύτερη διακύμανση, η δεύτερη κύρια συνιστώσα αυτή με τη δεύτερη μεγαλύτερη διακύμανση που είναι και ορθογώνια προς την πρώτη κ.ο.κ. Θεωρητικά, υπάρχουν τόσες κύριες συνιστώσες όσος και ο αριθμός των αρχικών μεταβλητών αλλά για τα περισσότερα δεδομένα θεωρείται πως αρκεί να χρησιμοποιηθούν μερικές από τις κύριες συνιστώσες χωρίς απώλεια πληροφορίας. Είναι αυτονόητο πως στο ζήτημα των εγγράφων τα δεδομένα είναι τα έγγραφα και οι συνιστώσες οι λέξεις.

Η Λανθάνουσα Σημασιολογική Δεικτοδότηση (LSI) αρχικά δημιουργήθηκε για να αντιμετωπιστούν τα προβλήματα που προκύπτουν εξαιτίας της συνωνυμίας και της πολυσημίας των λέξεων. Αναλύοντας τις συσχετίσεις των όρων στα έγγραφα της συλλογής, η LSI προσπαθεί να μετασχηματίσει τα αρχικά διανύσματα αναπαράστασης σε ένα νέο, μικρότερο διανυσματικό χώρο (οι συνιστώσες του είναι οι "λανθάνουσες"). Με το μετασχηματισμό αυτό, έγγραφα που αναφέρονται στο ίδιο θέμα (πιθανώς με διαφορετικές λέξεις) τοποθετούνται "κοντά" στο νέο διανυσματικό χώρο.

Η LSI είναι μια εφαρμογή της μεθόδου SVD σε έναν πίνακα που αποτελείται από όλα τα διανύσματα των εγγράφων, είναι δηλαδή διάστασης $|n| \times |m|$ αν n είναι οι όροι και m τα έγγραφα. Μέσω της SVD ο αρχικός χώρος διάστασης n προβάλλεται σε ένα χώρο διάστασης k όπου $n >> k$ με την τιμή του k να επιλέγεται μεταξύ 100 και 150. Η διαφορά της LSI με την PCA είναι πως η τελευταία εφαρμόζεται μόνο σε τετραγωνικούς πίνακες ενώ η LSI εφαρμόζεται σε οποιονδήποτε πίνακα.

Ορίζοντας μαθηματικά το πρόβλημα, ο αρχικός πίνακας λέξεων-εγγράφων $\mathbf{X}_{n\times m}$ αναλύεται ως γινόμενο τριών πινάκων $\mathbf{T}$, $\mathbf{S}$ και $\mathbf{D}$ ως εξής:

$$\mathbf{X}_{n\times m} = \mathbf{T}_{n\times m}\mathbf{S}_{m\times m}(\mathbf{D}_{m\times m})^T \tag{4.3}$$

όπου n είναι ο αριθμός των λέξεων, m ο αριθμός των εγγράφων και οι πίνακες $\mathbf{T}$, $\mathbf{D}$ έχουν ορθογώνιες στήλες. Οι πίνακες $\mathbf{T}$ και $\mathbf{D}$ αναπαριστούν τις λέξεις και τα έγγραφα στο νέο χώρο. Ο διαγώνιος πίνακας $\mathbf{S}$ περιέχει τις ιδιάζουσες τιμές του πίνακα $\mathbf{X}$. Επιλέγοντας μόνο τις πρώτες k σειρές των πινάκων $\mathbf{T}$, $\mathbf{S}$, $\mathbf{D}$ τότε το γινόμενό τους είναι ο πίνακας $\hat{\mathbf{X}}$ στο νέο χώρο :

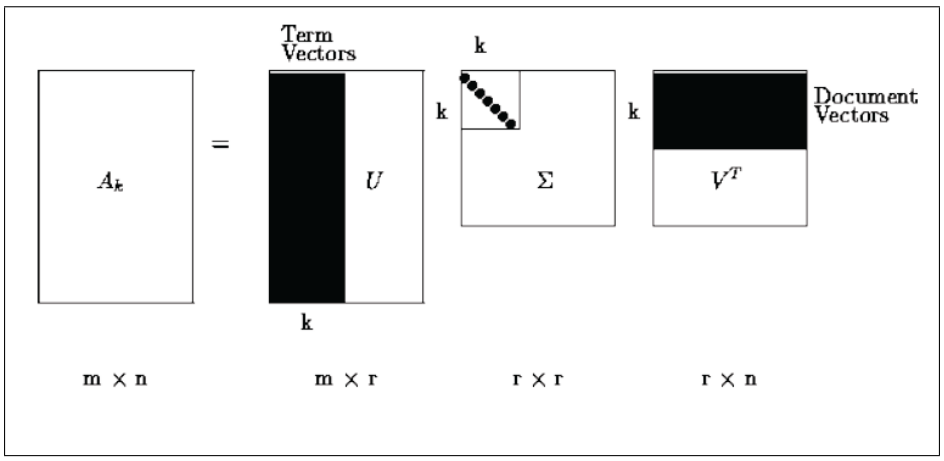
$$\hat{\mathbf{X}}_{n\times m} = \mathbf{T}_{n\times k}\mathbf{S}_{k\times k}(\mathbf{D}_{m\times k})^T \tag{4.4}$$

ο οποίος προφανώς είναι και η καλύτερη προσέγγιση βάσει της μικρότερης απόστασης Δ όπως ορίστηκε στην Εξίσωση 4.2.

Σε αυτή την αναπαράσταση οι στήλες του πίνακα $\mathbf{S}_k\mathbf{T}_k^T$ είναι οι όροι όπως "προβάλλονται" στο νέο χώρο (οι λανθάνοντες λέξεις ή συνιστώσες δηλαδή) και οι στήλες του πίνακα $\mathbf{D}_k$ είναι τα έγγραφα προβεβλημένα στο νέο χώρο. Ας σημειωθεί πως η νέα αναπαράσταση του εγγράφου j είναι το διάνυσμα $\mathbf{S}_k^{-1}\mathbf{T}_k^T\mathbf{X}(:,j)$ όπου το $\mathbf{X}(:,j)$ υποδηλώνει τη j-στήλη του πίνακα $\mathbf{X}$. Σχηματικά η λειτουργία της LSI φαίνεται στο Σχήμα 4.2.

Η υπόθεση που γίνεται στην LSI (και γενικά σε οποιαδήποτε παρόμοια τεχνική μείωσης διάστασης) είναι πως οι νέες διαστάσεις αποτελούν καλύτερη αναπαράσταση για τα έγγραφα. Μία κριτική στην SVD είναι πως -όπως και άλλες μέθοδοι που βασίζονται στη μείωση του τετραγωνικού σφάλματος- έχει σχεδιαστεί για ομοιόμορφα κατανεμημένα δεδομένα αλλά κάτι τέτοιο δεν ισχύει στα δεδομένα κειμένου. Ένα ακόμα προβληματικό χαρακτηριστικό της SVD είναι πως αφού η ανακατασκευή του πίνακα $\mathbf{X}$ βασίζεται σε κανονική κατανομή, μπορεί να οδηγήσει σε αρνητικές τιμές, το οποίο φυσικά δεν έχει φυσική σημασία στην περίπτωση των εγγράφων.

Η LSI θεωρεί πως οι νέες διαστάσεις αποτελούν τα "λανθάνοντα" θέματα τα οποία



Σχήμα 4.2: *Παράδειγμα λειτουργίας της LSI*

περιγράφονται από τη συλλογή των εγγράφων. Για την ακρίβεια, ομαδοποιούνται συνήθως διαστάσεις που αποτελούνται από λέξεις που μοιάζουν (εξ ου και προκύπτουν τα λανθάνοντα θέματα. Παρόλα αυτά, η φυσική ερμηνεία των νέων διαστάσεων μπορεί να είναι δύσκολη τις περισσότερες φορές.

Τέλος, υπάρχουν και παραλλαγές του κλασσικού μοντέλου της LSI, που υιοθετούν την πιθανοτική λογική (Probabilstic LSI, PLSI [Hofmann, 1999c]). Το αρχικό μοντέλο της PLSI ξεκινά από μία συλλογή εγγράφων και με μία γενίκευση του αλγορίθμου Μεγιστοποίησης Αναμονής (Expectation Maximization, EM) μπορεί να αντιμετωπίσει ζητήματα συνωνυμίας ή πολυσημίας. Σε αντίθεση με την κλασσική LSI, η πιθανοτική προσέγγιση έχει στέρεη στατιστική βάση και ορίζει πλήρως ένα μοντέλο δημιουργίας εγγράφων.

4.1.3 Παραλλαγές μοντέλου VSM

Στην αρχική μορφή του μοντέλου VSM χρησιμοποιήθηκε ως συνιστώσα κάθε διανύσματος η λέξη, η οποία αποτελεί και τη βασική μονάδα κάθε εγγράφου. Από εδώ προκύπτει και η ονομασία Bag-Of-Words για την αναπαράσταση των εγγράφων. Το πρόβλημα που δημιουργείται είναι πως αγνοείται η σημασιολογική συσχέτιση των λέξεων (π.χ. στην περίπτωση που σε ένα έγγραφο υπάρχει η φράση "data mining", το μοντέλο BOW θα την έσπαγε σε δύο αμφίσημες λέξεις: "data" και "mining". Είναι φανερό πως χάνεται η πραγματική έννοια της φράσης (που καλύπτεται μόνο αν οι δύο λέξεις κρατηθούν μαζί). Έτσι, κατά καιρούς έχουν προταθεί διάφορες παραλλαγές του μοντέλου BOW όπου αντί των απλών λέξεων, μπορούν να χρησιμοποιηθούν άλλες μονάδες (και αντίστοιχα να αποτελέσουν τις συνιστώσες των διανυσμάτων των εγγράφων).

4.1.3.1 Αναπαράσταση με βάση τα N-γράμματα

Τα N-γράμματα είναι μία τεχνική αναπαράστασης κειμένου που μετασχηματίζει τα έγγραφα σε διανύσματα μεγάλης διάστασης, όπου κάθε συνιστώσα του διανύσματος αντιστοιχεί σε μία διαδοχική ακολουθία χαρακτήρων μήκους N (π.χ. τα τριγράμματα περιλαμβάνουν όλες τις ακολουθίες με 3 χαρακτήρες). Είναι σαφές πως η διάσταση του διανύσματος αναπαράστασης με χρήση των N-γραμμάτων μπορεί να είναι πολύ μεγάλη για σχετικά χαμηλές τιμές του N, παρά την προεπεξεργασία που γίνεται (μετατροπή

κεφαλαίων σε πεζά, αφαίρεση ειδικών συμβόλων κλπ).

Η διαδικασία της εξαγωγής των N-γραμμάτων από ένα έγγραφο προσομοιάζεται με την κίνηση ενός παραθύρου διάστασης N από χαρακτήρα σε χαρακτήρα από την αρχή του εγγράφου έως το τέλος του. Κάθε πιθανή θέση του παραθύρου αυτού καλύπτει N χαρακτήρες και ορίζει ένα μοναδικό N-γραμμα. Σε αυτή τη διαδικασία, χαρακτήρες που δεν είναι γράμματα αντικαθίστανται από το διάστημα (space) και δύο ή περισσότερα διαδοχικά διαστήματα αντιμετωπίζονται ως ένα.

Σε σχέση με την αφαίρεση των κοινών λέξεων και την περιστολή των καταλήξεων, η αναπαράσταση με τα N-γράμματα πλεονεκτεί ως προς το ότι είναι λιγότερο ευαίσθητη σε γραμματικά ή τυπογραφικά σφάλματα και δεν απαιτεί γλωσσολογικές ή άλλες πληροφορίες με αποτέλεσμα να μπορεί να χρησιμοποιηθεί ανεξάρτητα από τη γλώσσα που είναι γραμμένη το κείμενο. Απ΄ την άλλη πλευρά όμως, η αναπαράσταση αυτή (όπως αναφέρθηκε) δε μειώνει τη διάσταση του διανύσματος αναπαράστασης.

4.1.3.2 Αναπαράσταση με βάση ομάδες λέξεων

Οι όροι που αποτελούνται από ομάδες λέξεων μπορούν να χρησιμοποιηθούν ως συνιστώσες των διανυσμάτων των εγγράφων. Σε περίπτωση που επιλεγεί αυτή η αναπαράσταση, υπάρχει η δυνατότητα για σημαντική μείωση της διάστασης των δεδομένων και γιαυτό θεωρείται πως είναι καλύτερη από την αναπαράσταση με βάση τις απλές λέξεις. Τα αποτελέσματα αρχικά ήταν ενθαρρυντικά αλλά όχι για όλες τις θεματικές κατηγορίες εγγράφων [Zhang et al., 2004], καθώς έπρεπε να βρεθεί ένας αποδοτικός τρόπος να εντοπιστούν οι λέξεις του κειμένου που αποτελούν έναν όρο. Για την επίτευξη του στόχου αυτού χρησιμοποιούνται γλωσσολογικές και στατιστικές πληροφορίες [Frantzi et al., 1998].

Οι γλωσσολογικές πληροφορίες περιλαμβάνουν κυρίως την επισήμανση των λέξεων με βάση το μέρος του λόγου το οποίο είναι και κατόπιν την επεξεργασία τους (φιλτράρισμα) ώστε να κρατηθούν οι συνήθεις συνδυασμοί μερών του λόγου που δίνουν όρους. Οι στατιστικές πληροφορίες έχουν να κάνουν με την ανάθεση σε υποψήφιους όρους ενός μέτρου που δείχνει κατά πόσο είναι πιθανό να αποτελούν έναν πραγματικό όρο βάσει χαρακτηριστικών των λέξεων που αποτελούν τους όρους. Για παράδειγμα μπορεί να αξιοποιηθεί η συχνότητα εμφάνισης του υποψήφιου όρου σε μια συλλογή εγγράφων, ο αριθμός των λέξεων κλπ

Στην κατεύθυνση του εντοπισμού των όρων που υπάρχουν σε ένα κείμενο συνεισφέρει και η εξέταση των γειτονικών λέξεων του όρου ("context"). Η ιδέα αυτή βασίζεται στο γεγονός πως λέξεις που χρησιμοποιούνται με παρόμοιους λεξιλογικούς τρόπους είναι υποψήφιες για να είναι συνώνυμες, επομένως οι λέξεις της γειτονιάς ενός όρου μπορούν να χρησιμοποιηθούν ως υπόδειξη της σημασίας τους. Η παρατήρηση αυτή επεκτεινόμενη, μπορεί να αξιοποιηθεί για τον εντοπισμό των ομάδων λέξεων που αποτελούν όρους. Για παράδειγμα, εάν κάποιο συγκεκριμένο επίθετο έχει την τάση να προηγείται όρων (π.χ. το επίθετο consistent να προηγείται ιατρικών όρων) τότε αυτό αξιοποιείται κατάλληλα. Αντίστοιχα, μπορεί να επεκταθεί πέραν από επίθετα και σε ουσιαστικά και ρήματα.

Οι συγκεκριμένες μέθοδοι παρότι μειώνουν τη διάσταση του διανύσματος αναπαράστασης έχουν το μειονέκτημα πως απαιτείται σημαντικός όγκος δουλειάς για τη δημιουργία των στατιστικών δεδομένων ενώ δεν υπάρχει κάποιος σίγουρος τρόπος να ελεγχθεί η ορθότητα των όρων που εξάγονται τελικά (πρέπει να γίνει χειροκίνητα).

4.1.3.3 Αναπαράσταση βάσει προτάσεων /φράσεων

Η λέξη αποτελεί τη βασική μονάδα οργάνωσης του κειμένου, παρόλα αυτά υπάρχουν αρκετές μεθοδολογίες που αξιοποιούν τη δομή του κειμένου σε προτάσεις (ή φράσεις) [Zamir & Etzioni, 1998], [Hammouda & Kamel, 2004].

Στις περισσότερες περιπτώσεις το έγγραφο αναπαρίσταται ως ένα διάνυσμα προτάσεων (και όχι λέξεων όπως στο κλασσικό μοντέλο BOW):

$$\mathbf{d}_i = \{\mathbf{s}_{ij} : j = 1, ..., p_i\} \tag{4.5}$$

$$\mathbf{s}_{ij} = \{t_{ijk} : k = 1, ..., l_{ij}; w_{ij}\} \tag{4.6}$$

όπου $\mathbf{d}_i$ είναι το έγγραφο i, $\mathbf{s}_{i,j}$ είναι η j πρόταση στο έγγραφο i, p_i είναι ο αριθμός των προτάσεων στο έγγραφο i, t_{ijk} είναι ο όρος k στην πρόταση s_{ij}, l_{ij} είναι το μήκος της πρότασης s_{ij} δηλαδή ο αριθμός των όρων που την αποτελούν, $w_{i,j}$ είναι το βάρος που ανατίθεται στην πρόταση s_{ij}.

Τα βάρη των προτάσεων ανατίθενται συνήθως λαμβάνοντας υπόψιν τη δομή του κειμένου (τίτλος, επικεφαλίδες, κείμενο με έντονη ή πλάγια γραφή κτλ). Ο παραπάνω ορισμός δε λαμβάνει υπόψιν του τη συχνότητα των προτάσεων (ή μέρους αυτών) σαν βάρος της πρότασης αλλά μπορεί να τροποποιηθεί για να ληφθεί ή συνηθέστερα να χρησιμοποιηθεί για το ταίριασμα φράσεων ανάμεσα σε έγγραφα. Η μεθοδολογία αυτή αναπαριστά τα έγγραφα διατηρώντας τη δομή των προτάσεων όπως παρουσιάζεται αρχικά (το μοντέλο BOW θεωρεί απλά τη συχνότητα των λέξεων). Έτσι, γίνεται αξιοποίηση των πληροφοριών της δομής αφενός των προτάσεων και αφετέρου του κειμένου.

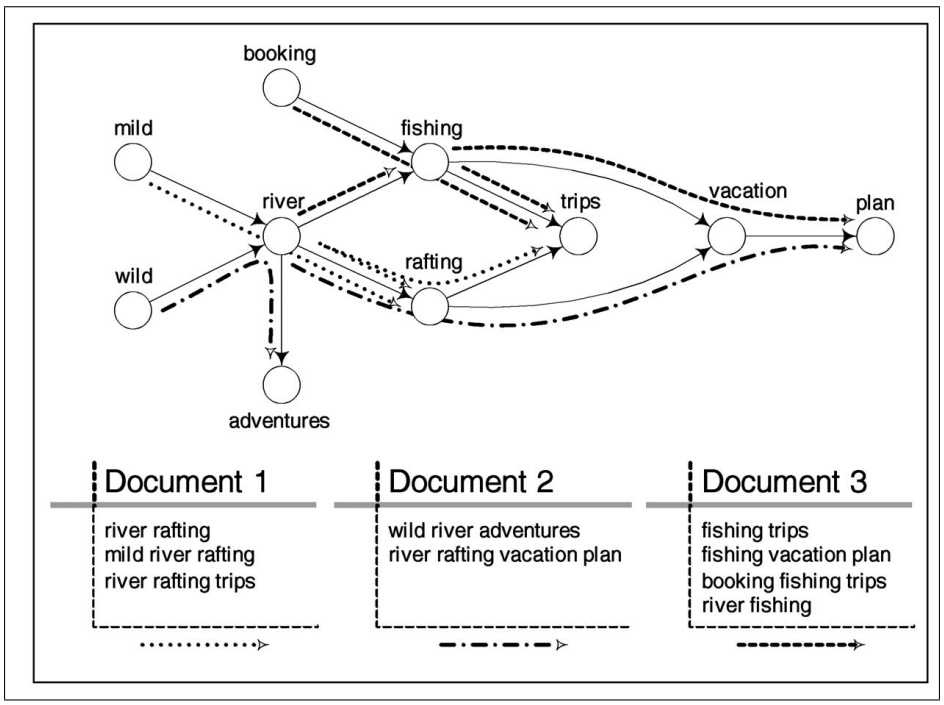
Αφού καθοριστούν οι φράσεις του κειμένου, το επόμενο βήμα είναι να βρεθεί μία ικανοποιητική μέθοδος αναπαράστασής τους. Δύο εναλλακτικές έχουν προταθεί: τα Suffix trees [Manber & Myers, 1990] ή οι γράφοι [Hammouda & Kamel, 2004]. Και στις δύο περιπτώσεις αναπτύσσονται τεχνικές υπολογισμού ομοιότητας εγγράφων βάσει των φράσεων τους. Ένα παράδειγμα αναπαράστασης με γράφο φαίνεται στο Σχήμα 4.3.

Το πλεονέκτημα μεθόδων αναπαράστασης με μονάδες μεγαλύτερες της λέξης (π.χ. προτάσεις) είναι πως οι ομάδες λέξεων εμπεριέχουν σημαντικότερο σημασιολογικό περιεχόμενο αλλά το μειονέκτημα είναι πως υπάρχει κίνδυνος όσο αυξάνεται η μοναδικότητα και η ιδιαιτερότητα των προτάσεων, τόσο να μειώνεται η δυνατότητα σύγκρισης αντίστοιχων ποσοτήτων μεταξύ κειμένων. Για το λόγο αυτό, επιλέγεται ως προσφορότερος τρόπος αναπαράστασης μία ενδιάμεση λύση μεταξύ των λέξεων και των προτάσεων που εξετάζεται στην επόμενη Παράγραφο και είναι οι ονοματικές φράσεις.

4.2 Αναπαράσταση με ονοματικές φράσεις (noun phrases)

4.2.1 Η σημασία των ονοματικών φράσεων στα κείμενα

Γραμματικά, οι ονοματικές φράσεις (noun phrases ή για συντομία NP) είναι φράσεις που αποτελούνται από ένα ουσιαστικό, αντωνυμία ή άλλη ονοματική λέξη που συνήθως συνοδεύεται από κάποια τροποποιητική λέξη όπως ένα επίθετο. Οι ονοματικές φράσεις θεωρείται πως είναι κοινές σε όλες τις γλώσσες (ενδεχόμενα με διαφορές στη σύνθεση των φράσεων ανάλογα με τις ιδιαιτερότητες της κάθε γλώσσας).



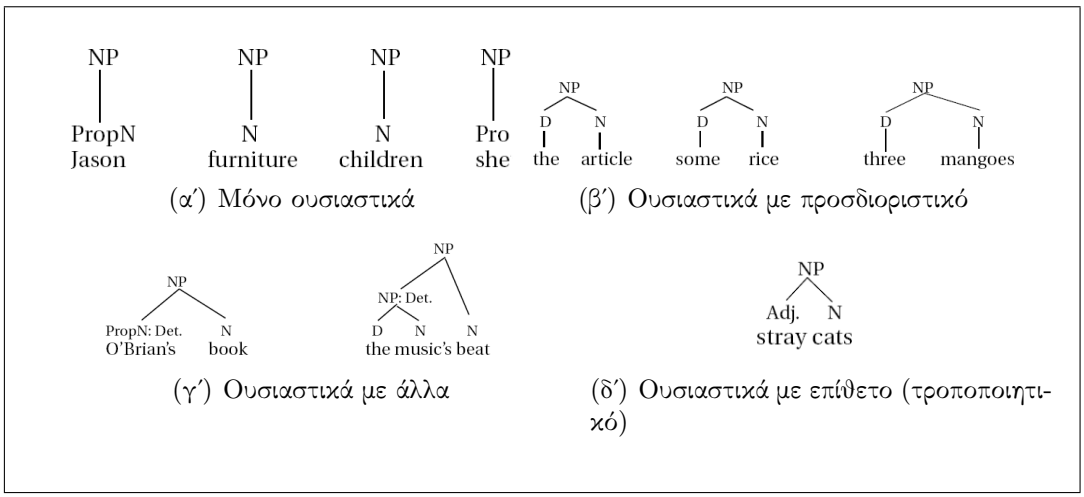
Σχήμα 4.3: *Παράδειγμα αναπαράστασης τριών εγγράφων χρησιμοποιώντας γράφους*

Παρακάτω παρουσιάζονται οι συνηθέστερες περιπτώσεις ονοματικών φράσεων (αφορούν στην αγγλική γλώσσα που χρησιμοποιείται στη διατριβή για λόγους σύγκρισης με διεθνείς εργασίες αλλά εύκολα μπορούν να γενικευτούν/προσαρμοστούν σε άλλες γλώσσες). Στο Σχήμα 4.4 αναλύονται ορισμένες συνήθεις περιπτώσεις ονοματικών φράσεων:

- Ουσιαστικά από μόνα τους μπορούν να αποτελούν ονοματική φράση (κύρια ονόματα, μαζικά ουσιαστικά, αριθμήσιμα ουσιαστικά και αντωνυμίες -που θεωρούνται περίπτωση ουσιαστικού-) (4.4(α′)),
- Ουσιαστικά μαζί με κάποιο προσδιορισμό (4.4(β′)),
- Ένα ή περισσότερα ουσιαστικά εμφανίζονται σε γενική πτώση (4.4(γ′)),
- Ουσιαστικό με κάποιο επίθετο (4.4(δ′))

Θεωρείται γενικώς πως τα ουσιαστικά διαδραματίζουν σημαντικό ρόλο σε πολλές γλώσσες (όπως η αγγλική). Έχει παρατηρηθεί [Algeo, 1995] πως μαζί με τα ρήματα αποτελούν το μεγαλύτερο μέρος των προτάσεων και πως το σημασιολογικό περιεχόμενο των προτάσεων υπάρχει κυρίως στα ουσιαστικά. Πιο συγκεκριμένα, η σημασία των NP (και κατ επέκταση και των ουσιαστικών) καταδεικνύεται και με πιο πρόσφατες εργασίες [Keizer, 2007] που αναδεικνύουν τις δομικές, λειτουργικές και κειμενικές διαστάσεις των NP και τις αλληλεπιδράσεις μεταξύ αυτών των διαστάσεων.

Ο ρόλος των ονοματικών φράσεων καταδεικνύεται περισσότερο αν προσπαθήσει κανείς να τις εντοπίσει σε ένα κείμενο. Για παράδειγμα στο Σχήμα 4.5 έχουν σημειωθεί με έντονη γραφή οι ονοματικές φράσεις και φαίνεται εύκολα πως καταλαμβάνουν το μεγαλύτερο μέρος του κειμένου (και άρα καθορίζουν έντονα και το περιεχόμενό του). Το επόμενο βήμα (που αναλύεται στην επόμενη Παράγραφο) είναι πως θα εντοπιστούν επαρκώς οι ονοματικές φράσεις ενός κειμένου.



Σχήμα 4.4: *Παραδείγματα ονοματικών φράσεων*

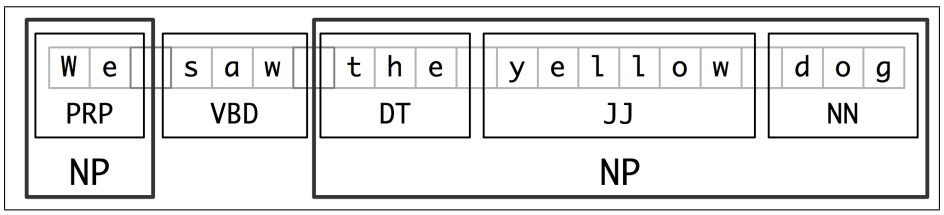
In **his social history of venereal disease**, **No Magic Bullet**, **Allan M. Brandt** describes **the controversy in the US military about preventing venereal disease among soldiers during World War I** . Should there be **a disease prevention effort that recognized that many young American men would succumb to the charms of French prostitutes** , or should there be **a more punitive approach to discourage sexual contact** ? Unlike **the New Zealand Expeditionary forces** , which gave **condoms** to **their soldiers** , **the United States** decided to give **American soldiers after-the-fact, and largely ineffective, chemical prophylaxis** . **American soldiers** also were subject to **court martial** if they contracted **a venereal disease** . **These measures** failed. **More than 383,000 soldiers** were diagnosed with **venereal diseases** between April 1917 and December 1919 and lost **seven million days of active duty** . **Only influenza** , which struck in **an epidemic** , was **a more common illness among servicemen**

Σχήμα 4.5: *Παράδειγμα κειμένου με σημειωμένες τις ονοματικές φράσεις*

4.2.2 Μεθοδολογίες εντοπισμού ονοματικών φράσεων

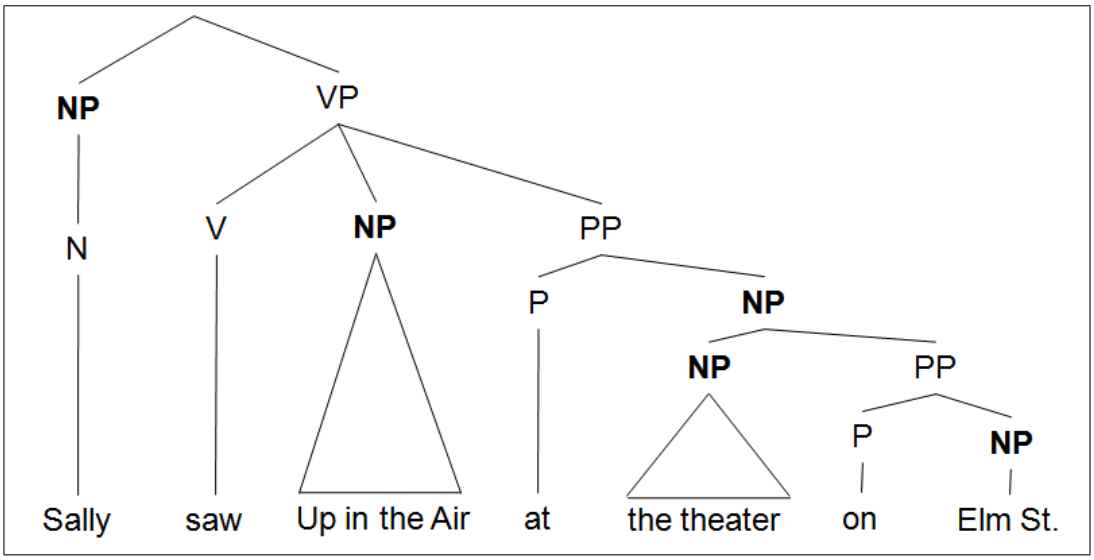
Ήδη από την περιγραφή των ονοματικών φράσεων, είναι φανερό πως απαιτείται μία διαδικασία επισήμανσης του κειμένου με τα μέρη του λόγου που αντιστοιχούν οι λέξεις του. Κατόπιν, παίρνοντας τους κατάλληλους συνδυασμούς μερών του λόγου προκύπτουν οι ονοματικές φράσεις. Η συνηθέστερη μεθοδολογία που χρησιμοποιείται για τον εντοπισμό των ονοματικών φράσεων είναι η κατάτμηση chunking, που χωρίζει και επισημαίνει ακολουθίες με πολλές λέξεις (όπως φαίνεται στο Σχήμα 4.6). Τη διαδικασία της κατάτμησης ακολουθεί η επισήμανση των μερών του λόγου που ανήκουν οι λέξεις της πρότασης (Part-Of-Speech (POS) tagging). Τα μικρά κουτιά δείχνουν την κατάτμηση στο επίπεδο της λέξης (και την επισήμανση των μερών του λόγου) ενώ τα μεγάλα κουτιά δείχνουν την κατάτμηση σε υψηλότερο επίπεδο. Κάθε ένα από τα μεγάλα κουτιά είναι ένα τμήμα (chunk).

Μετά την κατάτμηση, η διαδικασία της επιλογής τμημάτων (που δεν επικαλύπτονται) και αποτελούν ονοματικές φράσεις καλείται Κατάτμηση σε Ονοματικές Φράσεις (Noun Phrase Chunking). Για παράδειγμα, στην πρόταση του Σχήματος 4.7, είναι επιθυμητό



Σχήμα 4.6: *Παράδειγμα κατάτμησης κειμένου και επισήμανσης μερών του λόγου*

να εντοπιστούν οι ονοματικές φράσεις Sally, Up, the Air, the local theater, Elm Street.



Σχήμα 4.7: *Παράδειγμα συντακτικής ανάλυσης*

Υπάρχουν διάφορα συστήματα τα οποία έχουν αναπτυχθεί (ειδικά τον τελευταίο καιρό) και εντοπίζουν ονοματικές φράσεις σε κείμενο. Οι μεθοδολογίες που χρησιμοποιούνται θα μπορούσαν να χωριστούν σε δύο κατηγορίες:

- αναλυτές κειμένου βασισμένοι σε στατιστικούς κανόνες: Πρόκειται για την πιο απλή περίπτωση εντοπισμού ονοματικών φράσεων και βασίζεται στη χρήση απλών κανόνων (π.χ. *ένα άρθρο που προηγείται ουσιαστικού αποτελεί (μαζί με το ουσιαστικό) ονοματική φράση*) που περιγράφονται ως μορφή αφηρημένης γραμματικής με χρήση αυτομάτων.

- τεχνικές μηχανικής μάθησης: Προϋποθέτει την ύπαρξη πολλών παραδειγμάτων από τα οποία πρέπει το σύστημα να μάθει τους κανόνες και αντίστοιχα να τους ενημερώνει. Οι τεχνικές αυτές βασίζονται σε στατιστικές και πιθανοτικές μεθόδους. Στην κατηγορία αυτή εντάσσονται και τεχνικές όπως η μάθηση με μεγιστοποίηση της εντροπίας, με χρήση μοντέλων Markov ή με χρήση μηχανών διανυσμάτων υποστήριξης.

Τα πλεονεκτήματα και τα μειονεκτήματα των παραπάνω τεχνικών (ως επί το πλείστον συμπληρωματικά μεταξύ τους) συνοψίζονται στον Πίνακα 4.1.

Τα τελευταία χρόνια, αναπτύσσονται και συστήματα τα οποία προχωρούν ένα βήμα παραπέρα από τον εντοπισμό των ονοματικών φράσεων, καθώς τις κατατάσσουν και σε συγκεκριμένο είδος (π.χ. κύρια ονόματα, τοποθεσίες κτλ). Το ζήτημα αυτό καλείται Εντοπισμός Επώνυμων Οντοτήτων (Named Entity Tagging) και για παράδειγμα

Πίνακας 4.1: *Σύγκριση μεθόδων εξαγωγής ονοματικών φράσεων*

	Πλεονεκτήματα	**Μειονεκτήματα**
Απλές τεχνικές εξαγωγής βάσει κανόνων	Μεγάλη ευκολία υλοποίησης	Οι κανόνες δεν είναι πλήρεις
	Δεν απαιτείται σύνολο εκπαίδευσης	Δυσκολία εισαγωγής νέων κανόνων
Τεχνικές μηχανικής μάθησης	Διαχείριση πιο σύνθετων ή άγνωστων φαινομένων (εξαιρετικά χρήσιμο δεδομένων των άπειρων γλωσσικών συνδυασμών)	Απαιτείται μεγάλο σύνολο εκπαίδευσης
	Ευκολία προσαρμογής και αναθέσης βαρών στους κανόνες που δημιουργούνται αυτόματα	Σύνθετη υλοποίηση ανάλογα με το μοντέλο που επιλέγεται
		Μεγάλες απαιτήσεις μνήμης για την αποθήκευση των γραμματικών προτύπων

στην πρόταση του Σχήματος 4.7 ένα τέτοιο σύστημα είναι επιθυμητό να εντοπίσει ότι η λέξη Sally αποτελεί μια οντότητα-όνομα ενώ η φράση Up in the Air αποτελεί μια άλλη οντότητα-ταινία. Η κλασσική περίπτωση Εντοπισμού Ονομάτων Οντοτήτων περιλαμβάνει τον εντοπισμό 3 βασικών κατηγοριών λέξεων (ή ομάδων λέξεων) που είναι: Άτομα, Οργανισμοί και Τοποθεσίες (Person-Organization-Location).

Παραδείγματα τέτοιων συστημάτων υπάρχουν πολλά όπως το σύστημα OpenNLP NameFinder [OpenNLP, 2012], το σύστημα Illinois NER [Ratinov & Roth, 2009], το σύστημα Stanford NER [Stanford, 2009] και το σύστημα LingPipe [Alias-i, 2008]. Κοινό χαρακτηριστικό όλων αυτών των (πρόσφατα αναπτυχθέντων) συστημάτων είναι πως ξεφεύγουν από την παλαιότερη λογική χρήσης αλγορίθμων που βασίζονταν σε σταθερούς και χειρωνακτικά κατασκευασμένους κανόνες εξαγωγής και χρησιμοποιούν τεχνικές μηχανικής μάθησης. Επίσης, οι περισσότερες μέθοδοι καταλήγουν στο συμπέρασμα πως η χρήση εξωτερικής πηγής γνώσης και γενικά εξωτερικών γνωρισμάτων βοηθά καταλυτικά στη βελτίωση της απόδοσης ενός συστήματος αναγνώρισης οντοτήτων.

Ιδιαίτερη μνεία χρειάζεται σε συστήματα όπως το KNOWITALL [Etzioni et al., 2005] τα οποία χρησιμοποιούν τεχνικές μη-επιβλεπόμενης μάθησης, ανεξαρτήτως του είδους του περιεχομένου μεγέθους των συλλογών. Μεγάλο πλεονέκτημά τους είναι πως εξάγουν πληροφορία χωρίς να υπάρχουν προκαθορισμένα παραδείγματα εκπαίδευσης, παρόλα αυτά συνήθως χρειάζονται δύο στάδια για την εξαγωγή των οντοτήτων: στο πρώτο εντοπίζονται βάσει κάποιων συγκεκριμένων κανόνων ή προτύπων τα υποψήφια ονόματα προς εξαγωγή και στο δεύτερο στάδιο ελέγχεται η ορθότητα εξαγωγής κάθε οντότητας βάσει στατιστικών ή άλλων μεθόδων που αντιμετωπίζουν τον Παγκόσμιο Ιστό ως μια τεράστια αποθήκη γνώσης και την αξιοποιούν κατάλληλα.

Η κατακλείδα από τη μελέτη όλων των παραπάνω συστημάτων (εντοπισμού οντοτήτων και ονοματικών φράσεων) είναι πως αφενός οι ονοματικές φράσεις περιέχουν το σημαντικότερο σημασιολογικό περιεχόμενο των προτάσεων (και κατ επέκταση των κειμένων) και αφετέρου πως περαιτέρω σημασιολογική ενίσχυση του κειμένου μπορεί να γίνει μόνο μέσω εισαγωγής γνώσης από εξωτερική πηγή. Βασισμένη σε αυτούς τους

δύο άξονες (ονοματικές φράσεις και πληροφορία από εξωτερική πηγή γνώσης είναι και η μεθοδολογία που προτείνεται στην Παράγραφο 4.4.

4.3 Μεθοδολογίες αναπαράστασης εγγράφων με χρήση εξωτερικής γνώσης

Όπως έχει αναφερθεί ήδη στο Κεφάλαιο 2.5.2, το μοντέλο BOW εμφανίζει διάφορα μειονεκτήματα τα οποία το καθιστούν μη-αποδοτικό σε ζητήματα ανάκτησης και γενικά ανάλυσης εγγράφων. Κάθε σύστημα το οποίο αναπτύσσεται με στόχο να επεξεργαστεί (κατηγοριοποίηση, ομαδοποίηση, εξαγωγή θέματος κλπ) έγγραφα πρέπει να μπορεί να καλύψει κατά το δυνατόν τις σημασιολογικές σχέσεις και έννοιες των λέξεων που χρησιμοποιούνται σε αυτά.

Μία πρώιμη προσπάθεια για το θέμα, έγινε με τη χρήση της "περικείμενης πληροφορίας " ("contextual information") [Pullwitt, 2002]). Στη συγκεκριμένη εργασία εισάγεται ένα νέο διανυσματικό μοντέλο, το οποίο χρησιμοποιεί γνωρίσματα βασισμένα σε κατηγορίες προτάσεων που περιέχουν επιπλέον πληροφορία, ως εναλλακτική προσέγγιση στο κλασσικό μοντέλο BOW. Στο πρώτο στάδιο, το έγγραφο χωρίζεται σε προτάσεις, θεωρώντας πως κάθε πρόταση μπορεί να θεωρηθεί ως μία σχετικά αυτόνομη νοηματική ενότητα. Για την αναπαράσταση των προτάσεων χρησιμοποιείται το κλασσικό BOW μοντέλο, οπότε κάθε έγγραφο μπορεί να κωδικοποιηθεί ως το κεντροειδές των διανυσμάτων όλων των προτάσεων. Η παρατήρηση των συγγραφέων είναι πως ενώ σε μικρά έγγραφα οι λίγες προτάσεις παρουσιάζουν κανονική κατανομή, τα μεγαλύτερα έγγραφα παρουσιάζουν ανομοιόμορφη κατανομή έχοντας τα κεντροειδή μετατοπισμένα. Στη φάση αυτή εισάγονται οι κατηγορίες των προτάσεων (ως νευρώνες ενός μοντέλου αυτο-οργανούμενου χάρτη), στις οποίες γίνεται αντιστοίχιση όλων των προτάσεων βάσει κάποιας συνάρτησης ομοιότητας, το οποίο λαμβάνει χώρα στο δεύτερο στάδιο.

Διάφορες προσεγγίσεις έχουν γίνει ώστε να δημιουργηθούν οι κατάλληλες πηγές γνώσης (γενικού σκοπού) που θα εμπλουτίσουν το περιεχόμενο των εγγράφων. Στην πρώτη κατηγορία αυτών των προσπαθειών περιλαμβάνονται προσεγγίσεις που αξιοποιούν τεχνικές εξαγωγής πληροφορίας όπως ταίριασμα προτύπων (pattern matching), συντακτική ανάλυση φυσικής γλώσσας (natural language parsing), στατιστική μάθηση (statistical learning) κ.α. [Suchanek et al., 2006], [Agichtein et al., 2001], [Cafarella et al., 2005], [Cunningham et al., 2002], [Etzioni et al., 2004]. Κοινό χαρακτηριστικό όλων των παραπάνω μεθόδων είναι πως η επίτευξη ανάκλησης recall άνω του 90% για ένα συγκεκριμένο πεδίο συνήθως οδηγεί σε απώλεια της ακρίβειας precision και γενικά τα αποτελέσματά τους παρότι ικανοποιητικά και υποσχόμενα, δε συγκρίνονται με εκείνα που δημιουργούν πηγές γνώσης που κατασκευάζονται χειροκίνητα και όχι αυτόματα.

Οι τελευταίες ερευνητικές εργασίες στο πρόβλημα χρησιμοποιούν διάφορες μεθοδολογίες όπως ο εμπλουτισμός της ανάλυσης και της οπτικοποίησης των εγγράφων με σημασιολογικά γνωρίσματα από γνώση που προέρχεται από εξωτερικές πηγές [Breaux & Reed, 2005].

Τέτοιες πηγές γνώσης περιλαμβάνουν οντολογίες όπως το WordNet [Fellbaum, 1998], το Cyc ή OpenCyc [Matuszek et al., 2006], το SUMO [Niles & Pease, 2001], και πλήθος άλλων οντολογιών που περιορίζονται σε συγκεκριμένο πεδίο. (όπως το SNOMED [1] και το GeneOntology[2]. Το κύριο πλεονέκτημα αυτών των πηγών γνώσης

[1] http://www.snomed.org

[2] http://www.geneontology.org

είναι πως πληρούν υψηλές προδιαγραφές ποιότητας αλλά κατασκευάζονται χειροκίνητα, επομένως, υπάρχει το πρόβλημα της χαμηλής κάλυψης και του υψηλού κόστους συντήρησης. Στην κατηγορία των κατασκευασμένων πηγών γνώσης εντάσσεται και η Wikipedia, μία από τις μεγαλύτερες online διαθέσιμες εγκυκλοπαίδειες, η οποία όμως έχει το πλεονέκτημα πως συντηρείται από πληθώρα εκατομμυρίων χρηστών, με αποτέλεσμα οι ανανεώσεις του υλικού να είναι άμεσες (για παράδειγμα μπορεί κανείς να δει ποια είναι η τρέχουσα τελευταία έκδοση του αγαπημένου του λειτουργικού ή ποιος είναι ο τωρινός πρόεδρος των ΗΠΑ ακόμα και αν αυτά συνέβησαν μόλις μία μέρα πριν). Γιαυτό το λόγο, η Wikipedia (ή η εναλλακτική μορφή της DBPedia) χρησιμοποιείται ευρέως τον τελευταίο καιρό ως πηγή γνώσης εμπλουτισμού των εγγράφων ([Bizer et al., 2009], [Suchanek et al., 2008], [Navigli & Ponzetto, 2010]).

Οι [Bunescu & Pasca, 2007] στην εργασία τους χρησιμοποιούν μία μηχανή διανυσμάτων υποστήριξης που εκπαιδεύεται να εντοπίζει ονόματα ανθρώπων με βάση την ιδιότητά τους, ενώ γίνεται και αποσαφήνιση στην περίπτωση ονομάτων με άλλες ιδιότητες. Τα αποτελέσματα είναι ικανοποιητικά αλλά περιορίζονται μόνο σε κύρια ονόματα.

Το σύστημα Wikify! [Mihalcea & Csomai, 2007] ενοποιεί σε μία διαδικασία τον εντοπισμό λέξεων-κλειδιών (keywords) και της αποσαφήνισης εννοιών λέξεων βάσει του περιεχομένου της Wikipedia. Παρόλα αυτά η επισήμανση ενός μικρού ποσοστού (4.5% με 6%) των λέξεων των εγγράφων δε θεωρείται ικανοποιητική. Το ίδιο χαμηλό ποσοστό επισήμανσης ισχύει και για την εργασία του [Cucerzan, 2007] με την απόδοση όμως να θεωρείται καλύτερη λόγω της επιπλέον αξιοποίησης στατιστικών από τη Wikipedia αλλά και αποτελέσματα αναζήτησης στο WWW.

Στην εργασία των [Milne & Witten, 2008] γίνεται ένα βήμα παραπέρα και καταβάλλεται προσπάθεια να εισαχθεί η έννοια της οργάνωσης και της δομής σε ένα οποιοδήποτε κομμάτι κειμένου (που είναι αδόμητο).

Οι [Fader et al., 2009] επεκτείνουν την ιδέα των [Bunescu & Pasca, 2007] και [Cucerzan, 2007] χρησιμοποιώντας πληροφορίες για την αποσαφήνιση των εννοιών όχι μόνο από τα συμφραζόμενα αλλά και από εξωτερικά δεδομένα (όπως η σχέση μιας συμβολοσειράς με τις οντότητες που εμφανίζονται στη Wikipedia) ενώ η μέθοδός τους εφαρμόζεται σε πιο γενικά κείμενα (οι προηγούμενες εφαρμόζονται είτε σε άρθρα της Wikipedia είτε σε άρθρα ειδήσεων).

Η πιο πρόσφατη εργασία είναι αυτή των [Mendes et al., 2011] (Spotlight) οι οποίοι αξιοποιούν την εναλλακτική μορφή της Wikipedia τη DBpedia και δίνουν τη δυνατότητα στο χρήστη να προσαρμόσει τον εντοπισμό οντοτήτων βάσει των δικών του αναγκών αξιοποιώντας διάφορα μέτρα ποιότητας όπως τη σχετικότητα με τις υπόλοιπες οντότητες του κειμένου κλπ. Το σύστημα αποδίδει ικανοποιητικά αλλά η χρονική του απόδοση δεν είναι ιδιαίτερα ικανοποιητική όταν ο όγκος των κειμένων αυξάνει.

Στα πλαίσια της διατριβής στηρίζονται για τον εμπλουτισμό γνώσης για καλύτερη αναπαράσταση εγγράφων, η Wikipedia, η οποία παρουσιάζεται αμέσως μετά.

4.3.1 Η εγκυκλοπαίδεια Wikipedia και η εκτεταμένη χρήση της

Η Wikipedia έχει σαν στόχο τη δημιουργία μιας δωρεάν εγκυκλοπαίδειας σε πολλές γλώσσες. Σήμερα, θεωρείται η πιο μεγάλη και πιο ευρέως χρησιμοποιούμενη εγκυκλοπαίδεια σε χρήση. Η Wikipedia έχει γνωρίσει τέτοια άνθηση λόγω του βάρους που δίνεται στην οικειοθελή (και ελεύθερη) προσπάθεια και κρίση των συμμετεχόντων (θεωρητικά ο καθένας μπορεί να προσφέρει στη Wikipedia). Λόγω αυτής της ανάπτυ-

ξης, η Wikipedia έχει εξελιχτεί σε μία σημαντική πηγή γνώσης που χρησιμοποιήθηκε πρόσφατα σε πολλές εφαρμογές κειμένου όπως ανάκτηση κειμένου [Li et al., 2007], εξαγωγή οντοτήτων [Toral & Munoz, 2006], κατηγοριοποίηση εγγράφων [Gabrilovich & Markovitch, 2006], [Wang et al., 2009] και ομαδοποίηση εγγράφων [Hu et al., 2009], [Bloehdorn et al., 2006].

Η Wikipedia συγκεντρώνει όλα τα δομικά χαρακτηριστικά που έχουν και οι παραδοσιακές έντυπες εγκυκλοπαίδειες (οργάνωση των άρθρων με εσωτερικές αναφορές σε παρεμφερή άρθρα ή εξωτερικές αναφορές σε περαιτέρω βιβλιογραφία και κάποιο είδος ευρετηρίου βάσει θέματος) και παρουσιάζει και καινούρια γνωρίσματα τα οποία συνολικά συνοψίζονται παρακάτω.

- *Άρθρα:*

  Η βασική μονάδα οργάνωσης πληροφορίας της Wikipedia είναι το άρθρο. Κάθε άρθρο αντιστοιχεί σε μία σελίδα και σήμερα η Wikipedia περιέχει πάνω από 10 εκατομμύρια άρθρα σε σύνολο πάνω από 250 διαφορετικές γλώσσες. Η αγγλική έκδοση περιλαμβάνει περίπου 2.5 εκατομμύρια άρθρα (μη μετρώντας τις ανακατευθύνσεις και τις σελίδες αποσαφήνισης που αναλύονται παρακάτω).

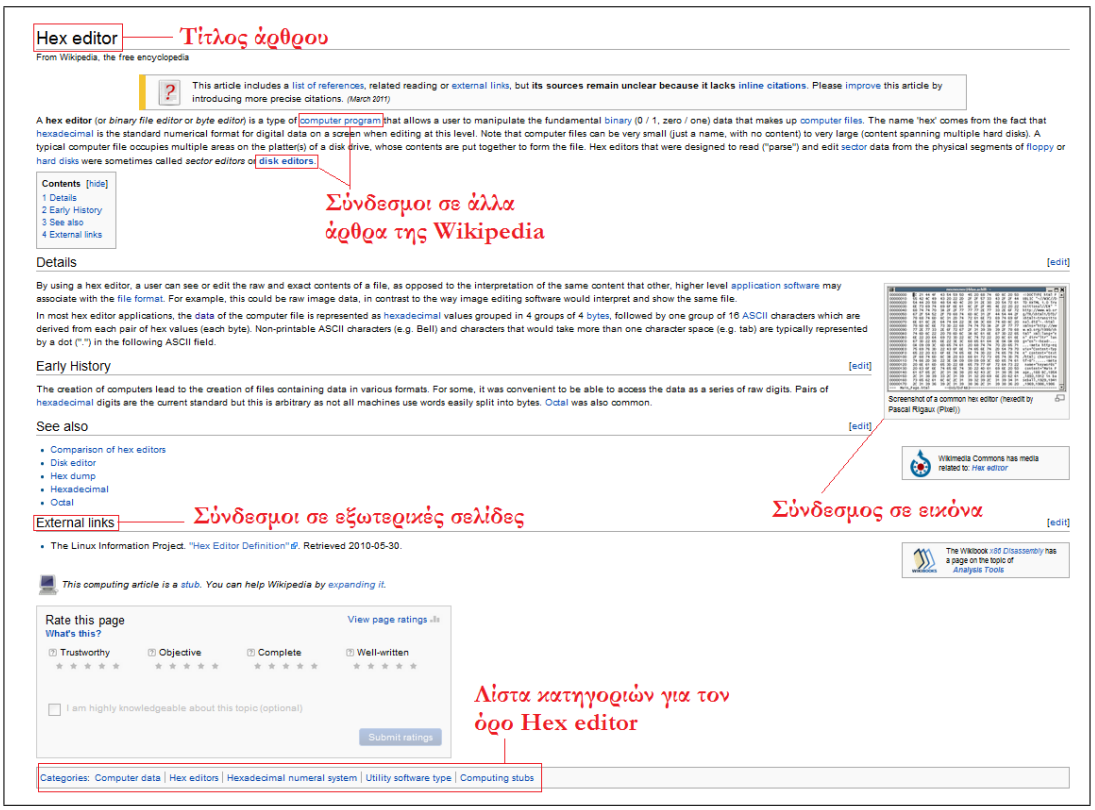
  Τα άρθρα είναι γραμμένα στη μορφή ελευθέρου κειμένου με καλά δομημένες προτάσεις βοηθώντας έτσι τη συνέπεια και τη συνοχή του άρθρου. Κάθε άρθρο υπακούει σε ορισμένες αρχές (που ελέγχονται από τους ίδιους τους χρήστες):

  1. Κάθε άρθρο περιγράφει μια συγκεκριμένη έννοια, ένα συγκεκριμένο αντικείμενο,
  2. Οι τίτλοι των άρθρων είναι λιτοί και περιεκτικοί (πάντα όμως περιγραφικοί) ώστε να μοιάζουν με όρους ενός συμβατικού θησαυρού (ή οντολογίας),
  3. Παρόμοιοι όροι συνδέονται μεταξύ τους μέσω των ανακατευθύνσεων (δες και παρακάτω),
  4. Οι σελίδες αποσαφήνισης παρέχουν τις διάφορες πιθανές ερμηνείες-έννοιες για τις πολύσημες έννοιες και οι χρήστες μπορούν να επιλέξουν το άρθρο που τους ενδιαφέρει (δες και παρακάτω),
  5. Τα άρθρα ξεκινούν με μία μικρή περιγραφή του θέματος και η πρώτη πρόταση συνήθως ορίζει τι ακριβώς περιγράφεται στο άρθρο,
  6. Τα άρθρα περιέχουν συνδέσμους που δείχνουν τη σχέση τους με άλλα άρθρα (δες και παρακάτω)

  Το Σχήμα 4.8 παρουσιάζει ένα τυπικό άρθρο που περιγράφει τον όρο Hex editor.

- *Ανακατευθύνσεις:*

  Μία σελίδα ανακατεύθυνσης δεν περιέχει κείμενο αλλά μόνο ένα σύνδεσμο σε ένα άρθρο (το οποίο θεωρείται παρόμοιο). Συνολικά υπάρχουν περίπου 3 εκατομμύρια τέτοιες ανακατευθύνσεις στην αγγλική Wikipedia. Για παράδειγμα, για τον όρο library υπάρχουν περίπου 12 τέτοιες σελίδες που καλύπτουν περιπτώσεις όπως: πληθυντικό αριθμό (libraries), τεχνικούς όρους (biblioteca), κοινά λάθη συλλαβισμού (libary) ή άλλες συνώνυμες μορφές (reading room, book stack). Ο σκοπός είναι σαφής: να υπάρχει ένα και μόνο άρθρο για μία συγκεκριμένη έννοια και όλες οι παρόμοιες να οδηγούν σε αυτό το άρθρο.



Σχήμα 4.8: *Παράδειγμα πληροφοριών από τη Wikipedia για τον όρο Hex editor*

- *Σελίδες αποσαφήνισης:*

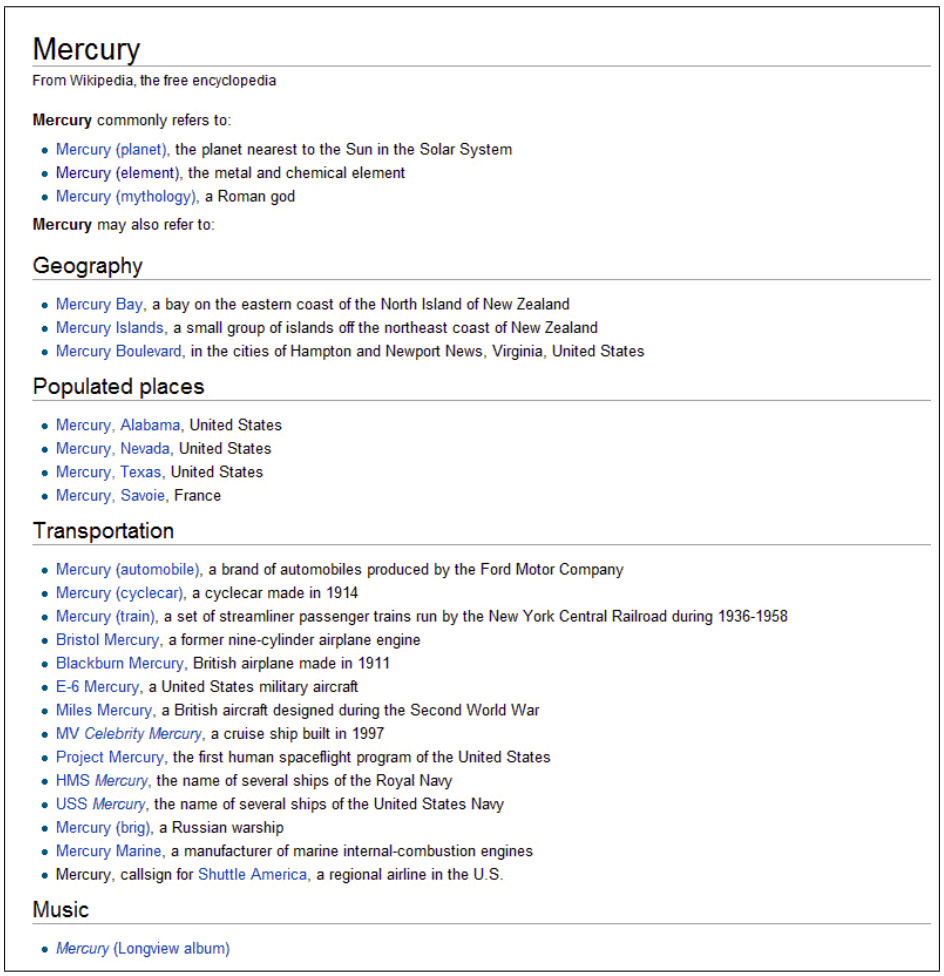
  Όταν ένας όρος είναι πολύσημος, τότε η αναζήτηση της Wikipedia παραπέμπει το χρήστη (στην πλειοψηφία των περιπτώσεων) σε μία σελίδα που οι χρήστες μπορούν να επιλέξουν την έννοια που τους ενδιαφέρει. Για τη διάκριση των διαφόρων άρθρων στον τίτλο προστίθεται (μέσα σε παρένθεση) και η αντίστοιχη θεματική περιοχή κάθε διαφορετικής έννοιας. Η αγγλική Wikipedia περιέχει περίπου 100.000 σελίδες αποσαφήνισης. Ένα παράδειγμα τέτοιας σελίδας φαίνεται στο Σχήμα 4.9.

- *Σύνδεσμοι:*

  Τα άρθρα της Wikipedia εμπλουτίζονται με συνδέσμους προς άλλα άρθρα (μεσοσταθμικά υπολογίζεται πως κάθε άρθρο έχει περίπου 25 συνδέσμους προς άλλα άρθρα). Η αγγλική Wikipedia περιέχει περίπου 60 εκατομμύρια συνδέσμους. Παρέχουν επεξηγήσεις θεμάτων που συζητούνται και εμπεριέχουν μία έννοια σχέσης του εξεταζόμενου άρθρου με το άρθρο που οδηγεί ο σύνδεσμος.

- *Κατηγορίες:*

  Τα άρθρα της Wikipedia είναι οργανωμένα ιεραρχικά σε μία πλούσια δομή κατηγοριών. Ένα άρθρο μπορεί να ανήκει σε περισσότερες της μιας κατηγορίες. Υπάρχουν περίπου 400.000 κατηγορίες στην αγγλική Wikipedia, καθεμιά περιέχει κατά μέσο όρο 19 άρθρα. Οι κατηγορίες δεν αποτελούν άρθρα, αλλά είναι σαν κόμβοι που χρησιμεύουν για την κατηγοριοποίηση των άρθρων. Αξίζει να σημειωθεί πως η ιεραρχική κατηγοριοποίηση της Wikipedia δεν είναι μία απλή δενδρική δομή αλλά περισσότερο ένας σύνθετος ακυκλικός κατευθυνόμενος γράφος.



Mercury

From Wikipedia, the free encyclopedia

Mercury commonly refers to:

- Mercury (planet), the planet nearest to the Sun in the Solar System
- Mercury (element), the metal and chemical element
- Mercury (mythology), a Roman god

Mercury may also refer to:

Geography

- Mercury Bay, a bay on the eastern coast of the North Island of New Zealand
- Mercury Islands, a small group of islands off the northeast coast of New Zealand
- Mercury Boulevard, in the cities of Hampton and Newport News, Virginia, United States

Populated places

- Mercury, Alabama, United States
- Mercury, Nevada, United States
- Mercury, Texas, United States
- Mercury, Savoie, France

Transportation

- Mercury (automobile), a brand of automobiles produced by the Ford Motor Company
- Mercury (cyclecar), a cyclecar made in 1914
- Mercury (train), a set of streamliner passenger trains run by the New York Central Railroad during 1936-1958
- Bristol Mercury, a former nine-cylinder airplane engine
- Blackburn Mercury, British airplane made in 1911
- E-6 Mercury, a United States military aircraft
- Miles Mercury, a British aircraft designed during the Second World War
- MV *Celebrity Mercury*, a cruise ship built in 1997
- Project Mercury, the first human spaceflight program of the United States
- HMS *Mercury*, the name of several ships of the Royal Navy
- USS *Mercury*, the name of several ships of the United States Navy
- Mercury (brig), a Russian warship
- Mercury Marine, a manufacturer of marine internal-combustion engines
- Mercury, callsign for Shuttle America, a regional airline in the U.S.

Music

- *Mercury* (Longview album)

Σχήμα 4.9: *Παράδειγμα σελίδας αποσαφήνισης της Wikipedia για τον όρο Mercury*

- *Άλλα δομικά στοιχεία:*

  Άλλα δομικά στοιχεία που συνθέτουν τα άρθρα της Wikipedia είναι μερικά πληροφοριακά πλαίσια (περιέχουν χρήσιμες πληροφορίες σε δομημένη μορφή βάσει κάποιων προτύπων (π.χ. περιέχονται πληροφορίες για την τοποθεσία, τον πληθυσμό κτλ πόλεων).

  Σε αρκετά άρθρα υπάρχουν επίσης και εικόνες (σύνδεσμος στη Wikipedia Commons Images).

Χάρη σε όλα τα παραπάνω χαρακτηριστικά (εύκολη ανανέωση άρθρων τα οποία είναι και καλά δομημένα, ιεραρχική δόμηση κατηγοριών, πυκνή διασύνδεση άρθρων κ.α.), η Wikipedia έχει χρησιμοποιηθεί ευρέως ως εξωτερική πηγή για εμπλουτισμό κειμένων. Επιπλέον, έχουν αναπτυχθεί διάφορες διεπαφές επικοινωνίας με τη βάση της Wikipedia, όπως το MediaWiki [3] το οποίο παρέχει δυνατότητα ανάκτησης πληροφοριών από τη βάση της Wikipedia σε πραγματικό χρόνο. Οι δυνατότητες που παρέχονται είναι πολλές και αφορούν στη δυνατότητα ανάκτησης κάθε δομικού στοιχείου της Wikipedia (τίτλοι, περιεχόμενα, σύνδεσμοι κτλ), σε πολλές πιθανές μορφές (ελεύθερο κείμενο, μορφή XML κλπ.)

Ας σημειωθεί πως η Wikipedia δεν παρέχει (στην αυθεντική της μορφή) μία πλήρως δομημένη βάση δεδομένων (οντολογία), αλλά πολλοί ερευνητές εργάζονται σε αυτή

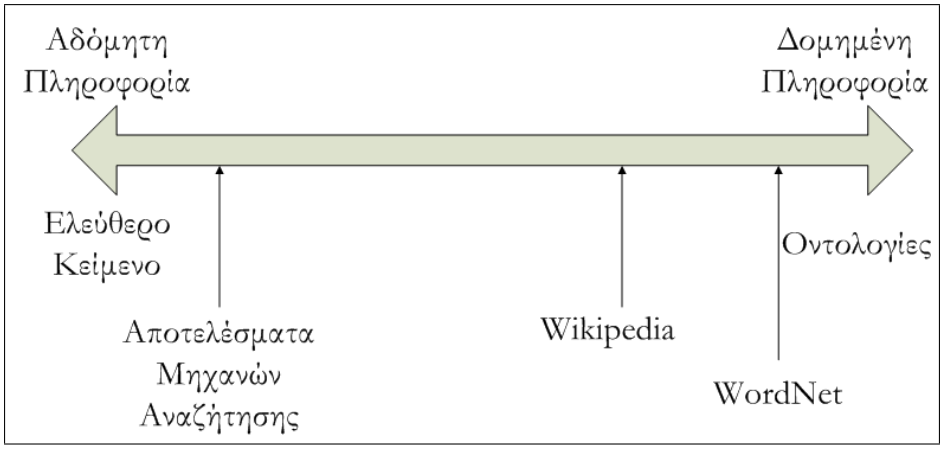
[3] http://www.mediawiki.org/wiki/API

Πίνακας 4.2: *Πίνακας χαρακτηριστικών των σημαντικότερων εξωτερικών πηγών γνώσης*

	Ιεραρχική Δομή	Θησαυρός	Ανανέωση	Κειμενικό Περιεχόμενο
WordNet	Ναι	Μεγάλος	Αργή	Λίγο
Wikipedia	Ναι	Περιορισμένος	Γρήγορη	Πολύ
Open Directory	Ναι	-	Αργή	-
Αποτελέσματα Μηχανών Αναζήτησης	Μερική	-	Γρήγορη	Λίγο

την κατεύθυνση, δηλαδή στη μετατροπή της σε μία πλήρη οντολογία [Bizer et al., 2009], [Suchanek et al., 2008], [Navigli & Ponzetto, 2010]. Με χρήση αυτών των οργανωμένων μορφών της Wikipedia ή της δομημένης ισοδύναμης οντολογίας τη DBpedia [Bizer et al., 2009], υπάρχει μεγάλο πεδίο έρευνας στον τομέα της αναγνώρισης ονομάτων οντοτήτων [Cucerzan, 2007], δίνοντας τη δυνατότητα να αντιστοιχιστούν οντότητες σε πρόσωπα, οργανισμούς, τοποθεσίες κ.τ.λ. Στα επόμενα συνοψίζονται τα κυριότερα συστήματα που έχουν αναπτυχθεί και χρησιμοποιούν τη Wikipedia ως πηγή γνώσης για εμπλουτισμό κειμένων και κατόπιν χρησιμοποιούν το αποτέλεσμα του εμπλουτισμού για κάποια εφαρμογή ανάλυσης κειμένων (π.χ. κατηγοριοποίηση).

Στον Πίνακα 4.2 φαίνονται συνοπτικά μερικές από τις πηγές εξωτερικής γνώσης που χρησιμοποιούνται μαζί με τα χαρακτηριστικά τους, ενώ στο Σχήμα 4.10 υπάρχει ο άξονας της "δομημένης πληροφορίας " στο ένα άκρο του οποίου υπάρχει το ελεύθερο κείμενο και στο άλλο οι οντολογίες και έχουν τοποθετηθεί σε αυτόν τρεις πηγές γνώσης σε σχέση με το βαθμό στον οποίο πλησιάζουν το ένα άκρο ή το άλλο.



Σχήμα 4.10: *Άξονας δόμησης της πληροφορίας και εξωτερικές πηγές γνώσης*

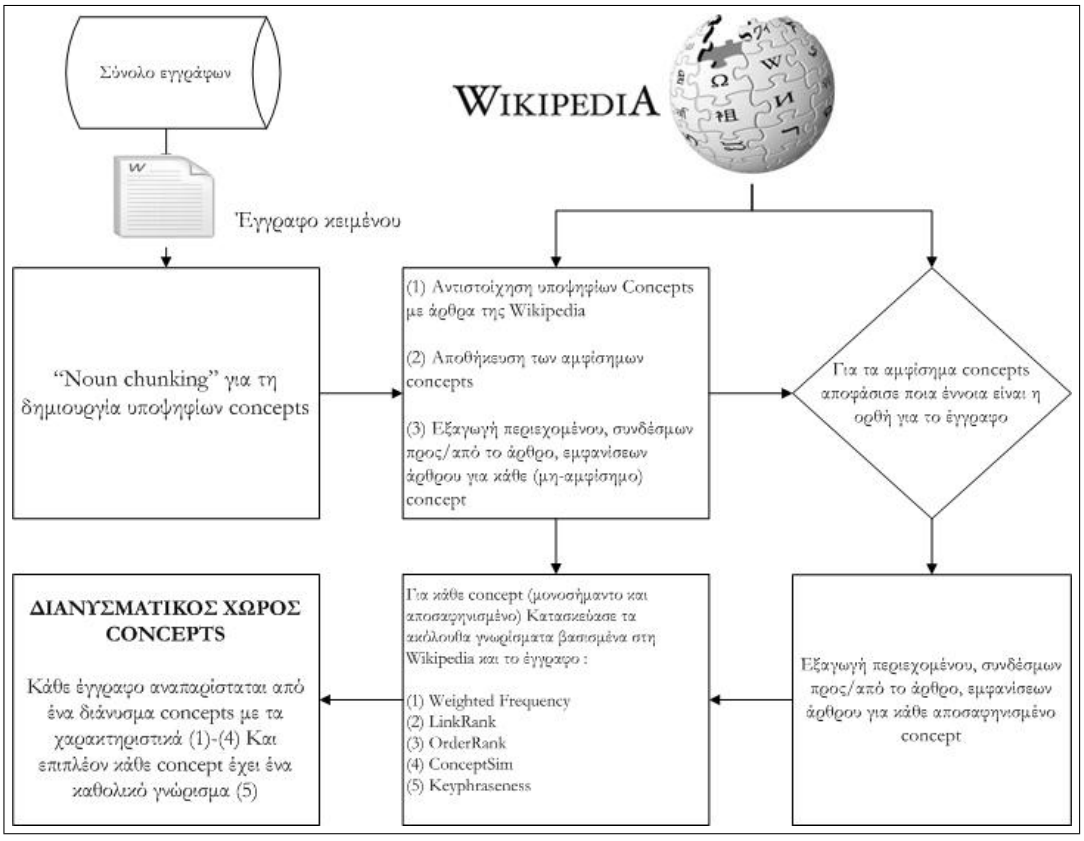
4.4 Μοντέλο αναπαράστασης εγγράφων με χρήση της Wikipedia

Στα προηγούμενα Κεφάλαια έγινε μια περιγραφή των προβλημάτων σημασιολογίας που δυσχεραίνουν την αναπαράσταση εγγράφων με ένα αποτελεσματικό και αποδοτικό μοντέλο. Οι λέξεις (μέσω του μοντέλου BOW) εξακολουθούν να αποτελούν τη βασική μονάδα αναπαράστασης εγγράφων, εξ ου και κυριαρχούν σε διαδικασίες π.χ. αναζήτη-

σης (οι περισσότεροι άνθρωποι αναζητούν ένα έγγραφο βάσει λέξεων-κλειδιών). Σε αυτή τη δυναμική της χρήσης των λέξεων βασίστηκε και η μεθοδολογία που αναπτύχθηκε στο Κεφάλαιο 3, όμως οι αδυναμίες που υπάρχουν από τη φύση της γλώσσας αλλά και τους περιορισμούς ενός μοντέλου βασισμένο σε λέξεις (και περιγράφηκαν στα προηγούμενα Κεφάλαια) αποτελούν στενωπό για την αποδοτική λειτουργία του μοντέλου.

Η εισαγωγή εξωτερικής γνώσης με στόχο το σημασιολογικό εμπλουτισμό του περιεχομένου των εγγράφων είναι μία πολλά υποσχόμενη κατεύθυνση και ακολουθείται στα πλαίσια της διατριβής για τη δημιουργία ενός μοντέλου αναπαράστασης εγγράφων που αφενός δε βασίζεται σε λέξεις αλλά σε έννοιες και αφετέρου εισάγει επιπλέον σημασιολογικό περιεχόμενο. Λόγω των χαρακτηριστικών και των πλεονεκτημάτων που αναφέρθηκαν στα προηγούμενα, χρησιμοποιείται η Wikipedia, βάσει της οποίας εντοπίζονται οι επώνυμες οντότητες - έννοιες κάθε εγγράφου και εξάγονται όλες οι χρήσιμες πληροφορίες για την εν λόγω έννοια από τη Wikipedia.

Η μεθοδολογία αξιοποίησης της Wikipedia με στόχο τον εμπλουτισμό της αναπαράστασης εγγράφων φαίνεται στο Σχήμα 4.4 και περιγράφεται αναλυτικά στις επόμενες Παραγράφους.



Σχήμα 4.11: *Μεθοδολογία αναπαράστασης εγγράφου με βάση άρθρα της Wikipedia*

4.4.1 Εξαγωγή επώνυμων οντοτήτων(named entities) - εννοιών(concepts) από τη Wikipedia

Ο στόχος της διαδικασίας είναι η εξαγωγή άρθρων της Wikipedia που περιγράφονται από μία ή περισσότερες διαδοχικές λέξεις του εγγράφου. Για παράδειγμα, εάν ένα έγγραφο περιέχει τη φράση "data mining", είναι επιθυμητό να εξαχθούν και οι δύο λέξεις

ως φράση και να αντιστοιχιστούν στο σχετικό άρθρο της Wikipedia, σχηματίζοντας μία επώνυμη οντότητα ή έννοια του εγγράφου. Το πρόβλημα που εγείρεται εδώ, είναι πως η πολυπλοκότητα της εξαγωγής όλων των πιθανών N-γραμμάτων (N-grams) (ώστε να ελεγχθούν όλες οι πιθανές περιπτώσεις άρθρων της Wikipedia που υπάρχουν στο έγγραφο) είναι πολύ μεγάλη [Wang & Domeniconi, 2008], και οι έως τώρα μέθοδοι βασίζονται μόνο στον περιορισμό της θεματολογίας σε συγκεκριμένο αντικείμενο [Wang et al., 2003]. Ο χώρος δηλαδή αναζήτησης για τη λύση του προβλήματος είναι πολύ μεγάλος, αλλά με χρήση τεχνικών γραμματικής ανάλυσης (επισήμανση μερών του λόγου κτλ), μπορούν να αναπτυχθούν μέθοδοι που εντοπίζουν ορθά οντότητες ονομάτων όπως ο Stanford Named Entity Recognizer [Stanford, 2009] ή ο LingPipe Exact Dictionary-Based Chunker [Alias-i, 2008] με αποτελέσματα αρκετά ενθαρρυντικά.

Στην προσέγγιση που αναπτύχθηκε, κάθε έγγραφο επισημαίνεται με πληροφορίες για το μέρος του λόγου (Part-of-Speech, POS) κάθε λέξης με χρήση του εργαλείου TreeTagger [Schmid, 1994]. Τα άρθρα της Wikipedia έχουν αρκετά περιγραφικούς τίτλους, οπότε δεν είναι απαραίτητο να γίνει περιστολή ή αφαίρεση λέξεων χωρίς νόημα. Για παράδειγμα στο Σχήμα 4.12 φαίνεται ένα κείμενο και στο Σχήμα 4.13 τα αποτελέσματα της επισήμανσης. Ο συμβολισμός των μερών του λόγου ακολουθεί τη σύμβαση του Penn Treebank Tagset [Marcus et al., 1993] και αναλυτικά όλες οι πιθανές τιμές του φαίνονται στο Παράρτημα Α′.

Depo Provera was developed in the 1960s and has been approved for contraception in many other countries.

The UpJohn Company of Kalamazoo, Mich., which will market the drug under the name, Depo Provera Contraceptive Injection, first submitted it for approval in the United States in the 1970s.

At that time, animal studies raised questions about its potential to cause breast cancer.

Σχήμα 4.12: *Παράδειγμα κειμένου από το έγγραφο* #59284 *του 20-NG*

Depo/NNP Provera/NNP was/VBD developed/VBN in/IN the/DT 1960s/NNS and/CC has/VBZ been/VBN approved/VBN for/IN contraception/NN in/IN many/JJ other/JJ countries/NNS

The/DT UpJohn/NNP Company/NNP of/IN Kalamazoo/NNP , Mich/NNP , which/WDT will/MD market/VB the/DT drug/NN under/IN the/DT name/NN , Depo/NNP Provera/NNP Contraceptive/NNP Injection/NNP first/RB submitted/VBD it/PRP for/IN approval/NN in/IN the/DT United/NNP States/NNPS in/IN the/DT 1970s/NNS

At/IN that/DT time/NN , animal/NN studies/NNS raised/VBD questions/NNS about/IN its/PRP$ potential/NN to/TO cause/VB breast/NN cancer/NN /

Σχήμα 4.13: *Παράδειγμα επισήμανσης μερών του λόγου σε κείμενο του εγγράφου* #59284 *του 20-NG*

Μετά από τη διαδικασία της επισήμανσης, επιλέγονται ως υποψήφιες έννοιες οι συνεχόμενες λέξεις που το POS τους είναι σε μία από τις ακόλουθες κατηγορίες :

- ουσιαστικά, ή κύρια ονόματα σε ενικό ή πληθυντικό (NN, NP, NNS, NPS)

- προθέσεις ή σύνδεσμοι (CC, IN),
- η λέξη to (TO)

με την υποσημείωση η πρώτη λέξη του συμπλέγματος να είναι στην 1η κατηγορία (NN,NP,NNS,NPS).

Ομαδοποιώντας τις συνεχόμενες λέξεις του εγγράφου που χαρακτηρίζονται από μία από τις προηγούμενες ετικέτες, πραγματοποείται πλήρης εξαγωγή των ονοματικών φράσεων (Noun Phrases) (για παράδειγμα με αυτό το σύστημα μπορεί να γίνει εξαγωγή τόσο του "Barack Obama" όσο και του "President of USA"), ενώ μειώνεται σημαντικά το υπολογιστικό κόστος του ελέγχου κάθε πιθανού N-γράμματος, συμπεριλαμβάνοντας ρήματα κ.τ.λ. Οι ονοματικές φράσεις που εξάγονται σχηματίζουν τις υποψήφιες επώνυμες οντότητες - έννοιες.

Κάθε υποψήφια έννοια, ελέγχεται αυτόματα εάν υπάρχει ή όχι σαν άρθρο της Wikipedia με χρήση του αντίστοιχου API [Wikipedia API, 2010]. Για να καλυφθεί η περίπτωση που κάποια πρόθεση ή σύνδεσμος δεν αποτελεί μέρος της έννοιας αλλά υπάρχει στο κείμενο προσδιοριστικά, γίνεται επιπλέον έλεγχος ξεχωριστά των ονοματικών φράσεων (χωρίς τις προθέσεις/συνδέσμους) εφόσον δεν έχει βρεθεί η έννοια στη Wikipedia με αυτές. Αν η έννοια έχει πολλαπλές ερμηνείες (επομένως υπάρχουν και πολλαπλά άρθρα της Wikipedia που αναφέρονται στην ίδια ονοματική φράση), εκτελείται αποσαφήνιση της έννοιας με τη διαδικασία που περιγράφεται στην Παράγραφο 4.4.2, για να επιλεγεί η πιο συναφής με το κείμενο έννοια. Μόλις επιτευχθεί μοναδική αντιστοίχιση μεταξύ της υποψήφιας έννοιας και της Wikipedia, η έννοια επιλέγεται ως μία συνιστώσα του διανύσματος του εγγράφου που θα σχηματιστεί. Για παράδειγμα, θεωρείστε το ακόλουθο απόσπασμα από ένα έγγραφο του συνόλου 20 Newsgroups (περιγράφεται αναλυτικά στο Παράρτημα Γ′). Στο ακόλουθο απόσπασμα (από το έγγραφο #59284 του 20-NG), οι έννοιες οι οποίες εξάγονται φαίνονται με **έντονη γραφή**.

Depo Provera *was developed in the* ***1960s*** *and has been approved for* ***contraception*** *many other* ***countries****. The* ***UpJohn Company*** *of* ***Kalamazoo, Mich.****, which will market the* ***drug*** *under the* ***name****, Depo Provera Contraceptive Injection, first submitted it for* ***approval*** *in the* ***United States*** *in the* ***1970s****. At that* ***time, animal studies*** *raised* ***questions*** *about its* ***potential*** *to cause* ***breast cancer****.*

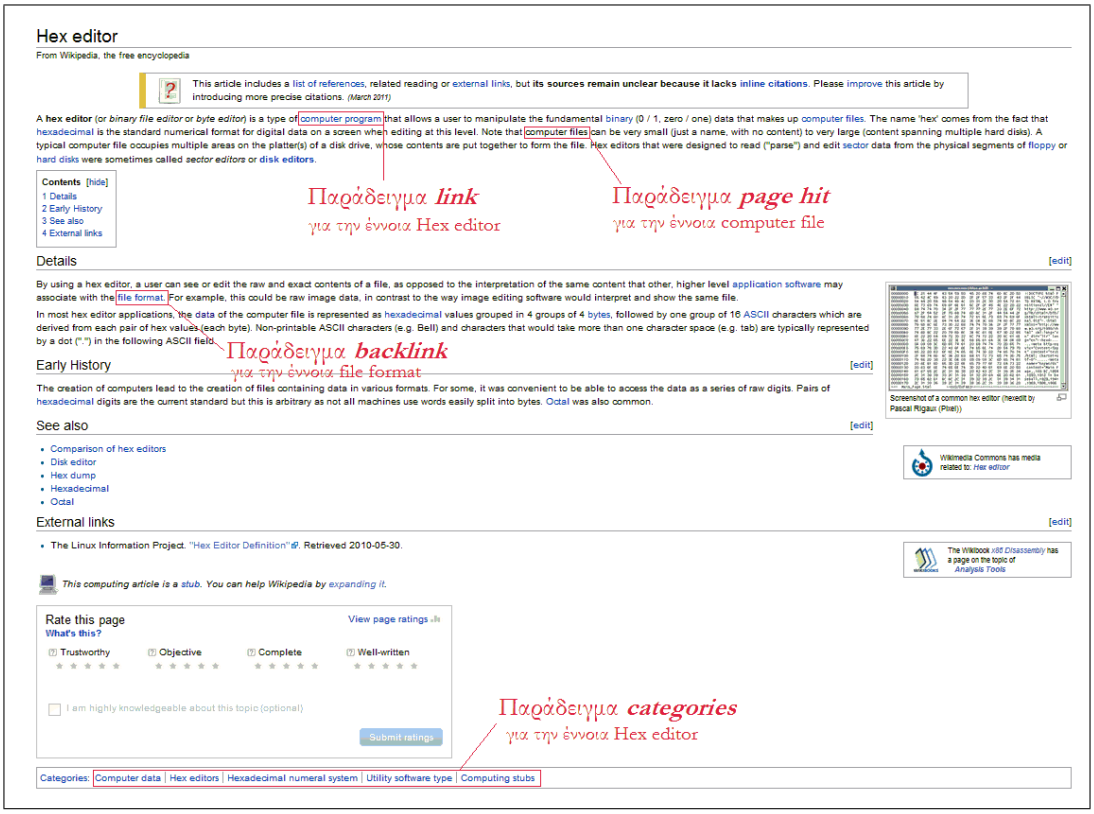
Είναι προφανές το πόσο σημαντικό είναι σε ένα σύστημα ανάκτησης πληροφοριών να υπάρχει η δυνατότητα να εντοπιστούν έννοιες όπως οι "Depo Provera", "UpJohn Company", "United States" οι οποίες υπό άλλες συνθήκες θα χωρίζονταν σε δύο λέξεις χωρίς σαφές περιεχόμενο (π.χ. "Depo" ή "Provera") ή με άλλη σημασία (π.χ. "United" ή "States").

Αφού γίνει ο έλεγχος και βρεθεί το σύνολο των επωνύμων οντοτήτων που αποτελούν τις έννοιες του εγγράφου, με χρήση του API της Wikipedia, για κάθε επιλεγμένη έννοια i, εξάγονται τα γνωρίσματα που φαίνονται στο Σχήμα 4.14. Ένα παράδειγμα των γνωρισμάτων που εξάγονται (για τον όρο "Hex Editor" που παρουσιάστηκε στο Σχήμα 4.8) φαίνεται στο Σχήμα 4.15.

Μετά την εξαγωγή των γνωρισμάτων που περιγράφηκαν στο Σχήμα 4.14 για κάθε έννοια σε ένα έγγραφο, γίνεται ένας συνδυασμός τους για τη δημιουργία χαρακτηριστικών τα οποία εμπλουτίζουν την αναπαράσταση του εγγράφου. Ο στόχος είναι η

- $Content_i$: το κειμενικό περιεχόμενο του άρθρου της Wikipedia
- $Links_i$: σύνδεσμοι από το εξεταζόμενο άρθρο σε άλλα άρθρα της Wikipedia
- $BackLinks_i$: άρθρα τα οποία έχουν σύνδεσμο στο εξεταζόμενο άρθρο
- $PageHits_i$: άρθρα στα οποία το εξεταζόμενο άρθρο (noun phrase) εμφανίζεται, είτε ως σύνδεσμος είτε όχι (απλή αναφορά)
- $Categories_i$: η λίστα κατηγοριών της Wikipedia που το εξεταζόμενο άρθρο ανήκει

Σχήμα 4.14: *Γνωρίσματα που εξάγονται με τη βοήθεια του API της Wikipedia*



Σχήμα 4.15: *Παράδειγμα γνωρισμάτων άρθρων της Wikipedia που εξάγονται*

δημιουργία ενός διανύσματος χαρακτηριστικών που θα δίνουν το βάρος μιας έννοιας i σε κάθε έγγραφο j και τα οποία περιγράφονται στις ακόλουθες εξισώσεις.

- Weighted Frequency ($Wfreq$) ορίζεται ως εξής :

$$WFreq_{j,i} = size_i * frequency_{j,i} \tag{4.7}$$

όπου :

– $size_i$ είναι ο αριθμός των λέξεων που σχηματίζουν την έννοια i,

– $frequency_{j,i}$ είναι ο αριθμός των φορών που η έννοια i εμφανίζεται σε ένα j.

Το συγκεκριμένο μέτρο μοιάζει με τον κλασσικό όρο tf που περιγράφηκε στα προηγούμενα, εκφράζει δηλαδή τη συχνότητα εμφάνισης μιας έννοιας στο έγγραφο, αλλά ενισχύεται και από τον αριθμό των λέξεων που απαρτίζουν την έννοια. Έτσι δίνεται βάση στο σημασιολογικό περιεχόμενο που περικλείουν ομάδες όρων.

- *LinkRank* είναι ένα μέτρο του πόσους συνδέσμους έχει κοινούς μία έννοια με το σύνολο των συνδέσμων που περιέχονται σε ένα έγγραφο, δηλαδή είναι ένα μέτρο της σημαντικότητας της έννοιας για το έγγραφο και ορίζεται τυπικά ως εξής :

$$LinkRank_{j,i} = \frac{|Links_i \bigcap Links_{Doc_j}|}{|Links_{Doc_j}|} \tag{4.8}$$

όπου :

– $Links_i$ είναι ο αριθμός των συνδέσμων της έννοιας i, όπως ορίστηκε στο σχήμα 4.14,

– $Links_{Doc_j}$ είναι το σύνολο των συνδέσμων του εγγράφου j, που ορίζεται ως οι σύνδεσμοι όλων των εννοιών που αναπαριστούν το έγγραφο j.

Με το μέτρο αυτό δίνεται έμφαση στην πλούσια διασύνδεση των άρθρων της Wikipedia και που υπονοεί πως υπάρχουν σχέσεις μεταξύ των διασυνδεδεμένων άρθρων. Έτσι, μία έννοια που περιέχει πολλούς κοινούς συνδέσμους με μία άλλη έννοια του ίδιου εγγράφου (πέραν του ότι μπορεί να συνάγει κανείς πως έχουν κοινά στοιχεία μεταξύ τους) έχει μεγάλες πιθανότητες να περιγράφει το θέμα του εγγράφου.

- *ConceptSim* είναι η ομοιότητα μεταξύ του εγγράφου και του περιεχομένου του άρθρου της Wikipedia που αντιστοιχεί στην έννοια, υπολογιζόμενη με κλασσικούς όρους του μοντέλου $tf - idf$ όπως παρουσιάστηκε στα προηγούμενα και υπολογίζεται από την ακόλουθη εξίσωση :

$$ConceptSim_{j,i} = \cos(\mathbf{v}_j, \mathbf{v}_i) \tag{4.9}$$

όπου :

– $\mathbf{v}_j$ είναι το διάνυσμα $tf - idf$ για το έγγραφο j,

– $\mathbf{v}_i$ είναι το διάνυσμα $tf - idf$ για το κείμενο του άρθρου της Wikipedia που αντιστοιχεί στην έννοια i,

– cos είναι η συνάρτηση συνημιτόνου που υπολογίζει την ομοιότητα ανάμεσα στα δύο διανύσματα.

Το μέτρο αυτό αξιοποιεί την παραδοσιακή διαδικασία ανάκτησης κειμένου (χρήση διανυσμάτων $tf - idf$) ανάμεσα στο αρχικό έγγραφο που εξετάζεται και στο κειμενικό περιεχόμενο που συνάγεται από το άρθρο της Wikipedia. Είναι προφανές πως πληθώρα κοινών λέξεων είναι ενδεικτική της σημαντικότητας της έννοιας στο έγγραφο.

- *OrderRank* είναι μία τιμή που παίρνει μεγαλύτερες τιμές για έννοιες οι οποίες εμφανίζονται στην αρχή του εγγράφου, μια ιδέα βασισμένη στην παρατήρηση πως σημαντικές έννοιες-λέξεις συχνά εμφανίζονται στην αρχή ενός κειμένου [Xue & Zhou, 2009]. Τυπικά ορίζεται ως εξής :

$$OrderRank_{j,i} = 1 - \frac{arraypos_i}{|j|} \tag{4.10}$$

όπου :

  - *arraypos* είναι ένας πίνακας που περιέχει όλες τις λέξεις του εγγράφου στη σειρά που εμφανίζονται και $arraypos_i$ αναπαριστά τη θέση της πρώτης εμφάνισης της έννοιας i στον πίνακα. Εάν μία έννοια αποτελείται από περισσότερες της μιας λέξης, τότε λαμβάνεται υπόψιν η θέση εμφάνισης της πρώτης λέξης της έννοιας,
  - $|j|$ είναι το μέγεθος του εγγράφου j, δηλαδή από πόσες λέξεις αποτελείται.

Επιπλέον, ορίζεται ένα καθολικό (ανεξάρτητο από τα έγγραφα) μέτρο για κάθε έννοια το οποίο ορίζεται ως ακολούθως:

- *Keyphraseness* είναι ένα μέτρο, προσαρμοσμένο από το [Mihalcea & Csomai, 2007], που έχει μια συγκεκριμένη τιμή για κάθε διαφορετική έννοια, ανεξαρτήτως του εγγράφου στο οποίο αναφερόμαστε, και είναι μία ένδειξη του κατά πόσον μία έννοια είναι περιγραφική και συγκεκριμένη για ένα θέμα. Ορίζεται ως εξής :

$$Keyphraseness(i) = \frac{BackLinks_i}{PageHits_i} \tag{4.11}$$

Μία έννοια με μεγάλη τιμή για το *Keyphraseness* (που σημαίνει πως οι περισσότερες εμφανίσεις του στη Wikipedia είναι με τη μορφή συνδέσμων στο αντίστοιχο άρθρο και όχι απλές εμφανίσεις) έχει μεγαλύτερη περιγραφική δύναμη από μία έννοια με μικρή τιμή *Keyphraseness*, ακόμα και αν η δεύτερη εμφανίζεται περισσότερες φορές στη Wikipedia (αλλά λιγότερες ως σύνδεσμος).

4.4.2 Αποσαφήνιση εννοιών (Sense Disambiguation)

Εάν μία υποψήφια έννοια έχει πολλαπλές σημασίες, τότε είναι απαραίτητη η διαδικασία της αποσαφήνισης για να βρεθεί η σημασία που ταιριάζει καλύτερα στο έγγραφο που εξετάζεται. Πολλές τεχνικές αποσαφήνισης (disambiguation) που βασίζονται στη Wikipedia έχουν παρουσιαστεί κατά καιρούς με ικανοποιητικά αποτελέσματα ([Milne & Witten, 2008], [Ratinov et al., 2011], [Bunescu & Pasca, 2007], [Mendes et al., 2011]). Παρόλα αυτά, για την επιτάχυνση της διαδικασίας της αντιστοίχησης εννοιών του εγγράφου σε άρθρα της Wikipedia, υιοθετείται εδώ μία πιο απλή και γρήγορη αλλά συνάμα αποτελεσματική τεχνική αποσαφήνισης.

Το *ConceptSim* (όπως παρουσιάστηκε στην εξίσωση 4.9) χρησιμοποιείται για τη διαδικασία της αποσαφήνισης. Υπενθυμίζεται πως το *ConceptSim* βασίζεται στην ομοιότητα μεταξύ δύο διανυσμάτων $tf - idf$ (του εγγράφου και της εξεταζόμενης έννοιας), επομένως όσο μεγαλύτερη τιμή έχει, τόσο μεγαλύτερη είναι και η ομοιότητα ανάμεσα στα αντίστοιχα κείμενα. Δηλαδή, η έννοια με τη μεγαλύτερη τιμή *ConceptSim* είναι η πιο σχετική με το εξεταζόμενο έγγραφο.

Για την επίτευξη μεγαλύτερης ακρίβειας στην αποσαφήνιση, γίνεται και χρήση των κατηγοριών της Wikipedia στις οποίες ανήκει η εξεταζόμενη έννοια και γίνεται ενσωμάτωση με το *ConceptSim*, οδηγώντας σε ένα πιο χαρακτηριστικό μέτρο που δείχνει πόσο σχετική είναι μία έννοια (c) με το εξεταζόμενο έγγραφο (j) και το οποίο καλείται $SenseSim_{j,c}$ και ορίζεται από την ακόλουθη εξίσωση :

$$SenseSim_{j,c} = \lambda * ConceptSim_{j,c} + (1 - \lambda) * Dice(Categories_c, Categories_{Doc_j}) \tag{4.12}$$

όπου :

- $ConceptSim_{j,c}$ δίνεται από την εξίσωση 4.9,
- $Categories_c$ δείχνει τις κατηγορίες του c όπως ορίζεται στο Σχήμα 4.14,
- $Categories_{Doc_j}$ δείχνει τις κατηγορίες του εγγράφου j, δηλαδή τις κατηγορίες όλων των μονοσήμαντων εννοιών που αναπαριστούν το έγγραφο j,
- λ είναι μία παράμετρος στο διάστημα $[0, 1]$ για την ανάθεση βαρών μεταξύ των δύο μετρικών,
- *Dice* είναι ο γνωστός συντελεστής που ορίζεται ως εξής :

$$Dice(A, B) = \frac{2 * |A \bigcap B|}{|A + B|} \tag{4.13}$$

Για παράδειγμα, το έγγραφο #67480 του συνόλου δεδομένων 20-NG ανήκει στην κατηγορία comp.windows.x, και η έννοια "Client" στη Wikipedia αναφέρεται σε διαφορετικές σημασίες της λέξης, όπως φαίνεται στον Πίνακα 4.3. Ένα μικρό απόσπασμα του περιεχομένου του εγγράφου #67480 στο οποίο συναντάται η λέξη "client" φαίνεται παρακάτω :

"Since the server end is (or was) always at this end (California) it is faster to remotely run the client via DESQview X and have a short hop to the server than running the client locally and having a long hop to the server."

Το μέτρο *SenseSim* υπολογίζεται για κάθε μία από τις σημασίες και αυτή με τη μεγαλύτερη τιμή επιλέγεται ως έννοια που είναι μέρος της αναπαράστασης του εγγράφου (στην περίπτωση αυτή είναι προφανές πως η λέξη "client" χρησιμοποιείται με την υπολογιστική της σημασία). Για πιο ακριβή αποτελέσματα και για να μην εισαχθούν στο έγγραφο λανθασμένες σημασίες, είναι δυνατό να εισαχθεί ένα κατώφλι στην ελάχιστη τιμή για το *SenseSim* κάτω από την οποία η έννοια θα αγνοείται στο έγγραφο. Στα πειράματα που πραγματοποιήθηκαν στο επόμενο Κεφάλαιο (Ομαδοποίηση Εγγράφων), το κατώφλι τέθηκε στο 0.05.

4.4.3 Αναπαράσταση εγγράφου

Μετά την ολοκλήρωση της διαδικασίας αποσαφήνισης, υπάρχει ένας αριθμός εννοιών που αναπαριστούν το έγγραφο. Ο σκοπός είναι η κατασκευή ενός διανύσματος όπου κάθε συνιστώσα θα ανταποκρίνεται στη σημαντικότητα κάθε έννοιας στο έγγραφο. Όπως προηγούμενα αναφέρθηκε, κάθε έννοια έχει τέσσερα γνωρίσματα σχετιζόμενα

Πίνακας 4.3: *Αποτελέσματα αποσαφήνισης για την έννοια "Client" στο έγγραφο* #67480 *του 20-NG*

Έννοιες "Client"	*SenseSim*
Client (computing)	0.0578
Client (ancient Rome)	0.0240
Client (band)	0.0170
Clients (album)	0.0168
Client (album)	0.0097

με το έγγραφο που περιγράφονται από τις εξισώσεις 4.7 έως 4.10 και ένα καθολικό γνώρισμα που περιγράφεται από την εξίσωση 4.11.

Για παράδειγμα, ένα μικρό τμήμα της αναπαράστασης με έννοιες του εγγράφου #67480 από το σύνολο 20-NG φαίνεται στον Πίνακα 4.4. Τα μέτρα $WFreq$, $OrderRank$, $LinkRank$, και $ConceptSim$ όλου του εγγράφου κανονικοποιούνται στο διάστημα $[0, 1]$, Στον Πίνακα αυτόν φαίνεται καθαρά η ικανότητα αναπαράστασης εννοιών όπως το "Word for Windows", το οποίο προφανώς αναφέρεται στο γνωστό κειμενογράφο, αλλά το οποίο το μοντέλο BOW θα χώριζε σε τρεις λέξεις οδηγώντας σε απώλεια της πραγματικής σημασίας των λέξεων.

Πίνακας 4.4: *Παράδειγμα αναπαράστασης εγγράφου με χρήση της Wikipedia*

Concept	$Wfreq$	*OrderRank*	*LinkRank*	*ConceptSim*	*Keyphraseness*
Hypertext Transfer Protocol	0.9333	0.8830	0.6628	0.0235	0.8932
Software versioning	0.9000	0.7395	0.6968	0.4718	0.8324
Software portability	0.9000	0.8601	0.7257	0.1549	0.8153
Web Ontology Language	0.9333	0.9572	0.7289	0.3967	0.7679
Application software	0.9000	0.9129	0.6494	0.1620	0.7290
Microsoft Access	0.6333	0.9466	0.9768	0.3685	0.7212
Network segment	0.3333	0.1055	0.7302	0.5041	0.7174
OpenLink ODBC Drivers	0.9333	0.8644	0.5957	0.2488	0.6479
File server	0.3333	0.1604	0.4529	1.0000	0.6338
ethernet	0.3333	0.2919	1.0000	0.9499	0.6320
Start menu	0.9667	0.7800	0.7365	0.2770	0.6065
World Wide Web	0.9333	0.9884	0.7781	0.0070	0.5956
btrieve	1.0000	0.8198	1.0000	0.2394	0.4237
borland	0.4000	0.9111	0.7772	0.3615	0.3343
xserver	0	0.2948	0.4432	0.2759	0.3077
source code	0.9667	0.8710	0.6031	0.1526	0.2689
codebase	1.0000	0.0000	0.3509	0.1315	0.2053
traffic flow	0.3333	0.4711	0.2958	0.7869	0.1045
Word for Windows	0.6667	0.4032	0.3576	0.7278	0.0833
mouse pointer	0.3333	0.3858	0.8342	0.7488	0.0460
Client (computing)	0.6667	0.4350	0.4246	0.6661	0.0426
Process (computing)	0	0.1647	0.4332	0.8365	0.0415
Extension (Mac OS)	0.9000	0.8163	0.6860	0.2700	0.0400
ftp	1.0000	0.6899	0.5943	0.2183	0.0349
Format (command)	0.9500	0.8812	0.3767	0.3615	0.0133
windows	1.0000	0.0000	0.7587	0.4460	0.0123
Function (computer science)	0.9000	0.5599	0.8021	0.3521	0.0116
Pascal (programming language)	0.4000	1.0000	0.9206	0.4484	0.0009
...	...	...	...	...	...

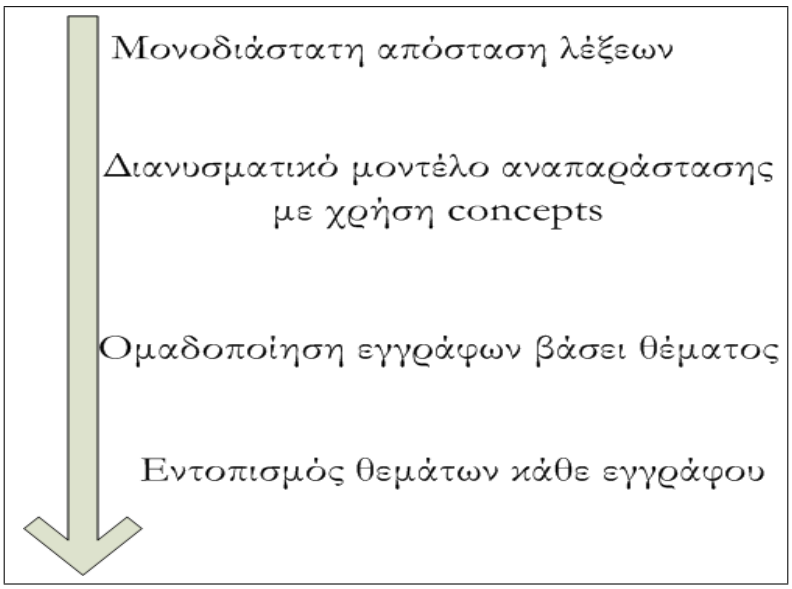
4.5 Από το μονοδιάστατο μέτρο απόστασης λέξεων στα πολυδιάστατα θέματα εγγράφων

Σημείο εκκίνησης της διατριβής ήταν η παρατήρηση πως τα σημασιολογικά προβλήματα της γλώσσας δυσχεραίνουν την αποδοτική αναπαράσταση και ανάλυση κειμενικών ζητημάτων στον υπολογιστή, προβλήματα τα οποία περιγράφηκαν αναλυτικά στο Κεφάλαιο 2. Με στόχο το σημασιολογικό εμπλουτισμό των εγγράφων ώστε να μπορέσουν να εφαρμοστούν ευφυείς τεχνικές επεξεργασίας και ανάλυσης εγγράφων, η πρώτη προσπάθεια που γίνεται και περιγράφηκε στο Κεφάλαιο 3 αξιοποίησε μεθόδους στατιστικής σημασιολογίας ώστε να δημιουργήσει ένα απλό βαθμωτό μέτρο απόστασης λέξεων. Οι λέξεις αποτελούν την πιο διαδεδομένη (και πιο φυσική ίσως για τον άνθρωπο) μονάδα αναπαράστασης εγγράφων, εξ ου και πολλές ιστοσελίδες αναζήτησης (π.χ. μηχανές αναζήτησης) βασίζονται σε αυτές (λέξεις-κλειδιά) για τη λειτουργία τους. Για την εύρεση συσχετίσεων ανάμεσα σε λέξεις, έγινε εισαγωγή εξωτερικής γνώσης (μηχανές αναζήτησης και WordNet), και σημαντικό της πλεονέκτημα είναι πως καταφέρνει να υπολογίσει μία αριθμητική τιμή ενδεικτική του πόσο σχετίζονται δύο λέξεις μεταξύ τους (βάσει όλων των δυνατών σχέσεων που μπορεί να έχουν). Ένα τέτοιο μέτρο μπορεί να αξιοποιηθεί γρήγορα και πρακτικά από πολλές εφαρμογές που ασχολούνται με τις συσχετίσεις λέξεων (αποσαφήνιση έννοιας λέξεων, διόρθωση συλλαβισμού, ερμηνεία συνθετικών ονομάτων κλπ). Όμως, τα προβλήματα από τη χρήση του μοντέλου BOW παραμένουν και έκαναν επιτακτική την ανάγκη επέκτασης του βαθμωτού μέτρου συσχέτισης λέξεων.

Προχωρώντας την ανάλυση παραπέρα, εύκολα παρατηρεί κανείς πως σε μεγάλες συλλογές εγγράφων ο αριθμός των διαφορετικών λέξεων (και εννοιών τους) αγγίζει μεγάλα νούμερα, επομένως η αναγωγή της εκτίμησης συσχέτισης εγγράφων σε εκτίμηση της συσχέτισης λέξεων οδηγεί μοιραία σε αύξηση της πολυπλοκότητας. Επιπροσθέτως, σε μεγάλες συλλογές εγγράφων εντοπίζονται όλο και πιο έντονα τα προβλήματα που περιγράφηκαν στα προηγούμενα Κεφάλαια (συνώνυμες/πολύσημες λέξεις, ομάδες λέξεων κλπ). Γίνεται φανερό πως η απλή προσέγγιση των λέξεων σε μεγάλα έγγραφα δε λειτουργεί ικανοποιητικά τόσο όσον αφορά στην αναπαράσταση, όσο και στη συσχέτιση εγγράφων. Χρειάζεται μια πλουσιότερη αναπαράσταση και καλύτερη δομή από την οργάνωση σε λέξεις και γιαυτό χρησιμοποιείται ένα επίπεδο παραπάνω, το επίπεδο των εννοιών. Η εισαγωγή του νέου μοντέλου αναπαράστασης, το οποίο βασίζεται όχι σε λέξεις αλλά σε έννοιες που προέρχονται από ονοματικές φράσεις εμπλουτίζοντας το έγγραφο με σημασιολογικό περιεχόμενο από τη Wikipedia. Έτσι, κάθε έγγραφο αναπαρίσταται από έννοιες οι οποίες δεν περιγράφονται μονοδιάστατα από κάποιο βάρος, αλλά υπάρχει ένα διάνυσμα που δείχνει το βάρος κάθε έννοιας για το έγγραφο και για τη συλλογή.

Το επόμενο βήμα στη διαδικασία της ανάλυσης είναι η εύρεση των σημασιολογικών σχέσεων ανάμεσα στα ίδια τα έγγραφα. Έχοντας ένα αρκετά λειτουργικό μοντέλο αναπαράστασης για τα έγγραφα, στο επόμενο Κεφάλαιο θα εξεταστούν δυνατότητες αλληλεπίδρασης εγγράφων ώστε να εντοπιστούν οι σχέσεις που υπάρχουν ανάμεσά τους και τρόποι ομαδοποίησής τους βάσει των θεμάτων που περιγράφουν. Στη διαδικασία αυτή καταλυτικό ρόλο παίζουν οι έννοιες οι οποίες εισάγονται με το μοντέλο αναπαράστασης που παρουσιάστηκε, καθώς εμπεριέχουν σημαντικό σημασιολογικό περιεχόμενο και τα χαρακτηριστικά που κατασκευάστηκαν (βλέπε Παράγραφο 4.4.1) καθορίζουν τόσο το βάρος της έννοιας στο έγγραφο, όσο και τη δυνατότητα της έννοιας να περιγράψει μια ομάδα εγγράφων. Στο Σχήμα 4.16, φαίνεται η πορεία αυτής της ανάλυσης από το

επίπεδο των λέξεων στο επίπεδο των εννοιών. Στο τελευταίο επίπεδο εντάσσεται ο εντοπισμός των διαφόρων θεμάτων που απαρτίζουν τα έγγραφα, μερικές μεθοδολογίες από τις οποίες παρουσιάζονται στην επόμενη Παράγραφο.



Σχήμα 4.16: *Πορεία ανάλυσης από το επίπεδο των λέξεων στο επίπεδο των εννοιών*

4.5.1 Εξαγωγή θέματος από έγγραφα

4.5.1.1 Χωρισμός κειμένου σε τμήματα

Στην εργασία των [Hearst, 1994] προτάθηκε μία απλή στατιστική μέθοδος χρησιμοποιώντας το μοντέλο BOW για την κατάτμηση κειμένου. Ο χωρισμός βασίζεται στη γεωμετρική απόσταση μεταξύ των διανυσματικών αναπαραστάσεων διαφόρων τμημάτων του κειμένου. Η συγκεκριμένη μέθοδος μπορεί να επεκταθεί προκειμένου να γίνει περισσότερο αποδοτική. Στην εργασία των [Ferret et al., 1998] προτάθηκε η επέκταση της έννοιας του υπολογισμού της απόστασης με τη χρήση εξωτερικής πηγής γνώσης. Ένα λεξιλογικό δίκτυο που περιλαμβάνει σχέσεις μεταξύ λέξεων μπορεί να χρησιμοποιηθεί για τον εμπλουτισμό του αρχικού διανύσματος του εγγράφου. Για παράδειγμα αν μία λέξη A υπάρχει σε ένα έγγραφο, τότε μία συγγενής λέξη B (που δεν υπάρχει στο έγγραφο) μπορεί να χρησιμοποιηθεί. Είναι σαφές πως μια τέτοια μέθοδος παρότι βελτιώνει την ακρίβεια της μεθόδου, εισάγει και την εξάρτηση από τη γλώσσα, το οποίο αντιμετωπίζεται στην εργασία των [Richmond & Smith, 1997] μέσα από την ανάπτυξη ενός μέτρου αποστάσεων ανεξαρτήτως γλώσσας. Η ιδέα είναι πως αφήνει έξω από την αναπαράσταση του εγγράφου λέξεις που δεν έχουν συγκεκριμένο περιεχόμενο.

Μία άλλη προσέγγιση χρησιμοποιεί λεξιλογικές αλυσίδες για την ανίχνευση ορίων στα έγγραφα [Manabu & Takeo, 1994]. Κατά τη διάσχιση του κειμένου δημιουργούνται αλυσίδες που περιέχουν γειτονικές λέξεις. Κάθε φορά που συναντάται μία λέξη, προστίθεται σε μία αλυσίδα ώστε να διατηρηθεί η μεγαλύτερη δυνατή συνοχή μεταξύ τους. Οι λεξιλογικές αλυσίδες είναι κατάλληλες για αυτή την εργασία διότι διατηρούν πληροφορίες για το πραγματικό περιεχόμενο και βοηθούν στην αποσαφήνιση όταν περισσότερες από μια πιθανότητες ανάθεσης υπάρχουν. Μετά την κατασκευή των αλυσίδων είναι εύκολο να εντοπιστούν τα όρια των διαφόρων τμημάτων: Σε κάθε θέση στο κείμενο, υπολογίζεται ο αριθμός των αλυσίδων που αρχίζουν και ο αριθμός των

αλυσίδων του τελειώνουν. Τα νούμερα αυτά δίνουν ένα μέτρο της πιθανότητας ενός διαχωρισμού στο συγκεκριμένο σημείο και βοηθούν στη σωστή επιλογή.

Όλες αυτές οι τεχνικές έχουν αποδειχθεί σχετικά αξιόπιστες αν και κατά περίπτωση κάποιες μπορεί να λειτουργούν πιο αποδοτικά.

4.5.1.2 Μέθοδοι εξαγωγής θέματος

Οι συνηθέστερες τεχνικές εξαγωγής θέματος από κείμενο ξεκινούν από την εξαγωγή σημαντικών λέξεων-κλειδιών (keywords). Μία κλασσική μέθοδος [Salton & Buckley, 1987] είναι η χρήση της εξίσωσης 2.3 ($tf - idf$) για την επιλογή των λέξεων που έχουν σημαντικό περιεχόμενο και έχουν μεγάλη σημασία για το κείμενο. Η απλή αυτή τεχνική έχει επεκταθεί με πολλούς τρόπους όπως περιγράφεται στις μεθόδους [Allan et al., 1998], [Nallapati, 2003] και [Ferret et al., 1998]. Η απλή εξαγωγή λέξεων μπορεί να είναι απαγορευτική για το χρήστη καθώς παρουσιάζει μόνο μερικές λέξεις (χωρίς άμεση συνεπαγωγή του θέματος του κειμένου) αλλά είναι ιδιαίτερα καλή τεχνική για την ευρετηρίαση (indexing) του εγγράφου.

Η χρήση λεξιλογικών αλυσίδων στα κείμενα βοηθά στην καλύτερη συνοχή λεξιλογικά και εννοιολογικά. Στην εργασία [Barzilay & Elhadad, 1997] οι αλυσίδες χρησιμοποιούνται για την εξαγωγή των σημαντικότερων προτάσεων, παρόλα αυτά η μέθοδος μπορεί να επεκταθεί κατάλληλα για την εξαγωγή λέξεων-κλειδιών που θα αποτελούσαν το θέμα του εγγράφου. Τέλος, πολλά υποσχόμενη κατεύθυνση και εδώ, είναι η εισαγωγή εξωτερικής γνώσης (όπως λεξιλογικά δίκτυα, οντολογίες) για τον υπολογισμό των ομοιοτήτων μεταξύ λέξεων.

4.5.1.3 Πιθανοτικά μοντέλα εξαγωγής θέματος

Στα πιθανοτικά μοντέλα, γίνεται η υπόθεση ύπαρξης ενός (πιθανοτικού) μοντέλου δημιουργίας εγγράφων και με χρήση του εκ των υστέρων πιθανοτικού συμπερασμού (posterior probabilistic inference) γίνεται προσπάθεια να βρεθούν οι παράμετροι αυτού του μοντέλου σε ήδη γνωστά (δημιουργημένα δηλαδή) έγγραφα. Παραδείγματα τέτοιων μοντέλων έχουν εφαρμογή σε διάφορους τομείς της μηχανικής μάθησης : Hidden Markov Models [Rabiner, 1989], φίλτρα Kalman [Kalman, 1960] , φυλλογενετικά δενδρικά μοντέλα [Mau et al., 1999], Mixture Models.

Το βασικό μοντέλο εξαγωγής θέματος είναι η Latent Dirichlet Allocation, LDA [Blei et al., 2003] και βασίζεται ακριβώς στην ιδέα πως τα έγγραφα αποτελούνται από λέξεις που προέρχονται από πολλά διαφορετικά θέματα, όπου κάθε θέμα (topic) ορίζεται ως μία κατανομή σε ένα συγκεκριμένο σύνολο όρων (το λεξιλόγιο της συλλογής εγγράφων). Συγκεκριμένα, θεωρείται πως έχουμε K θέματα που σχετίζονται με μία συλλογή εγγράφων και κάθε έγγραφο σχετίζεται με αυτά τα θέματα σε διαφορετικές αναλογίες. Η υπόθεση αυτή μπορεί να θεωρηθεί ασφαλής, καθώς τα περισσότερα έγγραφα τείνουν να είναι ετερογενή, συνδυάζοντας διάφορες ιδέες και σκέψεις που χαρακτηρίζουν ολόκληρη τη συλλογή εγγράφων που μας απασχολεί.

Για παράδειγμα έστω ένα έγγραφο που αναφέρεται σε *γενετική* και *τεχνολογία*, ένα άλλο που αναφέρεται σε *γενετική* και *νευροεπιστήμη* και ένα που αναφέρεται σε *τεχνολογία* και *νευροεπιστήμη*. Ένα μοντέλο που θα περιόριζε κάθε έγγραφο σε ένα θέμα δε θα μπορούσε να καλύψει με τον ίδιο τρόπο τη σημαντικότητα π.χ. της *νευροεπιστήμης* σαν θέμα, σε σχέση με ένα άλλο μοντέλο που καθορίζει θέματα ανά μέρος του εγγράφου. Φυσικά η πρόκληση στη συγκεκριμένη περίπτωση είναι πως τα θέματα αυτά δεν είναι εκ των προτέρων γνωστά και ο σκοπός είναι η εκμάθηση από τα δεδομένα. Η

LDA υλοποιεί στην ουσία αυτή τη "διαίσθηση" (ότι δηλαδή υπάρχουν διάφορα θέματα σε κάθε έγγραφο) σε ένα μοντέλο κρυφών μεταβλητών, στο οποίο τα γνωστά δεδομένα αλληλεπιδρούν με τις κρυφές τυχαίες μεταβλητές.

□

Κεφάλαιο 5

Ομαδοποίηση εγγράφων

5.1 Το ζήτημα της συσχέτισης και ομαδοποίησης εγγράφων

Το ζήτημα της σημασιολογικής ομοιότητας εγγράφων [Goldstone & Son, 2004] ασχολείται με την ανάθεση ενός μέτρου που καθορίζει πόσο όμοια είναι τα έγγραφα βάσει του περιεχομένου τους (που καθορίζεται από τους όρους που τα αποτελούν -συνηθέστερα τις λέξεις-). Αυτό επιτυγχάνεται με τον ορισμό μιας συνάρτησης τοπολογικής ομοιότητας, που κάνει χρήση είτε οντολογιών που επιτρέπουν τον υπολογισμό της απόστασης ανάμεσα σε λέξεις είτε στατιστικών μέσων όπως το μοντέλο VSM (που περιγράφηκε στο Κεφάλαιο 2.5.2) με στόχο τη συσχέτιση λέξεων βάσει της συνεμφάνισης. Αντί του μέτρου ομοιότητας εγγράφων μπορεί να χρησιμοποιηθεί (αντίστροφα) η έννοια της ανομοιότητας (ή της απόστασης) [Salton, 1989]

Οι αλγόριθμοι ομαδοποίησης (ή συσταδοποίησης) κειμένου τοποθετούν σε υποσύνολα (ή ομάδες) τα έγγραφα ενός συνόλου. Ο στόχος των αλγορίθμων είναι η δημιουργία ομάδων που αφενός να παρουσιάζουν (εσωτερική) συνοχή και αφετέρου να είναι διαφορετικές μεταξύ τους. Με άλλα λόγια, τα έγγραφα σε μία ομάδα πρέπει να είναι όσο το δυνατόν πιο όμοια μεταξύ τους και αντίστοιχα ένα έγγραφο μιας ομάδας πρέπει να διαφέρει σε σχέση με το έγγραφο μιας άλλης ομάδας.

Το χαρακτηριστικό της ομαδοποίησης είναι πως αποτελεί την πιο κλασσική μορφή *μη επιβλεπόμενης μάθησης*, δηλαδή δεν υπάρχει εκ των προτέρων γνώση για το που ανήκουν τα έγγραφα και πρέπει να ανακαλυφθεί από την παρατήρηση και την ανάλυση των εγγράφων. Αντίθετα, η κατηγοριοποίηση εγγράφων είναι μια περίπτωση *επιβλεπόμενης μάθησης*, η διαφορά τους δηλαδή έγκειται στο ότι στην κατηγοριοποίηση είναι γνωστές εκ των προτέρων οι κλάσεις στις οποίες είναι επιθυμητό να κατανεμηθούν τα έγγραφα (και επομένως το πρόβλημα είναι η πιστή αναπαράσταση μιας κατηγορικής διάκρισης που ένας άνθρωπος (ο "επιβλέπων" ή "δάσκαλος") επιβάλλει στα δεδομένα.

5.2 Υπάρχουσες τεχνικές ομαδοποίησης / κατηγοριοποίησης εγγράφων

Η συντριπτική πλειοψηφία των μεθόδων ομαδοποίησης εγγράφων χρησιμοποιούν το μοντέλο VSM (που περιγράφηκε αναλυτικά στο Κεφάλαιο 2.5.2) για να αναπαραστήσουν κάθε έγγραφο ως ένα διάνυσμα. Χρησιμοποιώντας τις εξισώσεις 2.1 έως 2.3 το σύνολο δεδομένων που είναι επιθυμητό να χωριστεί σε ομάδες αναπαρίσταται ως ένας

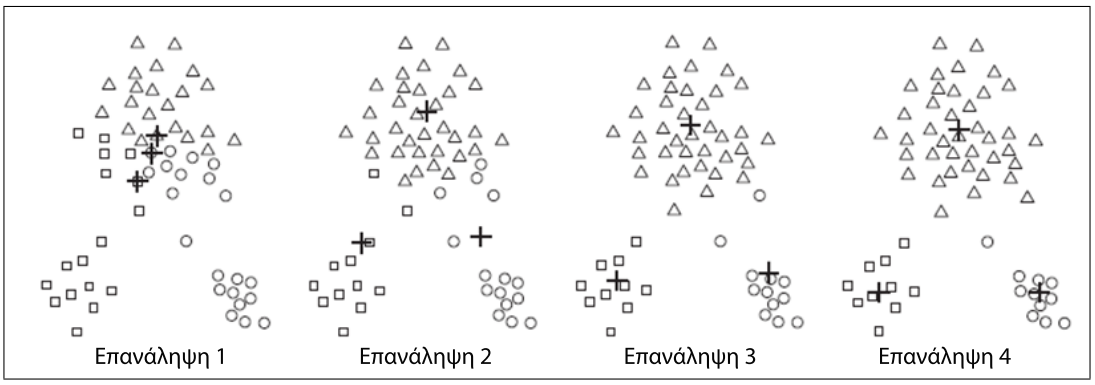
πίνακας $A : |D| \times |T|$ (πίνακας όρων-εγγράφων, term-by-document matrix) στον οποίο οι γραμμές απεικονίζουν τα διαφορετικά έγγραφα d και οι στήλες τους διαφορετικούς όρους t. Κάθε στοιχείο του πίνακα $a_{i,j}$ είναι η τιμή $tf - idf$ του όρου j στο έγγραφο i όπως προκύπτει από την εξίσωση 2.3. Το σημαντικό σε αυτή την αναπαράσταση είναι να εντοπιστούν οι γεωμετρικές σχέσεις μεταξύ των διανυσμάτων που συνθέτουν τον πίνακα ώστε να μοντελοποιηθούν οι ομοιότητες και διαφορές στο περιεχόμενο των εγγράφων.

5.2.1 Τεχνικές ομαδοποίησης που έχουν εφαρμοστεί σε έγγραφα

Οι περισσότερες μέθοδοι ομαδοποίησης εγγράφων βασίζονται σε υπάρχουσες μεθοδολογίες ομαδοποίησης δεδομένων, δεν έχουν αναπτυχθεί δηλαδή αμιγώς για έγγραφα, αλλά εφαρμόζονται και σε αυτά λόγω της διανυσματικής μορφής στην οποία αναπαρίστανται τα έγγραφα. Παρακάτω αναλύονται οι βασικοί αλγόριθμοι ομαδοποίησης (γενικά για οποιασδήποτε μορφής δεδομένα) που έχουν εφαρμοστεί στην περίπτωση των εγγράφων.

5.2.1.1 Ο αλγόριθμος των k-μέσων (k-means)

Η τεχνική του αλγορίθμου των k-μέσων (k-means) είναι αρκετά απλή. Αφού έχει προκύψει το σύνολο εγγράφων που θα χωριστεί σε ομάδες, επιλέγονται K διανύσματα που καλούνται κεντροειδή (βάσει του διανύσματος αναπαράστασης των εγγράφων), όπου το K είναι ο αριθμός των τελικών ομάδων που καθορίζεται από το χρήστη. Έπειτα, κάθε έγγραφο ανατίθεται στο πλησιέστερο κεντροειδές και η συλλογή εγγράφων που ανατίθεται σε κάθε κεντροειδές αποτελεί μία ομάδα. Το κεντροειδές κάθε ομάδας ανανεώνεται βάσει των εγγράφων που ανατέθηκαν στην ομάδα και η διαδικασία επαναλαμβάνεται μέχρι να μην υπάρχουν αλλαγές στα έγγραφα ή ισοδύναμα, τα κεντροειδή να παραμένουν τα ίδια. Ένα παράδειγμα φαίνεται στο Σχήμα 5.1.



Σχήμα 5.1: *Χρήση αλγορίθμου k-μέσων για τον εντοπισμό 3 ομάδων*

Ο αλγόριθμος συγκλίνει σχεδόν πάντα και μάλιστα από τα πρώτα βήματα, γιαυτό πολλές φορές η συνθήκη τερματισμού που αναφέρθηκε (δηλαδή να μην αλλάζουν τα κεντροειδή) αντικαθίσταται από κάποια άλλη πιο αδύναμη, ανάλογα και με το στόχο της ομαδοποίησης. Για την ανάθεση κάθε εγγράφου στο πλησιέστερο κεντροειδές χρειάζεται ένα μέτρο ομοιότητας (ή απόστασης). Μπορούν να χρησιμοποιηθούν διάφορες αποστάσεις όπως η Ευκλείδεια, η απόσταση Manhattan ή το μέτρο Jaccard. Ο επανυπολογισμός των κεντροειδών των ομάδων τίθεται γενικά αφού ανάλογα με το μέτρο

απόστασης που έχει επιλεγεί και με το στόχο της ομαδοποίησης. Ο στόχος της ομαδοποίησης συνήθως εκφράζεται μέσα από μία αντικειμενική συνάρτηση που εξαρτάται από τις αποστάσεις των εγγράφων μεταξύ τους ή από τα κεντροειδή (π.χ. ελαχιστοποίηση της απόστασης κάθε εγγράφου από το κεντροειδές της ομάδας του).

Από την περιγραφή του αλγορίθμου γίνεται σαφές πως μεγάλο ρόλο παίζει ο αρχικός καθορισμός των κεντροειδών: Αν επιλέγονται διαφορετικά αρχικά κεντροειδή με τυχαίο τρόπο, είναι πιθανό κάθε φορά να προκύπτουν διαφορετικά αποτελέσματα. Μία λύση στο πρόβλημα αυτό [Yuan et al., 2004] είναι να πραγματοποιηθούν διάφορες εκτελέσεις του αλγορίθμου κάτι το οποίο όμως είναι δαπανηρό από άποψη χρόνου. Μία αποδοτική τεχνική είναι η αρχική δειγματοληψία εγγράφων από τη συλλογή και η επιλογή των κεντροειδών από αυτά βάσει κάποιου πρόωρου διαχωρισμού τους. Η τακτική αυτή λειτουργεί ικανοποιητικά όταν (α) η συλλογή είναι σχετικά μικρή (μέχρι μερικές χιλιάδες έγγραφα) και (β) ο αριθμός των ομάδων K είναι μικρός σε σχέση με το μέγεθος της συλλογής.

Μία άλλη λύση στο πρόβλημα της αρχικής επιλογής δόθηκε με μία παραλλαγή του αλγορίθμου γνωστή ως διχοτόμος μέθοδος των k-μέσων (bisecting k-means) [Savaresi & Boley, 2001]. Η λογική και εδώ είναι απλή : Για τη δημιουργία των K επιθυμητών ομάδων, χωρίζονται αρχικά όλα τα έγγραφα της συλλογής σε δύο ομάδες, κατόπιν επιλέγεται μία νέα ομάδα και χωρίζεται αυτή. Η διαδικασία επαναλαμβάνεται μέχρι να παραχθούν K ομάδες. Υπάρχουν διάφοροι τρόποι για την επιλογή της ομάδας προς διαίρεση. Μπορεί να επιλέγεται η μεγαλύτερη ομάδα σε κάθε βήμα ή να χρησιμοποιείται κάποιο κριτήριο (με χρήση της αντικειμενικής συνάρτησης που έχει καθοριστεί στο πρόβλημα). Διαφορετικές επιλογές οδηγούν σε διαφορετικές ομάδες.

Οι αδυναμίες του αλγορίθμου συνοψίζονται στα ακόλουθα σημεία :

(1) Δε μπορεί να χειριστεί ομάδες διαφορετικών μεγεθών και πυκνοτήτων (αν και μπορεί να βρει ορθές υπο-ομάδες εφόσον καθοριστεί ένας μεγάλος αριθμός για το K),

(2) Παρουσιάζει πρόβλημα στο χειρισμό των εγγράφων που απέχουν πολύ (βάσει κάποιας απόστασης) από τα υπόλοιπα (συνήθως αυτά τα έγγραφα (outliers) πρέπει να εντοπιστούν και να απομακρυνθούν)

(3) Συνήθως είναι πιο αποδοτικός σε δεδομένα όπου υπάρχει πράγματι η έννοια του κέντρου (κεντροειδούς), δηλαδή αναπτύσσονται "σφαιρικά"

Οι απαιτήσεις του αλγορίθμου σε χώρο είναι μέτριες επειδή μόνο τα έγγραφα και τα κεντροειδή αποθηκεύονται ($O((m + K)n)$, όπου m είναι ο αριθμός των εγγράφων και n ο αριθμός των χαρακτηριστικών του διανύσματος αναπαράστασής τους). Οι απαιτήσεις του αλγορίθμου σε χρόνο είναι επίσης βασικά γραμμικές σε σχέση με τον αριθμό των εγγράφων. Πιο συγκεκριμένα, ο χρόνος που απαιτείται είναι της τάξης του $O(IKmn)$ όπου I είναι ο αριθμός των επαναλήψεων μέχρι τη σύγκλιση. Ο αριθμός των επαναλήψεων (όπως αναφέρθηκε) θεωρείται πως είναι μικρός καθώς οι περισσότερες αλλαγές συμβαίνουν στα πρώτα βήματα, επομένως ο αλγόριθμος είναι γραμμικός ως προς m με δεδομένο πως ο αριθμός των ομάδων K είναι σημαντικά μικρότερος από το m.

5.2.1.2 Ιεραρχική συσσωρευτική ομαδοποίηση (Hierarchical Agglomerative Clustering, HAC)

Η ιεραρχική συσσωρευτική ομαδοποίηση αρχίζει θεωρώντας όλα τα έγγραφα σαν ξεχωριστές ομάδες, και σε κάθε βήμα συνενώνει το ζευγάρι ομάδων που βρίσκονται πιο "κοντά" [Defays, 1977]. Προφανώς απαιτείται και πάλι κάποιο μέτρο απόστασης όχι μόνο εγγράφων αλλά ομάδων εν γένει. Η μέθοδος σταμάτα όταν προκύψει μία ομάδα (που περιέχει όλα τα έγγραφα). Η ομοιότητα μεταξύ ομάδων είναι βασικό σημείο του αλγορίθμου πάνω στο οποίο έχουν προταθεί διάφορα μέτρα όπως τα ακόλουθα:

(1) MIN : Λαμβάνει την ομοιότητα ομάδων ως την ομοιότητα μεταξύ των δύο εγγράφων των ομάδων που βρίσκονται πιο "κοντά" (δηλαδή έχουν τη μικρότερη απόσταση ή ισοδύναμα, τη μεγαλύτερη ομοιότητα),

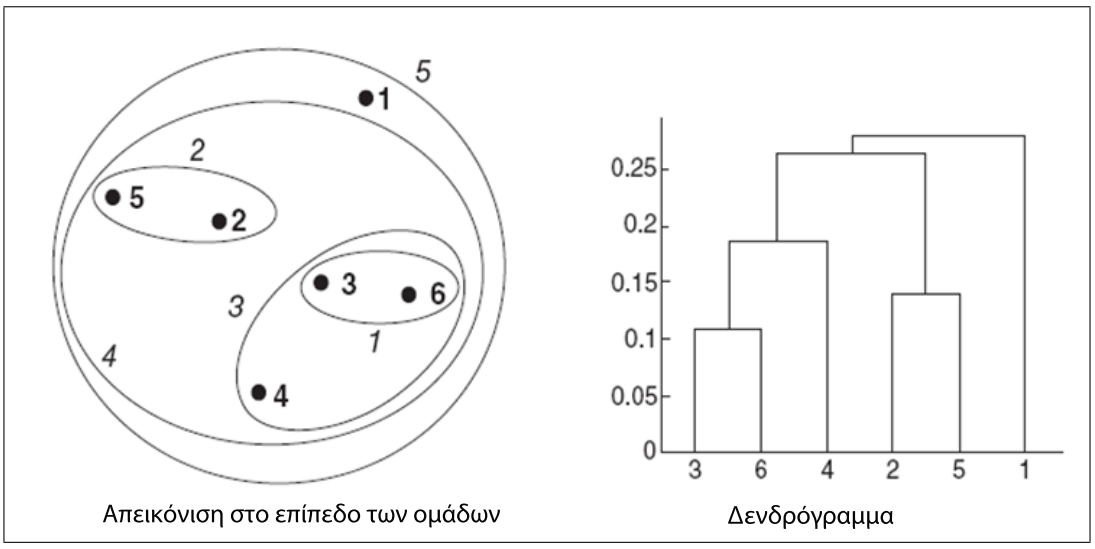
(2) MAX : Λαμβάνει την ομοιότητα ομάδων ως την ομοιότητα μεταξύ των δύο εγγράφων των ομάδων που βρίσκονται πιο "μακρυά",

(3) $GROUP$: Λαμβάνει το μέσο όρο των αποστάσεων όλων των εγγράφων των δύο ομάδων,

(4) $Ward$: Η ομοιότητα μεταξύ ομάδων λαμβάνεται ως η αύξηση του κόστους (σε όρους κάποιας αντικειμενικής συνάρτησης όπως το άθροισμα των τετραγώνων των αποστάσεων των εγγράφων από το κεντροειδές της ομάδας τους) από τη συνένωση των δύο ομάδων

Ο αλγόριθμος χρησιμοποιεί έναν πίνακα αποστάσεων ο οποίος απαιτεί αποθηκευτικό χώρο της τάξης του $\frac{1}{2}m^2$ (εφόσον θεωρείται συμμετρικός), όπου m είναι ο αριθμός των εγγράφων. Ο χώρος που χρειάζεται για την παρακολούθηση της εξέλιξης των ομάδων (έπειτα από τις συγχωνεύσεις) είναι $m-1$, επομένως η πολυπλοκότητα σε όρους χώρου είναι $O(m^2)$. Σε όρους χρόνου, ο αλγόριθμος καταρχάς απαιτεί χρόνο $O(m^2)$ για τον υπολογισμό του πίνακα αποστάσεων. Μετά από αυτό το βήμα απαιτούνται $m-1$ επαναλήψεις (επειδή αρχικά υπάρχουν m ομάδες και σε κάθε βήμα συγχωνεύονται 2). Εάν γίνεται γραμμική αναζήτηση στον πίνακα αποστάσεων τότε για την i επανάληψη το βήμα της συγχώνευσης των δύο ομάδων απαιτεί χρόνο $O((m-i+1)^2)$, που είναι ανάλογος με τον τρέχοντα αριθμό ομάδων ενώ το βήμα της ανανέωσης του πίνακα αποστάσεων απαιτεί χρόνο $O(m-i+1)$ καθώς μία συγχώνευση ομάδων επηρεάζει μόνο $m-i+1$ αποστάσεις. Χωρίς οποιαδήποτε βελτίωση, ο αλγόριθμος παρουσιάζει πολυπλοκότητα χρόνου $O(m^3)$, η οποία μπορεί να βελτιωθεί σε $O(m^2 log m)$ εφόσον χρησιμοποιείται μία ταξινομημένη λίστα (ή σωρός) για τον εντοπισμό των ομάδων που βρίσκονται πιο κοντά και επομένως θα συγχωνευθούν. Η πολυπλοκότητα του αλγορίθμου (σε χώρο και χρόνο) θέτει περιορισμούς για το μέγεθος των δεδομένων που μπορούν να χρησιμοποιηθούν, πόσο μάλλον για κειμενικά δεδομένα που οι διαστάσεις είναι εκ των πραγμάτων μεγάλες.

Παρόλα αυτά, το μεγάλο πλεονέκτημα των μεθόδων αυτών σε σχέση με τον αλγόριθμο των k-μέσων είναι η παραγωγή ενός δενδρογράμματος, το οποίο παρουσιάζει τόσο τις σχέσεις ομάδων-υποομάδων, όσο και τη σειρά με την οποία οι ομάδες ενώθηκαν. Ένα παράδειγμα φαίνεται στο Σχήμα 5.2. Η τιμή στον κάθετο άξονα του δενδρογράμματος αναπαριστά την τιμή απόστασης που προκύπτει από τον πίνακα αποστάσεων.



Σχήμα 5.2: *Παραγωγή δενδρογράμματος με χρήση ιεραρχικής συσσωρευτικής ομαδοποίησης*

Επίσης, οι αλγόριθμοι της κατηγορίας αυτής μπορούν να τροποποιηθούν ώστε να χειρίζονται ομάδες διαφορετικών μεγεθών. Υπάρχουν δύο προσεγγίσεις : η *μη-σταθμισμένη*, που αντιμετωπίζει όλες τις ομάδες ισότιμα και η *σταθμισμένη*, που λαμβάνει υπόψιν τον αριθμό των εγγράφων σε κάθε ομάδα (Οι όροι σταθμισμένος και μη-σταθμισμένος αναφέρονται στα έγγραφα και όχι στις ομάδες). Η παραλλαγή του σταθμισμένου αλγορίθμου (που χρησιμοποιεί την απόσταση $GROUP$ που αναφέρθηκαν πριν, στη βιβλιογραφία αναφέρεται ως Unweighted Pair Group Method using Arithmetic averages (UPGMA) [Murtagh, 1984].

Οι αδυναμίες του αλγορίθμου συνοψίζονται στα ακόλουθα σημεία :

(1) Έλλειψη μιας καθολικής αντικειμενικής συνάρτησης : Οι συσσωρευτικές μέθοδοι χρησιμοποιούν διάφορα τοπικά κριτήρια για να αποφασίσουν ποιες ομάδες θα συγχωνεύσουν δηλαδή δεν μπορεί να καθοριστεί ένας στόχος (π.χ. ελαχιστοποίηση της απόστασης των εγγράφων από τα κεντροειδή των ομάδων τους) ο οποίος θα βελτιστοποιηθεί (το πρόβλημα είναι υπολογιστικά αδύνατο),

(2) Παρότι ο αλγόριθμος αποφασίζει τοπικά για το ποιες ομάδες είναι καλύτερο να συγχωνευθούν, η απόφαση αυτή είναι τελική και δε μπορεί να αναστραφεί σε επόμενο βήμα. Αυτό εμποδίζει και ένα τοπικό κριτήριο βελτιστοποίησης να γίνει καθολικό,

(3) Είναι αρκετά απαιτητικός σε χώρο και χρόνο (όπως αναφέρθηκε ήδη).

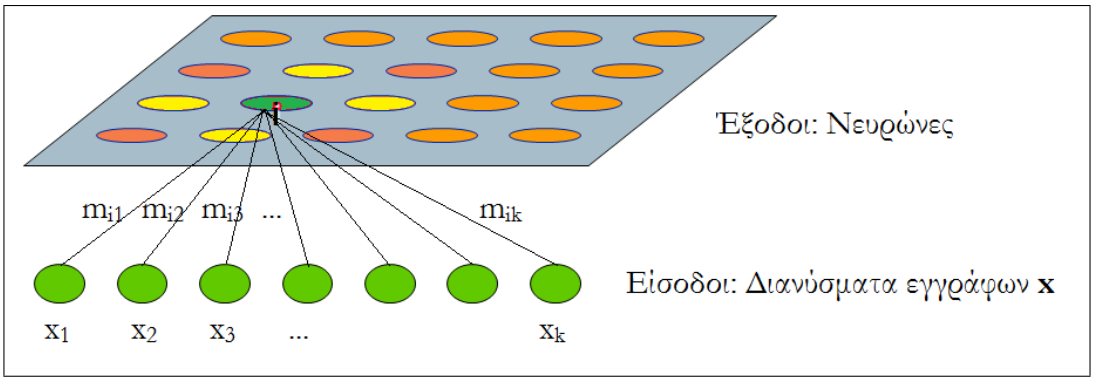
5.2.1.3 Χρήση αυτο-οργανούμενων χαρτών (Self-Organizing Maps) στην ομαδοποίηση εγγράφων

Μοντέλα νευρωνικών δικτύων που έχουν εφαρμοστεί σε ομαδοποίηση εγγράφων περιλαμβάνουν την ασαφή προσαρμοστική συντονισμένη θεωρία (fuzzy adaptive resonance theory, ART) [Carpenter et al., 1991] και τους αυτο-οργανούμενους χάρτες (Self-Organizing Maps, SOM) [Kohonen et al., 2001]. Τα δίκτυα ART είναι κατάλληλα για ομαδοποίηση καθώς προσαρμόζονται στην αντιμετώπιση προβλημάτων όπως τα μη-σταθερά δεδομένα (non stationary data) και η εμφάνιση καινούριων κλάσεων-ομάδων.

Παρόλα αυτά, η απόδοσή τους εξαρτάται από τη σειρά με την οποία εμφανίζονται τα πρότυπα και είναι ευαίσθητα σε υπερ-εκπαίδευση και θόρυβο [He et al., 2002].

Οι αυτο-οργανούμενοι χάρτες (SOM), όπως προτάθηκαν από τον [Kohonen, 1989] και περιγράφηκαν αναλυτικά στα [Kangas et al., 1990], [Kohonen et al., 2001] είναι μια κατηγορία τεχνητών νευρωνικών δικτύων που εκπαιδεύονται μέσω μη-επιβλεπόμενης μάθησης με στόχο την κατασκευή ενός χάρτη που θα αναπαριστά σε λίγες διαστάσεις (συνήθως δύο) το χώρο εισόδου των προτύπων (εγγράφων στην περίπτωσή που εξετάζεται εδώ) που παρουσιάζονται στην είσοδο. Οι αυτο-οργανούμενοι χάρτες διαφέρουν από άλλα νευρωνικά δίκτυα γιατί εισάγουν την έννοια της συνάρτησης γειτονιάς για να διατηρήσουν τις τοπολογικές ιδιότητες του χώρου εισόδου, δηλαδή τα όμοια πρότυπα να ομαδοποιηθούν το ένα κοντά στο άλλο. Επίσης, διαφέρουν από άλλες μεθόδους μείωσης της διάστασης των δεδομένων εισόδου (π.χ. SVD) διότι βασίζονται στην εκπαίδευση ενός νευρωνικού δικτύου ενώ οι άλλες μέθοδοι είναι αλγεβρικές.

Το SOM αποτελείται από έναν αριθμό υπολογιστικών στοιχείων (που καλούνται νευρώνες) και ανάλογα με την επιλεγμένη τοπολογία έχουν μία συγκεκριμένη θέση (πιο συχνές τοπολογίες είναι ένα δισδιάστατο τετραγωνικό ή εξαγωνικό πλέγμα). Επιπλέον, σε κάθε έναν από τους νευρώνες i ανατίθεται ένα διάνυσμα m_i. Είναι σημαντικό να αναφερθεί τα συγκεκριμένα διανύσματα έχουν την ίδια διάσταση με τα διανύσματα των εγγράφων που παρουσιάζονται στην είσοδο. Ένα παράδειγμα δομής SOM φαίνεται στο Σχήμα 5.3: Οι νευρώνες οργανώνονται σε ένα τετραγωνικό πλέγμα $5x4$ ενώ φαίνεται, η πλήρης διασύνδεση των διανυσμάτων εισόδου με τους νευρώνες της εξόδου και η έννοια της γειτονιάς (ο "πράσινος" νευρώνας έχει άμεσα γειτονικούς τους "κίτρινους", πιο μακρινούς τους "κόκκινους" και ακόμα πιο μακρινούς τους "πορτοκαλί").



Σχήμα 5.3: *Παράδειγμα αυτο-οργανούμενου χάρτη διάστασης* $5x4$

Η εκπαιδευτική διαδικασία του SOM περιλαμβάνει δύο φάσεις: την παρουσίαση των διανυσμάτων των εγγράφων (φάση ανταγωνισμού) και την προσαρμογή των διανυσμάτων των νευρώνων (φάση ανανέωσης). Κάθε επαναληπτικό βήμα της εκπαίδευσης (εποχή) ξεκινά με την τυχαία επιλογή ενός διανύσματος εγγράφου εισόδου $\mathbf{x}$. Αυτό το έγγραφο παρουσιάζεται στο SOM και κατόπιν καθορίζεται η ενεργοποίηση κάθε νευρώνα. Συνήθως, η ευκλείδεια απόσταση (ή οποιαδήποτε άλλη μετρική) ανάμεσα στο έγγραφο και στο διάνυσμα του νευρώνα χρησιμοποιείται για τον υπολογισμό της ενεργοποίησης κάθε νευρώνα. Έπειτα, ο νευρώνας που το διάνυσμά του έχει τη μικρότερη Ευκλείδεια απόσταση από το έγγραφο στην είσοδο καθορίζεται ως ο νευρώνας-νικητής για το πρότυπο αυτό. Αν χρησιμοποιηθεί ο δείκτης c για το συμβολισμό του νικητή και ο δείκτης t για το συμβολισμό της τρέχουσας εποχής έχουμε την εξής εξίσωση για τον καθορισμό του νευρώνα-νικητή :

$$c(t) = arg\min_i \{\|x(t) - m_i(t)\|\} \tag{5.1}$$

Στη φάση της ανανέωσης, το διάνυσμα του νευρώνα-νικητή, καθώς και τα διανύσματα των νευρώνων που βρίσκονται κοντά του ανανεώνονται. Η ανανέωση αυτή υλοποιείται ως σταδιακή μείωση της διαφοράς μεταξύ των αντίστοιχων συνιστωσών του διανύσματος του εγγράφου εισόδου και του διανύσματος του νευρώνα όπως φαίνεται στην παρακάτω εξίσωση:

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)] \tag{5.2}$$

Γεωμετρικά μιλώντας, τα διανύσματα των επηρεαζόμενων νευρώνων (m_i) "μετακινούνται" προς το διάνυσμα του εγγράφου εισόδου. Το μέγεθος αυτής της μετακίνησης καθορίζεται από έναν ρυθμό μάθησης ($\alpha(t)$), που φθίνει με το χρόνο. Ο αριθμός των νευρώνων που επηρεάζονται από την ανανέωση, καθώς και το μέγεθος αυτής της ανανέωσης βάσει της απόστασης από το νευρώνα-νικητή καθορίζεται από μία συνάρτηση γειτονιάς.

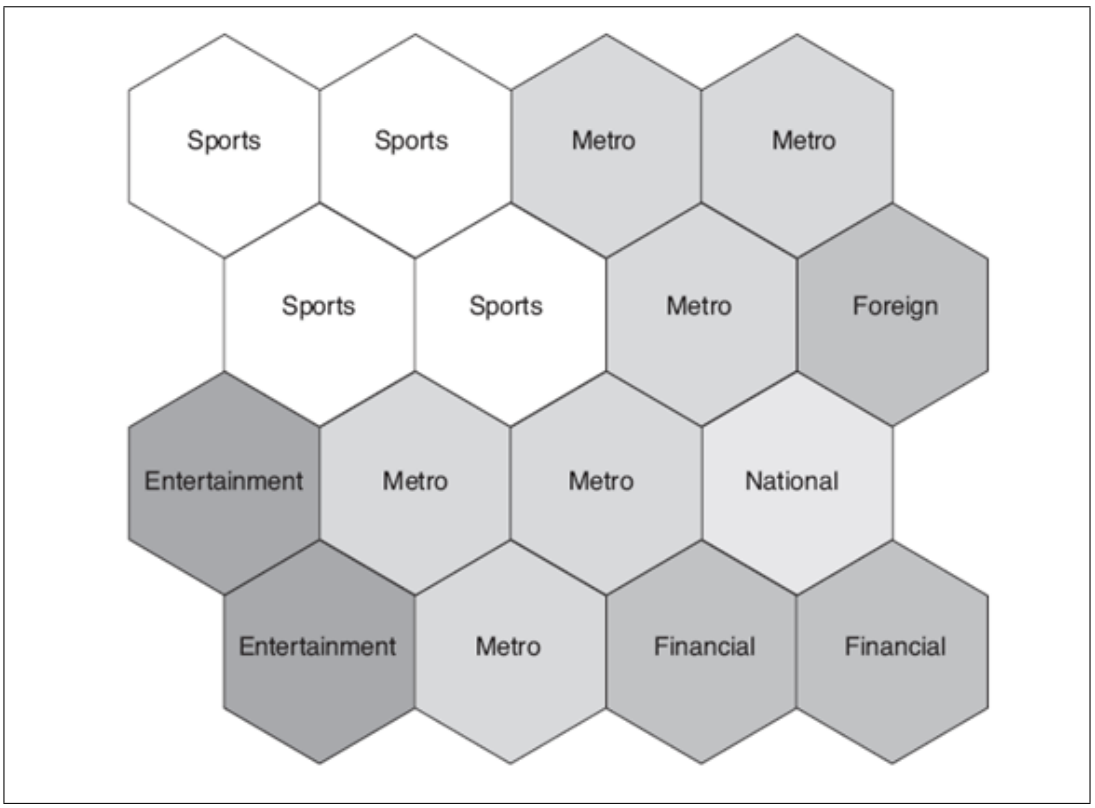
Συνήθως, η συνάρτηση γειτονιάς είναι μια συνάρτηση συμμετρική γύρω από το νικητή που φθίνει μονοτονικά καθώς αυξάνεται η απόσταση από το νικητή. Μια τυπική συνάρτηση γειτονιάς είναι η Γκαουσσιανή:

$$h_{ci}(t) = \exp\left(-\frac{\|r_c - r_i\|^2}{2 \cdot \sigma(t)^2}\right) \tag{5.3}$$

Σε αυτή την έκφραση, η ποσότητα $\|r_c - r_i\|$ δείχνει την απόσταση μεταξύ δύο νευρώνων c και i στο χώρο εξόδου με το r_i να αναπαριστά το διάνυσμα θέσης του νευρώνα στο χάρτη i στο πλέγμα νευρώνων στο χάρτη. Η παράμετρος σ που μεταβάλλεται με το χρόνο, φθίνει καθώς προχωρά η εκπαίδευση περιορίζοντας τη γειτονιά. Κοινή πρακτική είναι ο πυρήνας της γειτονιάς να επιλέγεται αρχικά αρκετά μεγάλος ώστε να καλύπτει μια μεγάλη περιοχή του χάρτη. Το χωρικό εύρος του πυρήνα αυτού περιορίζεται σταδιακά κατά τη διάρκεια της εκπαίδευσης ώστε προς το τέλος της μόνο το διάνυσμα του νευρώνα-νικητή να ανανεώνεται.

Οι μέθοδοι που βασίζονται στα SOM έχουν δύο σημαντικά πλεονεκτήματα έναντι άλλων μεθόδων. Πρώτον, πραγματοποιούν μη-γραμμική μείωση της διάστασης των δεδομένων, έτσι ώστε ο χώρος των προτύπων εισόδου να αποτυπώνεται σε ένα χώρο χαμηλότερων διαστάσεων στην έξοδο με ελάχιστη μείωση της πληροφορίας με σημαντικά μειωμένο φόρτο υπολογισμού σε σχέση με άλλες μεθόδους. Δεύτερον, η χωρική οργάνωση του χάρτη γνωρισμάτων επιτυγχάνεται μετά τη διαδικασία της μάθησης. Αυτό σημαίνει πως η δημιουργούμενη τοπολογία επιτρέπει στα παρόμοια έγγραφα ή θέματα να βρίσκονται κοντά στο χάρτη. Η έξοδος συνήθως αποτυπώνεται (μέσω κάποιας διαδικασίας οπτικοποίησης όπως περιγράφεται παρακάτω) με ένα δισδιάστατο χάρτη πάνω στον οποίο είναι εμφανείς οι διάφορες ομάδες, βοηθώντας έτσι την πλοήγηση και ανάκτηση των παρόμοιων εγγράφων. Τα SOM έχουν εφαρμοστεί με επιτυχία σε διάφορες εφαρμογές κειμένου [Merkl, 1998], [Lin et al., 1991], [Kohonen et al., 2000]. Στο Σχήμα 5.4 φαίνεται ένα παράδειγμα ομαδοποίησης σε ένα σύνολο δεδομένων κειμένου που περιέχει ειδήσεις από εφημερίδα. Με διαφορετική απόχρωση απεικονίζονται οι νευρώνες που ανήκουν σε διαφορετικές ομάδες.

Οι μεθοδολογίες που οπτικοποιούν το αποτέλεσμα του SOM (μετά τη διαδικασία της εκπαίδευσης), επιτρέποντας την καλύτερη εποπτεία του χάρτη που δημιουργείται



Σχήμα 5.4: *Παράδειγμα χρήσης SOM για ομαδοποίηση εγγράφων*

περιγράφονται αμέσως μετά. Μία γνωστή μεθοδολογία είναι ο Πίνακας Ενοποιημένων Αποστάσεων (Unified Distance Matrix, (U-matrix) [Ultsch & Siemon, 1990]. Η μεθοδολογία αυτή χρησιμοποιεί ένα χρωματικό κώδικα ώστε να επισημανθούν οι σχετικές αποστάσεις των διανυσμάτων γειτονικών νευρώνων. Επειδή ακριβώς οι ομάδες (clusters) είναι σύνολα διανυσμάτων που είναι κοντά μεταξύ τους, οι υψηλές τιμές στον πίνακα U-matrix δείχνουν ανομοιότητα μεταξύ των νευρώνων και επί της ουσίας δείχνουν τα σύνορα των ομάδων. Στο μοτίβο αυτό, υπάρχουν διάφορες παραλλαγές: Για παράδειγμα στο [Kraaijveld, 1992] ένα πλέγμα (πίνακας) χρησιμοποιείται για την αναπαράσταση του SOM, στο οποίο τα κελιά (στοιχεία) αναπαριστούν νευρώνες. Κάθε κελί χρωματίζεται ανάλογα με τη μέση απόσταση από τους γειτονικούς νευρώνες με το λευκό χρώμα να δείχνει τη μηδενική απόσταση και το μαύρο τη μεγαλύτερη δυνατή απόσταση.

Ένας διαφορετικός τρόπος οπτικοποίησης είναι να προβληθούν τα διανύσματα των νευρώνων σε έναν κατάλληλο μειωμένο σε διάσταση χώρο. Τέτοιες μη-γραμμικές μέθοδοι προβολής είναι για παράδειγμα οι: Multi-Dimensional Scaling (MDS) [Davison, 1983], ISOMAP [Tenenbaum et al., 2000], LLE [Roweis & Saul, 2000], Curvilinear Component Analysis (CCA) [Demartines & Herault, 1997] και η απεικόνιση Sammon [Sammon, 1969]. Τόσο η MDS όσο και η απεικόνιση Sammon στοχεύουν στη μείωση μιας τιμής σφάλματος ώστε μία συνάρτηση διαφορών μεταξύ των αρχικών αποστάσεων και των προβεβλημένων αποστάσεων να έχει τη μικρότερη δυνατή τιμή. Η μέθοδος ISOMAP [Tenenbaum et al., 2000] αρχικά σχεδιάστηκε ως γενίκευση της MDS, και υπολογίζει τις γεωδαιτικές αποστάσεις στον αρχικό χώρο και τις χρησιμοποιεί για την προβολή.

Μία άλλη μέθοδος οπτικοποίησης είναι να εμφανιστούν οι αποκρίσεις των διανυσμάτων των νευρώνων στα δεδομένα εισόδου. Ο συνήθης τρόπος να γίνει αυτό είναι

να επισημανθεί ο νευρώνας που ταιριάζει καλύτερα στο κάθε έγγραφο στο χάρτη. Για πολλά έγγραφα το αποτέλεσμα θα είναι ένα ιστόγραμμα με τους νευρώνες-νικητές που θα δείχνει την κατανομή στο χώρο εισόδου. Οι νευρώνες που βρίσκονται στα σύνορα των ομάδων συνήθως αντιστοιχίζονται σε πολύ λίγα έγγραφα, επομένως θα έχουν πολύ χαμηλές τιμές στο ιστόγραμμα. Τα συγκεκριμένα ιστογράμματα αποδίδουν γενικά ικανοποιητικά, αλλά λαμβάνουν υπόψιν τους μόνο το νευρώνα-νικητή για κάθε έγγραφο, ενώ στα περισσότερα πραγματικά προβλήματα (όπως αυτά της ομαδοποίησης εγγράφων) τα πρότυπα εισόδου αντιπροσωπεύονται καλύτερα από περισσότερους του ενός νευρώνες. Διάφορες παραλλαγές του ιστογράμματος αυτού έχουν προταθεί όπως το smoothed data histogram (SDH) [Pampalk et al., 2002], που μετρά τη σχετικότητα κάθε προτύπου (εγγράφου) με περισσότερους από έναν νευρώνες του χάρτη. Επί της ουσίας, η μέθοδος SDH επιλέγει και το νευρώνα-νικητή αλλά και μερικούς από τους λιγότερο κατάλληλους νευρώνες, ανάλογα με την κατάταξη των αποστάσεων μεταξύ του εγγράφου και των αντίστοιχων διανυσμάτων των νευρώνων.

Ένας τελείως διαφορετικός τρόπος προβολής των διανυσμάτων των νευρώνων είναι οι λεγόμενες "προσαρμοζόμενες συντεταγμένες " (adaptive coordinates) [Merkl & Rauber, 1997], που προτάθηκαν με στόχο την αντιστοίχηση της κίνησης των διανυσμάτων σε ένα δισδιάστατο χώρο κατά τη διάρκεια της εκπαίδευσης του SOM. Τέλος, το μοντέλο ViSOM (visualization-induced SOM) [Yin, 2002] χρησιμοποιείται για την ακριβή διατήρηση της πληροφορίας των αποστάσεων μαζί με την τοπολογία στο χάρτη.

Για να επιτευχθούν πιο αποδοτικά και διαφορετικά επίπεδα λεπτομέρειας, έχει προταθεί και το ιεραρχικό μοντέλο SOM [Miikkulainen, 1990], το οποίο επίσης χρησιμεύει ώστε να γίνονται λιγότεροι υπολογισμοί. Παρόλα αυτά, τα μεγέθη των χαρτών σε όλα τα επίπεδα πρέπει να καθοριστούν και επομένως και ο αριθμός των επιθυμητών ομάδων. Αυτό μπορεί να οδηγήσει είτε σε μικρούς χάρτες (μη δυνατότητα διαχωρισμού προτύπων), είτε σε μεγάλους χάρτες (με υποχρησιμοποιούμενους πόρους). Για το λόγο αυτό, έχουν δημιουργηθεί παραλλαγές όπως τα SOM μεταβλητού μεγέθους που συνδυάζονται με ιεραρχικές δομές όπως ο Μεταβλητός Ιεραρχικός Αυτο-Οργανούμενος Χάρτης (Growing Hierarchical SOM, (GH-SOM)) [Rauber et al., 2002] για την παραγωγή καλών αποτελεσμάτων με δυναμικά μεγέθη χάρτη. Και πάλι όμως, το πόσο θα μεγαλώσει ο χάρτης εξαρτάται από ευαίσθητες παραμέτρους που καθορίζονται από την αρχή της διαδικασίας.

Οι μέθοδοι οπτικοποίησης και εξερεύνησης των μεθόδων SOM θέτουν σαν στόχο και την αυτόματη οριοθέτηση και περιγραφή των περιοχών ενδιαφέροντος (επί του προκειμένου των ομάδων που δημιουργούνται). Τα SOM πολλαπλών στρωμάτων χρησιμοποιούνται για την οργάνωση εγγράφων (όπως το ET-MAP ([Chen et al., 1996]), στο οποίο σε κάθε κόμβο ανατίθεται μια ετικέτα που προσδιορίζεται από τους όρους που ταιριάζουν καλύτερα στον κόμβο και κατόπιν παρόμοιοι κόμβοι συγχωνεύονται για το σχηματισμό περιοχών στο χάρτη. Η μεθοδολογία αυτή εφαρμόστηκε για τη βελτίωση της κατηγοριοποίησης και αναζήτησης στο WWW μέσω της ανάλυσης λέξεων-κλειδιών και περιγραφητών (descriptors) των ιστοσελίδων με στόχο την αναπαράσταση (μέσω του χάρτη του SOM) βασικών θεματικών κατηγοριών. Η κύρια αδυναμία του μοντέλου είναι πως στην ιεραρχία που δημιουργείται (με τους αντίστοιχους όρους) παρατηρούνται προβλήματα που έχουν να κάνουν με συσχετίσεις όρων (προβλήματα συνωνυμίας κτλ) κάτι το οποίο δυσχεραίνει την πλοήγηση (δεδομένου πως πρόκειται και για ένα πολύπλοκο πολυεπίπεδο χάρτη).

Η προσέγγιση αυτή είναι παρόμοια με αυτή που υιοθετείται στο LABELSOM [Rauber, 1999], το οποίο χρησιμοποιεί το σφάλμα κβαντοποίησης για τον καθορισμό των

καλύτερων όρων για την ονοματοδοσία των κόμβων χωρίς να γίνονται συγχωνεύσεις μεταξύ τους. Το μειονέκτημα του LABELSOM είναι πως οι όροι που χρησιμεύουν ως ετικέτες δε μπορούν να θεωρηθούν ακριβείς περιγραφές των ομάδων που δημιουργούνται αλλά θεωρούνται περισσότερο ως τα σημαντικότερα χαρακτηριστικά κάθε νευρώνα.

Το WEBSOM χρησιμοποιεί ένα μέτρο βασισμένο στη συχνότητα των όρων για τον καθορισμό των ονομάτων των κόμβων [Lagus et al., 2004]. Η μέθοδος αυτή λαμβάνει υπόψιν τις συχνότητες των πιο ικανών για διαχωρισμό από τους υπόλοιπους όρους σε μια ομάδα. Το WEBSOM χρησιμοποιεί περισσότερους κόμβους από άλλες μεθόδους που δεν είναι όλοι ονοματισμένοι, αλλά υπάρχει μία σταθερή ακτίνα ελαχίστων αποστάσεων μεταξύ ονομάτων των κόμβων. Στο χάρτη, οι πυκνότητες των ομάδων χρωματίζονται και ομαλοποιούνται ώστε το αποτέλεσμα να είναι αισθητικά πιο καλό από τις μεθόδους των ET-MAP και GH-SOM. Παρόλα αυτά, ένα σημαντικό μειονέκτημα του WEBSOM είναι πως για να υπάρχουν ικανοποιητικά αποτελέσματα απαιτείται ένα αρκετά μεγάλο λεξιλόγιο (δυσανάλογα μεγάλο με τη συλλογή εγγράφων) και άλλες δομικές αδυναμίες που έχουν αναφερθεί [Georgakis et al., 2001].

Παρά τον όγκο εργασιών που έχει δημοσιευτεί, υπάρχουν ανοιχτά ερευνητικά ζητήματα που αφορούν στον αλγόριθμο του SOM. Πρώτον, το SOM χρησιμοποιεί μία δεδομένη αρχιτεκτονική που βασίζεται στον εκ των προτέρων καθορισμό της αρχιτεκτονικής (αριθμός και τοποθέτηση των νευρώνων-κόμβων). Προφανώς, ο καθορισμός της βέλτιστης αρχιτεκτονικής του δικτύου παραμένει ένα πρόβλημα καθόλου τετριμμένο. Δεύτερον, οι ιεραρχικές σχέσεις μεταξύ των δεδομένων εισόδου δεν αντικατοπτρίζονται άμεσα στους χάρτες εξόδου και συνήθως αναπαρίστανται όλοι οι νευρώνες στο ίδιο επίπεδο.

5.2.1.4 Άλλοι αλγόριθμοι που έχουν εφαρμοστεί σε ομαδοποίηση εγγράφων

Μία προσπάθεια για ομαδοποίηση εγγράφων με αλγόριθμο που βασιζόταν σε συχνά εμφανιζόμενες ομάδες λέξεων έγινε από τους [Fung et al., 2003]. Η μεθοδολογία ομαδοποίησης δε βασίζεται σε κάποιον υπάρχοντα αλγόριθμο, είναι αρκετά γρήγορη γιατί λαμβάνει υπόψιν μόνο τις συχνές εμφανίσεις λέξεων, οδηγεί σε ιεραρχική ομαδοποίηση, αλλά μια βασική της αδυναμία είναι πως ενδεχόμενα χρειάζονται αρκετά διαθέσιμα έγγραφα και ότι δε χρησιμοποιείται κάποια εξωτερική πηγή γνώσης για εμπλουτισμό των εγγράφων ή ανάθεση βαρών σε κάθε λέξη.

Επίσης, στην ομαδοποίηση εγγράφων έχει εφαρμοστεί και η στατιστική ανάλυση. Μέθοδοι της θεωρίας πληροφορίας [Slonim et al., 2002] και πιθανοτικές προσεγγίσεις [Liu et al., 2002], [Hofmann, 1999a] έχουν χρησιμοποιηθεί για να καθοριστεί ο κατάλληλος αριθμός ομάδων και έχουν παρουσιαστεί υποσχόμενα αποτελέσματα. Το κύριο μειονέκτημα των συγκεκριμένων μεθόδων είναι πως συχνά υποθέτουν μία συγκεκριμένη κατανομή λέξεων ή ένα συγκεκριμένο μοντέλο για κάθε ομάδα και ενδέχεται να χρειαστούν πολλές επαναλήψεις για να επιτευχθεί ένα σταθερό αποτέλεσμα. Ακόμη, σε άλλες τεχνικές έχουν χρησιμοποιηθεί πιθανοτικές έννοιες αντί των λογικών αναπαραστάσεων [Talavera & Bejar, 2001].

Τέλος, άλλες μέθοδοι περιλαμβάνουν: συλλογή δεδομένων σε υπο-ομάδες και κατόπιν χρήση Γκαουσσιανών μειγμάτων ("gaussian mixtures") για τις υπο-ομάδες [Jin et al., 2005], δημιουργία μιας σταθερής βάσης για την εφαρμογή επαγωγικού λογικού προγραμματισμού [Junker et al., 2000] και αυτόματη εξαγωγή κανόνων για ανάκτηση κειμένου [Soderland, 1999].

5.2.1.5 Αλγόριθμοι ομαδοποίησης εγγράφων με χρήση εξωτερικής γνώσης

Έχουν χρησιμοποιηθεί αρκετές εξωτερικές πηγές γνώσης (όπως το WordNet και η Wikipedia που αναφέρθηκαν σε προηγούμενα Κεφάλαια) για τη βελτίωση των αποτελεσμάτων ομαδοποίησης εγγράφων. Εδώ συγκεντρώνονται οι κυριότερες μεθοδολογίες που χρησιμοποιούν εξωτερική γνώση και αναφέρονται τα αποτελέσματα που έχουν στη βελτίωση του αποτελέσματος ομαδοποίησης εγγράφων. Αξίζει να σημειωθεί πως οι αλγόριθμοι αυτοί ενισχύουν την αναπαράσταση των εγγράφων και δεν επηρεάζουν τον αλγόριθμο της ομαδοποίησης (που παραμένει αμετάβλητος π.χ. k-means).

Η πλούσια ιεραρχική δομή του WordNet χρησιμοποιήθηκε για να πραγματοποιηθεί αποσαφήνιση των εννοιών βασισμένη στη σύνταξη αναθέτοντας σε κάθε λέξη ένα μέρος του λόγου (part-of-speech, POS) και εμπλουτίζοντας την διανυσματική αναπαράσταση (BOW) με συνώνυμα και υπερώνυμα. Παρόλα αυτά, ο θόρυβος που εισήχθη από μη-σωστές έννοιες που επιλέγονταν από το WordNet φαίνεται πως λειτούργησε ως στενωπός για το σωστό εμπλουτισμό του εγγράφου [Sedding & Kazakov, 2004]. Στην εργασία [Hotho et al., 2003] υπήρξε σημαντική βελτίωση των αποτελεσμάτων ομαδοποίησης με χρησιμοποίηση της των σχέσεων υπερωνυμίας του WordNet και την αποσαφήνιση της έννοιας των λέξεων σε συνδυασμό με τη στάθμιση βαρών.

Στις εργασίες [Gabrilovich & Markovitch, 2006], [Gabrilovich & Markovitch, 2007] προτείνεται μία μέθοδος ώστε να βελτιωθεί η απόδοση της κατηγοριοποίησης εγγράφων με εμπλουτισμό της αναπαράστασης εγγράφων από τη Wikipedia. Η αντιστοίχηση μεταξύ κάθε εγγράφου και των άρθρων της Wikipedia γίνεται μέσα από μία γεννήτρια παραγωγής γνωρισμάτων, η οποία λειτουργεί ως μία μηχανή ανάκτησης κειμένου: Λαμβάνει σαν είσοδο ένα απόσπασμα κειμένου, το οποίο μπορεί να είναι λέξη, λέξεις, πρόταση, ολόκληρη παράγραφος ή και ολόκληρο το έγγραφο και δίνει σαν έξοδο τα πιο σχετικά άρθρα της Wikipedia. Οι τίτλοι των άρθρων που επιστρέφονται φιλτράρονται και αυτοί που ξεχωρίζονται ως έχοντες υψηλή αναγνωριστική ικανότητα χρησιμοποιούνται ως επιπλέον γνωρίσματα για τον εμπλουτισμό της αναπαράστασης του εξεταζόμενου εγγράφου. Η εμπειρική επαλήθευση της μεθόδου έδειξε πως μπορεί να βελτιώσει σημαντικά την απόδοση της κατηγοριοποίησης.

Η Wikipedia έχει εφαρμοστεί επίσης για ομαδοποίηση εγγράφων. Στην εργασία [Banerjee et al., 2007] χρησιμοποιείται μία μέθοδος παρόμοια με αυτή των [Gabrilovich & Markovitch, 2006] για την ομαδοποίηση μικρών κειμένων. Η διαφορά της μεθόδου συνίσταται στο ότι χρησιμοποιούνται συμβολοσειρές αναζήτησης που δημιουργούνται από κείμενο του εγγράφου για την ανάκτηση των σχετικών άρθρων από τη Wikipedia. Οι τίτλοι των πιο δημοφιλών άρθρων που επιστρέφονται από τη Wikipedia χρησιμεύουν ως επιπλέον γνωρίσματα για την ομαδοποίηση ειδήσεων της Google. Αξίζει να σημειωθεί πως και οι δύο προαναφερθέντες μέθοδοι ([Banerjee et al., 2007] και [Gabrilovich & Markovitch, 2006]) μόνο ενισχύουν την αναπαράσταση των εγγράφων με άρθρα της Wikipedia χωρίς να λαμβάνουν υπόψιν την ιεραρχική δομή της Wikipedia ή άλλα χαρακτηριστικά που μπορούν να εξαχθούν.

Η δομή κατηγοριών της Wikipedia χρησιμοποιήθηκε στο [Wang & Domeniconi, 2008] για κατηγοριοποίηση εγγράφων και στο [Hu et al., 2008] για ομαδοποίηση εγγράφων. Αυτές οι μέθοδοι αφού εντοπίσουν άρθρα της Wikipedia σε κάθε έγγραφο, επεκτείνουν την αναζήτηση και σε παρόμοια άρθρα Wikipedia θεωρώντας συνώνυμες λέξεις και σχετικά άρθρα βάσει των συνδέσμων ανακατεύθυνσης ("redirect links") και υπερσυνδέσμων ("hyperlinks") της Wikipedia.

Στην εργασία [Huang et al., 2009] αρχικά εντοπίζονται οι όροι και οι φράσεις

της Wikipedia που υπάρχουν σε κάθε έγγραφο (και καλούνται *έννοιες* ή concepts). Κατόπιν αναπτύχθηκε ένα μέτρο ομοιότητας που αξιολογεί τη σημασιολογική σχέση ανάμεσα στα σύνολα εννοιών δύο εγγράφων και βάσει της θεώρησης πως τα έγγραφα δε συνδέονται μόνο μέσα από την από κοινού ύπαρξη σε αυτά των εννοιών αλλά και λόγω της επιμέρους συσχέτισης αυτών, βρίσκεται η σχέση ανάμεσα σε δύο οποιαδήποτε έγγραφα. Το μέτρο ομοιότητας που αναπτύχθηκε λαμβάνει υπόψιν πληροφορίες από τη Wikipedia όπως το περιεχόμενο και οι σύνδεσμοι που εξέρχονται από κάθε άρθρο.

Στην εργασία [Kiran, 2010] η αναπαράσταση εγγράφων εμπλουτίζεται με βάση αρκετές πηγές γνώσης (Wikipedia, dmoz, social bookmarks) και κατόπιν εξάγονται θέματα (ομάδες) από τα έγγραφα με διάφορες τεχνικές ανίχνευσης θεμάτων (π.χ. Latent Dirchlet Allocation (LDA) [Blei et al., 2003], Latent Semantic Analysis (LSA) [Hofmann, 1999b]).

Οι γενικές τεχνικές ομαδοποίησης έχουν το μειονέκτημα πως όταν εφαρμόζονται σε κειμενικά δεδομένα, δεν παρέχουν άμεσες περιγραφές των ομάδων που δημιουργούνται λόγω του ότι οι αλγόριθμοι αυτοί δεν σχεδιάστηκαν ειδικά για κείμενα. Οι περιγραφές των δημιουργούμενων ομάδων αποτελεί σημαντικό κομμάτι της ομαδοποίησης εγγράφων, καθώς παρέχει πληροφορία για το χωρισμό της συλλογής εγγράφων που γίνεται και πραγματοποιείται έτσι εξαγωγή θέματος από μια συλλογή εγγράφων (πρβλ και μεθοδολογίες της Παραγράφου 4.5.1). Παρόλα αυτά, οι υπάρχουσες τεχνικές ομαδοποίησης με αυτόματη εξαγωγή πληροφορίας για τις ομάδες ("conceptual clustering") είναι είτε ιδιαίτερα αργές [Stumme et al., 2002], είτε απαιτούν επιπλέον βήματα για τη μείωση των αριθμών των ομάδων [Hotho & Stumme, 2002]. Στις επόμενες Παραγράφους προτείνονται και αξιολογούνται δύο μεθοδολογίες ομαδοποίησης εγγράφων που βασίζονται στο μοντέλο αναπαράστασης που παρουσιάστηκε στο προηγούμενο Κεφάλαιο.

5.3 Ιεραρχική Ομαδοποίηση εγγράφων με χρήση των συχνών και σημαντικών εννοιών (Conceptual Hierarchical Clustering, CHC)

Για τη συνέχεια, είναι απαραίτητοι οι ακόλουθοι ορισμοί που βασίζονται στο μοντέλο εννοιών (και τα χαρακτηριστικά του) όπως παρουσιάστηκαν στο προηγούμενο Κεφάλαιο:
(α) Μια *καθολικά σημαντική έννοια* είναι μία έννοια η οποία :

- έχει τιμή *Keyphraseness* μεγαλύτερη από ένα συγκεκριμένο κατώφλι, που ορίζεται ως η ελάχιστη τιμή *Keyphraseness* (*minimum keyphraseness threshold* ή $MinKeyph$), και
- εμφανίζεται σε περισσότερα από ένα ποσοστό των εγγράφων της συλλογής, που ορίζεται ως η ελάχιστη τιμή συχνότητας ένα ελάχιστο (*minimum global frequency threshold* ή $MinFreq$).

Ένα *καθολικά σημαντικό k-σύνολο-εννοιών* είναι ένα σύνολο από k καθολικές σημαντικές έννοιες οι οποίες εμφανίζονται μαζί σε ένα ποσοστό των εγγράφων της συλλογής μεγαλύτερο από το $MinFreq$.
(β) Μία καθολικά σημαντική έννοια είναι *συχνή στην ομάδα* C_m, αν η έννοια περιέχεται σε ένα ελάχιστο ποσοστό εγγράφων της ομάδας που καλείται *ελάχιστη υποστήριξη*

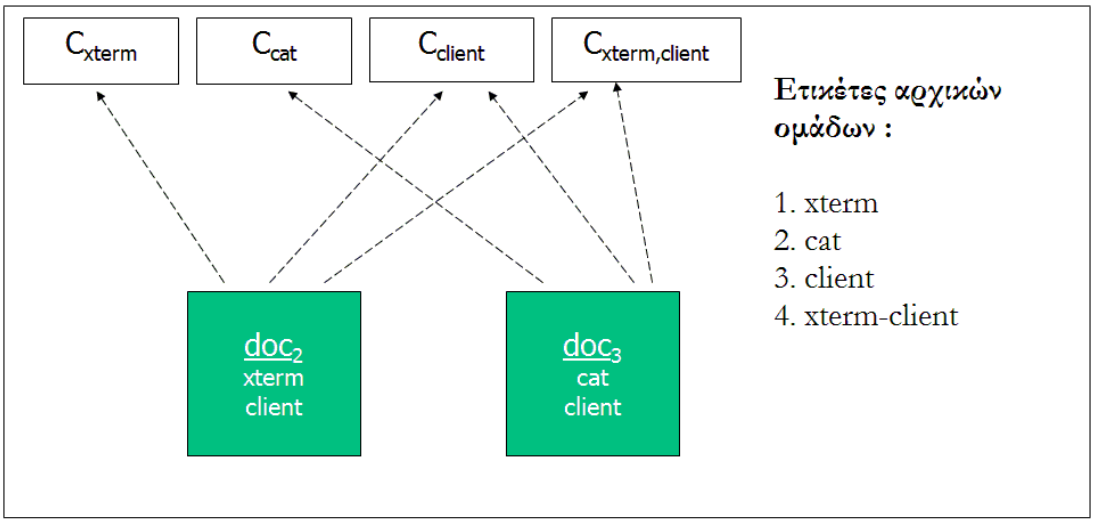
ομάδας.
(γ) Η *υποστήριξη ομάδας (cluster support)* για μία έννοια σε μια ομάδα C_m είναι το ποσοστό των εγγράφων της ομάδας C_mπου περιέχουν τη συγκεκριμένη έννοια.

Για να βρεθεί το συνολικό βάρος κάθε έννοιας σε κάθε έγγραφο γίνεται γραμμικός συνδυασμός των γνωρισμάτων που περιγράφονται από τις εξισώσεις 4.7 έως 4.10. Το τελικό βάρος της έννοιας i στο έγγραφο j δίνεται από την ακόλουθη εξίσωση:

$$\begin{aligned} Weight(j,i) =& \alpha * WFreq_{j,i} + +\beta * LinkRank_{j,i} + \\ &+ \gamma * OrderRank_{j,i} + (1-\alpha-\beta-\gamma) * ConceptSim_{j,i} \end{aligned} \tag{5.4}$$

Οι παράμετροι α, β και γ καθορίζονται πειραματικά και το διάστημα τιμών τους είναι το $[0,1]$. Με αυτόν τον τρόπο αντικαθίσταται το αραιό μοντέλο BOW από ένα πιο συμπιεσμένο και πλούσιο σε περιεχόμενο μοντέλο εννοιών, το οποίο αφενός μειώνει το μέγεθος του διανύσματος αναπαράστασης (σημαντικός παράγοντας όταν η επεξεργασία αφορά μεγάλο αριθμό εγγράφων) και εμπλουτίζει τα γνωρίσματα του εγγράφου με εξωτερική γνώση από τη Wikipedia.

Δεδομένων των ορισμών (α) έως (γ), κατασκευάζονται οι αρχικές ομάδες της συλλογής. Για κάθε σύνολο εννοιών που ανταποκρίνεται στους περιορισμούς του ορισμού (α), κατασκευάζεται μία ομάδα που αποτελείται από όλα τα έγγραφα που περιέχουν αυτό το σύνολο εννοιών. Είναι φανερό πως σε αυτή τη φάση, ένα έγγραφο μπορεί να ανήκει σε περισσότερες της μιας ομάδας. Στο επόμενο βήμα γίνεται η αποσύνδεση εγγράφων/ομάδων καθορίζοντας την πιο κατάλληλη ομάδα για κάθε έγγραφο: Η *ετικέτα της κάθε ομάδας* καθορίζεται από το καθολικά σημαντικό σύνολο εννοιών που περιλαμβάνεται σε όλα τα έγγραφα που ανατίθενται στη συγκεκριμένη ομάδα. Για παράδειγμα, στο Σχήμα 5.5 φαίνονται δύο έγγραφα που περιέχουν ως *καθολικά σημαντικά* σύνολα εννοιών τα xterm, client, cat και το συνδυασμό xterm-client. Έτσι δημιουργούνται 4 αρχικές ομάδες (αντίστοιχες των 4 συνόλων εννοιών), στις οποίες αρχικά τοποθετείται κάθε έγγραφο το οποίο περιέχει έστω και μία έννοια από την ετικέτα της κάθε ομάδας (π.χ. το έγγραφο 2 ανατίθεται σε 3 ομάδες: xterm, client, xterm-client.



Σχήμα 5.5: *Παράδειγμα δημιουργίας αρχικών ομάδων της συλλογής εγγράφων με τη μέθοδο CHC*

5.3.1 Διαχωρισμός αρχικών ομάδων

Για κάθε έγγραφο, καθορίζεται η πιο κατάλληλη αρχική ομάδα και το έγγραφο πλέον ανατίθεται μόνο σε αυτήν. Για το σκοπό αυτό, εισάγεται ένα μέτρο του πόσο "κατάλληλη" είναι μια ομάδα C_m για ένα έγγραφο Doc_j ορίζοντας μία συνάρτηση $Score(C_m \leftarrow Doc_j)$. Η ομάδα με την υψηλότερη τιμή για το μέτρο αυτό επιλέγεται ως η μοναδική για το εξεταζόμενο έγγραφο. Αν υπάρχουν περισσότερες της μιας ομάδας που μεγιστοποιούν τη συνάρτηση $Score$, επιλέγεται αυτή με το μεγαλύτερο αριθμό συνόλων εννοιών στην ετικέτα της. Ο τυπικός ορισμός της συνάρτησης $Score$ είναι ο ακόλουθος :

$$\begin{aligned} Score(C_m \leftarrow Doc_j) = [\sum_{x} Weight(j,x) \cdot cluster\_support(x)] \\ - [\sum_{x'} Weight(j,x') \cdot Keyphraseness(x')] \end{aligned} \tag{5.5}$$

όπου:

x αναπαριστά μία καθολικά σημαντικό έννοια στο Doc_j, που είναι "συχνή" και στην ομάδα C_m,

x' αναπαριστά μία καθολικά σημαντική έννοια στο Doc_j, που δεν είναι "συχνή" και στην ομάδα C_m,

$Weight(j,x)$ είναι το βάρος του x στο Doc_j όπως ορίζεται από την εξίσωση 5.4,

$Weight(j,x')$ ομοίως με τον παραπάνω ορισμό,

$cluster\_support(x)$ δίνεται από τον ορισμό (γ),

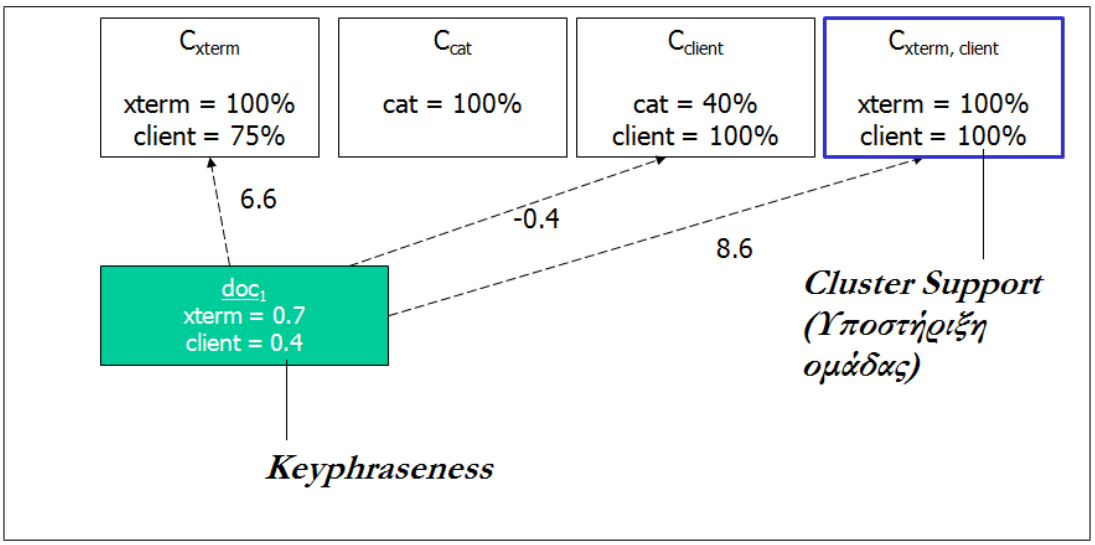
$keyphraseness(x')$ δίνεται από την εξίσωση 4.11.

Ο πρώτος όρος της συνάρτησης ανταμείβει την ομάδα C_m, εάν μία καθολικά σημαντική έννοια x στο Doc_j είναι συχνή στην ομάδα C_m. Η σημαντικότητα της έννοιας x σε διαφορετικές ομάδες καλύπτεται με τον πολλαπλασιασμό του βάρους της στο έγγραφο Doc_j με το βάρος της στην ομάδα C_m. Ο δεύτερος όρος της εξίσωσης "τιμωρεί" την ομάδα C_m εάν μία καθολικά σημαντική έννοια x' στο Doc_j δεν είναι συχνή στην ομάδα C_m. Το βάρος του x' στο Doc_j πολλαπλασιάζεται με την τιμή $Keyphraseness$ που εκφράζει πόσο σημαντική είναι η έννοια γενικά.

Για το παράδειγμα του Σχήματος 5.5, φαίνεται σχηματικά στο Σχήμα 5.6 ο υπολογισμός της πιο κατάλληλης ομάδας. Οι τιμές που φαίνονται στο έγγραφο αντιστοιχούν στο $Keyphraseness$ κάθε έννοιας ενώ οι τιμές που φαίνονται στις ομάδες είναι η *υποστήριξη ομάδας (cluster support)* για κάθε έννοια. Τη μεγαλύτερη τιμή δίνει η ομάδα $C_{xterm,client}$ και τελικά αυτή επιλέγεται ως καταλληλότερη για το έγγραφο που εξετάζεται.

5.3.2 Δημιουργία του δέντρου ομάδων

Σε αυτή τη φάση, ένα δέντρο ομάδων (ή ισοδύναμα θεμάτων) κατασκευάζεται βασισμένο στην ομοιότητα μεταξύ ομάδων. Στο δέντρο αυτό, κάθε ομάδα (εκτός από την ομάδα που είναι κενή και βρίσκεται στη ρίζα του δέντρου) έχει ακριβώς ένα γονέα.



Σχήμα 5.6: *Παράδειγμα επιλογής μιας ομάδας για ένα έγγραφο από τη μέθοδο CHC*

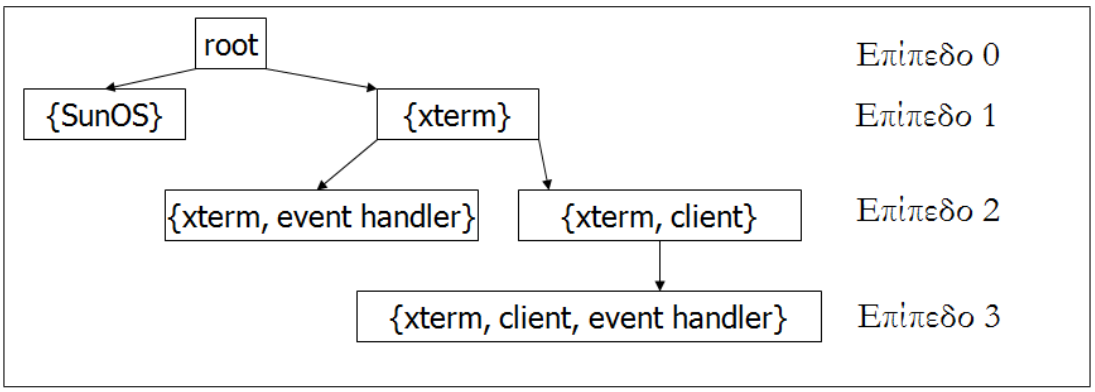
Στο σημείο αυτό, είναι χρήσιμη η υπενθύμιση πως κάθε ομάδα χρησιμοποιεί ένα καθολικά σημαντικό k-σύνολο εννοιών ως την ετικέτα της. Τέτοιες ομάδες θα καλούνται k-ομάδες από εδώ και κάτω. Στο δέντρο ομάδων, ο κόμβος της ρίζας εμφανίζεται στο επίπεδο 0, και ανταποκρίνεται στον κενό κόμβο και επί της ουσίας συλλέγει όλα τα έγγραφα που δεν ταξινομούνται πουθενά. Οι 1-ομάδες βρίσκονται στο επίπεδο 1 και κ.ο.κ. για κάθε επίπεδο. Το βάθος του δέντρου είναι ίσο με το μέγιστο μέγεθος των καθολικά σημαντικών συνόλων εννοιών.

Το δέντρο κατασκευάζεται "από κάτω προς τα πάνω" επιλέγοντας κάθε φορά έναν γονέα στο επίπεδο $k-1$ για κάθε ομάδα που βρίσκεται στο επίπεδο k. Για κάθε k-ομάδα C_m στο επίπεδο k, πρώτα βρίσκονται όλοι οι πιθανοί γονείς που είναι $(k-1)$-ομάδες και έχουν σαν ετικέτα ένα υποσύνολο της ετικέτας της ομάδας C_m. Δυνητικά υπάρχουν k τέτοιες ομάδες. Στο επόμενο βήμα, επιλέγεται ο καλύτερος γονέας. Η συνάρτηση $Score$ (εξίσωση 5.5) χρησιμοποιείται για αυτή την επιλογή αλλά σε αυτή τη φάση, συγχωνεύονται όλα τα έγγραφα στο υποδέντρο της ομάδας C_m σε ένα έγγραφο $(Doc(C_m)$, και έπειτα υπολογίζεται η τιμή για το έγγραφο $Doc(C_m)$ σε σχέση με κάθε πιθανό γονέα. Ο γονέας με την υψηλότερη τιμή για τη συνάρτηση $Score$ γίνεται ο γονέας για την ομάδα C_m. Ένα παράδειγμα δημιουργίας δέντρου φαίνεται στο Σχήμα 5.7. Η ομάδα στο επίπεδο-3 με ετικέτα xterm-client-event handler έχει δύο πιθανούς γονείς, εκ των οποίων επιλέγεται ο καταλληλότερος κ.ο.κ. μέχρι να τελειώσουν όλα τα επίπεδα (επίπεδο-0 είναι η ρίζα του δέντρου).

5.3.3 Κλάδεμα δέντρου

Ένα δέντρο μπορεί να είναι βαθύ και φαρδύ, ανάλογα με τις τιμές που έχουν επιλεγεί για τις ελάχιστες καθολικές συχνότητες και για την ελάχιστη τιμή $Keyphraseness$. Επομένως, είναι πιθανό έγγραφα να έχουν ανατεθεί σε ένα μεγάλο αριθμό μικρών ομάδων, που οδηγεί σε μικρή ακρίβεια. Ο σκοπός του κλαδέματος είναι να ενωθούν παρόμοιες ομάδες με στόχο να δημιουργηθεί μία πιο φυσική ιεραρχία για περιήγηση και να αυξηθεί η ακρίβεια. Πριν παρουσιαστούν οι τεχνικές κλαδέματος, είναι αναγκαίος ο ορισμός μιας συνάρτησης ομοιότητας ομάδων.

Για τη μέτρηση της ομοιότητας μεταξύ δύο ομάδων C_a και C_b, αντιμετωπίζεται η μία ομάδα ως έγγραφο (συνδυάζοντας όλα τα έγγραφα σε μία ομάδα) και μετριέται η



Σχήμα 5.7: *Παράδειγμα δημιουργίας ιεραρχίας δέντρου ομάδων από τη μέθοδο CHC*

τιμή της συνάρτησης $Score$ από την εξίσωση 5.5. Η διαφορά εδώ είναι πως το αποτέλεσμα πρέπει να κανονικοποιηθεί ώστε να εξαλειφθεί η επίδραση του μεταβαλλόμενου μεγέθους των εγγράφων και πως πρέπει να υπολογιστούν τόσο η ομοιότητα της ομάδας C_a με τη C_b όσο και η ομοιότητα της ομάδας C_b με τη C_a. Τυπικά, η ομοιότητα της ομάδας C_b με την ομάδα C_a ορίζεται ως εξής :

$$Sim(C_a \leftarrow C_b) = \frac{Score(C_a \leftarrow Doc(C_b))}{\sum_x Weight(Doc(C_b), x) + \sum_{x'} Weight(Doc(C_b), x')} + 1 \quad (5.6)$$

όπου:

$Doc(C_b)$ αντιπροσωπεύει όλα τα έγγραφα στο υποδένδρο της ομάδας C_b ενωμένα σε ένα έγγραφο,

x μία καθολικά σημαντική έννοια στο έγγραφο $Doc(C_b)$, που είναι επίσης συχνή στην ομάδα C_a,

x' μία καθολικά σημαντική έννοια στο έγγραφο $Doc(C_b)$, που δεν είναι συχνή στην ομάδα C_a,

$Weight(Doc(C_b), x)$, $Weight(Doc(C_b), x)$ είναι τα βάρη των εννοιών x και x' αντίστοιχα στο έγγραφο $Doc(C_b)$

Η κανονικοποίηση με τον παρανομαστή στην εξίσωση 5.6 εξηγείται ως εξής : Στη συνάρτηση $Score$, οι παράμετροι $Cluster\_Support$ και $Keyphraseness$ παίρνουν τιμές στο διάστημα $[0, 1]$, επομένως η μέγιστη τιμή της συνάρτησης $Score$ θα είναι $\sum_x Weight(j, x)$ και η ελάχιστη $-\sum_{x'} Weight(j, x')$. Άρα, μετά την προτεινόμενη κανονικοποίηση, η τιμή της Sim θα είναι στο διάστημα $[-1, 1]$. Για την αποφυγή αρνητικών τιμών για την ομοιότητα, προστίθεται ο όρος $+1$ και έτσι προκύπτει η παραπάνω εξίσωση. Προφανώς, το πεδίο τιμών της συνάρτησης Sim είναι το $[0, 2]$.

Η (τελική) ομοιότητα μεταξύ των ομάδων C_a και C_b υπολογίζεται ως ο γεωμετρικός μέσος των δύο κανονικοποιημένων τιμών που προκύπτουν από την εξίσωση 5.6 :

$$Similarity(C_a \longleftrightarrow C_b) = sqrtSim(C_a \leftarrow C_b) \times Sim(C_b \leftarrow C_a) \quad (5.7)$$

Το πλεονέκτημα του γεωμετρικού μέσου είναι πως δύο ομάδες για να θεωρηθούν όμοιες θα πρέπει να έχουν τιμές και για το $Sim(C_a \leftarrow C_b)$ και για το $Sim(C_b \leftarrow C_a)$ αρκετά μεγάλες. Η συνάρτηση $Similarity$ έχει το ίδιο πεδίο τιμών με την Sim, δηλαδή

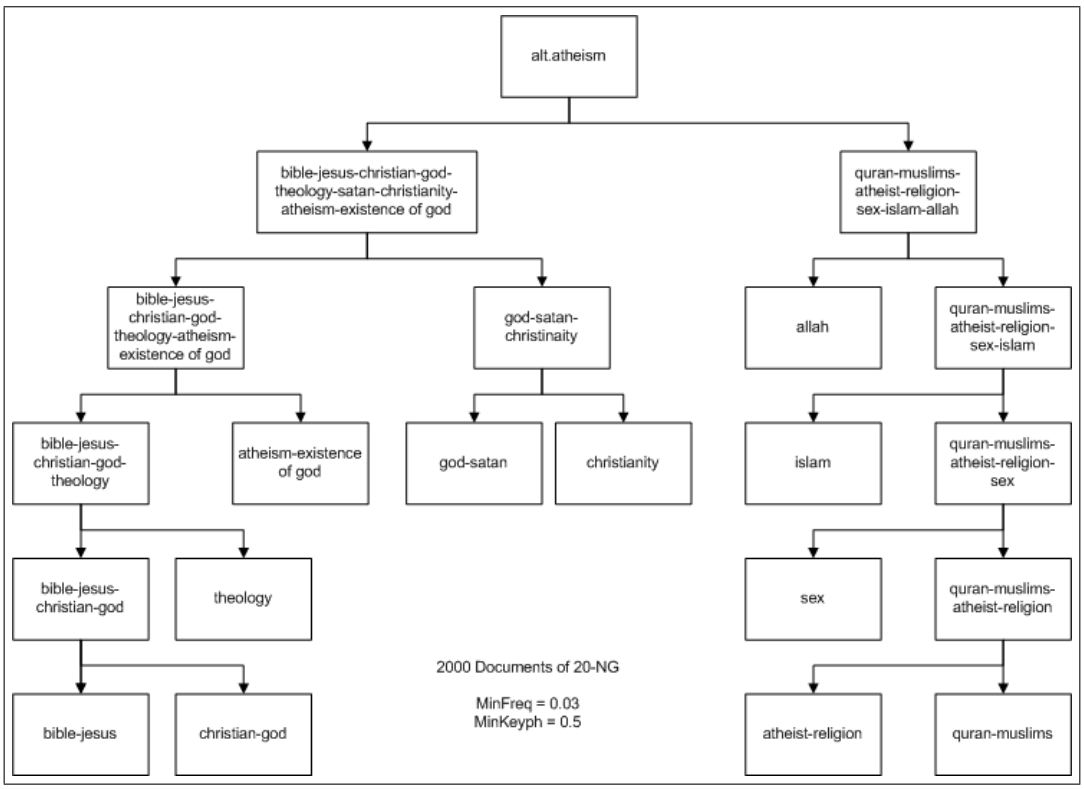
το $[0, 2]$. Στη μεθοδολογία που αναπτύχθηκε, η τιμή 1 για τη συνάρτηση *Similarity* θεωρείται ως το κατώφλι πέραν από το οποίο δύο ομάδες θεωρούνται όμοιες, αν και προφανώς αυτή η τιμή μπορεί να μεταβληθεί ανάλογα με τις ανάγκες του συστήματος.

Μετά τον ορισμό της συνάρτησης ομοιότητας ομάδων, ορίζονται τα κριτήρια κλαδέματος και συγχώνευσης ώστε το δέντρο να γίνει πιο αποδοτικό. Το κριτήριο κλαδέματος υπολογίζει την τιμή της *Similarity* μεταξύ ενός παιδιού και του γονέα του και ενεργοποιείται εφόσον η τιμή αυτή είναι μεγαλύτερη από 1, δηλαδή το παιδί είναι όμοιο με το γονέα του. Η διαίσθηση πίσω από αυτό το κριτήριο είναι πως εάν ένα υπο-θέμα (π.χ. god-atheism) είναι πολύ όμοιο με το γονικό του (π.χ. atheism), το υποθέμα είναι πιθανώς πολύ συγκεκριμένο και μπορεί να αφαιρεθεί. Ο έλεγχος του δέντρου γίνεται από κάτω προς τα πάνω (έως το επίπεδο 2, αφού η ρίζα συλλέγει μόνο έγγραφα που δεν εντάσσονται πουθενά) και εφόσον μια ομάδα κλαδευτεί, τότε τα παιδιά της γίνονται παιδιά του προπάππου τους. Για παράδειγμα, στο Σχήμα 5.7 θα μπορούσε να κλαδευτεί η ομάδα με ετικέτα xterm-client-event handler και τα έγγραφά της να ανατεθούν στη γενικότερη ομάδα με ετικέτα xterm-client.

Το κριτήριο συγχώνευσης παιδιών είναι μια διαδικασία που εφαρμόζεται στις ομάδες του επιπέδου 1 (το κλάδεμα που περιγράφηκε πριν δεν εφαρμόζεται στο επίπεδο αυτό). Κάθε φορά που η τιμή της *Similarity* είναι μεγαλύτερη από 1 για κάποιο ζευγάρι ομάδων στο επίπεδο 1, τότε συγχωνεύεται το ζευγάρι ομάδων με τη μεγαλύτερη τιμή. Αν κάποια ομάδα έχει παιδιά, τότε αυτά αποκτούν έναν γονέα πλέον. Για παράδειγμα, στο Σχήμα 5.7 θα μπορούσαν να συγχωνευθούν οι ομάδες SunOS και xterm δημιουργώντας μία μεγάλη ομάδα με παιδιά τις υπόλοιπες ομάδες του επιπέδου-2 και κάτω.

Ένα παράδειγμα ιεραρχίας για την ομάδα alt.atheism του συνόλου 20-Newsgroup φαίνεται στο Σχήμα 5.8. Το Σχήμα δείχνει πως οι μικρές ομάδες συγχωνεύονται μέχρι να σχηματιστεί η τελική ομάδα, που τελικά θα περιγράφεται από τις ετικέτες όλων των μικρών ομάδων. Ομάδες με μία έννοια σαν ετικέτα και χωρίς καθόλου παιδιά (π.χ. theology, sex, islam κ.α. είναι προφανώς μέρος του αρχικού συνόλου ομάδων. Το κλάδεμα των παιδιών έχει γίνει σε ομάδες χωρίς παιδιά αλλά με ετικέτα με περισσότερα της μιας έννοιας (π.χ. bible-jesus, christian-god κ.α.). Μια παρατήρηση στην ιεραρχία είναι αρκετή για να καταλάβει κανείς πως τα έγγραφα οργανώνονται σε 3 υποκατηγορίες με διακριτά υπο-θέματα : (α) bible, jesus, theology, existence of god (β) god, satan και (γ) quran, muslims, islam. Επομένως, φαίνεται πως η διαδικασία της ομαδοποίησης με την αυτόματη ετικετοποίηση που γίνεται, οδηγεί σε μιας μορφής εξαγωγή θέματος (η συλλογή εγγράφων χωρίζεται σε τρεις μεγάλες θεματικές κατηγορίες).

Επίσης, στον Πίνακα 5.1 φαίνονται οι ετικέτες 5 κατηγοριών του συνόλου δεδομένων 20-NG όπως παράγονται από τη μέθοδο. Το σημαντικό σε αυτό το παράδειγμα, είναι πως ενώ υπάρχει μεγάλη επικάλυψη των σημαντικών εννοιών (αυτά που έχουν μεγάλη τιμή για το *Keyphraseness* και μεγάλη συχνότητα) κάθε κατηγορίας, δεν υπάρχει επικάλυψη μεταξύ των ετικετών των ομάδων, λόγω της αρκετά υψηλής τιμής που απαιτείται για το *Keyphraseness*. Ιδιαίτερο ενδιαφέρουν έχουν οι ετικέτες που είναι έννοιες με πολλαπλές λέξεις (όπως τα "Existence of God" ή "windows manager") και ετικέτες που περιλαμβάνουν ή είναι ακρωνύμια (όπως "Windows NT" ή "SCSI controller").



Σχήμα 5.8: *Παράδειγμα ιεραρχικής δομής για την κατηγορία alt.atheism από τη μέθοδο CHC*

Πίνακας 5.1: *Ετικέτες ομάδων όπως δημιουργούνται από τη μέθοδο CHC και οι 5 πιο σημαντικές έννοιες για 5 κατηγορίες του συνόλου 20-NG*

Κατηγορία	Παράδειγμα ετικετών	Top-5 σημαντικές έννοιες σε συχνότητα και τιμή *Keyphraseness*
alt.atheism	atheism, Islam, existence of God, Quran	God, evidence, religion, atheist, bible, morality, Jesus, Satan, peace of God, death penalty
talk.religion.misc	Christianity, Branch Davidians	God, Jesus, Bible, Christians, Jews, Mormons, deity, Israelites, AMORC, Sermon
comp.os.ms-windows.misc	IBM, MSDOS, Windows NT, Unix	Windows, OS2, Microsoft, DOS, IBM, Memory, Mouse (computing), MSWindows, NDW, RAM
comp.sys.ibm.pc.hardware	SCSI controller, CMOS, Maxtor	motherboard, SCSI, bus (computing), DOS, IRQ, ISA bus, CDROM, OS2, CPU, power supply
comp.windows.x	window manager, SunOS, OpenWindows, xterm	SunOS, xterm, X-server, HP, bitmap, pixmap, event handler, source code, Xview, Xlib

5.4 Έλεγχος μεθοδολογίας CHC

Η σύγκριση της αποτελεσματικότητας των μεθόδων ομαδοποίησης γίνεται σε σχέση με τους πιο συνηθισμένους και ακριβείς αλγορίθμους ομαδοποίησης εγγράφων [Steinbach et al., 2000]: Την Ιεραρχική Συσσωρευτική Ομαδοποίηση (Hierarchical Agglomerative Clustering, HAC) και τη μέθοδο των k-μέσων(k-means), οι οποίες περιγράφηκαν στο Κεφάλαιο 5.2.1. Για την παραγωγή αποτελεσμάτων από τις δύο αυτές μεθόδους γίνεται χρήση του εργαλείου CLUTO-2.0 [Karypis, 2002].

5.4.1 Σύνολα δεδομένων εγγράφων

Τρία πολύ γνωστά σύνολα δεδομένων εγγράφων χρησιμοποιούνται για τον έλεγχο της μεθοδολογίας : το 20-NG, το Reuters-21578 (και τα δύο διαθέσιμα από το [UCI, 2011]) και η συλλογή Brown Corpus (διαθέσιμη από το [ICAME, 2011]). Και τα τρία αυτά σύνολα δεδομένων παρουσιάζονται αναλυτικά στο Παράρτημα Γ′.

5.4.2 Κριτήρια αξιολόγησης αποτελέσματος ομαδοποίησης εγγράφων

Για την αξιολόγηση της ποιότητας των μεθόδων ομαδοποίησης, θα χρησιμοποιηθούν τα ακόλουθα δημοφιλή μέτρα ποιότητας που που έχουν προταθεί από τη διεθνή βιβλιογραφία [Steinbach et al., 2000]: τα *F1-μέτρα* (μικροσκοπικό και μακροσκοπικό) (F-measure), η *καθαρότητα* (*Purity*), η *Εντροπία* (*Entropy*) και τα μέτρα *R, J* και *FM* που βασίζονται στην στατιστική.

Το $F1$-μέτρο συνδυάζει την *Ακρίβεια* (*Precision*) και την *Ανάκληση* (*Recall*) από το πεδίο της ανάκτησης πληροφοριών. Πιο συγκεκριμένα, η ακρίβεια, η ανάκληση και το F-μέτρο μιας ομάδας m σε σχέση με μία κλάση l ορίζονται ως εξής :

$$\begin{aligned} P = Precision(l,m) &= \frac{N_{l,m}}{N_m} \\ R = Recall(l,m) &= \frac{N_{l,m}}{N_l} \\ F1(l,m) &= \frac{2 \cdot P \cdot R}{P+R} \end{aligned} \tag{5.8}$$

όπου :
$F1(l,m)$ είναι το F-μέτρο για την κλάση l και την ομάδα m, $N_{l,m}$ είναι ο αριθμός των μελών της κλάσης l στην ομάδα m, N_l είναι ο αριθμός των μελών της κλάσης l και N_m ο αριθμός των μελών της ομάδας m.

Έπειτα, τα F1 μέτρα ορίζονται ως εξής :

$$\begin{aligned} Macro - F1 &= \frac{1}{|L|} \sum_{l \in L} F1(l, \sigma(l)) \\ Micro - F1 &= 2 \cdot \frac{microP \cdot microR}{microP + microR} \end{aligned} \tag{5.9}$$

όπου:
L είναι το σύνολο των κλάσεων, $\sigma(l) = \arg\max_m(F1(l,m))$ και

$$\begin{aligned} microP &= \frac{1}{|L|} \sum_{l \in L} \frac{N_{l,\sigma(l)}}{N_{\sigma(l)}} \\ microR &= \frac{1}{|L|} \sum_{l \in L} \frac{N_{l,\sigma(l)}}{N_l} \end{aligned}$$

Η *καθαρότητα* βασίζεται στο μέτρο της ακρίβειας. Κάθε ομάδα (cluster) m που υπάρχει διαθέσιμη στο τελικό αποτέλεσμα από ένα σύνολο C ομάδων αντιμετωπίζεται σαν να είναι το αποτέλεσμα μιας ερώτησης. Αντίστοιχα, κάθε σύνολο l εγγράφων από μία κλάση L αντιμετωπίζεται ως το επιθυμητό αποτέλεσμα για μία αναζήτηση. Τα μέτρα της καθαρότητας (Purity) και της αντίστροφης καθαρότητας (Inverse Purity)ορίζονται ως εξής:

$$\begin{aligned} Purity(C,L) &= \sum_{m \in C} \left(\frac{|C|}{|L|} \max_{l \in L} Precision(l,m) \right) \\ InversePurity(C,L) &= \sum_{l \in L} \left(\frac{|L|}{|C|} \max_{m \in C} Precision(l,m) \right) \end{aligned} \tag{5.10}$$

Το πρώτο μέτρο μετράει την καθαρότητα των ομάδων που έχουν δημιουργηθεί σε σχέση με την προκαθορισμένη κατηγοριοποίηση σε κλάσεις ενώ το δεύτερο μετράει πόσο σταθερές παραμένουν οι συγκεκριμένες κατηγορίες όταν κατανέμονται στα έγγραφα.

Η *εντροπία* μετράει το πόσο "κατάλληλη" είναι μια ομάδα ανάλογα με το πόσο ομογενής είναι. Όσο πιο ομογενής είναι μια ομάδα τόσο πιο χαμηλή είναι η εντροπία της ομάδας και ανάποδα. Για κάθε ομάδα m υπολογίζεται η πιθανότητα $p_{l,m}$ που είναι η πιθανότητα να ανήκει ένα μέλος της ομάδας m στην κλάση l. Η εντροπία ύστερα υπολογίζεται από την ακόλουθη εξίσωση :

$$E_m = -\sum_l p_{l,m} \cdot log(p_{l,m}) \tag{5.11}$$

όπου η άθροιση γίνεται για όλες τις κλάσεις l. Η συνολική εντροπία για το τελικό αποτέλεσμα της ομαδοποίησης C υπολογίζεται ως το άθροισμα όλων των εντροπιών κάθε ομάδας λαμβάνοντας υπόψιν και το μέγεθος κάθε ομάδας :

$$E_C = -\sum_{m \in C} (\frac{N_m}{N} \cdot E_m) \tag{5.12}$$

όπου :

C είναι το σύνολο των ομάδων,

N_m είναι ο αριθμός των εγγράφων στην ομάδα m,

N είναι ο συνολικός αριθμός των εγγράφων

Τέλος, εισάγονται μέτρα τα οποία βασίζονται στα στατιστικά στοιχεία μεταξύ ζευγών. Συμβολίζονται με SS ο αριθμός των ζευγών των αντικειμένων που ανήκουν στην ίδια ομάδα και στην ίδια κατηγορία, με DD ο αριθμός των ζευγών που ανήκουν σε διαφορετική κατηγορία και ομάδα, με SD ο αριθμός των ζευγών που ανήκουν στην ίδια ομάδα αλλά σε διαφορετική κατηγορία και με DS ο αριθμός των ζευγών που ανήκουν στην ίδια κατηγορία αλλά σε διαφορετική ομάδα. "Καλά" παραδείγματα είναι τα SS και DD, τα οποία είναι επιθυμητό να έχουν τις μεγαλύτερες δυνατές τιμές. Τρία συχνά χρησιμοποιούμενα μέτρα βάσει αυτών των αριθμών είναι *Rand Statistic (R)*, *Jaccard Coefficient (J)* και *Folkes and Mallows metric (FM)* :

$$R = \frac{SS + DD}{SS + SD + DS + DD} \tag{5.13}$$

$$J = \frac{SS}{SS + SD + DS} \tag{5.14}$$

$$FM = \sqrt{\frac{SS}{SS + SD} \frac{SS}{SS + DS}} \tag{5.15}$$

5.4.3 Αποτελέσματα για τη μέθοδο των πιο σημαντικών εννοιών (CHC)

Για κάθε έγγραφο κάθε συνόλου δεδομένων ακολουθείται η διαδικασία που περιγράφεται στο Σχήμα 4.4 για να αναπαρασταθεί με γνώση από τη Wikipedia. Για κάθε έννοια της Wikipedia, χρησιμοποιείται η εξίσωση 5.4 για τον υπολογισμό τους βάρους του στο έγγραφο και διατηρείται σε ένα καθολικό ευρετήριο η τιμή για το *Keyphraseness*.

5.4.3.1 Επιλογή παραμέτρων για τη μέθοδο CHC και αποτελέσματα

Έγινε πειραματισμός με διάφορες τιμές για τις παραμέτρους α, β και γ της εξίσωσης 5.4 για τον καθορισμό της επίδρασης των $WFreq$, $LinkRank$, $OrderRank$ και $ConceptSim$ στην αναπαράσταση του εγγράφου. Πιο συγκεκριμένα, κάθε φορά κρατείτο σταθερή η τιμή δύο παραμέτρων και μεταβαλόταν η τρίτη από 0 έως 1 (με βήμα 0.1). Με τη διαδικασία αυτή, βρέθηκαν οι βέλτιστες τιμές για τις παραμέτρους που παρήγαγαν τα καλύτερα αποτελέσματα ομαδοποίησης σε σχέση με τις τιμές για το F-μέτρο και την εντροπία. Οι παράμετροι $ConceptSim$ και $LinkRank$ φαίνεται να έχουν τη μεγαλύτερη επίδραση ενώ η παράμετρος $OrderRank$ τη μικρότερη. Οι βέλτιστες τιμές παραμέτρων της εξίσωσης 5.4 φαίνονται στον Πίνακα 5.2.

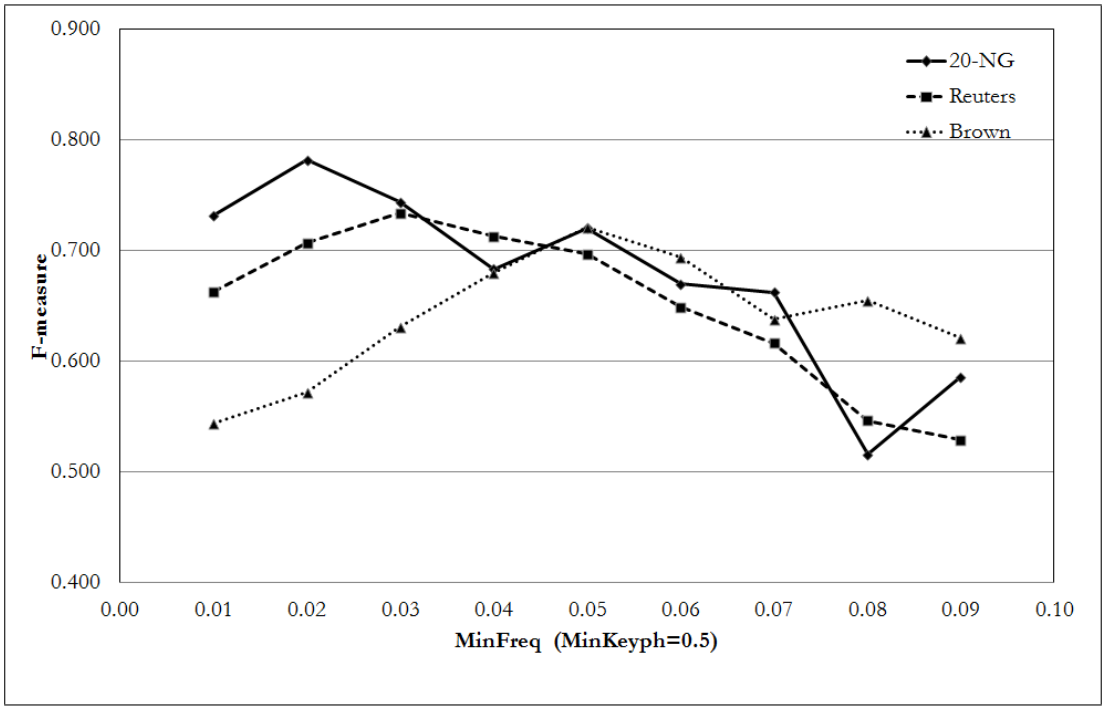
Πίνακας 5.2: *Βέλτιστες παράμετροι μεθόδου CHC*

Παράμετρος		**Τιμή**
$Wfreq$	α	0.2
$LinkRank$	β	0.4
$OrderRank$	γ	0.1
$ConceptSim$	1-α-β-γ	0.3
MinFreq		0.001 ÷ 0.03
MinKeyph		0.5

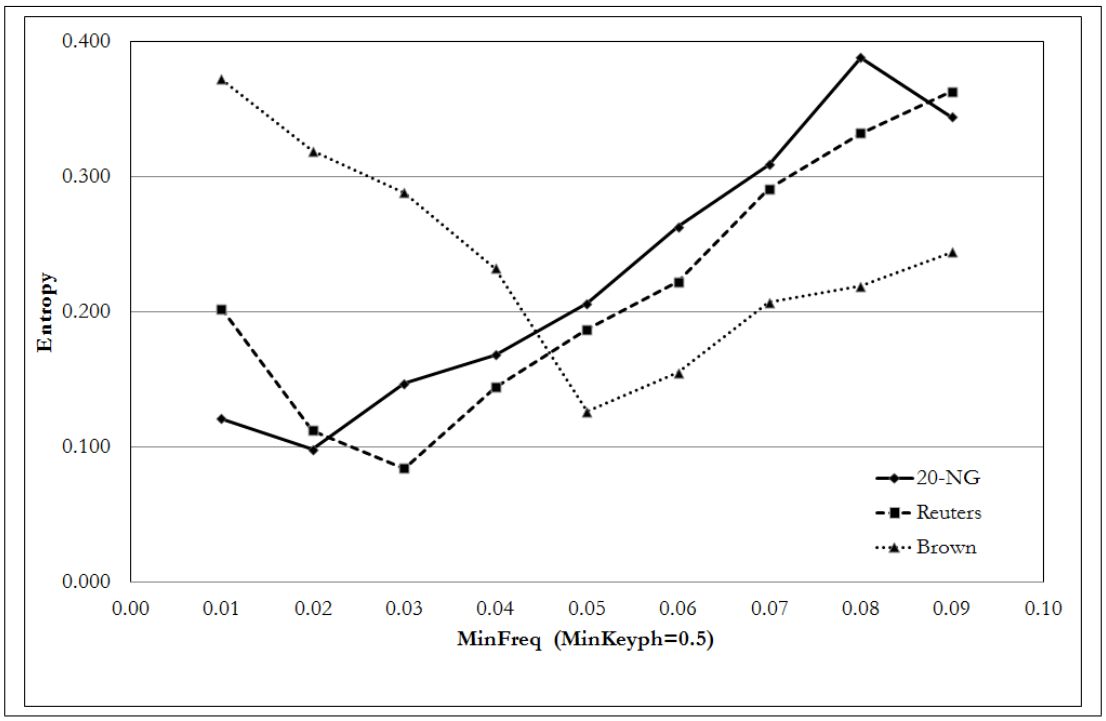
Μετά την ολοκλήρωση της αναπαράστασης των εγγράφων στο χώρο των εννοιών και τον υπολογισμό των βαρών τους, αρχίζει η διαδικασία της ομαδοποίησης όπως περιγράφεται στην Παράγραφο 5.3. Επιλέγονται οι αρχικές ομάδες θέτοντας την ελάχιστη τιμή κατωφλίου για το *Keyphraseness* (MinKeyph) και την ελάχιστη τιμή καθολικής συχνότητας της έννοιας(MinFreq) σε τιμές που οδηγούν σε περιγραφικές ετικέτες για τις ομάδες. Αυτό επιτυγχάνεται μέσω μιας αρκετά μεγάλης τιμής για το MinKeyph (το οποίο εξαλείφει αρκετές γενικές έννοιες) και μιας σχετικά χαμηλής-μέσης τιμής για το MinFreq (ανάλογα και με το πόσα έγγραφα είναι διαθέσιμα). Τα πειράματα δείχνουν πως μία τιμή για το MinKeyph γύρω στο 0.5 πάντα έχει καλά αποτελέσματα στα διάφορα σύνολα δεδομένων, εφόσον υπάρχουν τουλάχιστον μερικές εκατοντάδες έγγραφα διαθέσιμα. Τα Σχήματα 5.9 και 5.10 αναπαριστούν τις τιμές για το F-μέτρο και την εντροπία της μεθόδου αντίστοιχα, σε σχέση με την επιλεγμένη τιμή MinFreq (το MinKeyph τέθηκε στο 0.5 για αυτά τα πειράματα).

Επίσης, αρκετά μεγάλος αριθμός πειραμάτων έδειξε πως εφόσον ένα σύνολο δεδομένων περιέχει λιγότερα από 5000 έγγραφα τότε το MinFreq μπορεί να τεθεί μεταξύ 0.03 και 0.06, διαφορετικά αρκεί μία τιμή μεταξύ 0.005 και 0.03. Όλες οι βέλτιστες παράμετροι φαίνονται στον Πίνακα 5.2.

Ο προσδιορισμός των ομάδων γίνεται με τη διαδικασία που περιγράφηκε στην Παράγραφο 5.3.3. Τα τελικά αποτελέσματα ομαδοποίησης συγκρίνονται με εκείνα που



Σχήμα 5.9: *Μέθοδος CHC: Ευαισθησία F-μέτρου σε σχέση με το MinFreq*



Σχήμα 5.10: *Μέθοδος CHC: Ευαισθησία Εντροπίας σε σχέση με το MinFreq*

προκύπτουν από τους αλγορίθμους HAC και k-means για τα σύνολα δεδομένων 20-NG, Reuters και Brown και φαίνονται στον Πίνακα 5.3. Οι βελτιώσεις που αναφέρονται επιτεύχθηκαν με χρήση των παραμέτρων του Πίνακα 5.2. Για τη μέθοδο HAC υλοποιήθηκε η παραλλαγή UPGMA ενώ στη μέθοδο k-means η παραλλαγή bisecting με το k στο 8.

Πίνακας 5.3: *Πειραματικά αποτελέσματα μεθόδου CHC*

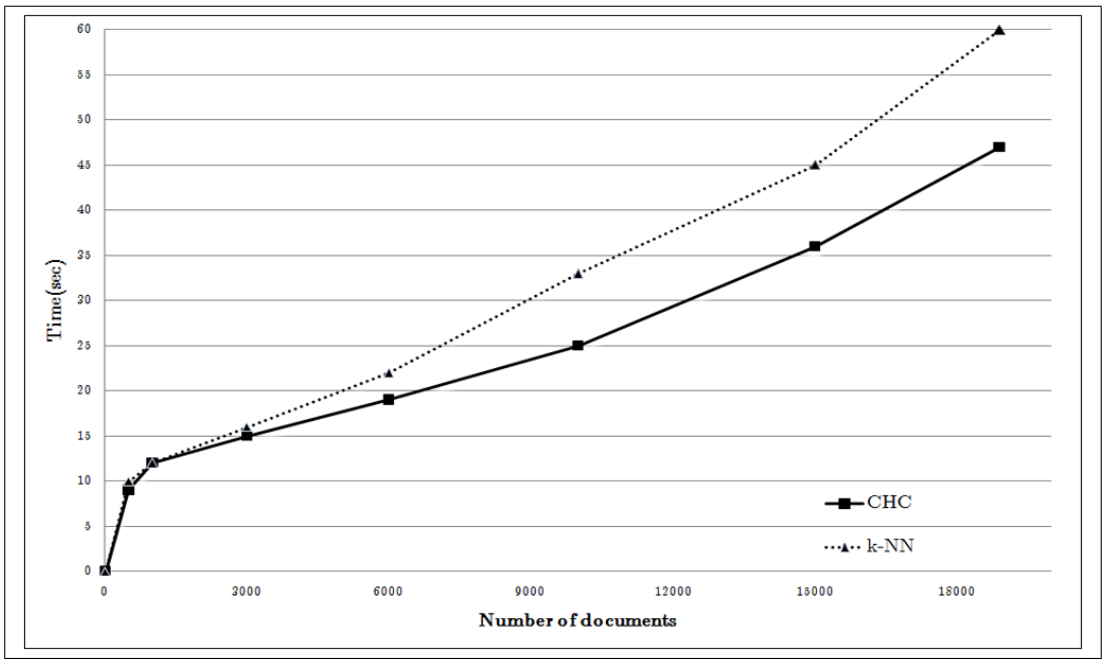
Σύνολο δεδομένων	Αλγόριθμος	F-measure	Βελτίωση	Εντροπία	Βελτίωση
20-Newsgroups	HAC	0.551	41.93%	0.205	-52.20%
	k-means	0.609	28.41%	0.114	-14.04%
	CHC	**0.782**		**0.098**	
Reuters	HAC	0.422	73.93%	0.196	-57.14%
	k-means	0.417	76.02%	0.092	-8.70%
	CHC	**0.734**		**0.084**	
Brown	HAC	0.482	49.59%	0.231	-45.45%
	k-means	0.475	51.79%	0.191	-34.03%
	CHC	**0.721**		**0.126**	

5.4.3.2 Πολυπλοκότητα αλγορίθμου CHC

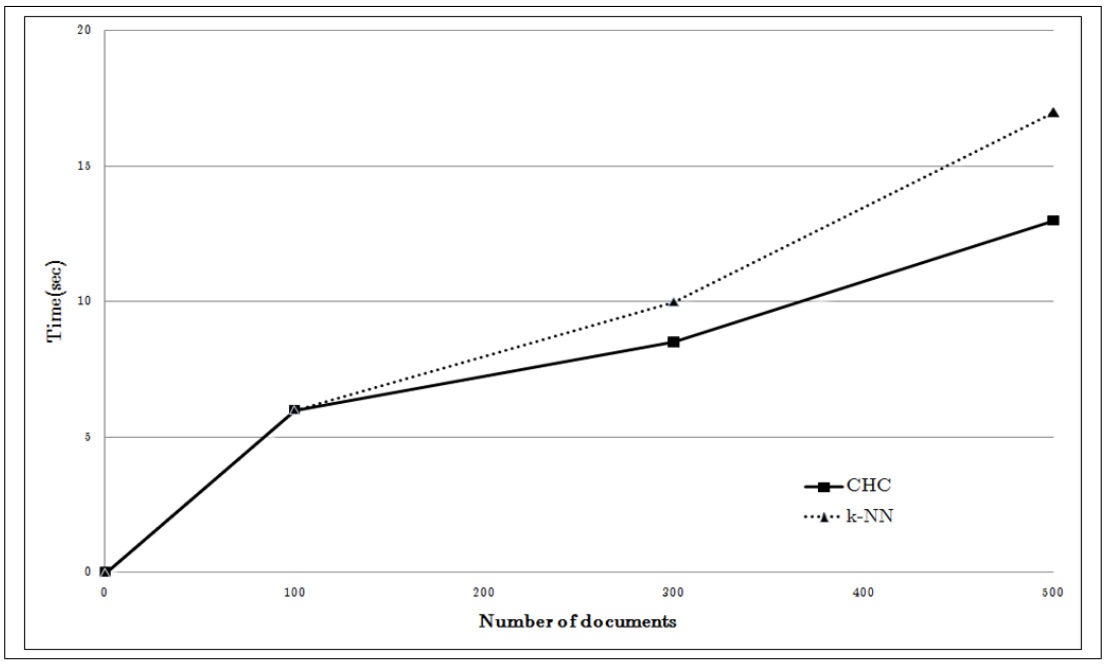
Μελετήθηκε ο χρόνος εκτέλεσης του αλγορίθμου σε σχέση με αυτόν της μεθόδου k-means. (Η επίδοση της μεθόδου HAC είναι αρκετά χαμηλή και δεν περιλήφθηκε σε αυτό το πείραμα). Για το πείραμα αυτό χρησιμοποιήθηκε το σύνολο 20-NG (έχει το μεγαλύτερο αριθμό εγγράφων ανάμεσα στα 3 σύνολα που χρησιμοποιήθηκαν) και το σύνολο Brown (έχει τα έγγραφα με τις περισσότερες λέξεις). Τα αποτελέσματα του πειράματος αυτού φαίνονται στα Σχήματα 5.11 και 5.12. Η προτεινόμενη μέθοδος βελτιώνει σημαντικά το χρόνο εκτέλεσης σε σχέση με τη μέθοδο k-means και στα δύο σύνολα δεδομένων, κυρίως λόγω της μετάβασης από το μεγάλο και αραιό χώρο BOW στον διαστατικά μειωμένο και σημασιολογικά πλουσιότερο χώρο των εννοιών. Και οι δύο μέθοδοι αποδίδουν καλά σε μικρότερα σύνολα δεδομένων αλλά όταν ο αριθμός των εγγράφων μεγαλώνει, η προτεινόμενη μέθοδος αποδίδει καλύτερα. Η βελτίωση είναι σημαντικότερη στο σύνολο 20-NG λόγω του ότι το σύνολο Brown περιέχει πιο πλούσια σε λέξεις έγγραφα, επομένως η απόδοση του μοντέλου BOW κρίνεται επαρκής.

Η προτεινόμενη μεθοδολογία περιλαμβάνει δύο φάσεις: Τη μετάβαση από το μοντέλο BOW στο χώρο των εννοιών και τη μεθοδολογία ομαδοποίησης (CHC). Η διαδικασία του ταιριάσματος των λέξεων των εγγράφων με έννοιες της Wikipedia μπορεί να γίνεται δυναμικά, κάτι το οποίο είναι σημαντικό λόγω των σχεδόν καθημερινών αλλαγών του περιεχομένου της Wikipedia, αν και η συγκεκριμένη διαδικασία μπορεί να γίνεται και off-line αφού το περιεχόμενο της Wikipedia είναι διαθέσιμο για μεταφόρτωση και επεξεργασία. Επίσης, η διαδικασία αποσαφήνισης (περιπου 20% των εννοιών κάθε εγγράφου είναι πολύσημες) ολοκληρώνει την εξαγωγή των εννοιών. Το επόμενο βήμα είναι ο υπολογισμός των βαρών των εννοιών σε κάθε έγγραφο (βάσει της εξίσωσης 5.4), το οποίο γίνεται off-line αφού τα χαρακτηριστικά που κατασκευάζονται βάσει της Wikipedia αποθηκεύονται τοπικά.

Είναι σημαντικό να τονιστεί το διάνυσμα των εγγράφων στο χώρο των εννοιών δεν είναι τόσο μεγάλο και τόσο αραιό όσο στον κλασσικό χώρο BOW. Για παράδειγμα, στο σύνολο 20-NG, η διάσταση του διανύσματος είναι 61174, ενώ στο χώρο των εννοιών είναι μόλις 22510.



Σχήμα 5.11: *Χρονική απόδοση μεθόδου CHC στο σύνολο 20-NG*



Σχήμα 5.12: *Χρονική απόδοση μεθόδου CHC στο σύνολο Brown*

Η αρχική διαδικασία ομαδοποίησης απαιτεί να ελεγχθούν δύο φορές τα διανύσματα των εγγράφων, μία φορά για εντοπιστούν οι καθολικά σημαντικά έννοιες και μία για τον διαχωρισμό των ομάδων, πάντα όμως υπό το πρίσμα ότι λόγω της μεγάλης μείωσης της διάστασης η διαδικασία είναι πολύ πιο γρήγορη απότι στο μοντέλο BOW.

Τέλος, όσον αφορά στην κατασκευή του δέντρου, η διαδικασία εντοπισμού ενός γονέα για μία k-ομάδα C στο επίπεδο $k-1$ απαιτεί τον έλεγχο το πολύ k ομάδων, αφού οι πιθανοί γονείς του C έχουν ετικέτες που είναι υποσύνολο της ετικέτας του C. Η διαδικασία απαιτεί τον υπολογισμό ομοιοτήτων συγκεκριμένων ομάδων αλλά ο αριθμός k είναι συνήθως πολύ μικρός, οπότε ο χρόνος εκτέλεσης παραμένει γραμμικός. Τέλος, το κλάδεμα παιδιών απαιτεί έλεγχο των ομάδων μόνο μία φορά ενώ η συγχώνευση παιδιών γίνεται μόνο στο υψηλότερο επίπεδο του δέντρου (επίπεδο 1).

5.5 Μεθοδολογία ομαδοποίησης εγγράφων : Document Self-Organizer (DoSO)

Σαν βελτίωση της μεθόδου που περιγράφηκε στην Παράγραφο 5.3, μελετάται η ενσωμάτωσή της με τους αυτο-οργανούμενους χάρτες (Self-Organizing Maps, SOM). Ο στόχος είναι η αυτόματη εξαγωγή μιας καλής περιγραφής των ομάδων βασισμένη σε έννοιες της Wikipedia αλλά και η τοποθέτηση των ομάδων στο χώρο με τοπολογική ορθότητα ώστε να ικανοποιούνται οι μεταξύ τους σημασιολογικές σχέσεις. Σε κάθε νευρώνα αντιστοιχίζεται μία ετικέτα κατά τη διάρκεια της εκπαίδευσης του SOM χρησιμοποιώντας το γνώρισμα *Keyphraseness* που έχει εξαχθεί από τη Wikipedia και το οποίο είναι χαρακτηριστικό της περιγραφικότητας της ετικέτας. Ο ορισμός (α) της Παραγράφου 5.3 παραμένει ο ίδιος αλλά τροποποιούνται ή εισάγονται νέοι ορισμοί για να περιγράψουν τους νευρώνες του μοντέλου SOM. Αναλυτικά οι ορισμοί που τροποποιούνται ή εισάγονται είναι:
(α) Μια *καθολικά σημαντική έννοια* είναι μια έννοια η οποία:

- έχει τιμή *Keyphraseness* μεγαλύτερη από ένα συγκεκριμένο κατώφλι, που ορίζεται ως η ελάχιστη τιμή *Keyphraseness* (*minimum keyphraseness threshold* ή *MinKeyph*), και
- εμφανίζεται σε περισσότερα από ένα ποσοστό των εγγράφων της συλλογής, που ορίζεται ως η ελάχιστη τιμή συχνότητας ένα ελάχιστο (*minimum global frequency threshold* ή *MinFreq*).

Ένα *καθολικά σημαντικό k-σύνολο εννοιών* είναι ένα σύνολο από k καθολικά σημαντικές έννοιες οι οποίες εμφανίζονται μαζί σε ένα ποσοστό των εγγράφων της συλλογής μεγαλύτερο από το *MinFreq*.
(β) Το προτεινόμενο μοντέλο χρησιμοποιεί νευρώνες (όπως ακριβώς το κλασσικό SOM) που περιγράφονται από :

- ένα διάνυσμα βαρών (*διάνυσμα νευρώνα*), το οποίο έχει την ίδια διάσταση με το διάνυσμα που περιγράφει τα έγγραφα και κάθε συνιστώσα του αντιστοιχεί στο βάρος της συγκεκριμένης έννοιας στο νευρώνα,
- μία ετικέτα (*ετικέτα νευρώνα*), η οποία καθορίζεται από τα *καθολικά σημαντικά k-σύνολα-εννοιών* που περιέχονται στα έγγραφα τα οποία ανατίθενται στο νευρώνα με τη διαδικασία που θα περιγραφεί στη φάση της αρχικοποίησης του μοντέλου αργότερα,

- μία θέση στο δισδιάστατο επίπεδο (*θέση νευρώνα*), που μιλώντας με όρους του κλασσικού SOM ισοδυναμεί με τη θέση κάθε νευρώνα στο χώρο εξόδου (πλέγμα εξόδου) κατάλληλα προσαρμοσμένου στο νέο μοντέλο που εισάγεται και θα επεξηγηθεί αργότερα.

(γ) Μια καθολικά σημαντική έννοια είναι και *σημαντική στο νευρώνα* N_m, αν το βάρος της έννοιας στο νευρώνα είναι μεγαλύτερο από ένα συγκεκριμένο κατώφλι που ορίζεται ως το *ελάχιστη υποστήριξη νευρώνα* (MinWeight).

Η μέθοδος αποτελείται από 3 βήματα. Στο πρώτο βήμα, οι νευρώνες επιλέγονται και τα έγγραφα ανατίθενται σε αυτούς (ανάλογα με την τιμή $Keyphraseness$ των εννοιών και της συχνότητας των εννοιών και συνόλων εννοιών). Επίσης, γίνεται προβολή των νευρώνων βάσει της μεθόδου ISOMAP [Tenenbaum et al., 2000] ώστε να επιτευχθεί οπτικοποίηση του αποτελέσματος στις δύο διαστάσεις. Στο δεύτερο βήμα, εκτελείται η φάση ανταγωνισμού του SOM, γίνεται η ανανέωση των βαρών και η συνεργασία των νευρώνων. Τέλος, στο τρίτο βήμα ακολουθείται μια διαδικασία για τον προσδιορισμό των ομάδων της εξεταζόμενης συλλογής εγγράφων και κατασκευάζεται μία ιεραρχική δομή της συλλογής αυτής.

5.5.1 Φάση 1: Αρχική επιλογή νευρώνων και αρχικοποίηση

Δεδομένων των ορισμών (α) και (β) (όπως παρουσιάστηκαν παραπάνω) καθορίζεται ο αριθμός και οι ετικέτες των νευρώνων. Για κάθε σύνολο εννοιών που πληρεί τις προϋποθέσεις του ορισμού (α), κατασκευάζεται ένας νευρώνας που περιέχει όλα τα έγγραφα τα οποία περιέχουν αυτό το σύνολο εννοιών. Παραλείπονται τα (*k-1*)-σύνολα-εννοιών εάν οι έννοιές τους εμφανίζονται ως k-σύνολα-εννοιών (για παράδειγμα εάν οι έννοιες "astronomy" και "*comet*" είναι καθολικά σημαντικές έννοιες αλλά και το "*astronomy-comet*" είναι επίσης καθολικά σημαντικό σύνολο εννοιών τότε δημιουργείται μόνο ένας νευρώνας για το σύνολο "*astronomy-comet*". Είναι προφανές πως σε αυτή τη φάση, ένα έγγραφο μπορεί να ανήκει σε περισσότερους τους ενός νευρώνες. Η αποσύνδεση εγγράφων/νευρώνων καθορίζοντας τον πιο κατάλληλο νευρώνα για κάθε έγγραφο γίνεται στο επόμενο βήμα.

Η ετικέτα κάθε νευρώνα καθορίζεται από το καθολικά σημαντικό σύνολο εννοιών που περιλαμβάνεται σε όλα τα έγγραφα που ανατίθενται στο νευρώνα. Στο σημείο αυτό, τα διανύσματα βαρών του νευρώνα αρχικοποιούνται σε σχέση με τα έγγραφα που τους έχουν ανατεθεί βάσει της ακόλουθης εξίσωσης :

$$NW(k,i) = \sum_{j \in M_{ki}} \frac{Weight(j,i)}{|M|} \tag{5.16}$$

όπου :

- $NW(k,i)$ είναι το βάρος της έννοιας i στο νευρώνα k,
- M_{ki} είναι το σύνολο εγγράφων που αρχικά ανατίθενται στο νευρώνα k και περιέχουν την έννοια i,
- $Weight(j,i)$ είναι το βάρος της έννοιας i στο έγγραφο j όπως ορίζεται στην εξίσωση 5.4,

Έννοιες οι οποίες έχουν τιμή NW μεγαλύτερη από ένα καθορισμένο κατώφλι που καθορίζεται από το χρήστη (πρβλ ορισμό (γ)) είναι σημαντικά για το νευρώνα. Ας επισημανθεί εδώ πως η αρχικοποίηση αυτή είναι σημαντική για την ταχύτερη εκπαίδευση του SOM.

Μετά τον καθορισμό του αριθμού των νευρώνων, ο πιο κατάλληλος νευρώνας για κάθε έγγραφο πρέπει να βρεθεί και το έγγραφο να ανατεθεί μόνο σε αυτόν. Για το σκοπό αυτό, τροποποιείται η εξίσωση 5.5 ώστε να μπορεί να εφαρμοστεί σε νευρώνες ως εξής :

$$\begin{aligned} Score(N_m \leftarrow Doc_j) = [\sum_x Weight(j,x) \cdot NW(m,x)] \\ - [\sum_{x'} Weight(j,x') \cdot Keyphraseness(x')] \end{aligned} \tag{5.17}$$

όπου :

- x αναπαριστά μια καθολικά σημαντική έννοια στο Doc_j, που είναι σημαντική και στο νευρώνα N_m,
- x' αναπαριστά μια καθολικά σημαντική έννοια στο Doc_j, που δεν είναι σημαντική και στο νευρώνα N_m,
- $Weight(j,x)$ είναι το βάρος του x στο Doc_j όπως ορίζεται από την εξίσωση 5.4,
- $Weight(j,x')$ ομοίως με το παραπάνω,
- $NW(m,x)$ δίνεται από την εξίσωση 5.16,
- $Keyphraseness(x')$ δίνεται από την εξίσωση 4.11.

Ο νευρώνας με τη μεγαλύτερη τιμή $Score$ είναι ο πιο κατάλληλος (ή ο "νικητής "), στη φάση της αρχικοποίησης :

$$winner = \arg\max_m \{Score(N_m \leftarrow Doc_j)\} \tag{5.18}$$

Αφού ολοκληρωθεί η ανάθεση κάθε εγγράφου σε ένα μόνο νευρώνα, υπολογίζονται οι αποστάσεις των νευρώνων στο χώρο εισόδου χρησιμοποιώντας την εξής εξίσωση :

$$D(N_m, N_n) = \sqrt{\sum_i \left[(NW(m,i) - NW(n,i))^2\right]} \tag{5.19}$$

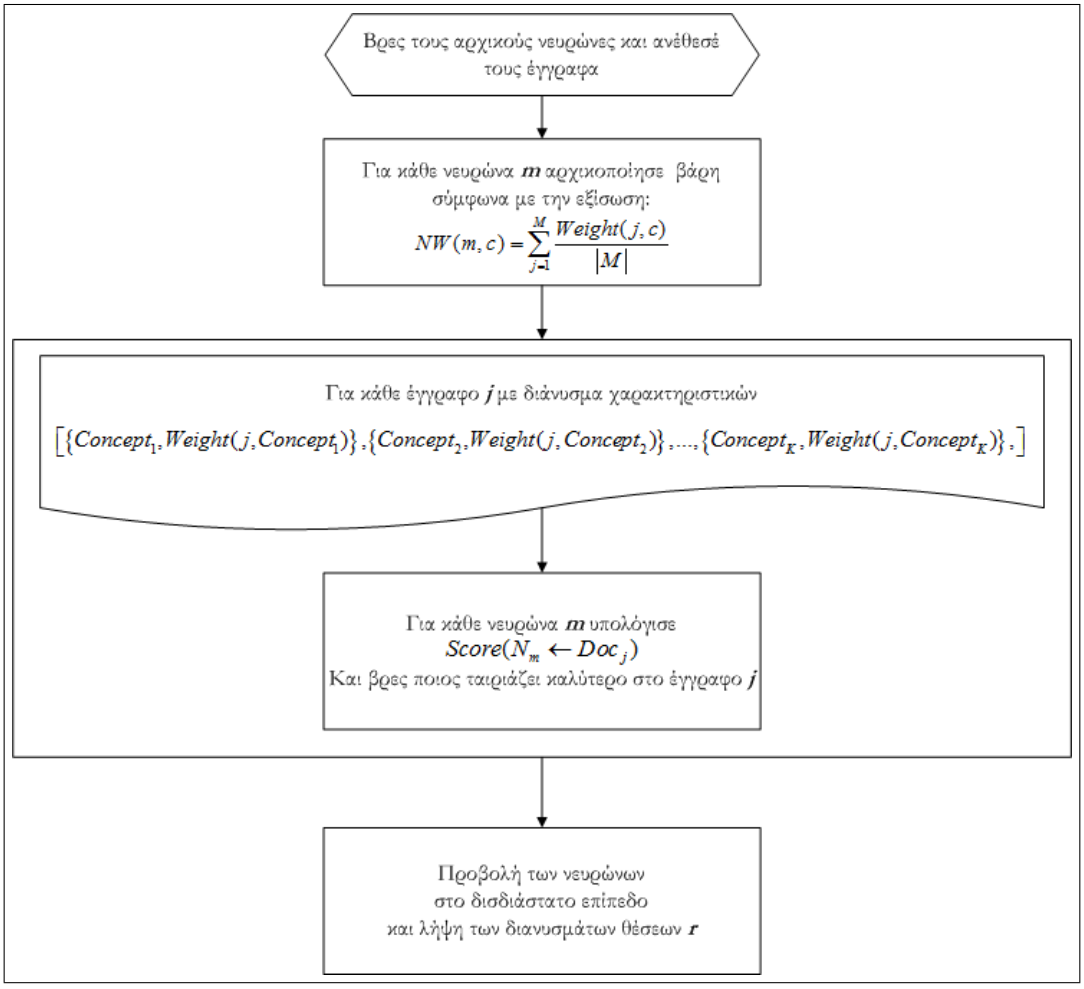
όπου N_m και N_n είναι οι δύο νευρώνες των οποίων η απόσταση υπολογίζεται και το NW αναφέρεται στο αντίστοιχο διάνυσμα του κάθε νευρώνα, ενώ η άθροιση λαμβάνει χώρα για όλες τις έννοιες i της συλλογής.

Η φάση 1 ολοκληρώνεται με τη χρήση της μεθόδου ISOMAP βάσει της οποίας εξάγεται μία οπτικοποίηση του μοντέλου σε 2 διαστάσεις. Η μέθοδος ISOMAP αρχικά ορίζει το γράφο γειτνίασης μεταξύ των νευρώνων και συνδέει κάθε έναν με τους k πλησιέστερους (το k ορίζεται από το χρήστη). Έπειτα, υπολογίζονται οι αποστάσεις των νευρώνων στο γράφο (οριζόμενες ως το άθροισμα των βαρών των ακμών πάνω στη συντομότερη διαδρομή μεταξύ των δύο κόμβων-νευρώνων). Τελικά, τα n ιδιοδιανύσματα του πίνακα αποστάσεων στο γράφο, αναπαριστούν τις συντεταγμένες στο νέο

n-δισδιάστατο Ευκλείδειο χώρο. Θέτοντας $n = 2$, οι νευρώνες προβάλλονται στο δισδιάστατο χώρο και ο αλγόριθμος του SOM μπορεί να χρησιμοποιηθεί για εκπαίδευση. Ένα παράδειγμα οπτικοποίησης φαίνεται στο Σχήμα 5.14 για 4 κατηγορίες. Μερικοί νευρώνες έχουν ήδη διαχωριστεί από τους υπόλοιπους σχηματίζοντας ανεξάρτητες ομάδες αλλά άλλοι χρειάζονται επιπλέον επεξεργασία (εκπαίδευση) για να διαχωριστούν καλύτερα. Ας σημειωθεί πως κάθε νευρώνας αντιστοιχεί σε μία θέση στο δισδιάστατο επίπεδο ($\mathbf{r}_m = (x_m, y_m)$) (πρβλ ορισμό (β)) και μπορεί να οριστεί η απόσταση μεταξύ νευρώνων στο δισδιάστατο χώρο με χρήση της επόμενης εξίσωσης :

$$\Delta(N_m, N_n) = \|\mathbf{r}_m - \mathbf{r}_n\| = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2} \tag{5.20}$$

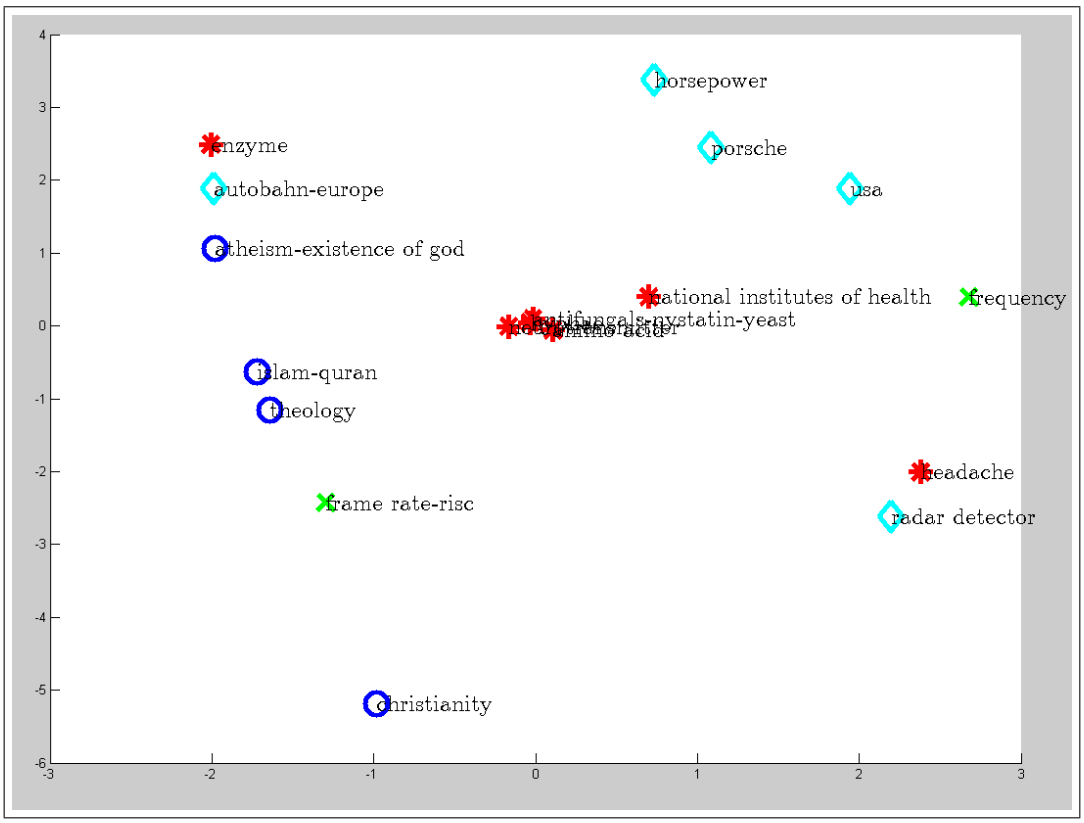
Η όλη διαδικασία της αρχικοποίησης φαίνεται στο Σχήμα 5.13



Σχήμα 5.13: *Περιγραφή φάσης αρχικοποίησης DoSO*

5.5.2 Φάση 2: Εκπαίδευση και ανταγωνισμός

Μετά την αρχικοποίηση των νευρώνων, λαμβάνει χώρα η διαδικασία εκπαίδευσης βάσει των εξισώσεων του SOM οι οποίες προσαρμόζονται για τα δεδομένα του συστήματος. Επιπλέον, στην προσέγγιση που ακολουθείται, οι θέσεις των νευρώνων (όπως υπολογίστηκαν στη φάση 1) αλλάζουν δυναμικά βάσει ενός κανόνα ανανέωσης παρόμοιου με τον κλασσικό κανόνα μάθησης SOM. Η *φάση ανταγωνισμού* περιλαμβάνει την εύρεση



Σχήμα 5.14: *Παράδειγμα οπτικοποίησης νευρώνων από τη μέθοδο DoSO για 4 κλάσεις (πριν την εκπαίδευση)*

του πιο κατάλληλου νευρώνα για κάθε έγγραφο j σύμφωνα με την ακόλουθη εξίσωση:

$$Sim(N_m, j) = \sum_i \{Weight(j,i) \times NW(m,i)\} \tag{5.21}$$

όπου η άθροιση γίνεται για όλες τις έννοιες i του εγγράφου j. Προφανώς, ο νευρώνας-νικητής m^* είναι αυτός που μεγιστοποιεί τη συνάρτηση Sim :

$$m^* = \arg\max_m \{Sim(N_m, j)\} \tag{5.22}$$

Η συνάρτηση γειτονιάς ορίζεται ως εξής :

$$h_{m,m^*}(t) = \exp\left(-\frac{\Delta\left(N_m, N_{m^*}\right)^2}{2 \cdot \sigma(t)^2}\right) \tag{5.23}$$

όπου :

- Δ είναι η απόσταση των νευρώνων στο δισδιάστατο χώρο (χώρο εξόδου), όπως ορίζεται από την εξίσωση 5.20,
- σ είναι μία γκαουσσιανή συνάρτηση που μειώνεται με το χρόνο (εποχές)

Μετά την επιλογή του νευρώνα-νικητή για το έγγραφο j, ανανεώνονται τα βάρη όλων των εννοιών i στους νευρώνες σύμφωνα με την ακόλουθη εξίσωση:

$$NW(m,i)^{t+1} = NW(m,i)^t + \eta^t \cdot h_{m,m^*}(t) \cdot \left[Weight(j,i) - NW(m,i)^t\right] \tag{5.24}$$

όπου :

- $NW(m,i)^{t+1}$ είναι το ανανεωμένο βάρος της έννοιας i στο νευρώνα m,
- $NW(m,i)^{t}$ είναι το παλιό βάρος της έννοιας i στο νευρώνα m,
- $Weight(j,i)$ δίνεται από την εξίσωση 5.4,
- η είναι ο ρυθμός μάθησης, που μειώνεται με τις εποχές,
- $h_{m,m^*}(t)$ είναι η συνάρτηση γειτονιάς όπως ορίζεται από την εξίσωση 5.23,

Επιπλέον, ανανεώνονται οι θέσεις των νευρώνων (στο δισδιάστατο χώρο) σύμφωνα με την ακόλουθη εξίσωση :

$$\mathbf{r}_m^{t+1} = \mathbf{r}_m^t + \zeta(t) \cdot H_{m,m^*}(t) \cdot \left[\mathbf{r}_{m^*}^t - \mathbf{r}_m^t\right] \tag{5.25}$$

όπου :

- $\mathbf{r}_m^{t+1}$ αντιστοιχεί στην ανανεωμένη θέση του νευρώνα m στο διδιάστατο χώρο,
- $\mathbf{r}_m^t$ αντιστοιχεία στην παλιά θέση του νευρώνα m στο διδιάστατο χώρο,
- ζ είναι ο ρυθμός μάθησης, που μειώνεται με τις εποχές,
- H_{m,m^*} είναι μία συνάρτηση γειτνίασης που ορίζεται από την ακόλουθη εξίσωση :

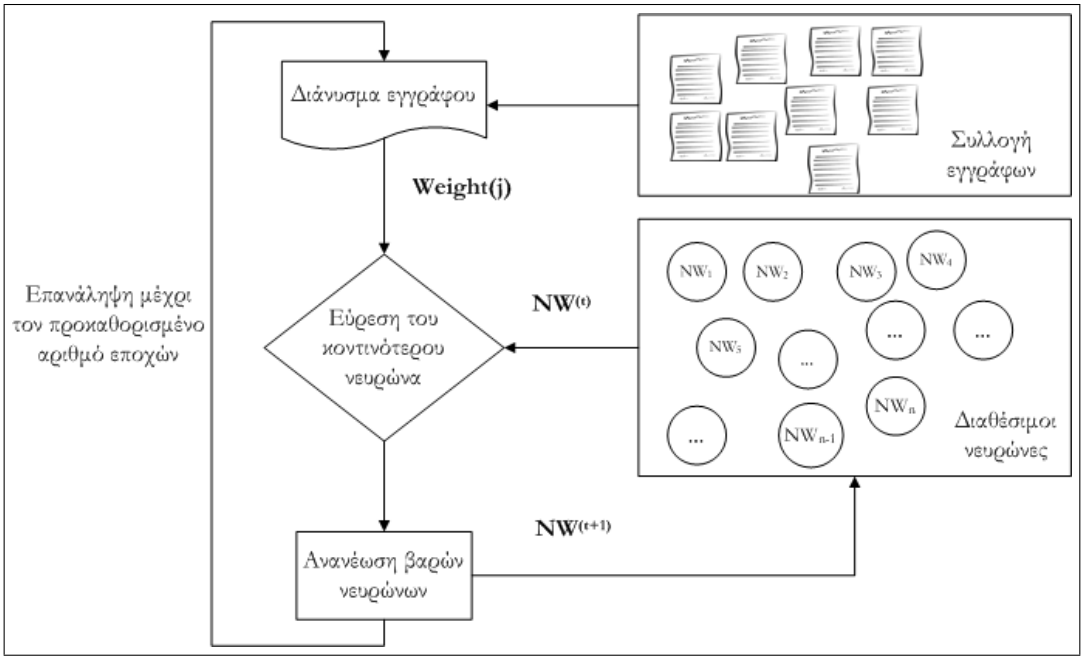
$$H_{m,m^*}(t) = \exp\left(-\frac{D\left(m,m^*\right)^2}{2 \cdot \sigma'(t)^2}\right) \tag{5.26}$$

όπου :

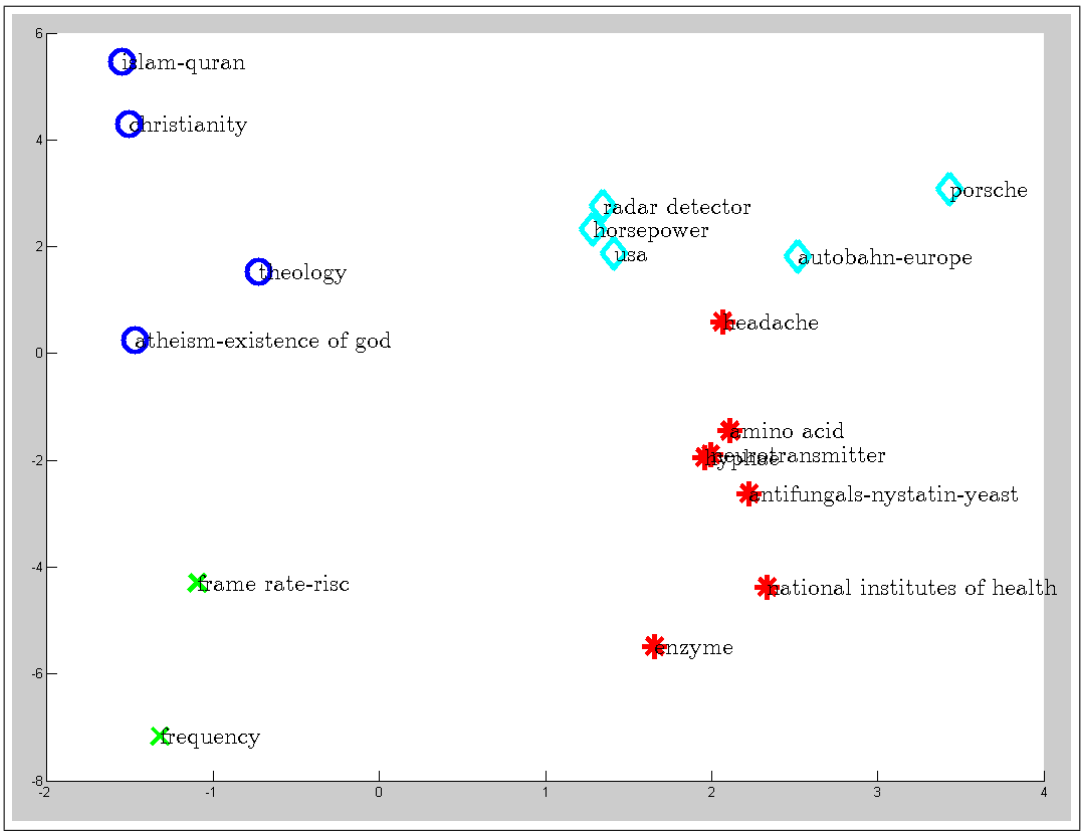
D είναι η απόσταση των νευρώνων (στο χώρο εισόδου) όπως ορίζεται από την εξίσωση 5.19,

σ' είναι μία γκαουσσιανή συνάρτηση που μειώνεται με το χρόνο (θα μπορούσε να είναι και ίδια με τη συνάρτηση σ της εξίσωσης 5.23)

Είναι εμφανής η ομοιότητα μεταξύ των εξισώσεων 5.25 και 5.24 που υλοποιούν την ανανέωση των θέσεων των νευρώνων και των διανυσμάτων τους αντίστοιχα. Η όλη διαδικασία της φάσης 2 φαίνεται στο Σχήμα 5.15. Με αυτή τη "διπλή" διαδικασία εκπαίδευσης, επιτυγχάνεται η ανανέωση τόσο των διανυσμάτων των νευρώνων όσο και των θέσεων στο δισδιάστατο χώρο, και έτσι μέχρι το τέλος της διαδικασίας εκπαίδευσης αναμένεται (α) τα βάρη των διανυσμάτων των νευρώνων να αντανακλούν τη σημαντικότητα κάθε έννοιας στον κάθε νευρώνα και (β) τη δημιουργία γειτονιών παρόμοιων νευρώνων (ανάλογα με το πόσο παρόμοια είναι τα διανύσματα βαρών τους) στο δισδιάστατο χώρο εξόδου. Οι τελικές θέσεις των νευρώνων τους Σχήματος 5.14 φαίνονται στο Σχήμα 5.16 στο οποίο γίνεται φανερό πως η εκπαίδευση οδήγησε σε πλήρη διαχωρισμό των νευρώνων (και μοιραία των κλάσεων των εγγράφων).



Σχήμα 5.15: *Περιγραφή φάσης εκπαίδευσης και ανταγωνισμού DoSO*



Σχήμα 5.16: *Παράδειγμα οπτικοποίησης νευρώνων από τη μέθοδο DoSO για 4 κλάσεις (μετά την εκπαίδευση)*

5.5.3 Τελική φάση: Εντοπισμός ομάδων και ιεραρχική δόμηση

Έπειτα από την εκπαίδευση ενός συνόλου δεδομένων με τη μέθοδο SOM είναι απαραίτητο να βρεθούν οι ομάδες στις οποίες χωρίζονται τα έγγραφα. Διάφορες τεχνικές έχουν προταθεί. Στις εργασίες των [Himberg, 2000] και [Vesanto & Alhoniemi, 2000] μία ομάδα ορίζεται ως ένα σύνολο κόμβων (νευρώνων) με μικρή απόσταση μεταξύ τους και μεγάλη απόσταση με τους υπόλοιπους, παρόλα αυτά δεν προτείνεται κάποια συγκεκριμένη μεθοδολογία για τον υπολογισμό. Στη μέθοδο των [Moutarde & Ultsch, 2005] γίνεται μία αναπαράσταση του SOM ως ένα πλέγμα-πίνακας [Kraaijveld, 1992]. Παρόλα αυτά, είναι απαραίτητο με αυτή τη μέθοδο να καθοριστούν κάποιοι νευρώνες ως σύνορα των ομάδων, κάτι το οποίο δεν είναι πολύ σαφές πως μπορεί να επιτευχθεί.

Εύρεση ομάδων με τη μέθοδο DoSO

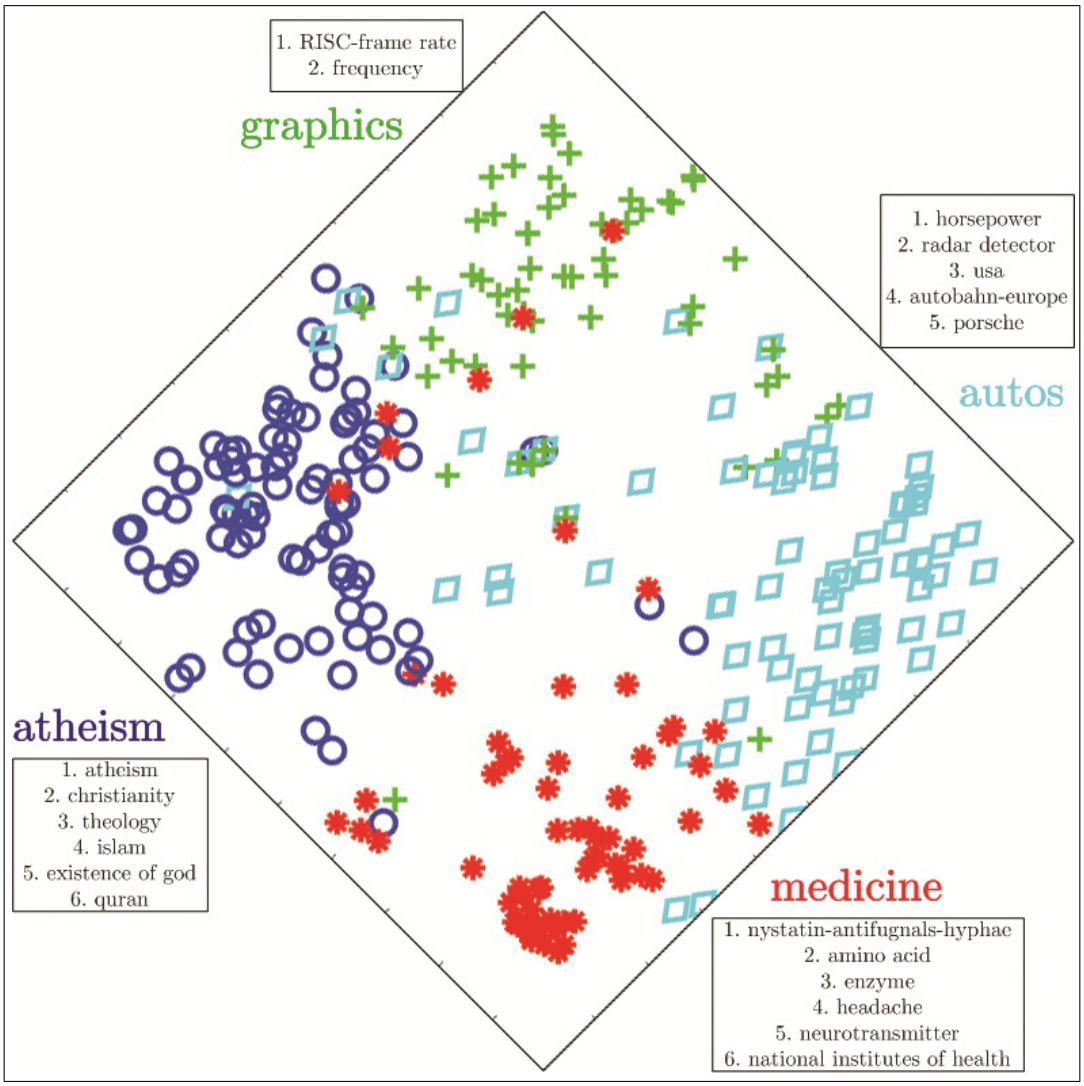
1: $Cluster\_Discovery(T)$
2: Σημείωσε όλους τους νευρώνες ως μη-επισκεφθέντες
$UN = \{Allneurons\}$
3: **while** $UN \neq \varnothing$ **do**
4: - Βρες ένα νευρώνα N_i από το UN
5: - Ξεκίνα μια καινούρια ομάδα C
6: - Κάλεσε $Cluster\_New(N, C, T)$
7: **end while**
8:
9: $Cluster\_New(C, N, T)$
10: $C = C \cup \{N\}$
11: $UN = UN - \{N\}$
12: **for all** $A \in UN$ γειτονικούς του N έτσι ώστε $Dist(N, A) \leq T$ **do**
13: Call $Cluster\_New(C, A, T)$
14: **end for**

Σχήμα 5.17: *Περιγραφή της μεθόδου εύρεσης των τελικών ομάδων με τη μέθοδο DoSO*

Η προσέγγιση που ακολουθείται για την ανακάλυψη των ομάδων περιγράφεται στο Σχήμα 5.17. Πιο συγκεκριμένα, επιλέγεται ένας αρχικός νευρώνας N από το σύνολο των μη επιλεγμένων νευρώνων (UN: αρχικά περιέχει το σύνολο των νευρώνων). Δημιουργείται μία καινούρια ομάδα C που αρχικά περιέχει το νευρώνα N (η ετικέτα του νευρώνα είναι επίσης η αρχική ετικέτα της ομάδας). Κατόπιν, αφαιρείται ο νευρώνας από το σύνολο UN και αναζητούνται εκείνοι οι νευρώνες A που ικανοποιούν δύο συνθήκες : (α) είναι γειτονικοί στον N σύμφωνα με την παράμετρο k που έχει επιλεγεί στο βήμα της οπτικοποίησης και (β) η απόστασή τους από τον N είναι μικρότερη από ένα συγκεκριμένο κατώφλι T, το οποίο καθορίζει τον αριθμό των ομάδων που τελικά θα δημιουργηθούν. Η κανονικοποίηση των αποστάσεων ώστε η μεγαλύτερη απόσταση μεταξύ δύο γειτονικών νευρώνων να είναι 1, δίνει τη δυνατότητα να παίρνει το κατώφλι T τιμές από 0 έως 1 ανεξάρτητα από την πραγματική τιμή. Αν ο νευρώνας A ανταποκρίνεται στις συνθήκες (α) και (β), τότε προστίθεται στην υπάρχουσα ομάδα C (και η ετικέτα του αντίστοιχα, προστίθεται στην ετικέτα της ομάδας) και η διαδικασία αυτή εφαρμόζεται αναδρομικά σε όλους τους γειτονικούς νευρώνες του A, αλλιώς (αν

δεν υπάρχουν άλλοι νευρώνες με απόσταση μικρότερη του κατωφλίου) επιλέγεται ένας καινούριος νευρώνας από το σύνολο UN και ξεκινά μια καινούρια ομάδα. Προφανώς, όσο πιο μεγάλη είναι η τιμή του T, τόσο μικρότερος είναι ο αριθμός των ομάδων. Με την επιλογή διαφορετικών τιμών για το T, μπορεί να βρεθεί κάθε φορά μία διαφορετική ομαδοποίηση με χρήση του αλγορίθμου του Σχήματος 5.17.

Αξίζει να τονιστεί πως οι αποστάσεις μεταξύ των νευρώνων μιας ομάδας δε χρειάζεται απαραίτητα να είναι μικρότερες από το κατώφλι που έχει καθοριστεί. Παρόλα αυτά, κάθε νευρώνας πρέπει να συνδέεται με οποιονδήποτε άλλο της ομάδας με ένα μονοπάτι που κάθε ακμή του αναπαριστά απόσταση μικρότερη από το κατώφλι. Ο αλγόριθμος που περιγράφηκε παραπάνω μπορεί να εφαρμοστεί είτε σε ένα πλήρως εκπαιδευμένο SOM, είτε ενδιάμεσα σαν μία διαδικασία παρακολούθησης της εξέλιξης της ομαδοποίησης ή και σαν έλεγχος παύσης της εκπαίδευσης.

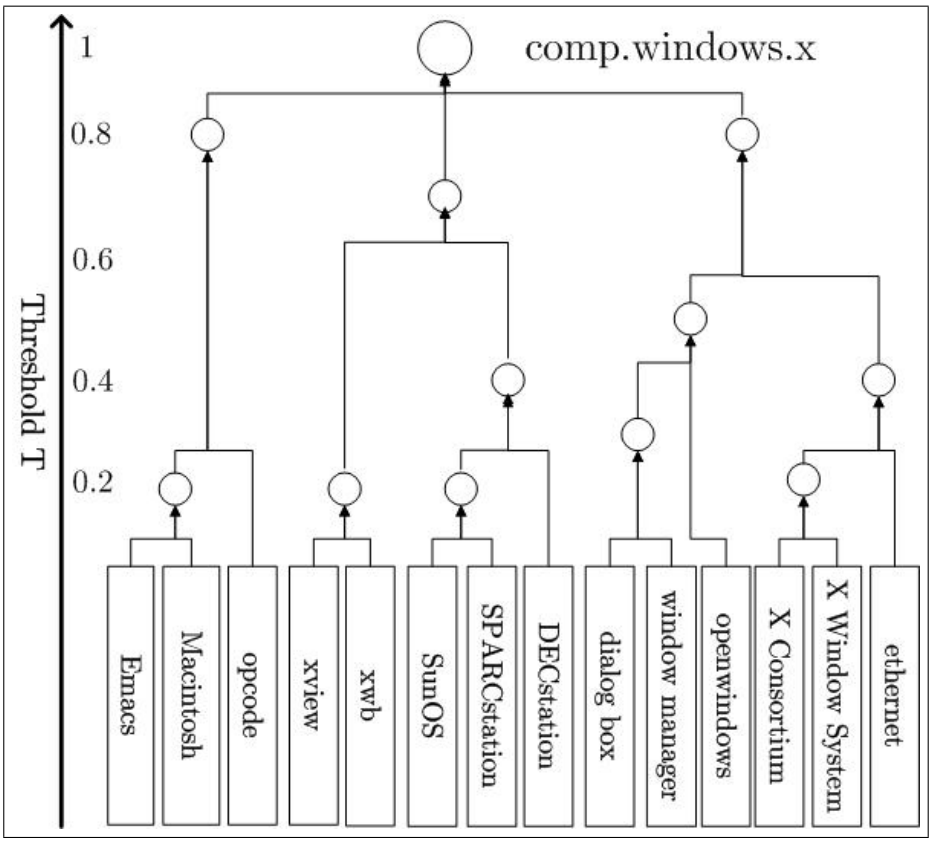


Σχήμα 5.18: *Παράδειγμα ομαδοποίησης εγγράφων με τη μέθοδο DoSO σε έναν τετραδιάστατο σημασιολογικό χώρο*

Επίσης, στο Σχήμα 5.18 έχει γίνει μια προσπάθεια αποτύπωσης ενός σημασιολογικού χώρου (simplex) κάνοντας την απλοποιητική παραδοχή πως υπάρχουν μόνο 4 κλάσεις (atheism,medicine,autos,graphics), οι οποίες και αποτελούν τις τέσσερις διαστάσεις του χώρου αυτού και καθεμιά καταλαμβάνει μια κορυφή. Στην πραγματικότητα,

ο σημασιολογικός χώρος έχει προφανώς πολύ μεγαλύτερη διάσταση (όσες και οι σημασίες) κάνοντας την απεικόνιση δύσκολη, όμως ακόμα και μέσα από αυτό το παράδειγμα φαίνεται πως λειτουργεί. Δίπλα σε κάθε κορυφή φαίνονται οι ετικέτες των νευρώνων που τελικά ανατίθενται στην κάθε κλάση (ή ομάδα) και μέσα στο simplex φαίνονται τα έγγραφα ανάλογα με την απόσταση από κάθε κορυφή (ή κλάση). Το Σχήμα αυτό δείχνει το γεγονός πως οι κλάσεις *atheism* και *medicine* είναι πιο συμπαγείς και ομογενείς (οι έννοιες των εγγράφων που τις αποτελούν είναι πολύ συγκεκριμένες και σπάνια εμφανίζονται σε έγγραφα άλλων κλάσεων) σε σχέση με τις κλάσεις *graphics* και *autos* που παρουσιάζουν αρκετές επικαλύψεις μεταξύ τους (λόγω του ότι έχουν αρκετά κοινές έννοιες.

Με την επιλογή διαφορετικών κατωφλίων T, είναι εύκολο να ανακτηθεί μία ιεραρχική δομή για τις ομάδες στη συλλογή εγγράφων που εξετάζεται. Η επιλογή μικρών τιμών T οδηγεί σε πολλές ομάδες (που περιέχουν λίγα έγγραφα) αλλά μεγαλύτερες τιμές οδηγούν σε μεγαλύτερες ομάδες (που εμπεριέχουν τις μικρότερες), επομένως δημιουργείται ένα ιεραρχικό δέντρο που είναι άμεσα διαθέσιμο. Ένα παράδειγμα της ιεραρχίας για την ομάδα εγγράφων windows.x του συνόλου 20-Newsgroup φαίνεται στο Σχήμα 5.19. Στο χαμηλότερο επίπεδο ($T = 0$) κάθε νευρώνας αντιστοιχεί σε μια διαφορετική ομάδα και καθώς ανεβαίνουμε στην ιεραρχία, η τιμή του κατωφλίου T αυξάνεται, οδηγώντας σε συγχωνεύσεις νευρώνων (σημειώνονται με τους κύκλους).



Σχήμα 5.19: *Παράδειγμα ιεραρχικής δομής για την κατηγορία windows.x που παράγεται με τη μέθοδο DoSO*

5.6 Αποτελέσματα ομαδοποίησης με τη μέθοδο DoSO

Όπως και στη μεθοδολογία CHC, έτσι και στη μέθοδο DoSO χρησιμοποιήθηκαν τα σύνολα δεδομένων και τα μέτρα επίδοσης που περιγράφονται στις Παραγράφους 5.4.1 και 5.4.2 αντίστοιχα. Ειδικά για τους αυτο-οργανούμενους χάρτες θα χρησιμοποιηθούν κάποια επιπλέον μέτρα επίδοσης που αξιολογούν τη λειτουργία τους.

5.6.1 Μέτρα αξιολόγησης αυτο-οργανούμενων χαρτών

Τα πρόσθετα μέτρα που επιλέγονται είναι τρία αρκετά δημοφιλή στη βιβλιογραφία της αξιολόγησης αυτο-οργανούμενων χαρτών [Pölzlbauer, 2004]: το *σφάλμα κβαντισμού (quantization error)*, το *τοπογραφικό γινόμενο* (*topographic product*) και το *μέτρο παραμόρφωσης του SOM* (*SOM distortion*).

Το *σφάλμα κβαντισμού* (QE) παραδοσιακά εφαρμόζεται σε όλες τις μορφές διανυσματικού κβαντισμού και σε αλγορίθμους ομαδοποίησης. Επομένως, το μέτρο αυτό αγνοεί την τοπολογία του χάρτη αλλά αξιολογεί την ποιότητα των ομάδων χωρίς τη χρήση των κατηγοριών του αρχικού συνόλου δεδομένων. Το QE υπολογίζεται λαμβάνοντας υπόψιν τη μέση απόσταση των διανυσμάτων των εγγράφων που έχουν ανατεθεί σε ένα νευρώνα σε σχέση με τα κεντροειδή που τους αναπαριστούν. Τυπικά, το σφάλμα κβαντισμού για κάθε νευρώνα i ορίζεται από την ακόλουθη εξίσωση :

$$QE_i = \frac{1}{M_i} \sum_{x_j \in i} ||m_i - x_j|| \tag{5.27}$$

όπου:

i αναφέρεται στο νευρώνα i,

M_i αναφέρεται σε όλα τα έγγραφα που ανατίθενται στο νευρώνα i,

x_j αναφέρεται στο διάνυσμα του εγγράφου j,

m_i αναφέρεται στο διάνυσμα του νευρώνα i

Μετά τον υπολογισμό του QE_i για κάθε νευρώνα του SOM, ο μέσος αυτών των τιμών δίνει το σφάλμα κβαντισμού QE για ολόκληρο το SOM.

Το *τοπογραφικό γινόμενο* είναι ένα από τα παλιότερα μέτρα που ποσοτικοποιούν τις αναλογίες διατήρησης της τοπολογίας από τον αυτο-οργανούμενο χάρτη. Το αποτέλεσμα του υπολογισμού του τοπογραφικού γινομένου δείχνει τελικά εάν ο χάρτης έχει το κατάλληλο μέγεθος. Ένα άλλο χαρακτηριστικό του είναι πως χρησιμοποιούνται μόνο τα διανύσματα νευρώνων και όχι το αρχικό σύνολο. Η βασική ιδέα του μέτρου αυτού είναι η σύγκριση των γειτονιών μεταξύ δύο νευρώνων σε σχέση με τις θέσεις τους στο χάρτη από τη μία ($Q_2(j,k)$) και στα διανύσματα που τους αναπαριστούν από την άλλη ($Q_1(j,k)$) :

$$Q_1(j,k) = \frac{dist(m_j, m_{n_k^A(j)})}{dist(m_j, m_{n_k^V(j)})} \tag{5.28}$$

$$Q_2(j,k) = \frac{dist(r_j, r_{n_k^A(j))}}{dist(r_j, r_{n_k^V(j)})} \tag{5.29}$$

όπου:

j αναφέρεται στο νευρώνα j,

m_j αναφέρεται στο διάνυσμα του νευρώνα j,

r_j αναφέρεται στο διάνυσμα θέσης του νευρώνα j,

$n_k^A(j)$ αναφέρεται στον k-πλησιέστερο γείτονα του j στο χώρο εισόδου V,

$n_k^V(j)$ αναφέρεται στον k-πλησιέστερο γείτονα του j στο χάρτη A,

$dist$ είναι μια συνάρτηση απόστασης (π.χ. Ευκλείδεια).

Η παράμετρος k πρέπει να επιλεγεί από το χρήστη και καθορίζει μέχρι σε ποιο βαθμό θα γίνουν οι υπολογισμοί των γειτόνων (π.χ. για $k = 4$, οι 4 πλησιέστεροι γείτονες ελέγχονται).

Συνδυάζοντας τις εξισώσεις 5.28 και 5.29 προκύπτει η τοπογραφική σχέση μεταξύ του νευρώνα j και των k-πλησιέστερων γειτόνων :

$$P_3(j,k) = \left(\prod_{l=1}^{k} Q_1(j,l) \cdot Q_2(j,l) \right)^{\frac{1}{2k}} \tag{5.30}$$

Το μέτρο επεκτείνεται σε κάθε νευρώνα του SOM και το τοπογραφικό γινόμενο P τελικά ορίζεται ως εξής :

$$P = \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{k=1}^{N-1} log(P_3(j,k)) \tag{5.31}$$

Η τιμή του P μπορεί εύκολα να ερμηνευθεί ως εξής: Εάν $P << 0$, ο χάρτης είναι πολύ μικρός (δηλαδή έχει λίγους νευρώνες) ενώ εάν $P >> 0$, ο χάρτης είναι πολύ μεγάλος για το σύνολο που εξετάζεται.

Τέλος, χρησιμοποιείται το *μέτρο παραμόρφωσης*. Έχει αποδειχθεί πως εάν η ακτίνα της γειτονιάς παραμένει σταθερή, τότε υπάρχει μια συνάρτηση κόστους την οποία το SOM ελαχιστοποιεί, κάτι το οποίο εκφράζεται από το μέτρο παραμόρφωσης. Αυτή η συνάρτηση μπορεί να χρησιμοποιηθεί για τον υπολογισμό μιας συνάρτησης σφάλματος για ολόκληρο το χάρτη. Επίσης, ένα πλεονέκτημα είναι πως το σφάλμα μπορεί να αποσυντεθεί με διάφορους τρόπους (ανά νευρώνα, ώστε το σφάλμα να περιορίζεται τοπικά στο χάρτη ή ανά συνιστώσα). Τυπικά ορίζεται ως εξής :

$$DM = \sum_{j=1}^{M} \sum_{i=1}^{N} h_{b_j,i} ||m_i - x_j|| \tag{5.32}$$

όπου:

M είναι ο αριθμός των εγγράφων,

N είναι ο αριθμός των νευρώνων,

m_i είναι το διάνυσμα του νευρώνα i,

x_j είναι το διάνυσμα του εγγράφου j,

b_j είναι ο νευρώνας-νικητής για το έγγραφο που αναπαρίσταται από το διάνυσμα x_j,

$h_{b_j,i}$ είναι η συνάρτηση γειτνίασης όπως ορίζεται από την εξίσωση 5.23

5.6.2 Αποτελέσματα για τη μέθοδο DoSO

Η μεθοδολογία DoSO ελέγχθηκε τόσο για το τελικό αποτέλεσμα της ομαδοποίησης (με βάση τα μέτρα F1, εντροπία και R,J,FM) όσο και για την αξιολόγηση της ως παραλλαγή μοντέλου SOM και μεθόδου οπτικοποίησης (με βάση τα μέτρα Σφάλμα Κβαντισμού, Τοπογραφικό Γινόμενο και Παραμόρφωση).

5.6.2.1 Επιλογή παραμέτρων για τη μέθοδο DoSO

Για την επιβεβαίωση των παραμέτρων του Πίνακα 5.2 έγινε ο ίδιος έλεγχος στα ίδια σύνολα δεδομένων και με τη μεθοδολογία DoSO, ώστε να επιβεβαιωθεί η βαρύτητα κάθε χαρακτηριστικού στη σημασία του εγγράφου. Λόγω του ότι δεν υπάρχει συγκεκριμένος τρόπος να προβλεφθεί η τιμή για καθεμία από τις παραμέτρους α, β, γ του μοντέλου αναπαράστασης εγγράφων, ο μοναδικός τρόπος καθορισμού τους είναι ο εμπειρικός πειραματισμός σε σχέση με κάποιο μέτρο επίδοσης [Guyon & Elisseeff, 2003]. Αρχικά διαχωρίζεται ένα ανεξάρτητο σύνολο δεδομένων. Τα υπόλοιπα δεδομένα χρησιμοποιούνται τόσο για εκπαίδευση όσο και για την επιλογή παραμέτρων. Στα πειράματα που έγιναν, υλοποιήθηκε διασταυρωμένη επικύρωση (k-fold cross-validation) (με το k να τίθεται στο 5) σε κάθε ένα από τα 3 σύνολα δεδομένων που περιγράφηκαν παραπάνω. Τα έγγραφα κάθε συνόλου δεδομένων χωρίζονται σε 5 (περίπου) ισομεγέθεις ομάδες. Έτσι, γίνονται 5 επαναλήψεις μάθησης και επικύρωσης έτσι ώστε κάθε φορά ένα διαφορετικό μέρος των δεδομένων να μένει εκτός εκπαίδευσης για επικύρωση, ενώ τα υπόλοιπα 4 να χρησιμοποιούνται για τη μάθηση. Είναι αυτονόητο πως τα έγγραφα χωρίστηκαν με τέτοιο τρόπο ώστε να εξασφαλίζεται πως σε κάθε ένα από τα 5 τμήματα να υπάρχουν αντιπροσωπευτικά δείγματα όλης της συλλογής εγγράφων και πως όλα τα έγγραφα χρησιμοποιούνται για επικύρωση. Τέλος, το ανεξάρτητο σύνολο που έμεινε εκτός αρχικά, χρησιμοποιείται για τη συνολική αξιολόγηση της επίδοσης του μοντέλου, ξανά με χρήση του πρωτοκόλλου της διασταυρωμένης επικύρωσης.

Οι βέλτιστες τιμές παραμέτρων της εξίσωσης 5.4 είναι αυτές που παράγουν τα καλύτερα αποτελέσματα ομαδοποίησης σε σχέση με τις τιμές για το μέτρο-F και την εντροπία (στα σύνολα εκπαίδευσης και επικύρωσης) και φαίνονται στον Πίνακα 5.4 για κάθε ένα από τα 3 σύνολα δεδομένων που χρησιμοποιήθηκαν. Η παράμετρος $LinkRank$ έχει τη μεγαλύτερη επίδραση στην αναπαράσταση των εγγράφων (επιβεβαιώνοντας τη βαθειά και σημασιολογικά πλούσια δομή της Wikipedia, ενώ η παράμετρος $OrderRank$ έχει τη μικρότερη. Η παράμετρος $ConceptSim$ είναι σημαντικότερη από την παράμετρο $WFreq$ εκτός από το σύνολο Brown (όπου λόγω του μεγάλου μεγέθους των εγγράφων το κλασσικό μοντέλο $tf - idf$ είναι αρκετά αποτελεσματικό).

Πίνακας 5.4: *Βέλτιστες παράμετροι μεθόδου DoSO*

Βέλτιστες τιμές και αποκλίσεις				
Parameter		**20-NG**	**Reuters**	**Brown**
Wfreq	a	0.2520 ± 0.0101	0.2320 ± 0.0031	0.256 ± 0.0051
LinkRank	b	0.3800 ± 0.0025	0.4120 ± 0.0036	0.3920 ± 0.0033
OrderRank	c	0.1000 ± 0.0000	0.1000 ± 0.0000	0.1000 ± 0.0000
ConceptSim	1-a-b-c	0.2680 ± 0.1114	0.2560 ± 0.0034	0.2520 ± 0.0026
MinFreq		0.001-0.01 (μεγάλα σύνολα), 0.01-0.03 (μικρά σύνολα)		
MinKeyph		0.5		

Μετά την ολοκλήρωση της αναπαράστασης εγγράφων στο χώρο των εννοιών και τον υπολογισμό των βαρών των παραμέτρων, εφαρμόζεται η κυρίως μεθοδολογία του DoSO όπως περιγράφτηκε στην Παράγραφο 5.5. Επιλέγονται οι αρχικοί νευρώνες βάσει των παραμέτρων (MinKeyph) και (MinFreq) ώστε να δημιουργηθούν περιγραφικές ετικέτες για τους νευρώνες (και άρα για τις τελικές ομάδες). Όπως και στη μέθοδο των πιο σημαντικών εννοιών έτσι και εδώ επιλέγεται μία σχετικά μεγάλη τιμή για την παράμετρο $Keyphraseness$ και μία σχετικά χαμηλή τιμή για την παράμετρο $MinKeyph$ βάσει των τιμών των παραμέτρων που φαίνονται στους Πίνακες 5.2 και 5.4.

Όπως περιγράφηκε στην Παράγραφο 5.5.1, η μέθοδος DoSO χρησιμοποιεί την προβολή ISOMAP ώστε να οπτικοποιηθεί η ομαδοποίηση των νευρώνων/εγγράφων. Η μέθοδος ISOMAP χρησιμοποιεί την παράμετρο k για τον καθορισμό των κοντινότερων γειτόνων στο δημιουργούμενο διδιάστατο γράφο. Η παράμετρος k συνιστά στον αριθμό των γειτόνων (βάσει Ευκλείδειας απόστασης που η μέθοδος ISOMAP θα χρησιμοποιήσει σαν συνδέσεις στο δημιουργούμενο γράφο γειτνίασης και γενικά μπορεί να τεθεί στην τιμή μεταξύ 4 (για την αναπαράσταση ενός τετραγωνικού πλέγματος) και 6 (για την αναπαράσταση εξαγωνικού πλέγματος). Ας σημειωθεί εδώ πως μετά την αρχικοποίηση των νευρώνων και την προβολή τους, εξάγεται μία τοπολογία νευρώνων (όπως με τον κλασσικό αλγόριθμο του SOM) και στο σημείο αυτό η παράμετρος k χρησιμοποιείται για τον καθορισμό του τύπου αυτού του πλέγματος.

5.6.2.2 Αποτελέσματα Ομαδοποίησης μεθόδου DoSO

Η μεθοδολογία εντοπισμού των ομάδων βάσει των νευρώνων περιγράφηκε στην Παράγραφο 5.5.3. Το τελικό αποτέλεσμα της ομαδοποίησης συγκρίνεται με αυτό που παράγουν οι αλγόριθμοι HAC και k-means (τόσο για το χώρο BOW όσο και τον χώρο των εννοιών για τα 3 σύνολα δεδομένων που αναφέρθηκαν και παρουσιάζεται στον Πίνακα 5.5. Αρχικά, η επίδραση της εισαγωγής του μοντέλου των εννοιών στους τυπικούς αλγορίθμους ομαδοποίησης (HAC και $k-means$) εξετάζεται συγκρίνοντας τα αποτελέσματα του κλασσικού μοντέλου BOW με (α) το χώρο των εννοιών και τη χρήση του απλού μοντέλου $tf-idf$ ($CS-TFIDF$) και (β) το χώρο των εννοιών μέσα από τη χρήση βαρών για τις παραμέτρους της εξίσωσης 5.4 ($CS-WEIGHTED$). Τέλος, η προτεινόμενη μεθοδολογία (DoSO), (που χρησιμοποιεί το χώρο των εννοιών) ελέγχθηκε για τα αποτελέσματα που παράγει. Για την καλύτερη παρουσίαση των αποτελεσμάτων του πίνακα 5.5, έχουν επισημανθεί με **έντονη γραφή** τα καλύτερα αποτελέσματα για όλους τους αλγορίθμους ενώ με *πλάγια γραφή* η καλύτερη τιμή για κάθε αλγόριθμο της βιβλιογραφίας (HAC και $k-means$).

Όπως φαίνεται από τον Πίνακα 5.5, η εισαγωγή του χώρου των εννοιών στους αλγορίθμους HAC και $k-means$ δίνει καλύτερα αποτελέσματα για τα μέτρα $F1, R, J$ και FM με χρήση του μοντέλου $CS-WEIGTHED$ (και όχι του $CS-TFIDF$), λόγω του γεγονότος πως ο χώρος των εννοιών είναι περισσότερο χρήσιμος μέσα από τη στάθμιση των βαρών που καθορίζουν τις έννοιες λόγω της μεγάλης διαθεσιμότητας πληροφορίας από τη Wikipedia και όχι απλά μετρώντας εμφανίσεις. Παρόλα αυτά, φαίνεται πως το μοντέλο BOW οδηγεί σε αποτελέσματα με χαμηλότερες τιμές εντροπίας (πιο ομογενείς ομάδες) για 2 από τα 3 σύνολα δεδομένων (20-NG και Reuters). Αυτό οφείλεται κυρίως στο γεγονός πως τα έγγραφα του συνόλου Brown είναι αρκετά μεγαλύτερα σε μέγεθος από αυτά των 20-NG και Reuters, το οποίο οδηγεί σε πιο ακριβείς αναπαραστάσεις στο χώρο των εννοιών. Επιπλέον, η τιμή της εντροπίας ευνοεί αλγορίθμους που παράγουν ομάδες με σχετικά ομοιόμορφα μεγέθη (όπως ο αλγόριθμος k-$means$) και τιμωρούν αλγορίθμους που παράγουν ομάδες με μεγάλη διαφορά

Πίνακας 5.5: *Αποτελέσματα μεθόδου DoSO βάσει μέτρων ομαδοποίησης*

Μέτρα				Βάσει συνόλων		Βάσει συνεμφανίσεων			Βάσει εντροπίας
Σύνολο	Αλγόριθμος	Μοντέλο αναπαράστασης	Διάσταση	F1-macro	F1-micro	R	J	FM	E
20-Newsgroup	HAC	BOW	61174	0.542	0.558	0.87	0.70	0.74	*0.205*
		CS-TFIDF	22510	0.486	0.514	0.91	0.75	*0.78*	0.391
		CS-WEIGHTED	22510	*0.557*	*0.569*	*0.92*	*0.78*	0.76	0.216
	k-means	BOW	61174	0.604	0.612	0.95	0.73	0.75	**0.114**
		CS-TFIDF	22510	0.473	0.481	0.92	0.71	*0.78*	0.314
		CS-WEIGHTED	22510	*0.613*	*0.627*	*0.95*	*0.79*	0.75	0.199
	DoSO	CS-WEIGHTED	22510	**0.799**	**0.831**	**0.96**	**0.81**	**0.84**	**0.114**
Reuters	HAC	BOW	18985	0.417	0.437	0.65	0.56	0.62	*0.196*
		CS-TFIDF	10714	0.402	0.440	0.67	0.61	0.68	0.283
		CS-WEIGHTED	10714	*0.424*	*0.471*	*0.74*	*0.63*	*0.69*	0.207
	k-means	BOW	18985	0.392	0.426	0.72	0.65	0.69	***0.092***
		CS-TFIDF	10714	0.377	0.412	0.62	0.50	0.57	0.236
		CS-WEIGHTED	10714	*0.441*	*0.485*	*0.77*	*0.69*	*0.73*	0.187
	DoSO	CS-WEIGHTED	10714	**0.792**	**0.828**	**0.92**	**0.83**	**0.83**	0.119
Brown	HAC	BOW	59601	0.479	0.484	0.87	*0.69*	0.64	0.231
		CS-TFIDF	17880	0.454	0.470	0.90	0.66	*0.65*	0.240
		CS-WEIGHTED	17880	*0.494*	*0.498*	*0.91*	0.62	0.61	*0.162*
	k-means	BOW	59601	0.461	0.476	*0.88*	0.65	0.62	0.191
		CS-TFIDF	17880	0.443	0.467	0.84	0.61	0.66	0.238
		CS-WEIGHTED	17880	*0.472*	*0.487*	0.87	*0.67*	*0.69*	*0.123*
	DoSO	CS-WEIGHTED	17880	**0.784**	**0.845**	**0.92**	**0.78**	**0.71**	**0.112**

στα μεγέθη τους [Xiong et al., 2004] (όπως ο αλγόριθμος DoSO). Τέλος, πρέπει να αναφερθεί πως η εισαγωγή του χώρου των εννοιών οδηγεί σε πολύ χαμηλότερη πολυπλοκότητα χρόνου, λόγω του γεγονότος πως ο νέος χώρος είναι πιο συμπιεσμένος από τον απλό χώρο. Η μείωση στο διάνυσμα αναπαράστασης φαίνεται επίσης στον Πίνακα 5.5. Για την προτεινόμενη μεθοδολογία χρησιμοποιήθηκαν οι βέλτιστες παράμετροι του Πίνακα 5.4. Για τον αλγόριθμο HAC, υλοποιήθηκε η παραλλαγή UPGMA ενώ για τον αλγόριθμο k-means η τιμή k τέθηκε στο 8 και το κατώφλι ομοιότητας στην τιμή 0.25.

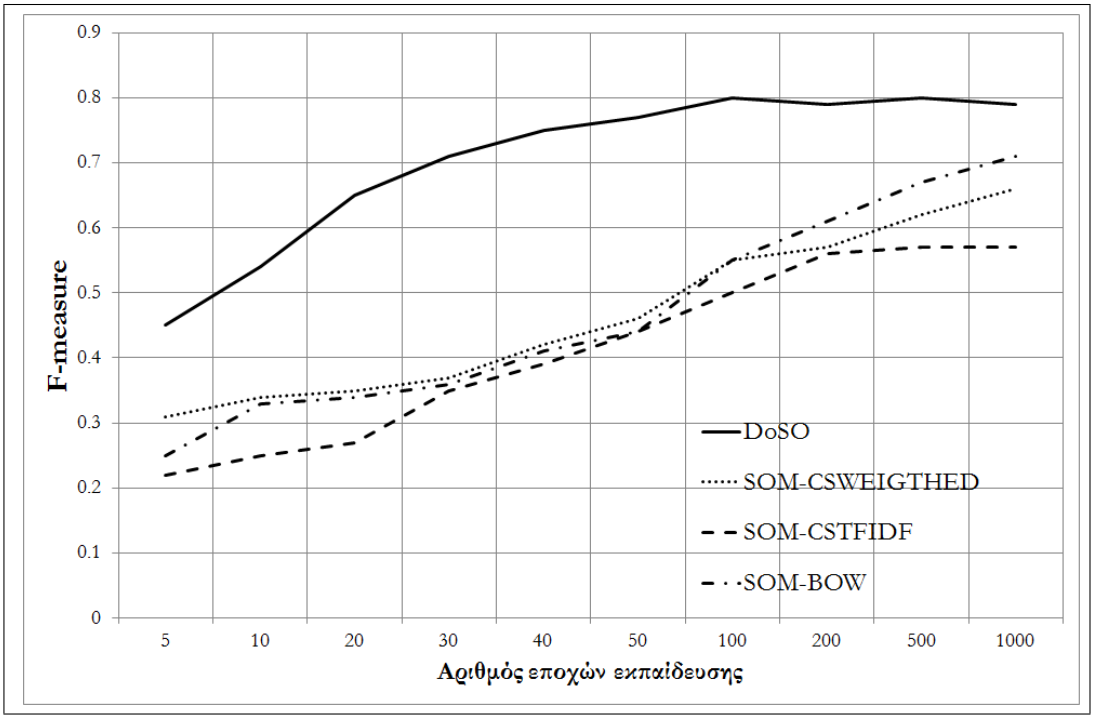
Η επίδοση του DoSO συγκρίνεται επίσης με τον κλασσικό αλγόριθμο του SOM που εφαρμόζεται και στα 3 σύνολα δεδομένων χρησιμοποιώντας τόσο το κλασσικό μοντέλο BOW όσο και το μοντέλο του χώρου των εννοιών, ώστε να αξιολογηθεί η μέθοδος βάσει των κριτηρίων για αυτο-οργανούμενους χάρτες. Τα αποτελέσματα φαίνονται στον Πίνακα 5.6. Έγιναν συγκρίσεις για τις 4 διαφορετικές προσεγγίσεις του SOM (απλό μοντέλο tf-idf, χώρος εννοιών tf-idf, χώρος εννοιών με στάθμιση βαρών και DoSO) με χρήση διαφορετικών μεγεθών χαρτών. Τα αποτελέσματα ήταν ανεξάρτητα από το μέγεθος του χάρτη. Ο Πίνακας 5.6 δείχνει τις μέσες τιμές των μέτρων για όλα τα μεγέθη. Το DoSO παρουσιάζει καλύτερα αποτελέσματα όσον αφορά στο Τοπογραφικό Γινόμενο το οποίο σημαίνει (σύμφωνα με την Παράγραφο 5.4.2) ότι το μέγεθος του χάρτη (δηλαδή ο αριθμός των νευρώνων) είναι κατάλληλο για το πρόβλημα. Ο κλασσικός αλγόριθμος του SOM αποδίδει καλύτερα από το DoSO σε κάποιες περιπτώσεις αλλά οι διαφορές θεωρούνται μικρές ειδικά αν ληφθεί υπόψιν και η βελτίωση στο χρόνο εκπαίδευσης του SOM λόγω της κατάλληλης αρχικοποίησης (όπως περιγράφηκε στην παράγραφο 5.5.1).

Επίσης, στο Σχήμα 5.20 φαίνεται η απόδοση των παραπάνω μεθόδων SOM (SOM/BOW, SOM/CS-TFIDF, SOM/CS-WEIGHTED, DoSO) σε σχέση με τον αριθμό των εποχών εκπαίδευσης και το καλύτερο αποτέλεσμα που επετεύχθη για αυτό τον αριθμό εποχών (όσον αφορά στο F-μέτρο αλλά με παρόμοια αποτελέσματα για τα υπόλοιπα μέτρα που αναφέρθηκαν) σε ένα υποσύνολο του Reuters. Πέραν από τη σαφή υπεροχή της προτεινόμενης μεθόδου DoSO που αποδίδει πολύ καλά επιτυγχάνοντας σχεδόν διπλάσιο μέτρο F ακόμα και με πολύ μικρό αριθμό εποχών (λόγω της ιδιαίτερης αρχικοποίησης που γίνεται στα βάρη των νευρώνων) φαίνεται πως η εισαγωγή του χώρου

Πίνακας 5.6: *Αποτελέσματα μεθόδου DoSO βάσει της δημιουργούμενης τοπολογίας*

Σύνολο	Αλγόριθμος	Μοντέλο αναπαράστασης	QE	Τοπογραφικό Γινόμενο	Παραμόρφωση SOM
20-Newsgroup	SOM	BOW	**0.1221**	0.0010	**0.3932**
		CS-TFIDF	1.2264	-0.0104	1.6130
		CS-WEIGHTED	1.1073	0.0023	1.2543
	DoSO	CS-WEIGHTED	0.1435	**0.0007**	0.7319
Reuters	SOM	BOW	0.2932	0.0044	1.3692
		CS-TFIDF	0.3339	-0.0081	1.6423
		CS-WEIGHTED	0.3431	0.0076	1.5219
	DoSO	CS-WEIGHTED	**0.1134**	**0.0027**	**1.0244**
Brown	SOM	BOW	0.3993	0.0082	**0.1423**
		CS-TFIDF	1.2342	-0.0092	7.3219
		CS-WEIGHTED	2.0081	0.0040	1.2108
	DoSO	CS-WEIGHTED	**0.1723**	**0.0025**	0.8492

των εννοιών οδηγεί το κλασσικό SOM σε πολύ ικανοποιητικά αποτελέσματα (βλέπε SOM/CS-WEIGHTED) με λίγες εποχές σε σχέση με το κλασσικό SOM/BOW. Είναι προφανές πως η σημασιολογία που εισάγεται με τις έννοιες έχει σημαντικά αποτελέσματα σε σχέση με τις απλές εμφανίσεις λέξεων του μοντέλου BOW. Καθώς μεγαλώνει ο αριθμός των εποχών, το κλασσικό SOM/BOW φαίνεται πως υπερισχύει του SOM/CS-WEIGHTED αλλά αφενός η διαφορά απόδοσης είναι μικρή και αφετέρου εισάγεται και η εξισορρόπηση (trade-off) μεταξύ χρόνου εκτέλεσης/απόδοσης.



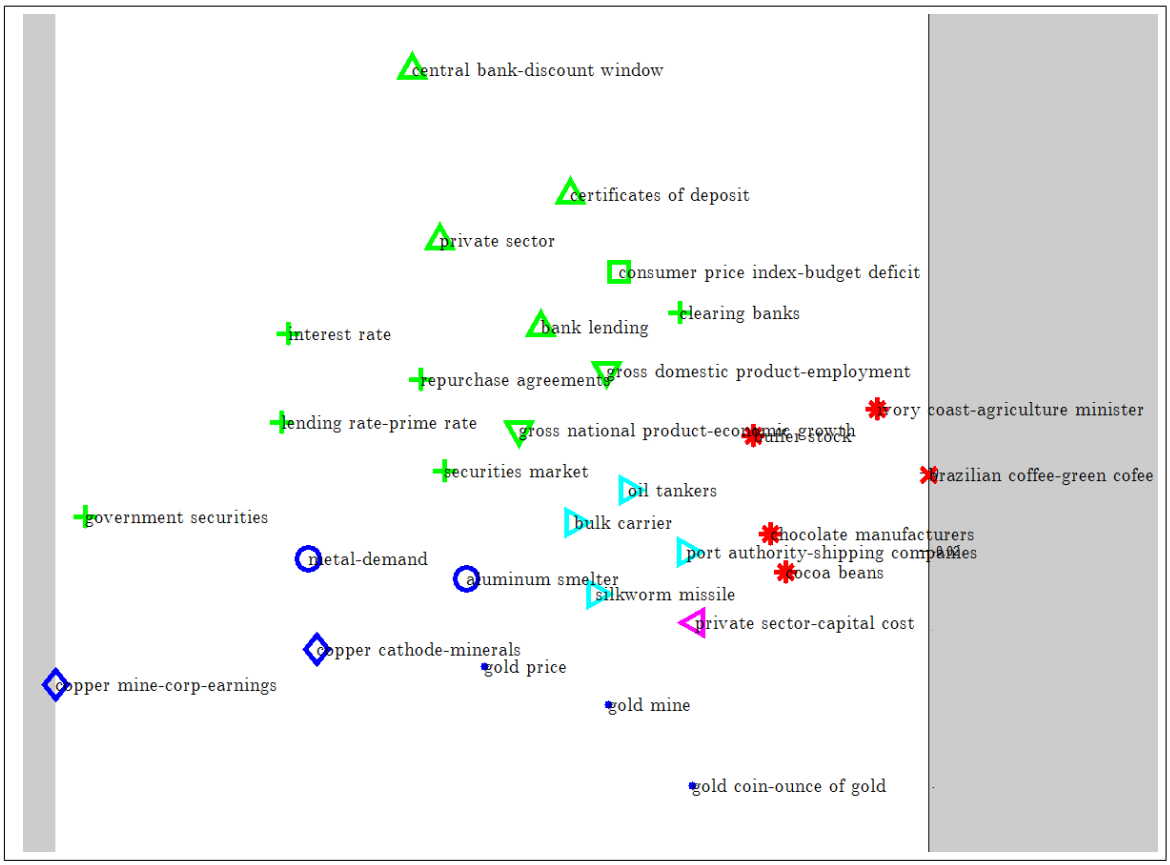
Σχήμα 5.20: *Συγκριτική απόδοση SOM/BOW, SOM/CS-TFIDF, SOM/CS-WEIGHTED, DoSO σε σχέση με τον αριθμό των εποχών εκπαίδευσης*

Τέλος, έγινε σύγκριση του DoSO με τους αλγορίθμους ομαδοποίησης που κάνουν χρήση εξωτερικής πηγής γνώσης και έχουν τα καλύτερα αποτελέσματα στη βιβλιογραφία. Το μοναδικό σύνολο δεδομένων που είναι κοινό και επομένως επιτρέπει την όποια σύγκριση είναι ένα υποσύνολο του συνόλου Reuters που περιέχει 1658 έγγραφα από 30 διαφορετικές κλάσεις (επιλέγονται κλάσεις με περισσότερα από 15 και λιγότερα από 200 έγγραφα). Τα αποτελέσματα φαίνονται στον Πίνακα 5.7 (προηγούμενα αποτελέσματα συλλέχθηκαν από την εργασία των Kiran et. al [Kiran, 2010]). Το DoSO παρουσιάζει καλύτερα αποτελέσματα από όλες τις προηγούμενες μεθόδους σε όλα τα μέτρα που χρησιμοποιήθηκαν.

Πίνακας 5.7: *Σύγκριση DoSO και άλλων μεθόδων ομαδοποίησης βάσει εξωτερικής γνώσης*

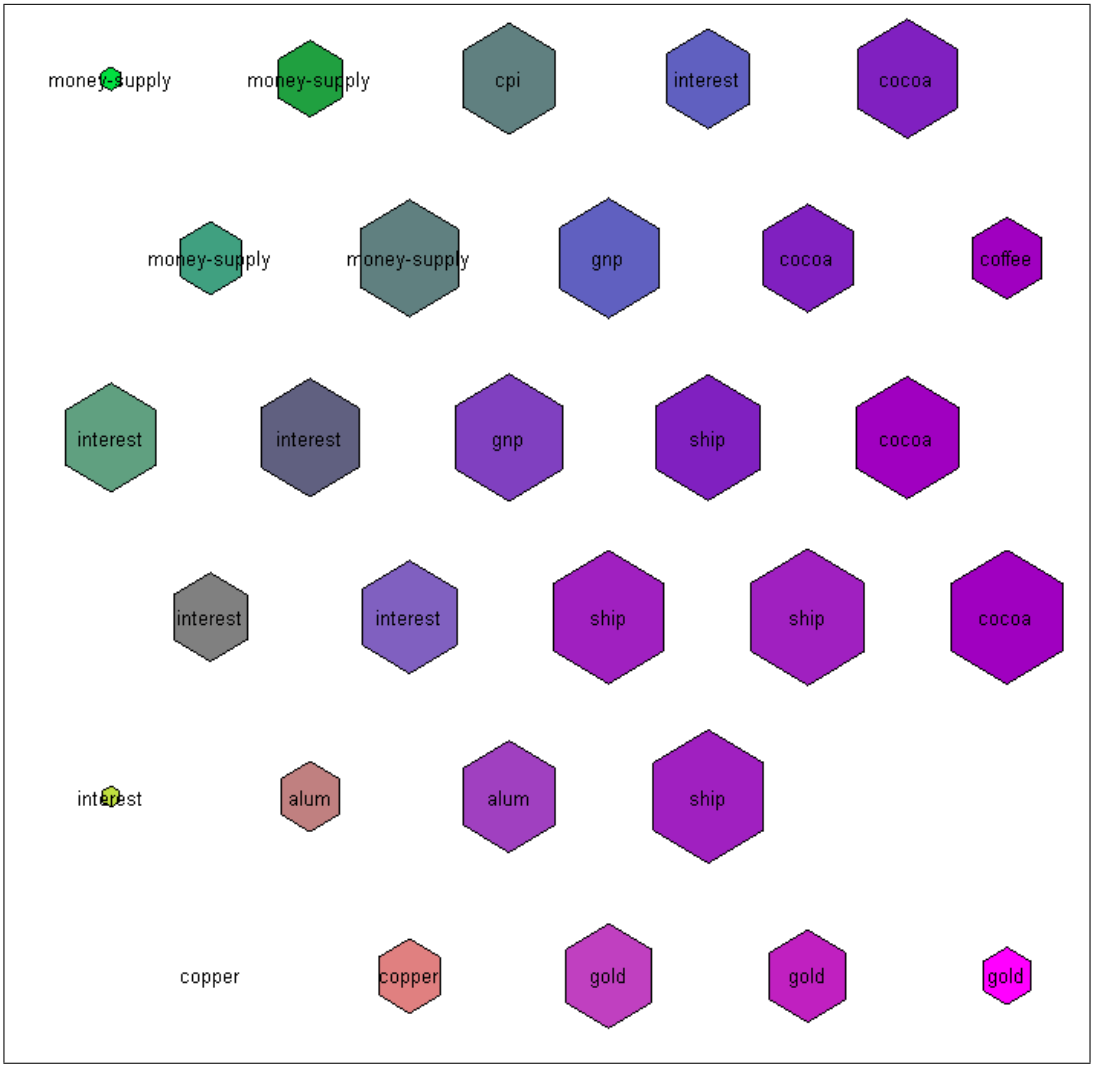
	F-measure	**Purity**	**Inv. Purity**	**Πηγή γνώσης**
BOW	0.618	0.603	0.544	-
Garbilovich et al.	-	0.605	0.548	Wikipedia
Hotho et al.	-	0.607	0.556	WordNet
Hu et al.	-	0.655	0.598	Wikipedia
Huang et al.	0.575	0.678	0.75	Wikipedia
Kiran et al.	0.732	0.684	0.778	Wikipedia, WordNet, dmoz, social bookmark
DoSO	**0.799**	**0.781**	**0.817**	Wikipedia

Ένα παράδειγμα οπτικοποίησης (μετά το πέρας της εκπαίδευσης) φαίνεται στο Σχήμα 5.21. Οι κατηγορίες είναι οι εξής: alum, cocoa, coffee, copper, cpi, gnp, gold, interest, money-supply, ship. Φαίνεται πως μετά το πέρας της εκπαίδευσης οι νευρώνες που αντιπροσωπεύουν γειτονικές κλάσεις βρίσκονται κοντά μεταξύ τους, εν αντιθέσει με νευρώνες άλλων κλάσεων. Πιο συγκεκριμένα, υπάρχει και τοπολογικό νόημα, καθώς έχουν τοποθετηθεί οι νευρώνες που ασχολούνται με "χρηματο-οικονομικά" ζητήματα στο επάνω μέρος, ενώ οι νευρώνες που ασχολούνται με "πρώτες ύλες κτλ" στο κάτω μέρος.



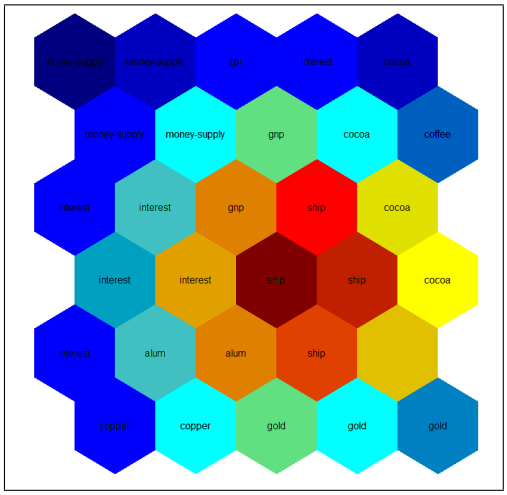
Σχήμα 5.21: *Παράδειγμα θέσεων νευρώνων DoSO στο δισδιάστατο επίπεδο (υποσύνολο Reuters)*

Επίσης, στο Σχήμα 5.22 έχει χρησιμοποιηθεί η κλασσική σύμβαση του U-matrix όπου σε κάθε νευρώνα έχει αποτυπωθεί το μέγεθος ανάλογα με το ποσοστό των εγγράφων που αναλαμβάνει ο νευρώνας (Σε κάθε νευρώνα έχει αποτυπωθεί και η κυρίαρχη κλάση).

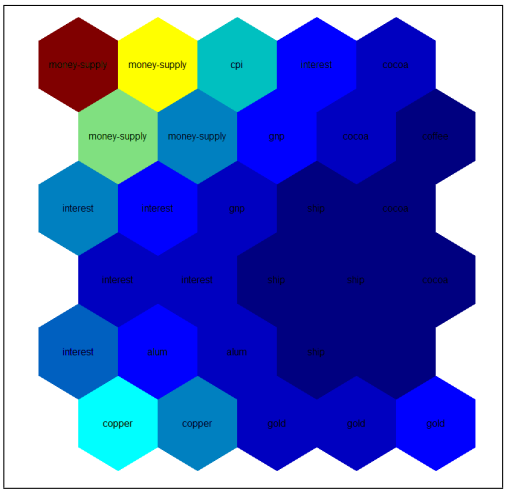


Σχήμα 5.22: *Παράδειγμα οπτικοποίησης DoSO σε υποσύνολο του συνόλου Reuters*

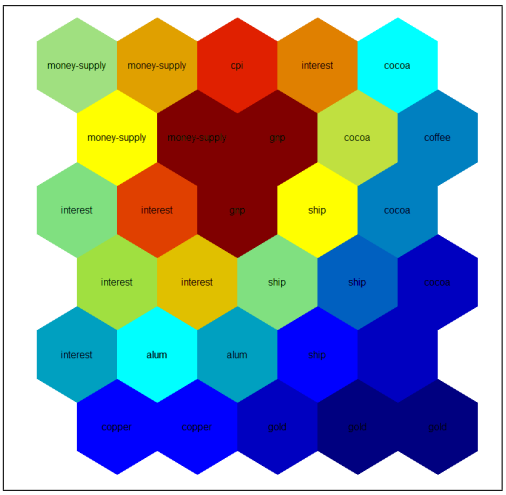
Τέλος, στα Σχήματα 5.23 έως 5.26 φαίνεται η σημαντικότητα 4 διαφορετικών εννοιών (bulk cargo, discount window, finance minister, L.M.E. (London Metal Exchange)) στους παραπάνω νευρώνες (σε αυτούς ακολουθείται η κλασσική σύμβαση των U-Matrix δηλαδή το ανοιχτό χρώμα αντιστοιχεί σε υψηλή τιμή βάρους ενώ το σκούρο χρώμα αντιστοιχεί σε χαμηλή τιμή βάρους).



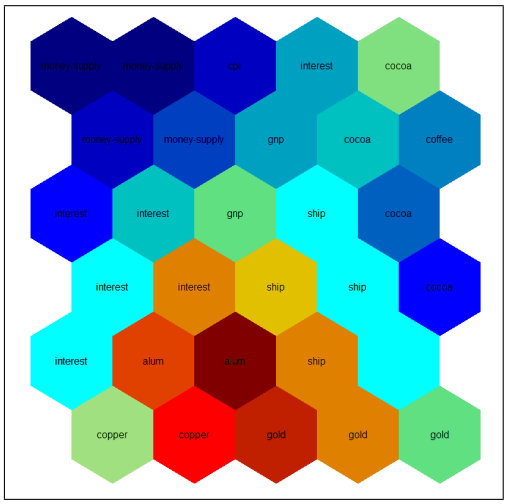
Σχήμα 5.23: *Παράδειγμα DoSO: Σημαντικότητα του όρου bulk cargo στους νευρώνες*



Σχήμα 5.24: *Παράδειγμα DoSO: Σημαντικότητα του όρου discount window στους νευρώνες*



Σχήμα 5.25: *Παράδειγμα DoSO: Σημαντικότητα του όρου finance minister στους νευρώνες*



Σχήμα 5.26: *Παράδειγμα DoSO: Σημαντικότητα του όρου L.M.E. στους νευρώνες*

□

Κεφάλαιο 6

Συνολικό Πόρισμα Διατριβής

6.1 Γενικά Συμπεράσματα

Η παρούσα διατριβή εστιάζει σε προβλήματα ανάλυσης και αναζήτησης κειμένου. Σημείο εκκίνησης είναι η παρατήρηση των πολλών πηγών κειμενικής πληροφορίας που υπάρχουν σήμερα (και οφείλονται εν πολλοίς και στην εκρηκτική ανάπτυξη του WWW): Ο Παγκόσμιος Ιστός αλλά και οι προσωπικοί υπολογιστές του καθενός κατακλύζονται από μεγάλες ποσότητες κειμένου (σε ελεύθερη μορφή) ενώ σπάνιες είναι οι περιπτώσεις που η πληροφορία αυτή παρέχεται σε δομημένη μορφή ή με επιπλέον πληροφορία (επισημάνσεις (annotations), ετικέτες (tags) κλπ). Ο κύριος τρόπος πρόσβασης σε αυτές τις μεγάλες ποσότητες κειμένων παραμένει η εξόρυξη κειμένου (text mining) με απλές και ευφυείς τεχνικές. Στην προσπάθεια αυτή οργάνωσης της πληροφορίας σημαντική είναι η βοήθεια από το Σημασιολογικό Ιστό (Semantic Web) που μέσα από τις οντολογίες παρέχει τη δυνατότητα καλύτερης οργάνωσης της διαθέσιμης πληροφορίας. Παρόλα αυτά, οι περισσότερες σελίδες στον Παγκόσμιο Ιστό παραμένουν σε ελεύθερη μορφή (χωρίς δηλαδή δυνατότητα άμεσης αξιοποίησης στα πλαίσια του Σημασιολογικού ιστού όπως και τα περισσότερα προσωπικά ψηφιακά έγγραφα. Έτσι, ανακύπτει η ανάγκη για ανάπτυξη αποτελεσματικών ευφυών τεχνικών που θα μπορούν να εφαρμοστούν σε κείμενα ελεύθερης μορφής ώστε να επιτευχθεί αποδοτική οργάνωση και αποτελεσματική αναζήτησή τους.

Στα δύο πρώτα Κεφάλαια της διατριβής έγινε επισκόπηση των δυσκολιών που αντιμετωπίζονται κατά την προσπάθεια αναπαράστασης εγγράφων με χρήση του υπολογιστή και οφείλονται στην πολυπλοκότητα και τη συνεχή εξέλιξη της γλώσσας. Τα ζητήματα που αντιμετωπίζει κανείς κατά την προσπάθεια αναπαράστασης εγγράφων στον υπολογιστή είναι κυρίως δύο: ποια μονάδα του κειμένου (χαρακτήρας, μόρφημα, λέξη, φράση, πρόταση κτλ) θα χρησιμοποιηθεί ως βάση για την αναπαράσταση και πως θα ενσωματωθεί η περισσότερη δυνατή πληροφορία από την αρχική μορφή του κειμένου. Η κύρια (και πιο φυσική ίσως στον άνθρωπο) βασική μονάδα αναπαράστασης εγγράφων παραμένει μέχρι και σήμερα η λέξη. Προβλήματα από την αναπαράσταση αυτή υπάρχουν πολλά: η μεταφορική χρήση λέξεων αλλάζει τα νοήματα, οι σημασιολογικές σχέσεις μεταξύ λέξεων (συνωνυμία, αντωνυμία κτλ) δεν καλύπτονται, η προηγούμενη γνώση που συνήθως έχουν οι άνθρωποι δεν είναι δυνατό να μεταφερθεί εύκολα στον υπολογιστή, η θέση της λέξης σε σχέση με την περικείμενη πληροφορία (context) αλλάζει τη σημασία της κτλ. Όμως, ακόμη και σήμερα η λογική των λέξεων κυριαρχεί στις κειμενικές αναπαραστάσεις κάτι το οποίο αποδεικνύεται από το ότι η αναζήτηση στον Παγκόσμιο Ιστό εξακολουθεί και γίνεται με λέξεις-κλειδιά (και όχι με ερωτήσεις

φυσικής γλώσσας ή άλλο τρόπο) αλλά και οι περισσότεροι άνθρωποι αναζητούν ένα έγγραφο βάσει των λέξεων που τους ενδιαφέρουν. Έτσι, δίνεται έμφαση στο βασικό μοντέλο αναπαράστασης που βασίζεται στις λέξεις και είναι το Μοντέλο Χώρου Διανυσμάτων (Vector Space Model, VSM ή Bag-of-Words, BOW) το οποίο εξετάζεται αναλυτικά.

Αναγνωρίζοντας το ρόλο των λέξεων στην αναπαράσταση εγγράφων, η διατριβή εξετάζει τη δυνατότητα εντοπισμού των σχέσεων μεταξύ λέξεων. Δύο λέξεις μπορεί να σχετίζονται με πολλούς τρόπους (συνώνυμες, αντώνυμες, υπερώνυμες, μερώνυμες κτλ). Στόχος της μεθοδολογίας που περιγράφεται στο Κεφάλαιο 3 είναι η κατασκευή ενός βαθμωτού μέτρου που θα υπολογίζει γρήγορα τη σχετικότητα δύο λέξεων λαμβάνοντας υπόψιν όλες τις δυνατές σχέσεις που μπορεί να έχουν αυτές. Γίνεται διάκριση δηλαδή ανάμεσα στην ομοιότητα λέξεων και στη σχετικότητα λέξεων, καθώς η δεύτερη καλύπτει μεγαλύτερο εύρος σχέσεων και επί της ουσίας αξιολογεί κατά πόσον δύο λέξεις είναι "κοντά" ή όχι μεταξύ τους. Για το σκοπό αυτό, αξιοποιούνται πληροφορίες που εξάγονται από τα αποτελέσματα μιας μηχανής αναζήτησης στο WWW για τις λέξεις των οποίων αναζητείται η σχετικότητα καθώς και το λεξικό WordNet, που παρέχει μεγάλο αριθμό σχετιζομένων λέξεων. Η μέθοδος είναι υβριδική αξιοποιώντας τόσο παραδοσιακές τεχνικές ανάκτησης κειμένου (βάσει του μοντέλου BOW) και μία μηχανή διανυσμάτων υποστήριξης (SVM) ποσοτικοποιώντας κατ΄ αυτό τον τρόπο γρήγορα και εύκολα τη σχέση δύο οποιωνδήποτε λέξεων. Τα αποτελέσματα δείχνουν την υπεροχή της μεθοδολογίας που προτείνεται έναντι των υπολοίπων της βιβλιογραφίας και τις δυνατότητες που υπάρχουν στην εκμετάλλευση της πληροφορίας που υπάρχει στα αποτελέσματα των μηχανών αναζήτησης αλλά και του WordNet.

Οι αδυναμίες του μοντέλου BOW παραμένουν παρά την εισαγωγή περαιτέρω γνώσης (μέσω των αποτελεσμάτων αναζήτησης και του WordNet) ενώ όπως είναι φυσικό η αναπαράσταση κειμένων με λέξεις που συνδέονται μεταξύ τους με ένα βαθμωτό μέγεθος περιορίζει τις δυνατότητες περαιτέρω βελτίωσης. Γιαυτό το λόγο, στο Κεφάλαιο 4 η διατριβή, αφού εξετάσει τις διάφορες δυνατότητες αναπαράστασης εγγράφων, παρουσιάζει ένα νέο μοντέλο αναπαράστασης που αξιοποιεί τη Wikipedia με στόχο να εντοπίσει επώνυμες οντότητες (named entities) που αποτελούν έννοιες (concepts) και υπάρχουν ως άρθρα της Wikipedia αλλά εμφανίζονται και στο έγγραφο. Με αυτή τη διαδικασία εισάγεται περισσότερη σημασιολογική πληροφορία σε κάθε έγγραφο, αφού για κάθε έννοια που εμφανίζεται στη Wikipedia αξιοποιούνται πλήρως οι πληροφορίες που παρέχονται (περιεχόμενο του άρθρου, σύνδεσμοι (εξερχόμενοι και εισερχόμενοι), ιεραρχική κατηγοριοποίηση άρθρων κτλ) και συνδυάζονται με τα κλασσικά γνωρίσματα της ανάκτησης κειμένου (συχνότητα, σειρά εμφάνισης κτλ) και κατασκευάζονται χαρακτηριστικά τα οποία περιγράφουν πλήρως τις έννοιες του εγγράφου αναθέτοντας βάρος σε καθεμιά. Ιδιαίτερη μνεία αξίζει να γίνει στο χαρακτηριστικό *Keyphraseness* το οποίο δείχνει την ικανότητα μιας έννοιας να περιγράψει αντιπροσωπευτικά ένα έγγραφο ή μια ομάδα εγγράφων. Το νέο αυτό μοντέλο δίνει δυνατότητες αναπαράστασης οντοτήτων που με διαφορετικό τρόπο δε θα ήταν δυνατό να κατανοηθούν ή να βρεθούν από τον υπολογιστή (ομάδες λέξεων, συντομογραφίες κτλ) μαζί με χαρακτηριστικά που αξιολογούν τη σημασία τους σε κάθε έγγραφο αλλά και γενικά ενώ συμπιέζει σημαντικά το μεγάλο χώρο που εισάγεται με το μοντέλο BOW, καταδεικνύοντας έτσι τα πλεονεκτήματα του μοντέλου αυτού.

Έχοντας ως εργαλείο τη νέα μορφή αναπαράστασης που παρουσιάστηκε, στη συνέχεια η διατριβή εξετάζει στο Κεφάλαιο 5 μετά τη σχετικότητα λέξεων και τη σχετικότητα εγγράφων. Επί της ουσίας, αναζητείται η δυνατότητα χωρισμού μιας μεγάλης

συλλογής εγγράφων σε ομάδες (ομαδοποίηση, clustering) μέσω αυτόματα καθοριζόμενων ετικετών που θα περιγράφουν (σημασιολογικά) τις ομάδες αυτές. Στον πυρήνα των δύο μεθοδολογιών ομαδοποίησης που αναπτύσσονται βρίσκεται το μοντέλο αναπαράστασης και οι έννοιες που εντοπίζονται βάσει της Wikipedia. Η πρώτη μεθοδολογία αξιοποιεί την ιδέα των πιο συχνών και πιο περιγραφικών εννοιών (βάσει της τιμής Keyphraseness) ώστε να χρησιμεύσουν ως ετικέτες που περιγράφουν τις ομάδες που εγγράφων δημιουργούνται ιεραρχικά. Η βασική παρατήρηση που προκύπτει από τη μεθοδολογία αυτή είναι πως οι ομάδες που δημιουργούνται (και ιεραρχικά καταλήγουν σε ολόκληρη τη συλλογή) πέραν του ότι έχουν αρκετά περιγραφικές ετικέτες (το οποίο αποτελεί σημαντικό πλεονέκτημα έναντι οποιωνδήποτε άλλων μεθόδων ομαδοποίησης), επί της ουσίας αντιστοιχούν στα ευρύτερα θέματα με τα οποία ασχολείται η συλλογή των εγγράφων (πραγματοποιείται δηλαδή εξαγωγή θέματος (topic extraction». Η δεύτερη μεθοδολογία αξιοποιεί τους Αυτο-Οργανούμενους Χάρτες (SOM) ώστε πέραν της δημιουργίας ομάδων εγγράφων με περιγραφικές ετικέτες να τοποθετηθούν οι ομάδες αυτές στο χώρο με τοπολογική ορθότητα λαμβάνοντας υπόψιν τις σημασιολογικές τους σχέσεις. Ο αλγόριθμος του SOM τροποποιείται δομικά ώστε να καλύψει την αρχικοποίηση των νευρώνων βάσει των σημαντικών εννοιών της συλλογής εγγράφων αλλά και να επιταχύνει την εκπαίδευση. Το αποτέλεσμα είναι να λαμβάνονται οπτικοποιήσεις των νευρώνων του SOM στο δισδιάστατο επίπεδο που απεικονίζουν πλήρως τις σημασιολογικές σχέσεις ανάμεσα στις ομάδες που χωρίζονται τα έγγραφα, εντοπίζοντας έτσι τα θέματα με τα οποία ασχολείται η συλλογή εγγράφων που εξετάζεται. Η εισαγωγή του μοντέλου των εννοιών βελτιώνει σημαντικά την απόδοση των υπαρχόντων αλγορίθμων (k-means, HAC, SOM), ενώ οι τροποποιημένοι αλγόριθμοι που προτείνονται (CHC, DoSO) επειδή ακριβώς σχεδιάστηκαν για κειμενικά δεδομένα παρουσιάζουν ακόμη καλύτερα αποτελέσματα (α) ως προς το κομμάτι της ορθής ομαδοποίησης των εγγράφων, (β) ως προς το κομμάτι της οπτικοποίησης παρέχοντας δυνατότητες καλύτερης πλοήγησης σε μία μεγάλη συλλογή εγγράφων και (γ) ως προς τις απαιτήσεις χρόνου για την αποδοτική λειτουργία των μεθοδολογιών.

Αξίζει να σημειωθεί πως η διατριβή παρουσιάστηκε μέσα από μία πορεία από το (απλό) επίπεδο των λέξεων στο πιο (πολύπλοκο) επίπεδο των εννοιών προσφέροντας έτσι τη δυνατότητα αξιοποίησης διαφορετικών δυνατοτήτων κάθε φορά, ανάλογα με το πρόβλημα που εξετάζεται (ή τις συνθήκες). Για παράδειγμα, πολλές εφαρμογές (όπως οι μηχανές αναζήτησης) εξακολουθούν να λειτουργούν με λέξεις-κλειδιά, επομένως ενδέχεται να είναι περισσότερο χρήσιμο ένα σύστημα καθορισμού της σημασιολογικής συσχέτισης λέξεων. Σε πιο σύνθετα προβλήματα (όπως αυτά που ασχολούνται με την οργάνωση εγγράφων) αλλά και για την ανάπτυξη πιο σύνθετων εφαρμογών (όπως ερωτήσεις σε φυσική γλώσσα) φαίνεται πως είναι μονόδρομος ένα σύστημα που θα βασίζεται σε πλουσιότερες σημασιολογικά μονάδες όπως οι έννοιες.

Ανακεφαλαιώνοντας, το γενικό συμπέρασμα, που αβίαστα εξάγεται από τη συνολική έρευνα, είναι πως στη σημερινή εποχή του διαρκώς αυξανόμενου όγκου πληροφοριών, είναι απαραίτητες νέες τεχνικές ευφυούς ανάλυσης, οργάνωσης και αναζήτησης κειμένων και εγγράφων. Η συνεισφορά της διατριβής στο συγκεκριμένο ζήτημα συνίσταται στην αξιοποίηση των δυνατοτήτων που παρέχονται από τις διαθέσιμες πηγές γνώσης για τον εμπλουτισμό των εγγράφων κάτι που δίνει τη δυνατότητα αναπαράστασής τους με τρόπο που ταιριάζει καλύτερα στον ανθρώπινο τρόπο σκέψης. Εξετάζονται οι σχέσεις σε διάφορα επίπεδα, ξεκινώντας από το πιο απλό (και συνηθισμένο) επίπεδο των λέξεων, συνεχίζοντας στο επίπεδο των εννοιών και καταλήγοντας στο επίπεδο των σχέσεων μεταξύ εγγράφων. Η συνεργασία αυτών των διαφόρων ενισχυμένων σημα-

σιολογικά μορφών αναπαράστασης που εισάγονται με ευφυή συστήματα, προσφέρουν λύσεις σε αρκετά ζητήματα που απασχολούν το χώρο της ανάκτησης και ανάλυσης εγγράφων και καταδεικνύουν τις δυνατότητες που υπάρχουν για αξιοποίηση των μεγάλων ποσοτήτων κειμένου που παρέχονται ελεύθερα τη σημερινή εποχή.

6.2 Μελλοντικές Επεκτάσεις

Από την παρούσα διατριβή προκύπτουν διάφορα ζητήματα τα οποία θα μπορούσαν να εξεταστούν στη συνέχεια. Το πεδίο έρευνας στο συγκεκριμένο τμήμα είναι ανοιχτό σε νέες εργασίες.

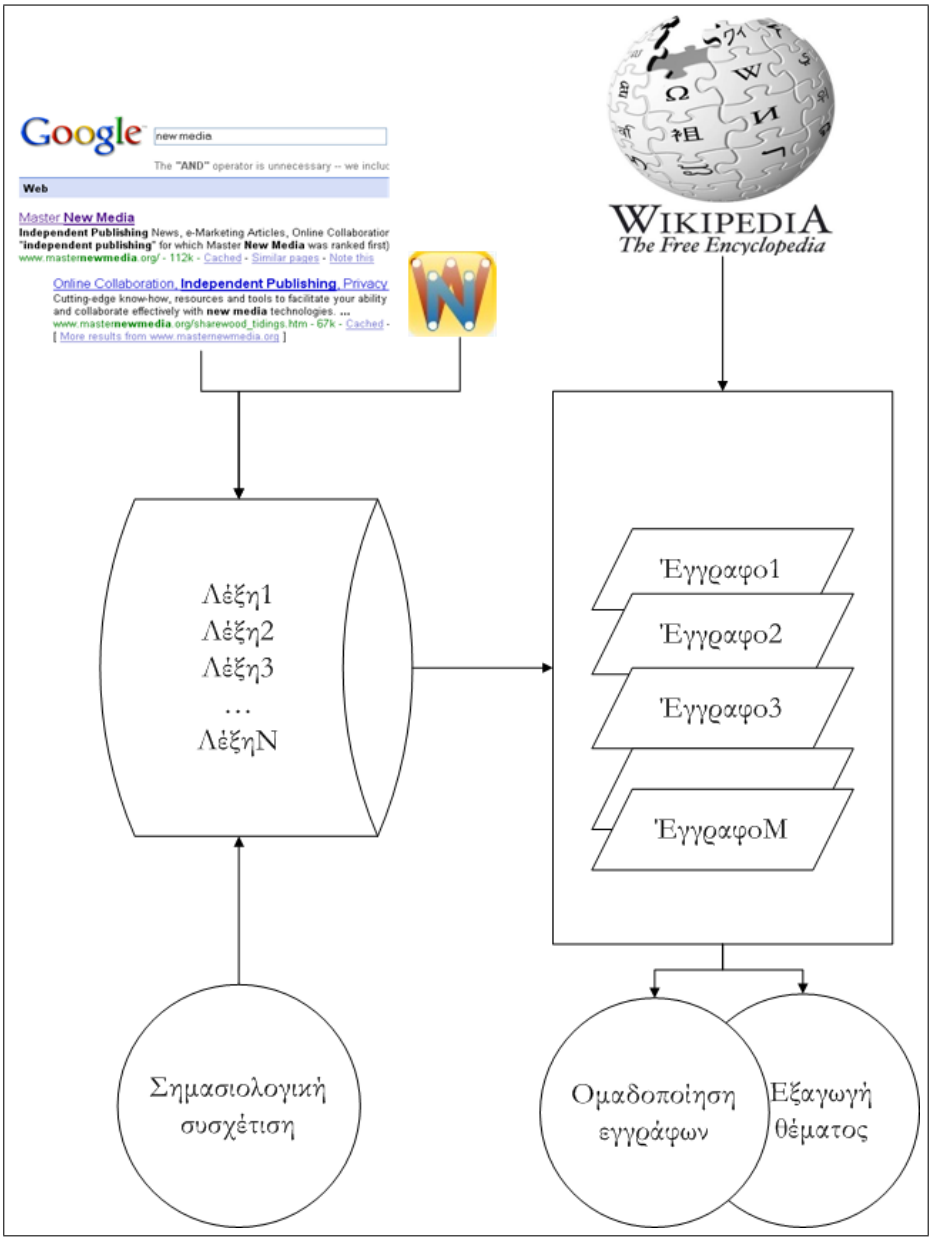
Η μεθοδολογία προσδιορισμού σημασιολογικής συσχέτισης λέξεων που παρουσιάστηκε στο Κεφάλαιο 3 μπορεί να επεκταθεί προκειμένου να περιλαμβάνει μικρά κείμενα (και όχι μόνο λέξεις) καθώς και να αξιοποιήσει πληροφορίες από τη Wikipedia και όχι μόνο από το WordNet.

Η μεθοδολογία DoSO μπορεί να επεκταθεί ώστε να αποφασίζει για το βαθμό που κάθε έγγραφο ανήκει σε κάποια ομάδα (και κατ΄ επέκταση σε κάποιο θέμα). Οι περισσότερες μέθοδοι ομαδοποίησης εγγράφων αναθέτουν κάθε έγγραφο ακριβώς σε μία ομάδα (κυρίως για τον έλεγχο της αποδοτικότητας της μεθόδου). Στην πραγματικότητα όμως, αυτά τα έγγραφα ενδέχεται να περιέχουν πληροφορία από δύο (ή περισσότερες) σημασιολογικά επικαλυπτόμενες ομάδες, κάτι το οποίο είναι επιθυμητό να καθορίζεται από ένα σύστημα, δηλαδή όχι μόνο να αποφασίζεται ποια ομάδα είναι η καλύτερη αλλά και ποιος ο βαθμός που κάθε έγγραφο ανήκει σε μία ομάδα. Πέραν του καλύτερου χωρισμού των εγγράφων σε ομάδες, έτσι είναι δυνατή και η αντίστροφη διαδικασία δηλαδή ο εντοπισμός των θεμάτων με τα οποία ασχολούνται τα έγγραφα, καθορίζοντας έτσι τις σημασιολογικές πτυχές κάθε έννοιας.

Στο σημείο αυτό θα πρέπει να αναφερθεί πως ο έλεγχος όλων των μεθοδολογιών πραγματοποιήθηκε με κλασσικά σύνολα κειμενικών δεδομένων της βιβλιογραφίας με κύρια γλώσσα την αγγλική, ώστε να είναι δυνατή η σύγκριση με προηγούμενες εργασίες και ο καλύτερος έλεγχος της επίδοσης των μέτρων που εξετάστηκαν. Μεγάλο ενδιαφέρον έχει η εφαρμογή των μεθόδων στην ελληνική γλώσσα, αξιοποιώντας πηγές πληροφορίας που υπάρχουν στην ελληνική και εξετάζοντας το αποτέλεσμα. Οι αντίστοιχες μεθοδολογίες στα ελληνικά εξάλλου έχουν μεγάλο περιθώριο βελτίωσης και πειραματισμού λόγω του μικρού όγκου εργασιών που έχουν δημοσιευτεί.

Τέλος, κάτι που αναμφισβήτητα θα είναι χρήσιμο θα ήταν ο έλεγχος της δυνατότητας ενσωμάτωσης όλων των μεθοδολογιών σε ένα ενιαίο σύστημα επεξεργασίας δεδομένων κειμένου. Για την υλοποίηση των διαφόρων μεθόδων (αλλά και για τον έλεγχο τους) ήταν αναγκαίο να αναπτυχθούν ως ξεχωριστές συνιστώσες, αποτελούν όμως μέρος ενός δυνητικά ενιαίου συστήματος. Επί της ουσίας, απαιτείται ένας συνδυασμός της μεθοδολογίας εξαγωγής σημασιολογικής συσχέτισης λέξεων με το σύστημα αναπαράστασης εγγράφων ώστε εντοπιστούν έννοιες (ή λέξεις) που είναι έντονα συσχετισμένες. Κάτι τέτοιο θα βοηθήσει καταλυτικά και στην περαιτέρω βελτίωση της ομοιότητας εγγράφων μεταξύ τους. Η συνέργεια αυτή των μεθοδολογιών φαίνεται στο Σχήμα 6.1.

□



Σχήμα 6.1: *Ολόκληρη η αλληλεπίδραση των μεθοδολογιών που αναπτύχθηκαν στα πλαίσια της διατριβής*

Παράρτημα Α΄

Σύνολο επισήμανσης μερών του λόγου **Penn Treebank**

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection

VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

□

Παράρτημα Β′

Σύνολα δεδομένων για υπολογισμό σημασιολογικής συσχέτισης

Β′.1 Σύνολο δεδομένων **Similarity-353**

Λέξη 1	Λέξη 2	Τιμή (μέση)	Λέξη 1	Λέξη 2	Τιμή (μέση)
love	sex	6.77	psychology	anxiety	7
tiger	cat	7.35	psychology	fear	6.85
tiger	tiger	10	psychology	depression	7.42
book	paper	7.46	psychology	clinic	6.58
computer	keyboard	7.62	psychology	doctor	6.42
computer	internet	7.58	psychology	Freud	8.21
plane	car	5.77	psychology	mind	7.69
train	car	6.31	psychology	health	7.23
telephone	communication	7.5	psychology	science	6.71
television	radio	6.77	psychology	discipline	5.58
media	radio	7.42	psychology	cognition	7.48
drug	abuse	6.85	planet	star	8.45
bread	butter	6.19	planet	constellation	8.06
cucumber	potato	5.92	planet	moon	8.08
doctor	nurse	7	planet	sun	8.02
professor	doctor	6.62	planet	galaxy	8.11
student	professor	6.81	planet	space	7.92
smart	student	4.62	planet	astronomer	7.94
smart	stupid	5.81	precedent	example	5.85
company	stock	7.08	precedent	information	3.85
stock	market	8.08	precedent	cognition	2.81
stock	phone	1.62	precedent	law	6.65
stock	CD	1.31	precedent	collection	2.5
stock	jaguar	0.92	precedent	group	1.77
stock	egg	1.81	precedent	antecedent	6.04

fertility	egg	6.69	cup	coffee	6.58
stock	live	3.73	cup	tableware	6.85
stock	life	0.92	cup	article	2.4
book	library	7.46	cup	artifact	2.92
bank	money	8.12	cup	object	3.69
wood	forest	7.73	cup	entity	2.15
money	cash	9.15	cup	drink	7.25
professor	cucumber	0.31	cup	food	5
king	cabbage	0.23	cup	substance	1.92
king	queen	8.58	cup	liquid	5.9
king	rook	5.92	jaguar	cat	7.42
bishop	rabbi	6.69	jaguar	car	7.27
Jerusalem	Israel	8.46	energy	secretary	1.81
Jerusalem	Palestinian	7.65	secretary	senate	5.06
holy	sex	1.62	energy	laboratory	5.09
fuck	sex	9.44	computer	laboratory	6.78
Maradona	football	8.62	weapon	secret	6.06
football	soccer	9.03	FBI	fingerprint	6.94
football	basketball	6.81	FBI	investigation	8.31
football	tennis	6.63	investigation	effort	4.59
tennis	racket	7.56	Mars	water	2.94
Arafat	peace	6.73	Mars	scientist	5.63
Arafat	terror	7.65	news	report	8.16
Arafat	Jackson	2.5	canyon	landscape	7.53
law	lawyer	8.38	image	surface	4.56
movie	star	7.38	discovery	space	6.34
movie	popcorn	6.19	water	seepage	6.56
movie	critic	6.73	sign	recess	2.38
movie	theater	7.92	Wednesday	news	2.22
physics	proton	8.12	mile	kilometer	8.66
physics	chemistry	7.35	computer	news	4.47
space	chemistry	4.88	territory	surface	5.34
alcohol	chemistry	5.54	atmosphere	landscape	3.69
vodka	gin	8.46	president	medal	3
vodka	brandy	8.13	war	troops	8.13
drink	car	3.04	record	number	6.31
drink	ear	1.31	skin	eye	6.22
drink	mouth	5.96	Japanese	American	6.5
drink	eat	6.87	theater	history	3.91
baby	mother	7.85	volunteer	motto	2.56
drink	mother	2.65	prejudice	recognition	3
car	automobile	8.94	decoration	valor	5.63
gem	jewel	8.96	century	year	7.59
journey	voyage	9.29	century	nation	3.16
boy	lad	8.83	delay	racism	1.19
coast	shore	9.1	delay	news	3.31

asylum	madhouse	8.87	minister	party	6.63
magician	wizard	9.02	peace	plan	4.75
midday	noon	9.29	minority	peace	3.69
furnace	stove	8.79	attempt	peace	4.25
food	fruit	7.52	government	crisis	6.56
bird	cock	7.1	deployment	departure	4.25
bird	crane	7.38	deployment	withdrawal	5.88
tool	implement	6.46	energy	crisis	5.94
brother	monk	6.27	announcement	news	7.56
crane	implement	2.69	announcement	effort	2.75
lad	brother	4.46	stroke	hospital	7.03
journey	car	5.85	disability	death	5.47
monk	oracle	5	victim	emergency	6.47
cemetery	woodland	2.08	treatment	recovery	7.91
food	rooster	4.42	journal	association	4.97
coast	hill	4.38	doctor	personnel	5
forest	graveyard	1.85	doctor	liability	5.19
shore	woodland	3.08	liability	insurance	7.03
monk	slave	0.92	school	center	3.44
coast	forest	3.15	reason	hypertension	2.31
lad	wizard	0.92	reason	criterion	5.91
chord	smile	0.54	hundred	percent	7.38
glass	magician	2.08	Harvard	Yale	8.13
noon	string	0.54	hospital	infrastructure	4.63
rooster	voyage	0.62	death	row	5.25
money	dollar	8.42	death	inmate	5.03
money	cash	9.08	lawyer	evidence	6.69
money	currency	9.04	life	death	7.88
money	wealth	8.27	life	term	4.5
money	property	7.57	word	similarity	4.75
money	possession	7.29	board	recommendation	4.47
money	bank	8.5	governor	interview	3.25
money	deposit	7.73	OPEC	country	5.63
money	withdrawal	6.88	peace	atmosphere	3.69
money	laundering	5.65	peace	insurance	2.94
money	operation	3.31	territory	kilometer	5.28
tiger	jaguar	8	travel	activity	5
tiger	feline	8	competition	price	6.44
tiger	carnivore	7.08	consumer	confidence	4.13
tiger	mammal	6.85	consumer	energy	4.75
tiger	animal	7	problem	airport	2.38
tiger	organism	4.77	car	flight	4.94
tiger	fauna	5.62	credit	card	8.06
tiger	zoo	5.87	credit	information	5.31
psychology	psychiatry	8.08	hotel	reservation	8.03
disaster	area	6.25	shower	flood	6.03

governor	office	6.34	weather	forecast	8.34
grocery	money	5.94	man	woman	8.3
registration	arrangement	6	man	governor	5.25
arrangement	accommodation	5.41	murder	manslaughter	8.53
month	hotel	1.81	soap	opera	7.94
type	kind	8.97	opera	performance	6.88
arrival	hotel	6	life	lesson	5.94
bed	closet	6.72	focus	life	4.06
closet	clothes	8	production	crew	6.25
situation	conclusion	4.81	television	film	7.72
situation	isolation	3.88	lover	quarrel	6.19
impartiality	interest	5.16	viewer	serial	2.97
direction	combination	2.25	possibility	girl	1.94
street	place	6.44	population	development	3.75
street	avenue	8.88	morality	importance	3.31
street	block	6.88	morality	marriage	3.69
street	children	4.94	Mexico	Brazil	7.44
listing	proximity	2.56	gender	equality	6.41
listing	category	6.38	change	attitude	5.44
cell	phone	7.81	family	planning	6.25
production	hike	1.75	opera	industry	2.63
benchmark	index	4.25	sugar	approach	0.88
media	trading	3.88	practice	institution	3.19
media	gain	2.88	ministry	culture	4.69
dividend	payment	7.63	problem	challenge	6.75
dividend	calculation	6.48	size	prominence	5.31
calculation	computation	8.44	country	citizen	7.31
currency	market	7.5	planet	people	5.75
OPEC	oil	8.59	development	issue	3.97
oil	stock	6.34	experience	music	3.47
announcement	production	3.38	music	project	3.63
announcement	warning	6	glass	metal	5.56
profit	warning	3.88	aluminum	metal	7.83
profit	loss	7.63	chance	credibility	3.88
dollar	yen	7.78	exhibit	memorabilia	5.31
dollar	buck	9.22	concert	virtuoso	6.81
dollar	profit	7.38	rock	jazz	7.59
dollar	loss	6.09	museum	theater	7.19
computer	software	8.5	observation	architecture	4.38
network	hardware	8.31	space	world	6.53
phone	equipment	7.13	preservation	world	6.19
equipment	maker	5.91	admission	ticket	7.69
luxury	car	6.47	shower	thunderstorm	6.31
five	month	3.38	architecture	century	3.78
report	gain	3.63	game	defeat	6.97
investor	earning	7.13	seven	series	3.56

liquid	water	7.89	seafood	sea	7.47
baseball	season	5.97	seafood	food	8.34
game	victory	7.03	seafood	lobster	8.7
game	team	7.69	lobster	food	7.81
marathon	sprint	7.47	lobster	wine	5.7
game	series	6.19	food	preparation	6.22
line	insurance	2.69	video	archive	6.34
day	summer	3.94	start	year	4.06
summer	drought	7.16	start	match	4.47
summer	nature	5.63	game	round	5.97
day	dawn	7.53	boxing	round	7.61
nature	environment	8.31	championship	tournament	8.36
environment	ecology	8.81	fighting	defeating	7.41
nature	man	6.25			

Β΄.2 Σύνολο δεδομένων Miller-Charles

Λέξη 1	Λέξη 2	Τιμή (μέση)
asylum	madhouse	3.61
automobile	car	3.92
bird	cock	3.05
bird	crane	2.97
boy	lad	3.76
brother	lad	1.66
brother	monk	2.82
car	journey	1.16
coast	forest	0.42
coast	hill	0.87
coast	shore	3.7
cord	smile	0.13
crane	implement	1.68
food	fruit	3.08
food	rooster	0.89
forest	graveyard	0.84
furnace	stove	3.11
gem	jewel	3.84
glass	magician	0.11
implement	tool	2.95
journey	voyage	3.84
lad	wizard	0.42
magician	wizard	3.5
midday	noon	3.42
monk	oracle	1.1
monk	slave	0.55
noon	string	0.08
rooster	voyage	0.08

□

Παράρτημα Γ′

Σύνολα δεδομένων εγγράφων

Γ′.1 Σύνολο δεδομένων **20-NewsGroup**

Το 20 Newsgroups, 20-NG [Lang, 1995] περιέχει περίπου 20000 έγγραφα από τη συλλογή των άρθρων του USENET news group. Κάθε newsgroup ανήκει σε διαφορετική κατηγορία, με διαφορετικούς βαθμούς επικάλυψης (υπολογίζεται πως περίπου 4% των εγγράφων ανήκουν σε περισσότερες της μιας κατηγορίας). Η πλήρης λίστα των 20 κατηγοριών φαίνεται παρακάτω (χωρισμένα σε 6 κατηγορίες ανάλογα με το ¨ευρύτερο' θέμα τους):

comp.graphics	rec.autos	sci.crypt
comp.os.ms-windows.misc	rec.motorcycles	sci.electronics
comp.sys.ibm.pc.hardware	rec.sport.baseball	sci.med
comp.sys.mac.hardware	rec.sport.hockey	sci.space
comp.windows.x		
misc.forsale	talk.politics.misc	talk.religion.misc
	talk.politics.guns	alt.atheism
	talk.politics.mideast	soc.religion.christian

Τα έγγραφα στην αρχική τους μορφή αποτελούν τυπικές δημοσιεύσεις και αποτελούνται από επικεφαλίδες (θέμα, υπογραφές και σχόλια από άλλα άρθρα). Ένα παράδειγμα από την κατηγορία sci.med φαίνεται παρακάτω:

From: noring@netcom.com (Jon Noring)
Subject: Need Reference: Multiple Personalities Disorders and Allergies
I heard third-hand (not the best form of information) that there was recently published results of a study on Multiple-Personality-Disorder Syndrome patients revealing some interesting clues that the root cause of allergy may have a psychological trigger or basis. What I heard about this study was that in one 'personality', a MPDS patient exhibited no observable or clinical signs of inhalant allergy (scratch tests were used, according to what I heard), while in other personalities they showed obvious allergy symptoms, including testing a full ++++ on scratch tests for particular inhalants.
If this is true, it is truly fascinating.
Jon Noring
–
Charter Member — INFJ Club.

Γ΄.2 Σύνολο δεδομένων Reuters

Το Reuters-21578 [Carnegie Group & Reuters, 1997] είναι το πιο ευρέως γνωστό σύνολο δεδομένων για κατηγοριοποίηση και ομαδοποίηση εγγράφων. Πρόκειται για συλλογή εγγράφων που παρουσιάστηκαν στο πρακτορείο Reuters το 1987 και έχουν κατηγοριοποιηθεί από διάφορους ανθρώπους. Για τις ανάγκες του ελέγχου της μεθοδολογίας της παρούσας διατριβής, επιλέχθηκαν τα άρθρα τα οποία ανήκουν ακριβώς σε μία κατηγορία (αφαιρώντας δεδομένα αταξινόμητα ή έγγραφα χωρίς περιεχόμενο) καταλήγοντας με περίπου 10.000 έγγραφα που ανήκουν σε πάνω από 100 κατηγορίες.

Οι ετικέτες που δίνονται στα έγγραφα αφορούν 5 διαφορετικές κλάσεις όπως 'people', 'places' και 'topics'. Οι συνολικές κατηγορίες είναι 672 αλλά αρκετές από αυτές έχουν πολύ λίγα έγγραφα. Μερικά έγγραφα ανήκουν σε περισσότερες της μιας κατηγορίες, άλλα ακριβώς σε μία και κάποια σε καμία. Ένα παράδειγμα εγγράφου (στην αρχική μορφή που διατίθεται) φαίνεται παρακάτω:

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="11055" NEWID="6142">
<DATE>17-MAR-1987 13:38:58.45</DATE>
<TOPICS><D>grain</D><D>corn</D></TOPICS>
<PLACES><D>south-africa</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<TEXT><TITLE>SOUTH AFRICA CROP WEATHER SUMMARY – USDA/NOAA</TITLE><DATELINE> WASHINGTON, March 17 - </DATELINE>
<BODY>Dry weather pushed further into South Africa's Orange Free State's Maize Triangle in the week ended March 14, the Joint Agricultural Weather Facility of the U.S. Agriculture and Commerce Departments said.
In a summary of its Weather and Crop Bulletin, the agency said scattered showers continued throughout Transvaal, but dry pockets persisted in the northeast and south.
Temperatures average one to four degrees C above normal throughout all grain areas, stressing grain-filling corn in areas receiving lightest rainfall, it said.
The agency said rainfall during February was near to above normal in most areas, but earlier periods of hot, dry weather reduced yield prospects in parts of the northern Transvaal and southern Orange Free State.
Reuters </BODY></TEXT></REUTERS>

Γ΄.3 Σύνολο δεδομένων Brown

Η συλλογή Brown [Francis & Kucera, 1964] περιέχει 500 έγγραφα που δημοσιεύτηκαν το 1961 και περιλαμβάνουν χαρακτηριστικά δείγματα γραφής Αμερικανικής Αγγλικής. Κάθε έγγραφο περιέχει περισσότερες από 2000 λέξεις και η συλλογή καλύπτει ένα εύρος 15 κατηγοριών (όπως βιβλία για τη θρησκεία, ακαδημαϊκά κείμενα, ικανότητες και χόμπι κτλ) που φαίνονται παρακάτω:

A. PRESS: REPORTAGE
B. PRESS: EDITORIAL
C. PRESS: REVIEWS
D. RELIGION
E. SKILL AND HOBBIES
F. POPULAR LORE
G. BELLES-LETTRES
H. MISCELLANEOUS: GOVERNMENT & HOUSE ORGANS
J. LEARNED
K: FICTION: GENERAL
L: FICTION: MYSTERY
M: FICTION: SCIENCE
N: FICTION: ADVENTURE
P. FICTION: ROMANCE
R. HUMOR

Ένα παράδειγμα εγγράφου φαίνεται παρακάτω:

Los Angeles in 1957 finally bowed to the skyscraper. And without high density in the core, rapid-transit systems cannot be maintained economically, let alone built from scratch at today's prices. However, the building of freeways and garages cannot continue forever. The new interchange among the four Los Angeles freeways, including the grade-constructed accesses, occupies by itself no less than eighty acres of downtown land, one-eighth of a square mile, an area about the size of Rockefeller Center in New York. It is hard to believe that this mass of intertwined concrete constitutes what the law calls the highest and best use of centrally located urban land. As it affects the city's fiscal situation, such an interchange is ruinous ; it removes forever from the tax rolls property which should be taxed to pay for the city services. Subways improved land values without taking away land ; freeways boost valuation less (because the garages they require are not prime buildings by a long shot), and reduce the acreage that can be taxed. Downtown Los Angeles is already two-thirds freeway, interchange, street, parking lot and garage – one of those preposterous statistics has already come to pass. The freeway with narrowly spaced interchanges concentrates and mitigates the access problem, but it also acts inevitably as an artificial, isolating boundary. City planners do not always use this boundary as effectively as they might. Less ambitious freeway plans may be more successful – especially when the roadways and interchanges are raised, allowing for cross access at many points and providing parking areas below the ramp. Meanwhile, the automobile and its friend the truck have cost the central city some of its industrial dominance.

□

Βιβλιογραφία

[Abbas, 2007] Abbas, O. A. (2007). Comparisons between data clustering algorithms.

[Agichtein et al., 2001] Agichtein, E., Gravano, L., Pavel, J., Sokolova, V., & Voskoboynik, A. (2001). Snowball: a prototype system for extracting relations from large text collections. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, SIGMOD '01 (pp. 612–). New York, NY, USA: ACM.

[Algeo, 1995] Algeo, J. (1995). Having a look at the expanded predicate. *B. Aarts and C. Meyer (1995) The verb in contemporary English: theory and description.)*, (pp. 203–217).

[Alias-i, 2008] Alias-i (2008). LingPipe 4.1.0.

[Aline et al., 2007] Aline, V., Valia, K., Yi, Z., Marco, I., & Carlos, R. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *EMNLP-CoNLL* (pp. 1034–1043).

[Allan et al., 1998] Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y., Umass, J. A., Cmu, B. A., Cmu, D. B., Cmu, A. B., Cmu, R. B., Dragon, I. C., Darpa, G. D., Cmu, A. H., Cmu, J. L., Umass, V. L., Cmu, X. L., Dragon, S. L., Dragon, P. V. M., Umass, R. P., Cmu, T. P., Umass, J. P., & Umass, M. S. (1998). Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop* (pp. 194–218).

[Baldwin et al., 2003] Baldwin, T., Bannard, C., Tanaka, T., & Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, MWE '03 (pp. 89–96). Stroudsburg, PA, USA: Association for Computational Linguistics.

[Banerjee et al., 2007] Banerjee, S., Ramanathan, K., & Gupta, A. (2007). Clustering short texts using wikipedia. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 787–788). New York, NY, USA: ACM.

[Bannard et al., 2003] Bannard, C., Baldwin, T., & Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume*

18, MWE '03 (pp. 65–72). Stroudsburg, PA, USA: Association for Computational Linguistics.

[Baraff, 1996] Baraff, D. (1996). Linear-time dynamics using Lagrange multipliers. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, SIGGRAPH '96 (pp. 137–146). New York, NY, USA: ACM.

[Barzilay & Elhadad, 1997] Barzilay, R. & Elhadad, M. (1997). Using lexical chains for text summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization* (pp. 10–17).

[Berry et al., 1999] Berry, M. W., Drmac, Z., & Jessup, E. R. (1999). Matrices, Vector Spaces, and Information Retrieval. *SIAM Rev.*, 41, 335–362.

[Bizer et al., 2009] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - a crystallization point for the web of data. *Journal Web Semantics*, 7(3), 154–165.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993–1022.

[Bloehdorn et al., 2006] Bloehdorn, S., Cimiano, P., & Hotho, A. (2006). Learning ontologies to improve text clustering and classification. In M. Spiliopoulou, R. Kruse, A. Nu"rnberger, C. Borgelt, & W. Gaul (Eds.), *From Data and Information Analysis to Knowledge Engineering: Proceedings of the 29th Annual Conference of the German Classification Society (GfKl 2005), March 9-11, 2005, Magdeburg, Germany*, volume 30 of *Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 334–341).: Springer, Berlin–Heidelberg, Germany.

[Bollegala et al., 2007] Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. In *WWW '07: Proceedings of the 16th international conference on World Wide Web* (pp. 757–766). New York, NY, USA: ACM.

[Breaux & Reed, 2005] Breaux, T. D. & Reed, J. W. (2005). Using ontology in hierarchical information clustering. In *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4* (pp. 111.2). Washington, DC, USA: IEEE Computer Society.

[Budanitsky & Hirst, 2006] Budanitsky, A. & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Comput. Linguist.*, 32(1), 13–47.

[Bunescu & Pasca, 2007] Bunescu, R. C. & Pasca, M. (2007). Using Encyclopedic Knowledge for Named entity Disambiguation. In *EACL*: The Association for Computer Linguistics.

[Cafarella et al., 2005] Cafarella, M. J., Downey, D., Soderland, S., & Etzioni, O. (2005). Knowitnow: fast, scalable information extraction from the web. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05 (pp. 563–570). Stroudsburg, PA, USA: Association for Computational Linguistics.

[Carnegie Group & Reuters, 1997] Carnegie Group, I. & Reuters, L. (1997). *Reuters-21578 text categorization test collection.*

[Carpenter et al., 1991] Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991). Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Netw.*, 4(6), 759–771.

[Chalmers et al., 1992] Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: a critique of artificial intelligence methodology. *J. Exp. Theor. Artif. Intell.*, 4(3), 185–211.

[Chen et al., 1996] Chen, H., Schuffels, C., & Orwig, R. (1996). Internet categorization and search: A self-organizing approach. *Journal of Visual Communication and Image Representation*, 7(1), 88 – 102.

[Chen et al., 2006] Chen, H.-H., Lin, M.-S., & Wei, Y.-C. (2006). Novel association measures using web search with double checking. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 1009–1016). Morristown, NJ, USA: Association for Computational Linguistics.

[Cimiano et al., 2004] Cimiano, P., Handschuh, S., & Staab, S. (2004). Towards the self-annotating web. In *WWW '04: Proceedings of the 13th international conference on World Wide Web* (pp. 462–471). New York, NY, USA: ACM.

[Collobert et al., 2001] Collobert, R., Bengio, S., & Williamson, C. (2001). Svmtorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1, 143–160.

[Cucerzan, 2007] Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *In Proc. 2007 Joint Conference on EMNLP and CNLL* (pp. 708–716).

[Cunningham et al., 2002] Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)* Philadelphia.

[Davison, 1983] Davison, M. L. (1983). *Multidimensional Scaling.* New York: Wiley.

[Defays, 1977] Defays, D. (1977). An efficient algorithm for a complete link method. 20(4), 364–366.

[Demartines & Herault, 1997] Demartines, P. & Herault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. Neural Netw.*, 8(1), 148–154.

[Etzioni et al., 2004] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2004). Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, WWW '04 (pp. 100–110). New York, NY, USA: ACM.

[Etzioni et al., 2005] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2005). Unsupervised named-entity extraction from the web: an experimental study. *Artif. Intell.*, 165(1), 91–134.

[Fader et al., 2009] Fader, A., Soderland, S., & Etzioni, O. (2009). Scaling Wikipedia-based Named Entity Disambiguation to Arbitrary Web Text. In *Proceedings of the WikiAI 09 - IJCAI Workshop: User Contributed Knowledge and Artificial Intelligence: An Evolving Synergy* Pasadena, CA, USA.

[Fellbaum, 1998] Fellbaum, C., Ed. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

[Ferret et al., 1998] Ferret, O., Grau, B., & Masson, N. (1998). Thematic segmentation of texts: two methods for two kinds of texts. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1* (pp. 392–396). Morristown, NJ, USA: Association for Computational Linguistics.

[Finkelstein et al., 2002] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1), 116–131.

[Fletcher & Linzie, 1998] Fletcher, C. R. & Linzie, B. (1998). Motive and opportunity: Some comments on lsa, hal, kdc, and principal components. *DISCOURSE PROCESSES*, 25, 355–361.

[Fodor, 2002] Fodor, I. K. (2002). *A survey of dimension reduction techniques*. Technical Report UCRL-ID-148494, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory.

[Forman, 2004] Forman, G. (2004). A pitfall and solution in multi-class feature selection for text classification. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning* (pp.38). New York, NY, USA: ACM.

[Francis & Kucera, 1964] Francis, W. N. & Kucera, H. (1964). *Manual of Information to accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Providence, Rhode Island.

[Frantzi et al., 1998] Frantzi, K. T., Ananiadou, S., & Tsujii, J.-i. (1998). The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '98 (pp. 585–604). London, UK: Springer-Verlag.

[French, 1988] French, R. (1988). Subcognitive probing: Hard questions for the turing test. In *Proceedings of the Tenth Annual Cognitive Science Society Conference* (pp. 361–367). Hillsdale, NJ: LEA.

[Fung et al., 2003] Fung, B. C. M., Wang, K., & Ester, M. (2003). Hierarchical document clustering using frequent itemsets. In *Proc. of the 3rd SIAM International Conference on Data Mining (SDM)* (pp. 59–70). San Francisco, CA: SIAM.

[Furnas et al., 1984] Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1984). : chapter Statistical semantics: analysis of the potential performance of keyword information systems, (pp. 187–242). Norwood, NJ, USA: Ablex Publishing Corp.

[Gabrilovich & Markovitch, 2006] Gabrilovich, E. & Markovitch, S. (2006). Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In *AAAI'06: proceedings of the 21st national conference on Artificial intelligence* (pp. 1301–1306).: AAAI Press.

[Gabrilovich & Markovitch, 2007] Gabrilovich, E. & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence* (pp. 1606–1611). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[Gentner, 1983] Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy*. *Cognitive Science*, 7(2), 155–170.

[Georgakis et al., 2001] Georgakis, A., Kotropoulos, C., Xafopoulos, A., & Pitas, I. (2001). MM-WEBSOM: A variant of websom based on order statistics. In *IEEE-EURASIP Workshop Nonlinear Signal and Image Processing* Baltimore, USA: IEEE Computer Society.

[Goldstone & Son, 2004] Goldstone, R. L. & Son, J. Y. (2004). Similarity. *Psychological Review*, 100, 254–278.

[Grefenstette, 1992] Grefenstette, G. (1992). Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of the 30st annual meeting of the Association for Computational Linguistics, ACL* (pp. 324–329).

[Griffiths & Steyvers, 2004] Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1), 5228–5235.

[Guyon & Elisseeff, 2003] Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3, 1157–1182.

[Hammouda & Kamel, 2004] Hammouda, K. M. & Kamel, M. S. (2004). Efficient phrase-based document indexing for web document clustering. *IEEE Trans. on Knowl. and Data Eng.*, 16, 1279–1296.

[He et al., 2002] He, J., hwee Tan, A., & lim Tan, C. (2002). ART-C: A Neural Architecture for Self-Organization Under Constraints. In *In Proceedings of International Joint Conference on Neural Networks (IJCNN* (pp. 2550–2555).

[Hearst, 1994] Hearst, M. A. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 9–16). Morristown, NJ, USA: Association for Computational Linguistics.

[Himberg, 2000] Himberg, J. (2000). A som based cluster visualization and its application for false coloring. In *IJCNN '00: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)-Volume 3* (pp. 3587). Washington, DC, USA: IEEE Computer Society.

[Hofmann, 1999a] Hofmann, T. (1999a). The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In *In IJCAI* (pp. 682–687).

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

[Hofmann, 1999c] Hofmann, T. (1999c). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99 (pp. 50–57). New York, NY, USA: ACM.

[Hotho et al., 2003] Hotho, A., Staab, S., & Stumme, G. (2003). Wordnet improves text document clustering. In Y. Ding, K. van Rijsbergen, I. Ounis, & J. Jose (Eds.), *Proceedings of the Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR 2003), August 1, 2003, Toronto Canada.*

[Hotho & Stumme, 2002] Hotho, A. & Stumme, G. (2002). Conceptual clustering of text clusters. In *Proceedings of FGML Workshop* (pp. 37–45).: Special Interest Group of German Informatics Society (FGML).

[Hu et al., 2008] Hu, J., Fang, L., Cao, Y., Zeng, H.-J., Li, H., Yang, Q., & Chen, Z. (2008). Enhancing text clustering by leveraging Wikipedia semantics. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 179–186). New York, NY, USA: ACM.

[Hu et al., 2009] Hu, X., Zhang, X., Lu, C., Park, E. K., & Zhou, X. (2009). Exploiting Wikipedia as external knowledge for document clustering. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 389–396). New York, NY, USA: ACM.

[Huang et al., 2009] Huang, A., Milne, D., Frank, E., & Witten, I. H. (2009). Clustering documents using a Wikipedia-based concept representation. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '09 (pp. 628–636). Berlin, Heidelberg: Springer-Verlag.

[ICAME, 2011] ICAME (2011). International Computer Archive of Modern and Medieval English http://icame.uib.no/ (Online accessed July 2011).

[Iosif & Potamianos, 2007] Iosif, E. & Potamianos, A. (2007). Unsupervised Semantic Similarity Computation using Web Search Engines. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 381–387). Washington, DC, USA: IEEE Computer Society.

[Jarmasz, 2003] Jarmasz, M. (2003). *Roget"s thesaurus as a lexical resource for natural language processing.* Technical report, University of Ottawa.

[Jiang & Conrath, 1997] Jiang, J. J. & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)* (pp. 9008+).

[Jin et al., 2005] Jin, H., Wong, M.-L., & Leung, K. S. (2005). Scalable model-based clustering for large databases based on data summarization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(11), 1710–1719.

[Joachims, 1999] Joachims, T. (1999). Making large-scale support vector machine learning practical. *Advances in kernel methods: support vector learning*, (pp. 169–184).

[Junker et al., 2000] Junker, M., Sintek, M., & Rinck, M. (2000). Learning for text categorization and information extraction with ilp. (pp. 247–258).

[Kalman, 1960] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D), 35–45.

[Kangas et al., 1990] Kangas, J., Kohonen, T., & Laaksonen, J. (1990). Variants of Self-Organizing Maps. *Neural Networks, IEEE Transactions on*, 1(1), 93 –99.

[Karypis, 2002] Karypis, G. (2002). CLUTO • A Clustering Toolkit. *Technical Report*, 02-017.

[Katz, 2006] Katz, G. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *In Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (pp. 12–19).

[Keizer, 2007] Keizer, E. (2007). *The English Noun Phrase: The Nature of Linguistic Categorization, Studies in English Language (series)*. Cambridge: Cambridge University Press.

[Kiran, 2010] Kiran, G V R, R. S. (2010). Enhancing document clustering using various external knowledge sources. In *Proceedings of the 15th Australasian Document Computing Symposium*.

[Kohonen, 1989] Kohonen, T. (1989). *Self-organization and associative memory: 3rd edition*. New York, NY, USA: Springer-Verlag New York, Inc.

[Kohonen et al., 2000] Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. *Neural Networks, IEEE Transactions on*, 11(3), 574 –585.

[Kohonen et al., 2001] Kohonen, T., Schroeder, M. R., & Huang, T. S., Eds. (2001). *Self-Organizing Maps*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

[Korkontzelos & Manandhar, 2010] Korkontzelos, I. & Manandhar, S. (2010). UoY: Graphs of unambiguous vertices for word sense induction and disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10 (pp. 355–358). Stroudsburg, PA, USA: Association for Computational Linguistics.

[Kraaijveld, 1992] Kraaijveld, M. (1992). A non-linear projection method based on Kohonen's topology preserving maps. In *Pattern Recognition, 1992. Vol.II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on* (pp. 41 –45).

[Kulkarni et al., 2009] Kulkarni, S., Singh, A., Ramakrishnan, G., & Chakrabarti, S. (2009). Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09 (pp. 457–466). New York, NY, USA: ACM.

[Lagus et al., 2004] Lagus, K., Kaski, S., & Kohonen, T. (2004). Mining massive document collections by the WEBSOM method. *Inf. Sci.*, 163(1-3), 135–156.

[Lang, 1995] Lang, K. (1995). Newsweeder : Learning to filter netnews. In *Proceedings of the International Conference on Machine Learning* Tahoe City, California, USA: Morgan Kaufmann.

[Li et al., 2007] Li, Y., Luk, W. P. R., Ho, K. S. E., & Chung, F. L. K. (2007). Improving weak ad-hoc queries using Wikipedia as external corpus. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 797–798). New York, NY, USA: ACM.

[Lin, 1998] Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics* (pp. 768–774). Morristown, NJ, USA: Association for Computational Linguistics.

[Lin, 1999] Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99 (pp. 317–324). Stroudsburg, PA, USA: Association for Computational Linguistics.

[Lin et al., 1991] Lin, X., Soergel, D., & Marchionini, G. (1991). A Self-Organizing Semantic Map for Information Retrieval. In *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 262–269). New York, NY, USA: ACM.

[Liu et al., 2002] Liu, X., Gong, Y., Xu, W., & Zhu, S. (2002). Document clustering with cluster refinement and model selection capabilities. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 191–198). New York, NY, USA: ACM.

[Lund & Burgess, 1996] Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28, 203–208.

[Manabu & Takeo, 1994] Manabu, O. & Takeo, H. (1994). Word sense disambiguation and text segmentation based on lexical cohesion. In *Proceedings of the 15th conference on Computational linguistics - Volume 2* (pp. 755–761). Morristown, NJ, USA: Association for Computational Linguistics.

[Manber & Myers, 1990] Manber, U. & Myers, G. (1990). Suffix arrays: a new method for on-line string searches. In *Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*, SODA '90 (pp. 319–327). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

[Manning & Schútze, 1999] Manning, C. D. & Schútze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

[Marcus et al., 1993] Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2), 313–330.

[Matuszek et al., 2006] Matuszek, C., Cabral, J., Witbrock, M., & DeOliveira, J. (2006). An Introduction to the Syntax and Content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*.

[Mau et al., 1999] Mau, B., A, L. M., & Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55, 1–12.

[Mccarthy et al., 2007] Mccarthy, D., Sussex, F. E., Venkatapathy, S., & Joshi, A. K. (2007). Detecting compositionality of verb-object combinations using selectional preferences. In *In Proceedings of the 200 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL* (pp. 369–379).

[Mendes et al., 2011] Mendes, P., Jakob, M., Garcna-Silva, A., & Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In *In the Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.

[Merkl, 1998] Merkl, D. (1998). Text classification with self-organizing maps: Some lessons learned. *Neurocomputing*, 21(1-3), 61–77.

[Merkl & Rauber, 1997] Merkl, D. & Rauber, A. (1997). Alternative Ways for Cluster Visualization in Self-Organizing Maps. In *In Proc. of the Workshop on Self-Organizing Maps (WSOM97* (pp. 106–111).

[Mihalcea & Csomai, 2007] Mihalcea, R. & Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 233–242). New York, NY, USA: ACM.

[Miikkulainen, 1990] Miikkulainen, R. (1990). Script recognition with hierarchical feature maps. *Connection Science*, 2, 83–101.

[Miller & Charles, 1991] Miller, G. A. & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28.

[Milne & Witten, 2008] Milne, D. & Witten, I. H. (2008). Learning to link with Wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08 (pp. 509–518). New York, NY, USA: ACM.

[Mitra et al., 2002] Mitra, P., Murthy, C. A., & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3), 301–312.

[Moutarde & Ultsch, 2005] Moutarde, F. & Ultsch, A. (2005). U*F clustering: a new performant "cluster-mining" method based on segmentation of Self-Organizing Maps. In *Workshop on Self-Organizing Maps (WSOM'2005)*.

[Murtagh, 1984] Murtagh, F. (1984). Complexities of hierarchic clustering algorithms: state of the art. 1, 101–113.

[Nallapati, 2003] Nallapati, R. (2003). Semantic language models for topic detection and tracking. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 student research workshop - Volume 3*, NAACL '03 (pp. 1–6). Morristown, NJ, USA: Association for Computational Linguistics.

[Navigli & Ponzetto, 2010] Navigli, R. & Ponzetto, S. P. (2010). BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10 (pp. 216–225). Stroudsburg, PA, USA: Association for Computational Linguistics.

[Niles & Pease, 2001] Niles, I. & Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01 (pp. 2–9). New York, NY, USA: ACM.

[OpenNLP, 2012] OpenNLP (2012). Name Finder http://http://opennlp.apache.org/ (accessed July 2011).

[Pampalk et al., 2002] Pampalk, E., Rauber, A., & Merkl, D. (2002). Using smoothed data histograms for cluster visualization in self-organizing maps. In *ICANN '02: Proceedings of the International Conference on Artificial Neural Networks* (pp. 871–876). London, UK: Springer-Verlag.

[Piao et al., 2006] Piao, S. S., Rayson, P., Mudraya, O., Wilson, A., & Garside, R. (2006). Measuring MWE Compositionality Using Semantic Annotation. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (pp. 2–11). Sydney, Australia: Association for Computational Linguistics.

[Pölzlbauer, 2004] Pölzlbauer, G. (2004). Survey and Comparison of Quality Measures for Self-Organizing Maps. In J. Paralič, G. Pölzlbauer, & A. Rauber (Eds.), *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)* (pp. 67–82). Sliezsky dom, Vysoké Tatry, Slovakia: Elfa Academic Press.

[Pullwitt, 2002] Pullwitt, D. (2002). Integrating contextual information to enhance SOM-based text document clustering. *Neural Netw.*, 15(8-9), 1099–1106.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257 –286.

[Ratinov & Roth, 2009] Ratinov, L. & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *CoNLL*.

[Ratinov et al., 2011] Ratinov, L., Roth, D., Downey, D., & Anderson, M. (2011). Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11 (pp. 1375–1384). Stroudsburg, PA, USA: Association for Computational Linguistics.

[Rauber, 1999] Rauber, A. (1999). LABELSOM: on the labeling of self-organizing maps. *Neural Networks, 1999. IJCNN '99. International Joint Conference on*, 5, 3527 –3532 vol.5.

[Rauber et al., 2002] Rauber, A., Merkl, D., & Dittenbach, M. (2002). The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks*, 13, 1331–1341.

[Resnik, 1995] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 448–45).

[Resnik, 1999] Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11, 95–130.

[Richmond & Smith, 1997] Richmond, K. & Smith, A. (1997). Detecting subject boundaries within text: A language independent statistical approach. In *Brown University, Providence, Rhode Island* (pp. 47–54).

[Roweis & Saul, 2000] Roweis, S. & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2323–2326.

[Saeed, 1997] Saeed, J. (1997). *Semantics*. Oxford: Blackwell Publishers.

[Sahami & Heilman, 2006] Sahami, M. & Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web* (pp. 377–386). New York, NY, USA: ACM.

[Salton, 1989] Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

[Salton & Buckley, 1987] Salton, G. & Buckley, C. (1987). *Term Weighting Approaches in Automatic Text Retrieval*. Technical report, Ithaca, NY, USA.

[Salton & McGill, 1983] Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York u.a.: McGraw-Hill.

[Sammon, 1969] Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, 18(5), 401–409.

[Savaresi & Boley, 2001] Savaresi, S. M. & Boley, D. L. (2001). On the performance of bisecting K-means and PDDP. In *Proceedings of the First SIAM International Conference on Data Mining (ICDM-2001* (pp. 1–14).

[Schmid, 1994] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* Manchester, UK.

[Schone & Jurafsky, 2001] Schone, P. & Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing* Pittsburgh, PA.

[Sebastiani & Ricerche, 2002] Sebastiani, F. & Ricerche, C. N. D. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34, 1–47.

[Sedding & Kazakov, 2004] Sedding, J. & Kazakov, D. (2004). WordNet-based text document clustering. In *ROMAND '04: Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data* (pp. 104–113). Morristown, NJ, USA: Association for Computational Linguistics.

[Shadbolt et al., 2006] Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3), 96–101.

[Slonim et al., 2002] Slonim, N., Friedman, N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 129–136). New York, NY, USA: ACM.

[Soderland, 1999] Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Mach. Learn.*, 34(1-3), 233–272.

[Stanford, 2009] Stanford (2009). Named Entity Recognizer http://www-nlp.stanford.edu/software/CRF-NER.shtml (accessed July 2011).

[Steinbach et al., 2000] Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In M. Grobelnik, D. Mladenic, & N. Milic-Frayling (Eds.), *KDD-2000 Workshop on Text Mining, August 20* (pp. 109–111). Boston, MA.

[Strube & Ponzetto, 2006] Strube, M. & Ponzetto, S. P. (2006). WikiRelate! computing semantic relatedness using Wikipedia. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06 (pp. 1419–1424).: AAAI Press.

[Stumme et al., 2002] Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., & Lakhal, L. (2002). Computing iceberg concept lattices with titanic. *Data Knowl. Eng.*, 42(2), 189–222.

[Suchanek et al., 2006] Suchanek, F. M., Ifrim, G., & Weikum, G. (2006). Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06 (pp. 712–717). New York, NY, USA: ACM.

[Suchanek et al., 2008] Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from Wikipedia and Wordnet. *Journal Web Semantics*, 6, 203–217.

[Talavera & Bejar, 2001] Talavera, L. & Bejar, J. (2001). Generality-based conceptual clustering with probabilistic concepts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2), 196 –206.

[Tenenbaum et al., 2000] Tenenbaum, J. B., Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500), 2319–2323.

[Toral & Munoz, 2006] Toral, A. & Munoz, R. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. In *EACL*: The Association for Computer Linguistics.

[Tsochantaridis et al., 2004] Tsochantaridis, I., Hofmann, T., Joachims, T., & Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning* (pp. 104). New York, NY, USA: ACM.

[Turney, 2001] Turney, P. D. (2001). Mining the web for synonyms: PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, 2167, 491–502.

[UCI, 2011] UCI (2011). Knowledge Discovery in Databases archive http://kdd.ics.uci.edu/ (Online accessed July 2011).

[Ultsch & Siemon, 1990] Ultsch, A. & Siemon, H. P. (1990). Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. In *Proceedings of International Neural Networks Conference (INNC)* (pp. 305–308).: Kluwer Academic Press.

[Vélez et al., 1997] Vélez, B., Weiss, R., Sheldon, M. A., & Gifford, D. K. (1997). Fast and effective query refinement. *SIGIR Forum*, 31(SI), 6–15.

[Vesanto & Alhoniemi, 2000] Vesanto, J. & Alhoniemi, E. (2000). Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 11(3), 586–600.

[Wang et al., 2003] Wang, B. B., Mckay, R. I. B., Abbass, H. A., & Barlow, M. (2003). A comparative study for domain ontology guided feature extraction. In *ACSC '03: Proceedings of the 26th Australasian computer science conference* (pp. 69–78). Darlinghurst, Australia, Australia: Australian Computer Society, Inc.

[Wang & Domeniconi, 2008] Wang, P. & Domeniconi, C. (2008). Building semantic kernels for text classification using Wikipedia. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 713–721). New York, NY, USA: ACM.

[Wang et al., 2009] Wang, P., Hu, J., Zeng, H.-J., & Chen, Z. (2009). Using Wikipedia knowledge to improve text classification. *Knowl. Inf. Syst.*, 19(3), 265–281.

[Wikipedia API, 2010] Wikipedia API (2010). http://en.wikipedia.org/w/api.php (Online accessed July 2011).

[Willett, 1988] Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. *Inf. Process. Manage.*, 24(5), 577–597.

[W.N. & D.A., 1955] W.N., L. & D.A., B. (1955). *Machine Translation of Languages*. Cambridge, MA: MIT Press.

[Xiong et al., 2004] Xiong, H., Steinbach, M., Tan, P., & Kumar, V. (2004). HICAP: Hierarchical clustering with pattern preservation. In *Proceedings of SIAM international conference on data mining* (pp. 279–290). Philadelphia, PA: SIAM.

[Xue & Zhou, 2009] Xue, X.-B. & Zhou, Z.-H. (2009). Distributional features for text categorization. *IEEE Trans. on Knowl. and Data Eng.*, 21(3), 428–442.

[Yang & Pedersen, 1997] Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97 (pp. 412–420). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[Yin, 2002] Yin, H. (2002). ViSOM - a novel method for multivariate data projection and structure visualization. *Neural Networks, IEEE Transactions on*, 13(1), 237 –243.

[Yuan et al., 2004] Yuan, F., Meng, Z.-H., Zhang, H.-X., & Dong, C.-R. (2004). A new algorithm to get the initial centroids. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 2 (pp. 1191 – 1193 vol.2).

[Zamir & Etzioni, 1998] Zamir, O. & Etzioni, O. (1998). Web document clustering: a feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98 (pp. 46–54). New York, NY, USA: ACM.

[Zhang et al., 2004] Zhang, Y., Zincir-Heywood, N., & Milios, E. (2004). Term-based clustering and summarization of web page collections. *Advances in Artificial Intelligence*, (pp. 60–74).

Κατάλογος δημοσιεύσεων του συγγραφέα

Δημοσιεύσεις σε διεθνή περιοδικά με κρίση

1. Gerasimos Spanakis, Georgios Siolas, Andreas Stafylopatis: **DoSO: A Document Self-Organizer"** *Journal of Intelligent Information Systems* (appeared online May 12th 2012)

2. Gerasimos Spanakis, Georgios Siolas, Andreas Stafylopatis : **Exploiting Wikipedia Knowledge for Conceptual Hierarchical Clustering of Documents"** (to appear) *The Computer Journal, Section C : Computational Intelligence, Vol.55, Issue 3, March 2012* (appeared online March 15th 2011)

Δημοσιεύσεις σε διεθνή συνέδρια με κρίση

3. Gerasimos Spanakis, Georgios Siolas, Andreas Stafylopatis : **A Conceptual Hierarchical Clustering of documents based on Wikipedia knowledge** *Proceedings of the 25th International Symposium on Computer and Information Sciences (ISCIS 2010)* London , September 2010

4. Gerasimos Spanakis, Georgios Siolas, Andreas Stafylopatis : **A hybrid web-based measure for computing semantic relatedness between words** *Proceedings of the 21st International Conference on Tools with Artificial Intelligence (ICTAI 2009)* Newark (NYC Metropolitan Aera), November 2009

Δημοσιεύσεις εκτός διατριβής

5. Christos A. Christodoulou, George Perantzakis, Gerasimos E. Spanakis, Panagiotis Karampelas : **Evaluation of lightning performance of transmission lines protected by metal oxide surge arresters using artificial intelligence techniques**. *Energy Systems* (δεκτό προς δημοσίευση)

6. C.A. Christodoulou, G.E. Spanakis, V. Vita, D.I. Karvouniari, L. Ekonomou : **Locating High Voltage Transmission Lines using ELECTRE I method** *International Journal on Power System Optimization, Vol.1, No.2*

□

Βιογραφικό Σημείωμα

Στοιχεία Επικοινωνίας

Εργαστήριο Ευφυών Συστημάτων
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο
Ηρώων Πολυτεχνείου 9, Ζωγράφου, 157 80, Αθήνα, Ελλάδα
Τηλέφωνο: (+30) 210 772 2504
Ηλεκτρονικό ταχυδρομείο (e-mail): gspana@ece.ntua.gr
Προσωπική Σελίδα: http://www.cc.ece.ntua.gr/~gspana

Σπουδές

- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (04/2007–σήμερα)
  Υποψήφιος Διδάκτωρ Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
  Επιβλέπων: καθ. Ανδρέας-Γεώργιος Σταφυλοπάτης

- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (10/2001–10/2006)
  Διπλωματούχος Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
  Βαθμός: 8.18/10
  Διπλωματική εργασία: Νευρωνικά Δίκτυα και Ενισχυτική Μάθηση
  Επιβλέπων: καθ. Ανδρέας-Γεώργιος Σταφυλοπάτης

Ερευνητικά Ενδιαφέροντα

- Αναγνώριση Προτύπων
- Ανάκτηση Πληροφοριών Κειμένου
- Εξαγωγή/Κατασκευή Γνωρισμάτων Ανάλυσης Κειμένου
- Εφαρμογή Ευφυών Τεχνικών σε Προβλήματα Βελτιστοποίησης

Διδακτική - Εργασιακή Εμπειρία

- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (2007-σήμερα)
  Βοηθός Διδασκαλίας
  - Προγραμματιστικές Τεχνικές
  - Νευρωνικά Δίκτυα-Ευφυή Συστήματα
- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (2004-2008)
  Συνεργάτης-Διαχειριστής του Δικτύου και Συστημάτων του Υπολογιστικού Κέντρου ΣΗΜΜΥ

Έργα Έρευνας και Ανάπτυξης

- ΣΑΜΣ : Σύστημα Αναγνώρισης Μαθηματικών Συμβόλων (2007-2008)

- Ανάπτυξη συστήματος καθαρισμού δεδομένων και ταιριάσματος για το Εθνικό Κέντρο Τεκμηρίωσης (ΕΚΤ) (2008-2009)

Δημοσιεύσεις

1. Gerasimos Spanakis, Georgios Siolas, Andreas Stafylopatis : **DoSO: A Document Self-Organizer"** *Journal of Intelligent Information Systems* (appeared online May 12th 2012)

2. Gerasimos Spanakis, Georgios Siolas, Andreas Stafylopatis : **Exploiting Wikipedia Knowledge for Conceptual Hierarchical Clustering of Documents"** (to appear) *The Computer Journal, Section C : Computational Intelligence, Vol.55, Issue 3, March 2012* (appeared online March 15th 2011)

3. Gerasimos Spanakis, Georgios Siolas, Andreas Stafylopatis : **A Conceptual Hierarchical Clustering of documents based on Wikipedia knowledge** *Proceedings of the 25th International Symposium on Computer and Information Sciences (ISCIS 2010)* London , September 2010

4. Gerasimos Spanakis, Georgios Siolas, Andreas Stafylopatis : **A hybrid web-based measure for computing semantic relatedness between words** *Proceedings of the 21st International Conference on Tools with Artificial Intelligence (ICTAI 2009)* Newark (NYC Metropolitan Aera), November 2009

5. Christos A. Christodoulou, George Perantzakis, Gerasimos E. Spanakis, Panagiotis Karampelas : **Evaluation of lightning performance of transmission lines protected by metal oxide surge arresters using artificial intelligence techniques**. *Energy Systems* (δεκτό προς δημοσίευση)

6. C.A. Christodoulou, G.E. Spanakis, V. Vita, D.I. Karvouniari, L. Ekonomou : **Locating High Voltage Transmission Lines using ELECTRE I method** *International Journal on Power System Optimization, Vol.1, No.2*

Ξένες γλώσσες

- Αγγλικά : The University of Michigan Certificate of Proficiency in English
- Γερμανικά : Zertificat Deutsch als Fremdsprach (ZDaF)
- Γαλλικά : Βασικές γνώσεις

Γνώσεις Υπολογιστή

- Γλώσσες προγραμματισμού: C, C++, Java, JavaScript, Pascal, Perl, PHP, Lisp, UNIX, shell scripting, SQL, FORTRAN, MATLAB, XML/XSL
- Λειτουργικά Συστήματα : Microsoft Windows(98,2003,Server,XP,7), Linux, Solaris
- Άλλες γνώσεις : Word,Excel,Internet

Διακρίσεις

- **Υπότροφος του Κληροδοτήματος Σπ. Αντύπα Υπέρ Κεφαλληνίας** (2007-2010)

Δραστηριότητες και προσωπικά ενδιαφέροντα

- Διαδίκτυο
- Ταξίδια
- Δημιουργός και διαχειριστής της φοιτητικής δικτυακής κοινότητας των φοιτητών ηλεκτρολόγων μηχανικών και μηχανικών υπολογιστών (`http://shmmy.ntua.gr`)
- Πολιτική Αεροπορία και παρατήρηση αεροσκαφών. Ιδρυτικό μέλος του Συλλόγου Φίλων Αεροπορίας «Απογείωση» (`http://www.air-born.gr`)

□