ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΤΑ ΠΛΑΙΣΙΑ ΤΟΥ
ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ (ΜΔΕ)
«ΕΠΙΣΤΗΜΗ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ»
ΤΟΥ ΤΜΗΜΑΤΟΣ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΡΟΣΩΠΟΠΟΙΗΜΕΝΗ ΠΡΟΒΟΛΗ ΠΕΡΙΕΧΟΜΕΝΟΥ ΤΟΥ ΔΙΑΔΙΚΤΥΟΥ ΜΕ ΤΕΧΝΙΚΕΣ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ, ΑΥΤΟΜΑΤΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΚΑΙ ΑΥΤΟΜΑΤΗΣ ΕΞΑΓΩΓΗΣ ΠΕΡΙΛΗΨΗΣ

Πουλόπουλος Βασίλειος
Α.Μ. 442

Επιβλέπων Καθηγητής
Χρήστος Μπούρας,
Αναπληρωτής Καθηγητής

Τριμελής Επιτροπή
Αθανάσιος Τσακαλίδης, Καθηγητής
Ιωάννης Γαροφαλλάκης, Αναπληρωτής Καθηγητής
Χρήστος Μπούρας, Αναπληρωτής Καθηγητής

Πάτρα, Σεπτέμβριος 2007

Σε αυτούς που όταν τα όνειρα μου φτάνουν σε ένα τέλος,
υπάρχουν για να μου θυμίζουν
πως θα είναι εκεί... στο επόμενο όνειρο

Πρόλογος

Η συγκεκριμένη εργασία αποτελεί κομμάτι μίας προσπάθειας δύο ετών Εργασίας και Έρευνας μέσα στο χώρο της Επιστήμης και της Τεχνολογίας. Μέσα στο χώρο του Πανεπιστημίου που μετά από 7 χρόνια σπουδών και εργασιών μπορώ να τον θεωρώ κομμάτι μου και φυσικά τιμή μου που με ανέχεται.

Η επιστήμη και η τεχνολογία είναι κομμάτια της ζωής μας και η επίδρασή τους στο καθημερινό γίγνεσθαι είναι μεγάλη. Ακόμα μεγαλύτερη ήταν και η δική μου δίψα να έρθω αντιμέτωπος με ένα πρόβλημα που ακόμα και εγώ ο ίδιος αντιμετωπίζω στην αέναη περιαγωγή στις εκατομμύρια σελίδες του διαδικτύου. Το πρόβλημα δεν ήταν άλλο από την αναζήτηση πληροφορίας στο χαοτικό τόπο που λέγεται παγκόσμιος ιστός.

Η συγκεκριμένη εργασία είχε σαν σκοπό να καταπιαστεί με ένα μικρό κομμάτι αναζήτησης στο χαοτικό διαδίκτυο που αφορά τον εντοπισμό ειδήσεων από μεγάλα ειδησεογραφικά πρακτορεία. Πολλοί είναι αυτοί που ουσιαστικά χάνονται στην προσπάθεια αναζήτησης μίας είδησης ενώ οι μηχανές αναζήτησης δεν είναι σε θέση να προσφέρουν βοήθεια καθώς η αναζήτηση γενικών θεμάτων (π.χ. κρίση στη μέση ανατολή, πόλεμος στο Ιράκ, φωτιές στην Ελλάδα) επιστρέφει εκατομμύρια αποτελέσματα.

Ο λόγος αυτός οδήγησε στην εκπόνηση αυτής της εργασίας και πιο συγκεκριμένα στην κατασκευή ενός μηχανισμού που θα είναι σε θέση να συλλέγει όλες τις ειδήσεις για το χρήστη και θα μπορεί να τις παρουσιάσει με τον τρόπο που ο χρήστης επιθυμεί.

Για το λόγο αυτό αναπτύχθηκαν 6 υποσυστήματα:

- Συλλογή σελίδων με άρθρα από το διαδίκτυο
- Απομόνωση χρήσιμου κειμένου
- Εξαγωγή λέξεων κλειδιών
- Κατηγοριοποίηση άρθρων
- Εξαγωγή Περίληψης από άρθρα
- Παρουσίαση προσωποποιημένης πληροφορίας στο χρήστη

Η κατάληξη ήταν επιτυχής και είμαι σε θέση σήμερα να παρουσιάσω το μηχανισμό με την ονομασία peRSSonal που περικλείει όλα τα παραπάνω.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της εργασίας μου, Αναπληρωτή Καθηγητή κ. Χρήστο Μπούρα για πολύτιμη βοήθειά του, και την υποστήριξή του για την εκπόνηση της συγκεκριμένης εργασίας και για τον τρόπο με τον οποίο μου συμπαραστέκεται όλα τα χρόνια που βρίσκομαι στο χώρο του Πανεπιστημίου.

Όμοίως, θέλω να ευχαριστήσω τους καθηγητές του ΤΜΗΥΠ για την τιμή που μου έκαναν να είναι μέλη της τριμελούς επιτροπής, τον Καθηγητή Αθανάσιο Τσακαλίδη και τον Αναπληρωτή Καθηγητή Ιωάννη Γαροφαλλάκη.

Επίσης θα ήθελα να ευχαριστήσω τον αδερφό μου Λεωνίδα που ανέχτηκε τις ανάγκες μου σε όλη τη διάρκεια εκπόνησης της εργασίας και τον καλό μου φίλο Βασίλη Τσόγκα, χωρίς τη βοήθεια του οποίου η εργασία θα ήταν ακόμα ένα όνειρο.

Κλείνοντας θέλω να αναφέρω κάποιους ανθρώπους που έδωσαν πάτημα στα όνειρά μου να ανέβω το δύσκολο δρόμο του Ακαδημαϊκού χώρου και να τους θυμίσω πως κι εγώ είμαι κοντά τους αν το θελήσουν: το Νίκο, την Ελένη, τη Λίνα, το Δημήτρη, το Γιάννη, την Ευγενία, την Έρη και τον Αντώνη.

Βασίλης Πουλόπουλος
Πάτρα, Σεπτέμβριος 2007

Επιτελική Σύνοψη

Σκοπός της Μεταπτυχιακής Εργασίας είναι η επέκταση και αναβάθμιση του μηχανισμού που είχε δημιουργηθεί στα πλαίσια της Διπλωματικής Εργασίας που εκπόνησα με τίτλο «Δημιουργία Πύλης Προσωποποιημένης Πρόσβασης σε Περιεχόμενο του WWW».

Η παραπάνω Διπλωματική εργασία περιλάμβανε τη δημιουργία ενός μηχανισμού που ξεκινούσε με ανάκτηση πληροφορίας από το Διαδίκτυο (HTML σελίδες από news portals), εξαγωγή χρήσιμου κειμένου και προεπεξεργασία της πληροφορίας, αυτόματη κατηγοριοποίηση της πληροφορίας και τέλος παρουσίαση στον τελικό χρήστη με προσωποποίηση με στοιχεία που εντοπίζονταν στις επιλογές του χρήστη.

Στην παραπάνω εργασία εξετάστηκαν διεξοδικά θέματα που είχαν να κάνουν με τον τρόπο προεπεξεργασίας της πληροφορίας καθώς και με τον τρόπο αυτόματης κατηγοριοποίησης ενώ υλοποιήθηκαν αλγόριθμοι προεπεξεργασίας πληροφορίας τεσσάρων σταδίων και αλγόριθμος αυτόματης κατηγοριοποίησης βασισμένος σε πρότυπες κατηγορίες.

Τέλος υλοποιήθηκε portal το οποίο εκμεταλλευόμενο την επεξεργασία που έχει πραγματοποιηθεί στην πληροφορία παρουσιάζει το περιεχόμενο στους χρήστες προσωποποιημένο βάσει των επιλογών που αυτοί πραγματοποιούν.

Σκοπός της μεταπτυχιακής εργασίας είναι η εξέταση περισσοτέρων αλγορίθμων για την πραγματοποίηση της παραπάνω διαδικασίας αλλά και η υλοποίησή τους προκειμένου να γίνει σύγκριση αλγορίθμων και παραγωγή ποιοτικότερου αποτελέσματος.

Πιο συγκεκριμένα αναβαθμίζονται όλα τα στάδια λειτουργίας του μηχανισμού. Έτσι, το στάδιο λήψης πληροφορίας βασίζεται σε έναν απλό crawler λήψης HTML σελίδων από αγγλόφωνα news portals. Η διαδικασία βασίζεται στο γεγονός πως για κάθε σελίδα υπάρχουν RSS feeds. Διαβάζοντας τα τελευταία νέα που προκύπτουν από τις εγγραφές στα RSS feeds μπορούμε να εντοπίσουμε όλα τα URL που περιέχουν HTML σελίδες με τα άρθρα. Οι HTML σελίδες φιλτράρονται προκειμένου από αυτές να γίνει εξαγωγή μόνο του κειμένου και πιο αναλυτικά του χρήσιμου κειμένου ούτως ώστε το κείμενο που εξάγεται να αφορά αποκλειστικά άρθρα. Η τεχνική εξαγωγής χρήσιμου κειμένου βασίζεται στην τεχνική web clipping. Ένας parser, ελέγχει την HTML δομή προκειμένου να εντοπίσει τους κόμβους που περιέχουν μεγάλη ποσότητα κειμένου και βρίσκονται κοντά σε άλλους κόμβους που επίσης περιέχουν μεγάλες ποσότητες κειμένου.

Στα εξαγόμενα άρθρα πραγματοποιείται προεπεξεργασία πέντε σταδίων με σκοπό να προκύψουν οι λέξεις κλειδιά που είναι αντιπροσωπευτικές του άρθρου. Πιο αναλυτικά, αφαιρούνται όλα τα σημεία στίξης, όλοι οι αριθμοί, μετατρέπονται όλα τα γράμματα σε πεζά, αφαιρούνται όλες οι λέξεις που έχουν λιγότερους από 4 χαρακτήρες, αφαιρούνται όλες οι κοινότυπες λέξεις και τέλος εφαρμόζονται αλγόριθμοι εύρεσης της ρίζας μίας λέξεις. Οι λέξεις κλειδιά που απομένουν είναι stemmed το οποίο σημαίνει πως από τις λέξεις διατηρείται μόνο η ρίζα.

Από τις λέξεις κλειδιά ο μηχανισμός οδηγείται σε δύο διαφορετικά στάδια ανάλυσης. Στο πρώτο στάδιο υπάρχει μηχανισμός ο οποίος αναλαμβάνει να δημιουργήσει μία αντιπροσωπευτική περίληψη του κειμένου ενώ στο δεύτερο στάδιο πραγματοποιείται αυτόματη κατηγοριοποίηση του κειμένου βασισμένη σε πρότυπες κατηγορίες που έχουν δημιουργηθεί από επιλεγμένα άρθρα που συλλέγονται καθ' όλη τη διάρκεια υλοποίησης του μηχανισμού. Η εξαγωγή περίληψης βασίζεται σε ευρεστικούς αλγορίθμους. Πιο συγκεκριμένα προσπαθούμε χρησιμοποιώντας λεξικολογική ανάλυση του κειμένου αλλά και γεγονότα για τις λέξεις του κειμένου αν δημιουργήσουμε βάρη για τις προτάσεις του κειμένου. Οι προτάσεις με τα μεγαλύτερη βάρη μετά το πέρας της διαδικασίας είναι αυτές που επιλέγονται για να διαμορφώσουν την περίληψη. Όπως θα δούμε και στη συνέχεια για κάθε άρθρο υπάρχει μία γενική περίληψη αλλά το σύστημα είναι σε θέση να

δημιουργήσει προσωποποιημένες περιλήψεις για κάθε χρήστη. Η διαδικασία κατηγοριοποίησης βασίζεται στη συσχέτιση συνημίτονου συγκριτικά με τις πρότυπες κατηγορίες. Η κατηγοριοποίηση δεν τοποθετεί μία ταμπέλα σε κάθε άρθρο αλλά μας δίνει τα αποτελέσματα συσχέτισης του άρθρου με κάθε κατηγορία.

Ο συνδυασμός των δύο παραπάνω σταδίων δίνει την πληροφορία που εμφανίζεται σε πρώτη φάση στο χρήστη που επισκέπτεται το προσωποποιημένο portal. Η προσωποποίηση στο portal βασίζεται στις επιλογές που κάνουν οι χρήστες, στο χρόνο που παραμένουν σε μία σελίδα αλλά και στις επιλογές που δεν πραγματοποιούν προκειμένου να δημιουργηθεί προφίλ χρήστη και να είναι εφικτό με την πάροδο του χρόνου να παρουσιάζεται στους χρήστες μόνο πληροφορία που μπορεί να τους ενδιαφέρει.

Executive Summary

The scope of this MsC thesis is the extension and upgrade of the mechanism that was constructed during my undergraduate studies under my undergraduate thesis entitled "Construction of a Web Portal with Personalized Access to WWW content".

The aforementioned thesis included the construction of a mechanism that would begin with information retrieval from the WWW and would conclude to representation of information through a portal after applying useful text extraction, text pre-processing and text categorization techniques.

The scope of the MsC thesis is to locate the problematic parts of the system and correct them with better algorithms and also include more modules on the complete mechanism.

More precisely, all the modules are upgraded while more of them are constructed in every aspect of the mechanism. The information retrieval module is based on a simple crawler. The procedure is based on the fact that all the major news portals include RSS feeds. By locating the latest articles that are added to the RSS feeds we are able to locate all the URLs of the HTML pages that include articles. The crawler then visits every simple URL and downloads the HTML page. These pages are filtered by the useful text extraction mechanism in order to extract only the body of the article from the HTML page. This procedure is based on the web-clipping technique. An HTML parser analyzes the DOM model of HTML and locates the nodes (leafs) that include large amounts of text and are close to nodes with large amounts of text. These nodes are considered to include the useful text.

In the extracted useful text we apply a 5 level preprocessing technique in order to extract the keywords of the article. More analytically, we remove the punctuation, the numbers, the words that are smaller than 4 letters, the stopwords and finally we apply a stemming algorithm in order to produce the root of the word.

The keywords are utilized into two different interconnected levels. The first is the categorization subsystem and the second is the summarization subsystem. During the summarization stage the system constructs a summary of the article while the second stage tries to label the article. The labeling is not unique but the categorization applies multi-labeling techniques in order to detect the relation with each of the standard categories of the system. The summarization technique is based on heuristics. More specifically, we try, by utilizing language processing and facts that concern the keywords, to create a score for each of the sentences of the article. The more the score of a sentence, the more the probability of it to be included to the summary which consists of sentences of the text.

The combination of the categorization and summarization provides the information that is shown to our web portal called perssonal. The personalization issue of the portal is based on the selections of the user, on the non-selections of the user, on the time that the user remains on an article, on the time that spends reading similar or identical articles. After a short period of time, the system is able to adopt on the user's needs and is able to present articles that match the preferences of the user only.

Περιεχόμενα

Κατάλογος εικόνων

Κατάλογος κομματιών κώδικα

Γλωσσάρι

Association Pattern	Πρότυπο Συσχέτισης
Boolean	Δυαδική Λογική
Browser	Φυλλομετρητής Ιστού
Categorization	Κατηγοριοποίηση
Classification	Ταξινόμηση
Content	Περιεχόμενο
Corpus	Συλλογή κειμένων με συγκεκριμένες ιδιότητες
Crawler, Bot, Spider	Μηχανισμοί που πραγματοποιούν αυτόματη περιήγηση στις σελίδες του Διαδικτύου
Data Mining	Εξόρυξη Δεδομένων
Decision Tree	Δένδρο Απόφασης
E-Mail	Ηλεκτρονικό Ταχυδρομείο
Embedded Software	Ενσωματωμένο Λογισμικό
Flexible	Ευέλικτος
Format	Μορφοποίηση
Front-End	Περιβάλλον αλληλεπίδρασης χρήστη
Fuzzy	Ασαφές
Generic	Γενικού Περιεχομένου
HTML	Η βασική γλώσσα δομής του διαδικτύου (HyperText Markup Language)
Information Filtering	Φιλτράρισμα πληροφορίας
Information Retrieval	Ανάκτηση Πληροφορίας
Internet	Διαδίκτυο
Keywords	Λέξεις κλειδιά
Knowledge Mining	Εξόρυξη Γνώσης
Link	Σύνδεσμος (αναφέρεται σε ιστοσελίδα)
Machine Understandable	Κατανοητός από μηχανή
Metadata	Μεταδεδομένα
Module	Τμήμα, Κομμάτι
NLP	Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing)
News Portal	Σελίδες ειδησεογραφικού περιεχομένου
Ontology	Οντολογία, αντικείμενο
Portable	Φορητός
Portal	Δικτυακή Πύλη ενημερωτικού περιεχομένου
Preprocessing	Προεπεξεργασία
Punctuation	Στίξη
RSS / RSS Feed	Πρωτόκολλο που καθορίζει κανάλι επικοινωνίας με τη βοήθεια της γλώσσας XML
Search Engine	Μηχανή αναζήτησης
Semantic Web	Σημασιολογικός Ιστός
State of the Art	Οι τρέχουσες εξελίξεις στην επιστήμη
Stemmer	Πρόγραμμα που εφαρμόζει αλγόριθμο εξαγωγής της ρίζας μίας λέξης
Stemming	Η διαδικασία εξαγωγής της ρίζας μίας λέξης

Stopword	Πρόκειται για συγκεκριμένες λέξεις (ουσιαστικά) τα οποία είναι πολύ κοινότυπα στην καθομιλουμένη και συνεπώς δε μπορούν να αποτελέσουν τις λέξεις κλειδιά ενός κειμένου
Tag	Επικεφαλίδα. Ο όρος χρησιμοποιείται για τις δηλώσεις που χρησιμοποιούνται στη γλώσσα διαδικτύου HTML
Text Analysis	Ανάλυση κειμένου
Text Categorization	Κατηγοριοποίηση Κειμένου
Training Set	Σύνολο από κείμενα/λέξεις που μπορούν να χρησιμοποιηθούν για να αποκτήσει «γνώση» μία μηχανή.
User Profile	Προφίλ Χρήστη
Vector Space Model	Μοντέλο Κατηγοριοποίησης που βασίζεται στη χρήση διανυσμάτων και πινάκων
WWW	Παγκόσμιος Ιστός – Διαδίκτυο (World Wide Web)

Συντομογραφίες

DBMS	Database Management System
HTML	HyperText Mark-up Language
IF	Information Filtering
IR	Information Retrieval
LSI	Latent Semantic Indexing
RSS	Rich Site Summary
SVM	Support Vector Machine
URL	Uniform Resource Locator
VSM	Vector Space Model
WWW	World Wide Web
ΑΠ	Ανάκτηση Πληροφορίας
ΒΔ	Βάση Δεδομένων
ΠΣ	Πληροφοριακό Σύστημα

1

ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό υπάρχουν εισαγωγικά στοιχεία για την εργασία

1. ΕΙΣΑΓΩΓΗ

Σκοπός της Μεταπτυχιακής Εργασίας είναι η επέκταση και αναβάθμιση του μηχανισμού που είχε δημιουργηθεί στα πλαίσια της Διπλωματικής Εργασίας που εκπόνησα με τίτλο «Δημιουργία Πύλης Προσωποποιημένης Πρόσβασης σε Περιεχόμενο του WWW».

Η παραπάνω Διπλωματική εργασία περιλάμβανε τη δημιουργία ενός μηχανισμού που ξεκινούσε με ανάκτηση πληροφορίας από το Διαδίκτυο (HTML σελίδες από news portals), εξαγωγή χρήσιμου κειμένου και προεπεξεργασία της πληροφορίας, αυτόματη κατηγοριοποίηση της πληροφορίας και τέλος παρουσίαση στον τελικό χρήστη με προσωποποίηση με στοιχεία που εντοπίζονταν στις επιλογές του χρήστη.

Στην παραπάνω εργασία εξετάστηκαν διεξοδικά θέματα που είχαν να κάνουν με τον τρόπο προεπεξεργασίας της πληροφορίας καθώς και με τον τρόπο αυτόματης κατηγοριοποίησης ενώ υλοποιήθηκαν αλγόριθμοι προεπεξεργασίας πληροφορίας τεσσάρων σταδίων και αλγόριθμος αυτόματης κατηγοριοποίησης βασισμένος σε πρότυπες κατηγορίες.

Τέλος υλοποιήθηκε portal το οποίο εκμεταλλευόμενο την επεξεργασία που έχει πραγματοποιηθεί στην πληροφορία παρουσιάζει το περιεχόμενο στους χρήστες προσωποποιημένο βάσει των επιλογών που αυτοί πραγματοποιούν.

Σκοπός της μεταπτυχιακής εργασίας είναι η εξέταση περισσοτέρων αλγορίθμων για την πραγματοποίηση της παραπάνω διαδικασίας αλλά και η υλοποίησή τους προκειμένου να γίνει σύγκριση αλγορίθμων και παραγωγή ποιοτικότερου αποτελέσματος.

Πιο συγκεκριμένα το στάδιο λήψης πληροφορίας παραμένει ίδιο ενώ όλα τα υπόλοιπα στάδια αναβαθμίζονται. Το στάδιο λήψης πληροφορίας βασίζεται σε έναν απλό crawler λήψης HTML σελίδων από αγγλόφωνα news portals. Οι HTML σελίδες φιλτράρονται προκειμένου από αυτές να γίνει εξαγωγή μόνο του κειμένου και πιο αναλυτικά του χρήσιμου κειμένου ούτως ώστε το κείμενο που εξάγεται να αφορά αποκλειστικά άρθρα.

Στα εξαγόμενα άρθρα πραγματοποιείται προεπεξεργασία πέντε σταδίων με σκοπό να προκύψουν οι λέξεις κλειδιά που είναι αντιπροσωπευτικές του άρθρου. Οι λέξεις κλειδιά που απομένουν είναι stemmed το οποίο σημαίνει πως από τις λέξεις διατηρείται μόνο η ρίζα.

Από τις λέξεις κλειδιά ο μηχανισμός οδηγείται σε δύο διαφορετικά στάδια ανάλυσης. Στο πρώτο στάδιο υπάρχει μηχανισμός ο οποίος αναλαμβάνει να δημιουργήσει μία αντιπροσωπευτική περίληψη του κειμένου ενώ στο δεύτερο στάδιο πραγματοποιείται αυτόματη κατηγοριοποίηση του κειμένου βασισμένη σε πρότυπες κατηγορίες που έχουν δημιουργηθεί από επιλεγμένα άρθρα που συλλέγονται καθ' όλη τη διάρκεια υλοποίησης του μηχανισμού.

Ο συνδυασμός των δύο παραπάνω σταδίων δίνει την πληροφορία που εμφανίζεται σε πρώτη φάση στο χρήστη που επισκέπτεται το προσωποποιημένο portal. Η προσωποποίηση στο portal βασίζεται στις επιλογές που κάνουν οι χρήστες, στο χρόνο που παραμένουν σε μία σελίδα αλλά και στις επιλογές που δεν πραγματοποιούν προκειμένου να δημιουργηθεί προφίλ χρήστη και να είναι εφικτό με την πάροδο του χρόνου να παρουσιάζεται στους χρήστες μόνο πληροφορία που μπορεί να τους ενδιαφέρει.

Μέσα από την εργασία προέκυψαν αποτελέσματα που έχουν να κάνουν με σύγκριση αλγορίθμων σε όλα τα παραπάνω στάδια του μηχανισμού αλλά και ανταπόκριση του μηχανισμού στις ανάγκες του χρήστη.

Μέσα από την εργασία προέκυψαν αποτελέσματα που έχουν να κάνουν με σύγκριση αλγορίθμων σε όλα τα παραπάνω στάδια του μηχανισμού αλλά και ανταπόκριση του μηχανισμού στις ανάγκες του χρήστη.

Η ερευνητική διατριβή που έγινε στα πλαίσια της συγκεκριμένης εργασίας οδήγησε στις παρακάτω δημοσιεύσεις:

Διεθνή Περιοδικά

PeRSSonal's core functionality evaluation: Enhancing text labelling through personalized summaries. Data and Knowledge Engineering Journal, Elsevier Science, 2007, C. Bouras, V. Poulopoulos, V. Tsogkas, 2007, (to appear)

Abstract: Σε αυτή τη δημοσίευση παρουσιάζουμε τα υποσυστήματα κατηγοριοποίησης και περίληψης ενός μηχανισμού που ξεκινά από λήψη σελίδων από το διαδίκτυο και καταλήγει με αναπαράσταση των δεδομένων στον τελικό χρήστη μέσα από ένα δικτυακό τόπο που εφαρμόζει αναλυτικές διαδικασίες προσωποποίησης στο χρήστη. Το σύστημα σκοπεύει να συλλέξει άρθρα από μεγάλα ειδησεογραφικά πρακτορεία και, ακολουθώντας μία αλγοριθμική διαδικασία, να δημιουργήσει μία διαφορετική «εικόνα» των άρθρων προς τον τελικό χρήστη ώστε αυτά να ταιριάζουν στις ανάγκες του χρήστη. Πριν από την παρουσίαση της πληροφορίας στο χρήστη, ο πυρήνας του συστήματος κατηγοριοποιεί αυτόματα την πληροφορία και εξάγει προσωποποιημένες περιλήψεις. Εστιάζουμε την έρευνά μας στον πυρήνα του συστήματος και πιο συγκεκριμένα παρουσιάζουμε αλγορίθμους που χρησιμοποιούνται για κατηγοριοποίηση και για εξαγωγή αυτόματης περίληψης. Οι αλγόριθμοι δε χρησιμοποιούνται αποκλειστικά για την παραγωγή μεμονωμένων δεδομένων αλλά ένας συνδυασμός αλγορίθμων που επιτυγχάνει τη διασύνδεση των μηχανισμών παρουσιάζεται προκειμένου να ενισχυθεί η κατηγοριοποίηση με τη χρήση προσωποποιημένων περιλήψεων.

Διεθνή Συνέδρια

Efficient Summarization Based On Categorized Keywords. The 2007 International Conference on Data Mining (DMIN07), Las Vegas, Nevada, USA, C. Bouras, V. Poulopoulos, V. Tsogkas, 25 - 28 June 2007

Abstract: Η πληροφορία που υπάρχει στο διαδίκτυο είναι αρκετά μεγάλη ώστε να εκτρέπει τους χρήστες στην προσπάθεια αναζήτησης πληροφορίας. Προκειμένου να αποφευχθούν τα προβλήματα που δημιουργούνται από την πληθώρα δεδομένων του Διαδικτύου πολλοί μηχανισμοί προσωποποίησης δεδομένων και περίληψης δεδομένων έχουν προταθεί. Σε αυτή τη δημοσίευση παρουσιάζουμε ένα μηχανισμό όπου εφαρμόζουμε τεχνικές αυτόματης εξαγωγής περίληψης σε άρθρα που έχουν εξαχθεί από το Διαδίκτυο και βασιζόμαστε σε τεχνικές κατηγοριοποίησης προκειμένου να επιτύχουμε αποδοτικότερα αποτελέσματα. Μέσα από αναλυτικά πειράματα αποδεικνύουμε πως η διαδικασία αυτόματης εξαγωγής περίληψης μπορεί να επηρεάσει το μηχανισμό κατηγοριοποίησης και το αντίστροφο. Αυτό σημαίνει πως όταν τα αποτελέσματα της κατηγοριοποίησης δεν είναι σαφή τότε μπορούμε να εφαρμόσουμε τον αλγόριθμο αυτόματης εξαγωγής περίληψης προκειμένου να λάβουμε καλύτερα αποτελέσματα στην κατηγοριοποίηση και από την άλλη μεριά, αν ο μηχανισμός αυτόματης εξαγωγής περίληψης δεν είναι σε θέση να αναγνωρίσει σαφώς την περίληψη ενός κειμένου εφαρμόζουμε παράγοντες κατηγοριοποίησης προκειμένου να παράγουμε μία καλύτερη περίληψη. Παράλληλα, σε αυτή τη δημοσίευση παρουσιάζουμε τον τρόπο με τον οποίο ο συνδυασμός των παραπάνω μπορεί να οδηγήσει όχι μόνο σε καλύτερα αποτελέσματα μεταξύ των προαναφερθέντων αλλά και στην υποστήριξη μιας προσωποποιημένης πύλης. Τέλος, προτείνουμε έναν συνολικό μηχανισμό ο οποίος μπορεί να χρησιμοποιηθεί προκειμένου να παρέχουμε στους χρήστες με εργαλεία που θα τον βοηθήσουν στην ευκολότερη εύρεση πληροφορίας.

Personalizing text summarization based on sentence weighting. IADIS European First International Conference Data Mining (ECDM 2007), Lisbon, Portugal, C. Bouras, V. Poulopoulos, V. Tsogkas, 3 - 8 July 2007

Abstract: Η πληροφορία που υπάρχει στο Διαδίκτυο είναι τόσο μεγάλη ώστε να εμποδίζει τους χρήστες στην προσπάθεια εύρεσης χρήσιμης πληροφορίας. Παράλληλα, η μεγάλη ανάπτυξη της τεχνολογίας όσον αφορά τις συσκευές μικρού μεγέθους και η δυνατότητα αυτών να συνδέονται με το διαδίκτυο έχει οδηγήσει σε πολλά προβλήματα που αφορούν τόσο την εύρεση πληροφορίας όσο και την παρουσίαση πληροφορίας. Μία λύση σε αυτό το πρόβλημα είναι η προσωποποίηση του διαδικτύου και η προσπάθεια μείωσης της ποσότητας του κειμένου που παρουσιάζεται στο χρήστη με χρήση αλγορίθμων Πολλοί μηχανισμοί περίληψης κειμένου έχουν παρουσιαστεί προς αυτή την κατεύθυνση με σκοπό να μειώσουν την πληροφορία που εμφανίζεται στο χρήστη στο ελάχιστο και παράλληλα πολλοί δικτυακοί τόποι παρουσιάζουν μηχανισμούς προσωποποίησης στο χρήση. Ωστόσο αυτές οι τεχνικές δε χρησιμοποιούνται ακόμα από κοινού για την καλύτερη επίλυση του προβλήματος. Σε αυτή τη δημοσίευση παρουσιάζουμε ένα μηχανισμό που κατασκευάζει προσωποποιημένες περιλήψεις κειμένων για τους χρήστες ενός δικτυακού τόπου. Ο δικτυακός τόπο αναπαράγει άρθρα που έχει συλλέξει από το διαδίκτυο και τα παρουσιάζει στους χρήστες βάσει των αναγκών τους. Επίσης παρουσιάζουμε την αξιολόγηση των μηχανισμών του συστήματός μας και παρουσιάζουμε ένα σημαντικό στοιχείο για το δικτυακό τόπο που δεν είναι άλλο από την υποστήριξη συσκευών μικρού μεγέθους.

The importance of the difference in text types to keyword extraction: Evaluating a mechanism. 7th International Conference on Internet Computing 2006 (ICOMP 2006), Las Vegas, Nevada, USA, C. Bouras, C. Dimitriou, V. Poulopoulos, V. Tsogkas, 26 - 29 June 2006, pp. 43 - 49

Abstract: Η πληροφορία υπάρχει παντού γύρω μας. Η εξάπλωση του διαδικτύου έχει βοηθήσει σε αυτή την κατεύθυνση. Το διαδίκτυο μας τροφοδοτεί με εξωπραγματικές ποσότητες πληροφορίας και η εκτενής χρήση υπολογιστών και άλλων συσκευών έχει οδηγήσει σε μία κατάσταση όπου διαθέτουμε αρκετή πληροφορία στα χέρια μας αλλά τις περισσότερες φορές είναι άχρηστη. Ο άνθρωπος δε μπορεί να βρει πληροφορία που πραγματικά χρειάζεται ακόμα κι αν την έχουν ήδη στην κατοχή τους. Πόσες φορές έχει χρειαστεί να αναζητήσετε πληροφορίες για ένα συγκεκριμένο άρθρο, ένα συγκεκριμένο mail ή ακόμα και ένα SMS. Για το λόγο αυτό έχουν προταθεί πολλές τεχνικές ανάκτησης πληροφορίας από κάθε μέσο. Σε αυτή τη δημοσίευση παρουσιάζουμε την πειραματική αποτίμηση ενός μηχανισμού εξαγωγής λέξεων κλειδιών και παρουσιάζουμε πως αντιμετωπίζουμε τα διαφορετικά κείμενα που δίνουμε σαν είσοδο στο μηχανισμό μας. Ο μηχανισμός εξαγωγής των λέξεων κλειδιών είναι κομμάτι ενός συνολικού μηχανισμού που περιλαμβάνει ανάκτηση πληροφορίας, κατηγοριοποίηση και αυτόματη εξαγωγή περίληψης.

Scalability of text classification. 2nd International Conference on Web Information Systems and Technologies (WEBIST 2006), Setubal, Portugal, I. Antonellis, C. Bouras, V. Poulopoulos, A. Zouzias, 19 - 22 April 2006, pp. 408 - 413

Abstract: Σε αυτή τη δημοσίευση ανιχνεύουμε θέματα κλιμάκωσης που αφορούν την κατηγοριοποίηση κειμένου όπου χρησιμοποιώντας προκατηγοριοποιημένα κείμενο προσπαθούμε να χτίσουμε μηχανισμούς κατηγοριοποίησης που είναι σε θέση να ελέγχουν και να εντοπίζουν την κατηγορία ενός κειμένου όπως επίσης και να το κατηγοριοποιούν σε περισσότερες από μία κατηγορίες. Ένα νέο μοντέλο προβλημάτων κατηγοριοποίησης, που ονομάζεται κλιμακωτό παρουσιάζεται και μπορεί να μοντελοποιήσει πολλά προβλήματα στον τομέα της ανάκτησης πληροφορίας από το διαδίκτυο. Ως κλιμάκωση ορίζεται η δυνατότητα του κατηγοριοποιητή να διαμορφώνει τα αποτελέσματα της κατηγοριοποίησης ανά χρήστη. Επιπρόσθετα, ελέγχουμε διάφορους τρόπους

ανάλυσης της διαδικασίας προσωποποίησης σαν κομμάτι της κατηγοριοποίησης αναλύοντας γνωστά datasets και χρησιμοποιώντας υπάρχοντες κατηγοριοποιητές. Παρουσιάζουμε λύσεις για το πρόβλημα των κλιμακωτών προβλημάτων κατηγοριοποίησης στηριζόμενοι σε συγκεκριμένες τεχνικές κατηγοριοποίησης και παρουσιάζουμε έναν αλγόριθμο που βασίζεται σε σημασιολογική ανάλυση χρησιμοποιώντας αποδόμηση προτάσεων.

Personalized News Categorization through Scalable Text Classification. The Eight Asia Pacific Web Conference (APWeb – 06), Harbin, China, I. Antonellis, C. Bouras, V. Poulopoulos, 16 - 18 January 2006, pp. 391 - 401

Abstract: οι υπάρχοντες δικτυακοί τόποι ειδησεογραφικού περιεχομένου έχουν σαν σκοπό να παρέχουν στους χρήστες άρθρα συγκεκριμένων κατηγοριών. Αυτή η διαδικασία βελτιώνει την παρουσίαση της πληροφορίας στο χρήστη. Στη συγκεκριμένη δημοσίευση παρουσιάζουμε ένα βελτιστοποιημένο τρόπο παρουσίασης, κατηγοριοποίησης και προσωποποίησης των δεδομένων που χρησιμοποιεί τη γνώση του χρήστη για ένα συγκεκριμένο θέμα προτού το παρουσιάσει. Η διαδικασία κατηγοριοποίησης του συστήματος βασίζεται σε ανάλυση των προτάσεων του κειμένου. Ο κλασικός πίνακας term-to-term αντικαθίσταται από τον πίνακα term-to-sentence κάτι που μας επιτρέπει να ελέγχουμε περισσότερα στοιχεία που αφορούν κάθε κείμενο.

Παράλληλα με αυτές τις προηγούμενες εργασίες ακολουθούν όλες οι δημοσιεύσεις που έχω πραγματοποιήσει από το 2004 μέχρι σήμερα.

International Journals

Enhancing a Web Based Community: the case of SIG-GLUE

International Journal of Web Based Communities (IJWBC), Inderscience Publishers, Vol. 2, No 1, I. Antonellis, C. Bouras, V. Kapoulas, V. Poulopoulos, 2006, pp. 112 – 130

Abstract: Η ανάπτυξη του διαδικτύου έχει πάρει μεγάλες διαστάσεις με τον αριθμό των κοινοτήτων διαδικτύου που υπάρχουν και των αριθμό αυτών που δημιουργούνται καθημερινά να αυξάνεται δραματικά. Παράλληλα, αυτό το φαινόμενο είναι μόδα και στις υπηρεσίες που προσφέρονται μέσω κινητών τηλεφώνων. Μία χαρακτηριστική περίπτωση είναι αυτή της Ελλάδας που σε περίοδο πέντε ετών περίπου 5 εκατομμύρια χρήστες σύναψαν συμβόλαιο με μία συγκεκριμένη εταιρία. Οι κοινότητες του διαδικτύου δεν είναι στατικές, αλλά δυναμικές. Η φιλοσοφία είναι απλή: μία καθολική κοινότητα πρέπει να είναι κινητή. Σε αυτή τη δημοσίευση παρουσιάζουμε την επέκταση της κοινότητας του SIG-GLUE προκειμένου να υποστηρίζει κινητούς χρήστες σε όλες τις υπηρεσίες που προσφέρει.

International Conferences

Input here - Execute there through networks: the case of gaming

The 15th Workshop on Local and Metropolitan Area Networks (LANMAN 2007), Princeton, NJ, USA, C. Bouras, V. Poulopoulos, I. Sengounis, V. Tsogkas, 10 - 13 June 2007

Abstract: όσο η επιστήμη των υπολογιστών παρουσιάζει εξελίξεις στον τομέα των δικτύων τα online παιχνίδια γίνονται ολοένα και μία μεγαλύτερη τάση. Ακολουθώντας τις τάσεις της εποχής το ευρωπαϊκό έργο Games @ Large παρουσιάζει μία καινούρια πλατφόρμα για την εκτέλεση διαδραστικών εφαρμογών πάνω από ασύρματα τοπικά δίκτυα. Σκοπός του έργου είναι η κατασκευή μίας καινούριας αρχιτεκτονικής η οποία θα ενισχύσει υπάρχουσες συσκευές όπως set-top-box που δεν έχουν πολλούς πόρους προκειμένου να προσφέρει μεγαλύτερες εμπειρίες. Σε αυτή τη δημοσίευση παρουσιάζεται υποσύστημα διαχείρισης της εισόδου των συσκευών το οποίο θα κατασκευαστεί στο πλαίσιο του έργου. Αναλυτικά παρουσιάζουμε τη γενική αρχιτεκτονική του συνολικού μηχανισμού

εστιάζοντας στα κομμάτια που λαμβάνουν την πληροφορία από την τελική συσκευή και την εκτελούν στο κομμάτι του εξυπηρετητή. Το υποσύστημα πελάτη λαμβάνει την είσοδο, τη στέλνει πάνω από το ασύρματο δίκτυο στον εξυπηρετητή ο οποίος είναι υπεύθυνος για τη σωστή εκτέλεσή της.

A Unified Framework for Political Support e - Democracy Practices

IADIS International Conference WWW/Internet 2005, Lisbon, Portugal, Volume II, C. Bouras, E. Giannaka, T. Karounos, A. Priftis, V. Poulopoulos, T. Tsiatsos, 19 - 22 October 2005, pp. 119 - 123

Abstract: Η ηλεκτρονική διακυβέρνηση και η ηλεκτρονική δημοκρατία αποτελούν ένα κυρίαρχο θέμα σε όλα τα επίπεδα των πολιτικών της κοινωνίας της πληροφορίας. Σε αυτή την κατεύθυνση πληθώρα από προσπάθειες έχουν γίνει και πολλά συστήματα έχουν αναπτυχθεί. Σε αυτή τη δημοσίευση παρουσιάζουμε μία μεθοδολογία για το σχεδιασμό και την υλοποίηση υπηρεσιών διαδικτύου που θα υποστηρίζουν πρακτικές ηλεκτρονικής δημοκρατίας. Παράλληλα, παρουσιάζουμε τις προσπάθειες που γίνονται από ένα ελληνικό κόμμα σε αυτή την κατεύθυνση, το σχεδιασμό δηλαδή και την υλοποίηση ενός κοινού πλαισίου προκειμένου να υποστηρίζονται και να προσφέρονται υπηρεσίες ηλεκτρονικής δημοκρατίας.

Creating a Polite Adaptive and Selective Incremental Crawler

IADIS International Conference WWW/INTERNET 2005, Lisbon, Portugal, Volume I, C. Bouras, V. Poulopoulos, A. Thanou, 19 - 22 October 2005, pp. 307 - 314

Abstract: Σε αυτή τη δημοσίευση παρουσιάζουμε ένα μηχανισμό ανάκτησης δεδομένων από το διαδίκτυο ο οποίος έχει σχεδιαστεί για να υποστηρίζει συστήματα ανάλυσης πληροφορίας. Ένας τέτοιος μηχανισμός θα πρέπει να είναι αποδοτικός, φιλικός προς τις σελίδες που επισκέπτεται και προς το δίκτυο που φιλοξενεί τις σελίδες. Συνεπώς είναι μεγάλης σημασίας το να ακολουθηθούν συγκεκριμένες μέθοδοι και συγκεκριμένοι κανόνες. Παράλληλα, ο μηχανισμός έχει σχεδιαστεί με τέτοιο τρόπο ώστε να χρησιμοποιεί έναν ιδιαίτερο αλγόριθμό επιλεκτικής προσπέλασης των σελίδων ο οποίος χρησιμοποιείται προκειμένου να επιτευχθεί η αποδοτικότερη και δικαιότερη λειτουργία του μηχανισμού. Η δομή και ο σχεδιασμός του μηχανισμού είναι απλός αλλά τα αποτελέσματα μας δείχνουν πως αυτή η απλότητα κάνει το μηχανισμό μας ιδιαίτερα ισχυρό.

Design and Implementation of a Game Based Learning Related Community

IADIS International Conference, Web Based Communities 2005, Algarve, Portugal, I. Antonellis, C. Bouras, V. Kapoulas, V. Poulopoulos, 23 - 25 February 2005, pp. 215 - 222

Abstract: Μία κοινότητα του διαδικτύου έχει σαν σκοπό να προσφέρει εργαλεία επικοινωνίας και συμμετοχής σε ανθρώπους με κοινά ενδιαφέροντα. Αυτή η δημοσίευση περιγράφει την αρχιτεκτονική και τη λειτουργικότητα μίας κοινότητας που σκοπεύει να φέρει κοντά χρήστες που ενδιαφέρονται για το πεδίο της εκπαίδευσης μέσα από παιχνίδια και τις δια βίου εκπαίδευσης. Η αρχιτεκτονική και θέματα υλοποίησης της πλατφόρμας αναλύονται και η χρήση τους εξηγείται διεξοδικά. Αναλυτικά, οι χρήστες της κοινότητας έχουν στη διάθεσή τους εργαλεία για να μπορούν να μοιράζονται τη γνώση τους και τις εμπειρίες τους όσον αφορά τη μάθηση μέσα από παιχνίδια και αυτό επιτυγχάνεται με τη βοήθεια forum και chat όπως επίσης και ειδήσεων, συναντήσεων και χρήση κοινόχρηστων χώρων. Όλες αυτές οι υπηρεσίες εμπλουτίζονται προκειμένου να ταιριάξουν με τις ανάγκες της κοινότητας στην οποία απευθυνόμαστε.

Game Based learning for Mobile Users

The 6th International GAME - ON Conference on the theme: Computer Games: AI and Mobile Users, Louisville, Kentucky, USA, I. Antonellis, C. Bouras, V. Poulopoulos, 27 - 30 July 2005

Abstract: Η χρήση παιχνιδιών για μάθηση έχει μελετηθεί σαν ένα σημαντικό βοήθημα στην κλασσική εκπαιδευτική διαδικασία. Οι υπάρχουσες πλατφόρμες του διαδικτύου που χρησιμοποιούν τη δύναμη του διαδικτύου για να παρέχουν πρόσβαση σε πληροφορίες που αφορούν μάθηση μέσα από παιχνίδια έχουν σαν σκοπό να δημιουργήσουν κοινότητες συνεργαζόμενων ανθρώπων. Ωστόσο, αυτές οι προσπάθειες στοχεύουν αποκλειστικά στους χρήστες του διαδικτύου αφήνοντας έξω από το παιχνίδι του νέους παίκτες που δεν είναι άλλοι από τους χρήστες συσκευών μικρού μεγέθους οι οποίες πλέον έχουν άμεση πρόσβαση στο διαδίκτυο. Σε αυτή τη δημοσίευση παρουσιάζουμε τον τρόπο δόμησης των υπηρεσιών έτσι ώστε να είναι άμεσα διαθέσιμες στους χρήστες συσκευών μικρού μεγέθους χωρίς αυτοί να αντιμετωπίζουν προβλήματα χρήσης.

A Web Clipping Service's Information Extraction Mechanism

3rd International Conference on Universal Access in Human - Computer Interaction, Las Vegas, Nevada, USA, C. Bouras, G. Kounenis, I. Misedakis, V. Poulopoulos, 22 - 27 July 2005

Abstract: Η υπερβολική πληροφορία είναι ένα από τα μεγαλύτερα προβλήματα του Διαδικτύου. Οι χρήστες συχνά χάνονται στην πληθώρα πληροφορίας όταν αναζητούν κάποιο ακόμα και συγκεκριμένο θέμα. Παρά το γεγονός ότι οι χρήστες έχουν συγκεκριμένες ανάγκες σε πληροφορία, η χρήση των μηχανών αναζήτησης οδηγεί σε αναζήτηση σε δεκάδες χιλιάδες δικτυακούς τόπους από πολύ σχετικούς με το θέμα έως και εντελώς άσχετους. Μία λύση για το συγκεκριμένο πρόβλημα είναι η υπηρεσία "web-clipping". Μία τέτοια υπηρεσία ψάχνει συνεχώς στο διαδίκτυο για να εντοπίσει σελίδες που μπορεί να ενδιαφέρουν μία μερίδα χρηστών και τότε ενημερώνει αυτούς τους χρήστες πως πρέπει να επισκεφθούν τη συγκεκριμένη σελίδα. Αυτή η δημοσίευση παρουσιάζει το μηχανισμό εξαγωγής πληροφορίας μίας υπηρεσίας web-clipping η οποία σχεδιάζεται σαν ένα κομμάτι ενός συνολικού μηχανισμού αναζήτησης και διαχείρισης πληροφορίας. Ο μηχανισμός χρησιμοποιείται για να εξάγει πληροφορία από σελίδες του διαδικτύου και αναζητά ποιοι σύνδεσμοι πρέπει να ακολουθώνται.

Multi-Agent Cooperation Infrastructure to Support Patient-Oriented Telecare Services

4th International Conference on Information Technology: Research and Education. Poulopoulos, Vassilis; Gortzis, Lefteris; Bakettas, Ioannis; Nikiforidis, George; 16-19 Oct. 2006 Page(s):253 – 257

Abstract: Οι τεχνολογίες που βασίζονται σε agents έχουν αρχίσει να χρησιμοποιούνται από ολοένα και περισσότερες εφαρμογές e-commerce. Ωστόσο, οι κλασσικοί agents που έχουν συγκεκριμένες εφαρμογές δεν έχουν τη δυνατότητα να μεταβάλουν τη συμπεριφορά τους δυναμικά και είναι περιορισμένης χρήσης σε εφαρμογές ιατρική αφού δε μπορούν να αλλάξουν ρόλους και να υποστηρίξουν την πολυπλοκότητα των συστημάτων τηλεϊατρικής. Σε αυτή τη δημοσίευση παρουσιάζεται ο σχεδιασμός και η υλοποίηση ενός συστήματος που είναι σε θέση να συλλέγει και να αναλύει μεγάλη ποσότητα ετερογενών ιατρικών δεδομένων προκειμένου να υποστηρίξει υπηρεσίες τηλεϊατρικής. Στα πλαίσια της εργασίας μας κατασκευάσαμε ένα σύστημα που διαβάζει δεδομένα σε XML μορφή και επιτρέπει την προσαρμογή στα εκάστοτε δεδομένων. Τα βασικά χαρακτηριστικά του συστήματος είναι πως μπορεί να «ακολουθήσει» τον ασθενή σε τρία βασικά βήματα: ο ασθενής προτού εισαχθεί στο σύστημα, ο ασθενής όσο ελέγχεται από το σύστημα και πρόβλεψη για τον ασθενή στο μέλλον. Παρά το γεγονός ότι βρισκόμαστε σε πρωταρχικό στάδιο διαμόρφωσης του δυναμικού χαρακτήρα του agent έχουμε ήδη δημιουργήσει κλινικά αποδεκτά αποτελέσματα χρησιμοποιώντας δεδομένα από 86 ασθενείς του τμήματος Καρδιολογίας του Πανεπιστημιακού Νοσοκομείου της Πάτρας.

Articles in Encyclopedias

Content Transformation Techniques

Encyclopaedia of Mobile Computing & Commerce (EMCC), Vol. 1, IDEA GROUP Publishing, I. Antonellis, C. Bouras, V. Poulopoulos, 2007, pp. 119 – 123

2

ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Στο κεφάλαιο αυτό περιγράφονται τα προβλήματα που καλείται να λύσει ο μηχανισμός που αναπτύχθηκε στη συγκεκριμένη εργασία.

Αναλυτικά:

- Συλλογή Δεδομένων
- Φιλτράρισμα Δεδομένων
- Προεπεξεργασία Πληροφορίας
- Προσωποποίηση στο Χρήστη
- Συμμετοχή του χρήστη στις διαδικασίες του συστήματος

2. ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Η ανάπτυξη νέων τεχνολογιών, κυρίως τα τελευταία δέκα χρόνια, έχουν φέρει την τεχνολογία σε όλους τους τομείς της καθημερινής μας ζωής. Μιας καθημερινής ζωής που κάνει κυρίως τη νέα γενιά ανθρώπων – αλλά και τους παλαιότερους λάτρεις των υπολογιστών και των gadgets να βλέπουν το διαδίκτυο σαν ένα αγαθό. Το αγαθό που ονομάζεται διαδίκτυο αυξάνεται δραματικά τα τελευταία χρόνια με ραγδαίους ρυθμούς με αποτέλεσμα να μιλούμε για μία κοινωνία, για μία κοινότητα χρηστών, η οποία απαρτίζεται από περισσότερα από 11 εκατομμύρια «σπίτια». Τα «σπίτια» αυτά αποτελούν τους δικτυακούς τόπους, σημεία ενημέρωσης και συνάντησης των χρηστών του διαδικτύου. Οι χώροι αυτοί μπορούν να κατηγοριοποιηθούν βάση πληθώρας οντολογιών ανάλογα με το περιεχόμενό τους. Έτσι, αυτή τη στιγμή, το πολυπληθές διαδίκτυο περιέχει δικτυακούς τόπους που αποβλέπουν στην επικοινωνία, στην ενημέρωση, στη διαφήμιση, στην προβολή, στην αυτοπροβολή, στην πληρότητα επιχειρηματικών διαδικασιών, στη διαμόρφωση ενός νέου μοντέλου κοινωνίας βασισμένο σε ηλεκτρονικά πρότυπα με προσφορά δημοκρατικών ελεύθερων και μη υπηρεσιών. Σε γενικές γραμμές, πολλές επιχειρήσεις, είτε ατομικές, είτε ομαδικές διαθέτουν ένα δικτυακό τόπο σαν απόρροια ενός πλήρους συστήματος διαχείρισης διαδικασιών της εταιρίας, σαν ένα μέσο προβολής, σαν ένα τόπο συνάντησης, σαν ένα κομμάτι που είναι απαραίτητο για τη συμμόρφωση με τα νέα κοινωνικά πρότυπα που θέλουν το Διαδίκτυο να κυριαρχεί σε κάθε κομμάτι της ζωής μας.

Όπως κάθε κοινωνία, έτσι και το Διαδίκτυο, έχει τα δικά του προβλήματα. Πηγή αυτών των προβλημάτων μπορεί να θεωρηθεί η «άναρχη δόμησή του», η έλλειψη σαφούς νομοθεσίας αλλά και η αίσθηση ελευθερίας που αφήνει στους «κατοίκους» του να ενεργούν ουσιαστικά κατά βούληση, βρίσκοντας στο Διαδίκτυο μία επανάσταση που θέλουν στην πραγματική τους ζωή, έναν τρόπο έκφρασης ιδεών, έναν τρόπο έκφρασης της γνώσης και της μάθησης. Στη συγκεκριμένη εργασία δε θα αναλωθούμε με την καταγραφή των πολλών, αν μη τι άλλο, προβλημάτων του Διαδικτύου αλλά θα επικεντρωθούμε σε ένα κομμάτι των προβλημάτων που προκύπτουν από την αέναη, καθημερινή και καταιγιστική δημιουργία δεδομένων και πληροφοριών. Ακόμα περισσότερο θα εστιάσουμε την προσοχή μας στις πληροφορίες που δημιουργούνται σε καθημερινή βάση από την πληθώρα των ενημερωτικών δικτυακών πυλών που κατακλύζουν στην κυριολεξία το Διαδίκτυο. Ο λόγος για τα γνωστά news portals. Πρόκειται για Δικτυακούς τόπους που σαν στόχο έχουν την ενημέρωση των χρηστών του Διαδικτύου για τα φλέγοντα – κυρίως – νέα σε παγκόσμιο επίπεδο. Μερικά και πολύ σημαντικά από αυτά είναι το CNN [122], το BBC [123], το Reuters [124], το FoxNews [125], καθώς και οι υπηρεσίες που προσφέρονται από τους πολυπληθείς και από τους πλέον αναγνωρίσιμους δικτυακούς τόπους Google [119] και Yahoo [121].

Οι Δικτυακοί αυτοί τόποι εστιάζονται στο να ενημερώνουν τους χρήστες τους για ότι συμβαίνει καθημερινά στον πλανήτη. Τα νέα/άρθρα παρουσιάζονται με δομημένο τρόπο στις συγκεκριμένες σελίδες, ωστόσο το πλήθος τους είναι τέτοιο ώστε να είναι σχεδόν αδύνατο από κάποιον χρήστη να μπορέσει εντός του εικοσιτετραώρου να παρακολουθήσει όλες τις ειδήσεις που δημοσιεύονται στις πολλές διαφορετικές κατηγορίες. Ακόμα και η εστίαση σε μία συγκεκριμένη κατηγορία απαιτεί τη συνεχή και διαρκή παρακολούθηση κάθε δικτυακού τόπου προκειμένου να υπάρχει πλήρης ενημέρωση. Επίσης, πολλά από αυτά τα νέα παρουσιάζονται από την οπτική γωνία του αρθρογράφου καθώς σπάνια – πλέον – δημοσιεύονται ακέραια ακόμα και τα δελτία τύπου με αποτέλεσμα να χάνεται, συχνά, το κριτήριο της αντικειμενικότητας μίας είδησης. Απόρροια όλων των παραπάνω είναι το εξής: οι χρήστες του διαδικτύου δυσκολεύονται στον εντοπισμό μίας είδησης που τους ενδιαφέρει με αποτέλεσμα να αναλώνουν το χρόνο τους

στην αναζήτηση της είδησης, του νέου, του άρθρου, παρά στην ανάγνωση του ίδιου του άρθρου.

Η παρουσία των RSS (Rich Site Summary), που σε ελεύθερη μετάφραση θα μπορούσαμε να καθονομάσουμε «Περίληψη του Δικτυακού Τόπου», έρχεται να δώσει μία πρώτη λύση στο δυσβάσταχτο πρόβλημα της ανεύρεσης ενός ενδιαφέροντος άρθρου από τους αναγνώστες – χρήστες του Διαδικτύου. Η αρχή της χρήσης των RSS από τους διαχειριστές των δικτυακών τόπων φέρνει μία νέα επανάσταση και αλλάζει τα δεδομένα στην καθημερινή παγκόσμια ειδησεογραφία. Οι χρήστες έχουν ένα ακόμα κανάλι επικοινωνίας που τους προσφέρει το ελπιδοφόρο Internet. Το κανάλι είναι μία διεύθυνση – αυτή του RSS – η πρόσβαση στην οποία επιτρέπει στους χρήστες να «έρθουν σε επαφή» με την πληροφορία που επιθυμούν και μόνον και όχι με τα υπόλοιπα – άχρηστα για τους χρήστες – στοιχεία μίας ιστοσελίδας. Το μόνο που είναι απαραίτητο είναι ένα πρόγραμμα ανάγνωσης RSS Feeds (RSS Reader) ενώ στην πορεία ακόμα και αυτό δεν είναι αναγκαίο καθότι φυλλομετρητές του Διαδικτύου έχουν τη δυνατότητα ανάλυσης του XML εγγράφου και παρουσίασης αυτού με δομημένο και ευδιάκριτο τρόπο στους τελικούς χρήστες.

Παράλληλα με τα RSS μια καινούρια τάση ξεκινά να επικρατεί στο διαδίκτυο. Ο πετυχημένος επιχειρηματικά όρος "My" περνά και στο Διαδίκτυο. Οι τεχνολογίες του διαδικτύου επιτρέπουν στο χρήστη αυτό που ονομάζουμε προσωποποίηση. Ο χρήστης έχει τη δυνατότητα να επισκεφθεί ένα δικτυακό τόπο, να εγγραφεί σε αυτόν και να δημιουργήσει τη δική του σελίδα. Η σελίδα, φυσικά, δε γεμίζει με πληροφορίες του χρήστη. Ο χρήστης έχει το δικαίωμα να επιλέξει ποιο από το περιεχόμενο του δικτυακού τόπου επιθυμεί να βλέπει στη δική του σελίδα καθώς και με ποιον τρόπο. Οι χρήστες γίνονται κομμάτι των δικτυακών τόπων έστω και ιδεατά και καθορίζουν τον τρόπο παρουσίασης των δεδομένων τα οποία – δίνεται η αίσθηση – ότι τα καθορίζουν οι ίδιοι. Η πραγματικότητα, όμως, οδηγεί αλλού. Οι χρήστες γίνονται δέσμιοι αυτών των τεχνολογιών, που φαίνεται πως έρχονται, όχι μόνο εκμεταλλευόμενες την ανάπτυξη που παρουσιάζουν, αλλά για να πολεμήσουν τα «κανάλια επικοινωνίας» που απομάκρυναν τους χρήστες από τους δικτυακούς τόπους.

Η κατάσταση, λοιπόν, είναι η εξής: οι χρήστες έχοντας κουραστεί από την ανούσια πληροφορία που έρχεται εμπρός τους τη στιγμή που περιδιαβαίνουν έναν δικτυακό τόπο ειδησεογραφικού περιεχομένου καταφεύγουν στα RSS. Με αυτή την αλλαγή στον τρόπο «επίσκεψης» μίας σελίδας, τα μεγάλα ειδησεογραφικά πρακτορεία παρατηρούν τη χαμηλή επισκεψιμότητα συγκεκριμένων σελίδων του δικτυακού τους τόπου οι οποίες ουσιαστικά δεν προβάλλονται προς το χρήστη ο οποίος αρκείται στο κανάλι επικοινωνίας που έχει και αποφεύγει κάθε επίσκεψη στο δικτυακό τόπο. Παράλληλα, η ανάπτυξη νέων τεχνολογιών και υπηρεσιών για το Διαδίκτυο κάνει τους διαχειριστές δικτυακών τόπων να επιθυμούν ακόμα μεγαλύτερη επισκεψιμότητα στις σελίδες τους προσφέροντας διαδραστικές υπηρεσίες, υπηρεσίες πολυμέσων κ.α. Μία κρυφή διαμάχη έχει ξεκινήσει ανάμεσα στις υπηρεσίες που διώχνουν τους χρήστες από το δικτυακό τόπο και σε αυτές που τον φέρνουν ακόμα πιο κοντά σε αυτόν. Κάθε υπηρεσία προσπαθεί να υπερισχύσει της άλλης προσφέροντας ολοένα και περισσότερα στοιχεία. Το RSS έχει περιορισμένες δυνατότητες ενώ οι υπηρεσίες προσωποποιημένης πρόσβασης έχουν πολλά να προσφέρουν στους χρήστες. Είναι φανερό πως οι δικτυακοί τόποι, σαν μία κανονική επιχείρηση, επιθυμούν οι χρήστες να «έρχονται» στο δικτυακό τόπο, να επισκέπτονται όλες τις σελίδες, να βλέπουν τις διαφημίσεις, να αξιοποιούν τις νέες υπηρεσίες, να χρησιμοποιούν κάθε δεδομένο που τους προσφέρεται.

Όσο εντυπωσιακά και αν φαίνονται όλα αυτά οι σχεδιαστές των υπηρεσιών έχουν παραλείψει σημαντικά στοιχεία. Πόσο εξοικειωμένοι είναι οι χρήστες στη χρήση περίπλοκων συστημάτων; Έχουν όλοι οι χρήστες αρκετά μεγάλη ταχύτητα στην πρόσβαση στο διαδίκτυο προκειμένου να μπορούν να χρησιμοποιούν χωρίς πρόβλημα τις προσφερόμενες υπηρεσίες; Οι χρήστες έχουν ερωτηθεί για τις

πληροφορίες που θα επιθυμούσαν να τους διατίθενται; Αποτέλεσμα όλων των παραπάνω είναι: προσωποποιημένες σελίδες δικτυακών τόπων, όπου ο χρήστης αδυνατεί να τις σχεδιάσει όπως επιθυμεί καθότι «χάνεται» στην πληθώρα δεδομένων που τους παρουσιάζονται, υπερπολλαπλασιασμός των καναλιών RSS των δικτυακών τόπων με αποτέλεσμα ο χρήστης να αντιμετωπίζει το ίδιο χάος. Τρανταχτό παράδειγμα αποτελεί το RSS Feed του CNN που αποτελείται από περισσότερα από 20 επικαλυπτόμενα κανάλια. Τέλος κάτι πολύ σημαντικό, κανείς δεν επιχειρεί να συνδυάσει τις δύο υπηρεσίες οι οποίες δε φαίνεται να διαφέρουν μεταξύ τους. Κανένας δικτυακός τόπος δεν προσπαθεί να συνδυάσει προσωποποιημένες πληροφορίες και RSS feeds.

Αποδελτιώνοντας, λοιπόν, όλα τα παραπάνω καταλήγουμε στα παρακάτω:

- Προσωποποιημένες σελίδες
  - Δύσχρηστες – πολύπλοκες
  - Βασίζονται σε λέξεις κλειδιά ή ακόμα χειρότερα σε γενικές κατηγορίες μόνον
  - Ο χρήστης σε κάθε περίπτωση παραμένει εκτός της διαδικασίας κατηγοριοποίησης ή κατασκευής περίληψης που παρουσιάζεται στην προσωποποιημένη σελίδα
- RSS feeds
  - Ο αριθμός τους είναι υπερβολικά μεγάλος
  - Ο αριθμός των άρθρων που περιέχουν είναι υπερβολικά μεγάλος
  - Συνήθως δε χρησιμοποιούνται σωστά

Όλα τα παραπάνω έχουν ως αποτέλεσμα οι χρήστες να δυσκολεύονται στην αναζήτηση ειδήσεων και πιο συγκεκριμένα στην παρακολούθηση αποκλειστικά των ειδήσεων που τους ενδιαφέρουν. Ακόμα περισσότερο οι χρήστες θα πρέπει με κάποιον τρόπο να γίνουν κομμάτι του πυρήνα ενός τέτοιου συστήματος και να διαμορφώνουν τον τρόπο με τον οποίο πραγματοποιείται η κατηγοριοποίηση αλλά και τον τρόπο με τον οποίο παρουσιάζονται τα αποτελέσματα της αναζήτησης στους χρήστες. Στη συνέχεια θα παρακολουθήσουμε κάθε διαδικασία του συστήματος που παρουσιάζεται και θα αναλυθούν τα προβλήματα που εντοπίζονται σε καθένα από αυτά τα συστήματα. Για να κατανοήσουμε καλύτερα τα προβλήματα που παρουσιάζονται στη διαδικασία του συστήματός μας θα περιγράψουμε ακροθιγώς κάθε διαδικασία. Το σύστημά μας ακολουθεί μία διαδικασία σειριακά προκειμένου να παράγει το ζητούμενο αποτέλεσμα το οποίο είναι η παρουσίαση προσωποποιημένων, κατηγοριοποιημένων άρθρων στον τελικό χρήστη. Για να γίνει αυτό θα πρέπει το σύστημα να είναι σε θέση να συλλέγει συνεχώς άρθρα από μεγάλα ειδησεογραφικά πρακτορεία. Η συλλογή των άρθρων δεν είναι αρκετή. Αφού τα άρθρα συγκεντρωθούν θα πρέπει να εφαρμοστούν σε αυτά μία σειρά από αλγορίθμους προκειμένου να «καθαριστεί» το κείμενό τους από οποιαδήποτε περιττή πληροφορία. Εν συνεχεία θα πρέπει να εφαμοστούν αλγόριθμοι κατηγοριποίησης του κειμένου και εξαγωγής περίληψης. Τέλος θα πρέπει να υπάρχει ένας μηχανισμός ο οποίος θα πραγματοποιεί προσωποποίηση των πιο πρόσφατων άρθρων

2.1. Συλλογή δεδομένων

Η συλλογή των δεδομένων είναι ένα πολύ σημαντικό κομμάτι ενός μηχανισμού σαν αυτό που θέλουμε να κατασκευάσουμε αλλά και γενικότερα ένα πολύ σημαντικό κομμάτι των μηχανισμών αναζήτησης και των μηχανισμών που βασίζονται στη συλλογή πληροφορίας. Στην περίπτωσή μας η συλλογή δεδομένων περιορίζεται στη συλλογή άρθρων από μεγάλους ειδησεογραφικούς πληροφοριακούς κόμβους. Το πρόβλημα συλλογής των κυριότερων νέων είναι μεγάλο καθότι αν παρατηρήσουμε τη δομή και οργάνωση αυτών των σελίδων,

αποτελεί πρόβλημα ο εντοπισμός αυτών των σελίδων αλλά και η συλλογή των πιο πρόσφατων ειδήσεων που είναι και το ζητούμενο.

Η συλλογή δεδομένων βασίζεται σε μηχανισμούς που περιδιαβαίνουν ολόκληρους τους ειδησεογραφικούς κόμβους και εντοπίζουν τα σημεία εκείνα που περιέχουν αρκετό κείμενο συγκριτικά με άλλες σελίδες που αποτελούν κεντρικούς κόμβους πληροφοριών.

Ωστόσο, οι νέες τεχνολογίες και κυρίως τα κανάλια επικοινωνίας που χρησιμοποιούνται από τους σύγχρονους δικτυακούς τόπους μπορούν να διευκολύνουν το πρόβλημα της συλλογής δεδομένων. Οι μηχανισμοί δεν είναι υποχρεωμένοι να «ανακαλύπτουν» τις πολλαπλές δυναμικές σελίδες που ανανεώνονται καθημερινά στους δικτυακούς τόπους. Αρκεί η συλλογή πληροφοριών από τα κανάλια επικοινωνίας που υπάρχουν για τη συγκέντρωση των πιο σημαντικών αλλαγών που προκύπτουν καθημερινά και εν προκειμένω τα νέα άρθρα που προστίθενται στα ειδησεογραφικά portal.

2.2. Φιλτράρισμα δεδομένων

Η συλλογή πληροφοριών έχει σαν αποτέλεσμα σελίδες που περιέχουν κυρίως HTML κώδικα στον οποίο βεβαίως μπορούμε να εντοπίσουμε και το κείμενο το οποίο επιθυμούμε να εξάγουμε από τη σελίδα και το οποίο αποτελεί το κύριο σώμα του άρθρου. Για το φιλτράρισμα τέτοιου είδους δεδομένων έχουν γίνει πολλές προτάσεις, κυρίως για τον τρόπο με τον οποίο μπορεί να εξαχθεί και βασικά να εντοπιστεί μέσα στη σελίδα. Το πρόβλημα σε αυτή την περίπτωση είναι η απομόνωση του χρήσιμου μόνο κειμένου το οποίο στην περίπτωση που εξετάζουμε είναι το σώμα του άρθρου αλλά και ο τίτλος του. Στο εξής θα αναφερόμαστε στον τίτλο του κειμένου αλλά και στο κύριο σώμα του σαν ΧΚ (Χρήσιμο Κείμενο)

2.3. Προεπεξεργασία πληροφορίας

Η προεπεξεργασία πληροφορίας είναι μία διαδικασία κατά την οποία το ΧΚ υπόκειται σε διαδικασία αφαίρεσης των σημείων στίξης, των αριθμών που τυχόν περιέχει, αφαίρεση λέξεων οι οποίες δεν περικλείουν κάποιο νόημα και τέλος το πολύ σημαντικό κομμάτι του Stemming το οποίο είναι η διαδικασία εύρεσης της ρίζας μίας λέξης. Σαν αποτέλεσμα έχει την εξαγωγή των λέξεων κλειδιών που υπάρχουν στο κείμενο συνοδευμένα από τη συχνότητα την οποία παρουσιάζουν μέσα στο κείμενο αλλά και το σημείο του κειμένου στο οποίο εντοπίζονται. Για τους μηχανισμούς εξαγωγής κειμένου και απόρριψης οποιασδήποτε πληροφορίας δεν σχετίζεται με το κείμενο η προεπεξεργασία πληροφορίας είναι μία πρόκληση. Παρά το γεγονός ότι βασίζεται σε συγκεκριμένα και σταθερά βήματα, θα πρέπει να γίνει εκτενής ανάλυση του είδους της πληροφορίας που είναι επιθυμητή προκειμένου το βήμα της προεπεξεργασίας να καταλήξει σε σημαντικά αποτελέσματα και πιο συγκεκριμένη στην εξαγωγή των σωστών λέξεων κλειδιών

2.4. Προσωποποίηση στο χρήστη

Η προσωποποίηση στο χρήστη είναι διαδικασία κατά την οποία τα αποτελέσματα που εμφανίζονται τελικά στο χρήστη προσαρμόζονται προκειμένου να είναι προσαρμοσμένα στις ανάγκες του. Πιο συγκεκριμένα, τα στάδια της προσωποποίησης αφορούν τον εντοπισμό άρθρων τα οποία ενδιαφέρουν το χρήστη και παρουσίασή τους με τέτοιον τρόπο ώστε να ταιριάζουν στις ανάγκες του χρήστη. Το πρόβλημα που τίθεται είναι ένας «έξυπνος» αλγόριθμος ο οποίος θα μπορεί να αξιοποιεί όλες τις πληροφορίες που μπορούν να συγκεντρωθούν από την περιήγηση του χρήστη στο δικτυακό τόπο και αξιοποίηση αυτών των πληροφοριών προκειμένου να εμφανιστούν όσο το δυνατόν καλύτερα και πιο ποιοτικά αποτελέσματα.

2.5. Συμμετοχή του χρήστη στις διαδικασίες του συστήματος

Ο χρήστης είναι αυτός που δέχεται την τελική πληροφορία και αυτός που ουσιαστικά διαμορφώνει την πληροφορία για τον εαυτό του. Αυτό σημαίνει πως ο χρήστης θα πρέπει να είναι αναπόσπαστο κομμάτι του συστήματος. Θα πρέπει να είναι σε θέση να διαμορφώσει διαδικασίες του πυρήνα του συστήματος όπως είναι η κατηγοριοποίηση και η εξαγωγή περίληψης.

Στα περισσότερα συστήματα τα οποία αντιμετωπίστηκαν κατά τη διάρκεια της μελέτης για τη συγκεκριμένη εργασία, παρατηρήθηκε πως ο χρήστης συμμετέχει μόνο στα επιτελικά στάδια των συστημάτων ενώ έχουν ήδη εκτελεστεί τα βασικά βήματα του πυρήνα των μηχανισμών. Η συμμετοχή του χρήστη στις διαδικασίες πυρήνα ενός large scale συστήματος είναι επίπονη διαδικασία η οποία απαιτεί αλγορίθμους που θα μπορούν να εκτελούνται αποδοτικά σε πραγματικό χρόνο προκειμένου ο χρήστης να διαμορφώνει όχι μόνον τα τελικά αποτελέσματα που εμφανίζονται σε αυτόν αλλά και συγκεκριμένες διαδικασίες ολόκληρου του συστήματος.

3

STATE OF THE ART

Στο κεφάλαιο αυτό περιγράφεται το State of the Art για κάθε υποσύστημα του μηχανισμού που θα κατασκευάσουμε. Πιο συγκεκριμένα υπάρχουν στοιχεία για τις εξής θεματικές ενότητες:

- Σημασιολογικός Ιστός
- Εξόρυξη Πληροφορίας
- Ανάκτηση Πληροφορίας
- Προεπεξεργασία Πληροφορίας
- Κατηγοριοποίηση Πληροφορίας
- Περίληψη Πληροφορίας
- Αξιοποίηση Πληροφορίας
- Προφίλ Χρήστη σε Δυναμικά Περιβάλλοντα

3. STATE OF THE ART

Στο παρόν κεφάλαιο θα παρουσιάσουμε τα θέματα με τα οποία θα ασχοληθεί η συγκεκριμένη εργασία καθώς επίσης και μία μικρή ανάλυση καθενός από αυτά προκειμένου να δημιουργηθεί το κατάλληλο υπόβαθρο για να είναι εφικτή η κατανόηση των όρων που θα χρησιμοποιηθούν στα επόμενα κεφάλαια αλλά και για να παρουσιάσουμε τις κυρίαρχες τεχνολογίες σε αυτό τον τομέα.

Τα θέματα με τα οποία θα ασχοληθούμε έχουν να κάνουν με ανάκτηση πληροφορίας από το Διαδίκτυο, εξαγωγή χρήσιμης πληροφορίας από συγκεκριμένες σελίδες που έχουμε ανακτήσει, κατηγοριοποίηση των σελίδων που έχουν αναλυθεί και τέλος θέματα που ασχολούνται με δημιουργία δυναμικού προφίλ των χρηστών του Διαδικτύου.

Τα παραπάνω θέματα θα μπορούσαν να περιγράψουν πλήρως την αρχιτεκτονική του συστήματός μας και μάλιστα με τη σειρά που αναγράφονται. Για να γίνει σαφές, προκειμένου να δημιουργήσουμε μία Δικτυακή πύλη ποιοτικού περιεχομένου αρκούν τα παρακάτω βήματα:

- Ανάκτηση σελίδων από το Διαδίκτυο
- Αποθήκευση του κώδικα των σελίδων
- Ανάλυση των αποθηκευμένων σελίδων, προκειμένου να εξαχθεί από αυτές η χρήσιμη πληροφορία
- Κατηγοριοποίηση της πληροφορίας βασισμένη σε συγκεκριμένες κατηγορίες που αντιπροσωπεύουν τόσο το σύνολο της πληροφορίας όσο και το σύνολο των απαιτήσεων των χρηστών του συστήματος.
- Παρουσίαση της πληροφορίας στο χρήστη σύμφωνα με τις θεματικές ενότητες που ο ίδιος έχει επιλέξει να του παρουσιάζονται.
- Δυναμική αλλαγή του προφίλ των χρηστών σύμφωνα με τις επιλογές που κάνουν από την πληροφορία που τους εμφανίζεται.

3.1. Σημασιολογικός Ιστός και Μεταδεδομένα

Το Διαδίκτυο σήμερα αποτελεί τη μεγαλύτερη πηγή πληροφοριών. Μεγάλοι όγκοι δεδομένων αναζητούνται, ανταλλάσσονται και επεξεργάζονται μέσω του Παγκόσμιου Ιστού. Επειδή, όμως ο όγκος των δεδομένων του Ιστού έχει πάρει μεγάλες διαστάσεις χωρίς να υπάρχει ενιαίος τρόπος οργάνωσης, η ανταλλαγή και η επεξεργασία τους είναι πολύ δύσκολη. Ο Σημασιολογικός Ιστός έρχεται ακριβώς να εξυπηρετήσει την ανάγκη για ενιαία οργάνωση των δεδομένων, ώστε το Διαδίκτυο να γίνει μια αποδοτική παγκόσμια πλατφόρμα ανταλλαγής και επεξεργασίας από ετερογενείς πηγές πληροφορίας. Ένας γενικός ορισμός μας λέει ότι ο Σημασιολογικός Ιστός δίνει δομή, οργάνωση και σημασιολογία στα δεδομένα, ώστε να είναι, σε μεγάλο βαθμό, κατανοητά από μηχανές (machine understandable).

Ο όρος Σημασιολογικός Ιστός (Semantic Web) χρησιμοποιήθηκε για πρώτη φορά το 1998 από το δημιουργό του πρώτου φυλλομετρητή ιστοσελίδων και εξυπηρετητή διαδικτύου, Tim Berners-Lee. Από τότε καταβάλλεται μεγάλη προσπάθεια από την επιστημονική κοινότητα για την υλοποίησή του πάνω από τον Παγκόσμιο Ιστό. Στο βασικότερο επίπεδό του, ο Σημασιολογικός Ιστός αποτελεί μία συλλογή από συνοπτική πληροφορία για τη διακινούμενη πληροφορία, τα μεταδεδομένα, η οποία δεν είναι ορατή στον τελικό χρήστη. Τα μεταδεδομένα χρησιμοποιούνται για να περιγράψουν υπάρχοντα έγγραφα, ιστοσελίδες, βάσεις δεδομένων, προγράμματα που βρίσκονται στο διαδίκτυο. Οι εφαρμογές λογισμικού που κάνουν χρήση μεταδεδομένων αποκτούν καλύτερη κατανόηση της σημασιολογίας του περιεχομένου τους και άρα μπορούν να τα επεξεργαστούν με

πιο αποδοτικό τρόπο. Η κατανόηση των μεταδεδομένων από τις μηχανές είναι δυνατή μέσω της χρήσης ειδικών λεξικών (των οντολογιών) τα οποία παρέχουν κοινούς κανόνες και λεξιλόγια για την ερμηνεία των δεδομένων. Με αυτό τον τρόπο είναι δυνατή η κοινή κατανόηση όρων και εννοιών από εφαρμογές που προέρχονται από διαφορετικά πληροφοριακά συστήματα. Απώτερος στόχος της όλης προσπάθειας είναι η ικανοποίηση των απαιτήσεων των συμμετεχόντων στην Κοινωνία της Πληροφορία για αυξημένη ποιότητα υπηρεσιών. Αυτό συνίσταται κυρίως στη βελτιωμένη αναζήτηση, εκτέλεση σύνθετων διεργασιών μέσω του Διαδικτύου και στην εξατομίκευση της πληροφορίας σύμφωνα με τις ανάγκες του εκάστοτε χρήστη.

Ένα από τα σημαντικότερα προβλήματα που καλείται να λύσει ο Σημασιολογικός Ιστός είναι η πρόσβαση στην πληροφορία. Σύμφωνα με πρόσφατες μελέτες, η ανθρωπότητα έχει παράγει από το 1999 μέχρι το 2003, τόσες νέες πληροφορίες όσες παρήγαγε όλα τα προηγούμενα χρόνια της ιστορίας της. Σε αυτό το διάστημα των τριών τελευταίων ετών παρήχθησαν 12 exabytes πληροφορίας υπό τη μορφή έντυπου, οπτικού ή και ηχητικού υλικού. Η αυξανόμενη αυτή παραγωγή και η συνεχής βελτίωση των μεθόδων ψηφιοποίησης συμβάλλουν στην παραγωγή ενός ωκεανού ψηφιακών δεδομένων που προφανώς δύναται να δημιουργήσει μεγάλο αριθμό προβλημάτων. Το πιο σημαντικό ίσως από αυτά είναι ο τρόπος με τον οποίο θα μπορεί κανείς να διαχειριστεί όλη αυτή την πληροφορία. Δε θα πρέπει φυσικά να αμελούμε το γεγονός πως η ικανότητα παραγωγής, αποθήκευσης και μετάδοσης της πληροφορίας έχει ξεπεράσει κατά πολύ τις δυνατότητες αναζήτησης, πρόσβασης και παρουσίασης.

Λόγω του αυξανόμενου όγκου της πληροφορίας και των προβλημάτων αποτελεσματικής πρόσβασης, έχει γίνει τα τελευταία χρόνια ξεκάθαρο προς την επιστημονική κοινότητα ότι για την αύξηση της απόδοσης χρειάζονται νέες μέθοδοι υπολογισμού ικανές να προσαρμοστούν σε μία πληθώρα παραμέτρων τόσο αντικειμενικών όσο και υποκειμενικών. Η απόδοση ενός συστήματος πρόσβασης στην πληροφορία εκτιμάται μέσα από την ανάκληση και την ακρίβεια.

Η αναφορά στα προβλήματα που αντιμετωπίζουν τα σύγχρονα συστήματα πρόσβασης στην πληροφορία έχει άμεση σχέση με τον τύπο των ερωτήσεων που δέχονται ως είσοδο. Υπάρχουν δύο διαφορετικά είδη ερωτημάτων, οι ερωτήσεις γενικού περιεχομένου και ειδικού περιεχομένου. Το μέγεθος της απάντησης σε ερωτήσεις γενικού περιεχομένου είναι μεγάλο και παρουσιάζει εξαιρετικά μεγάλες αποκλίσεις ως προς τη σχετικότητα της ίδιας της ερώτησης. Το πρόβλημα εστιάζεται στην επιλογή ενός μικρού συνόλου από τις πιο σχετικές απαντήσεις, είναι δηλαδή πρόβλημα ακρίβειας. Αντίθετα, για τις ερωτήσεις ειδικού περιεχομένου, το διαθέσιμο σύνολο σχετικών απαντήσεων είναι μικρό και το πρόβλημα που προκύπτει είναι πρόβλημα ανάκτησης.

Εκτός από τα κλασσικά προβλήματα που αντιμετωπίζουν τα ΠΣ στον τομέα της πρόσβασης στην πληροφορία, αναδύονται και άλλα άμεσα συνδεδεμένα με το είδος της ίδιας της πληροφορίας:

- Συνωνυμία: ανάκτηση μη σχετικών απαντήσεων που περιέχουν όρους συνώνυμους με αυτούς της ερώτησης
- Ασάφεια / Διφορούμενες έννοιες: ανάκτηση μη σχετικών λόγω ασάφειας της ερώτησης ή λόγω ύπαρξης διφορούμενων εννοιών.
- Πειθώ των μηχανών αναζήτησης (search engine persuasion): ταξινόμηση των ανακτημένων εγγράφων με βάση το βαθμό σχετικότητας τους προς την ερώτηση έχοντας υπόψη τα προβλήματα της συνωνυμίας και της ασάφειας.

Τα τελευταία χρόνια, μια νέα ερευνητική προσπάθεια έχει επικεντρωθεί σε αυτό το πεδίο το οποίο ανήκει στην περιοχή που ονομάζεται Προσαρμοσμένη Πρόσβαση στην Πληροφορία. (Adaptive Information Access). Η πρόσβαση στην πληροφορία περιλαμβάνει αρκετές ερευνητικές περιοχές που θα μπορούσαν να

συνδυαστούν για την κατασκευή συστημάτων ικανών να ανταποκριθούν στις σύγχρονες ανάγκες. Τέτοιες περιοχές είναι η έξυπνη αναζήτηση πληροφορίας, μάθηση μηχανής και αλληλεπίδραση ανθρώπου υπολογιστή. Στην παρούσα εργασία θα ασχοληθούμε με ζητήματα που έχουν να κάνουν τόσο με έξυπνη ανάκτηση πληροφορίας, με μάθηση μηχανής όσο και με αλληλεπίδραση χρηστών με τον υπολογιστή.

3.2. Εξόρυξη πληροφορίας από το Διαδίκτυο

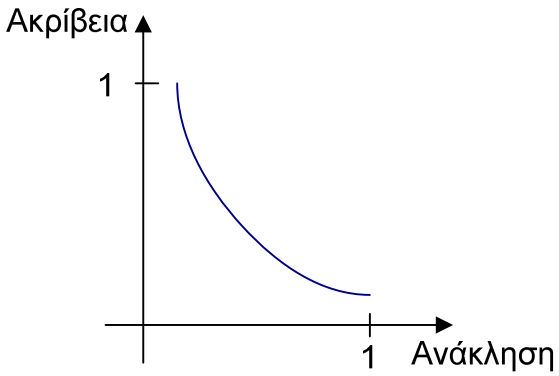
Εξόρυξη πληροφορίας από το Διαδίκτυο ονομάζεται κάθε διαδικασία που έχει σαν αποτέλεσμα ανάκτηση πληροφορίας (Information Retrieval) από τον παγκόσμιο ιστό. Στο εξής θα αναφερόμαστε στον όρο ανάκτηση πληροφορίας ως IR για συντομία. Η ανακτώμενη πληροφορία δεν περιορίζεται απλώς σε σελίδες HTML, αλλά μπορεί να είναι και αρχεία πολυμέσων ή οποιοδήποτε είδος αρχείου μπορεί να μεταφερθεί πάνω από το Διαδίκτυο. Η ανάγκη για ανάκτηση πληροφορίας πηγάζει από τις αρχές της δεκαετίας του 50 όταν ο Mooers [1] εξέφρασε ανοιχτά σε δημοσίευσή του την ανάγκη για ανάκτηση πληροφορίας. Αργότερα, στη δεκαετία του 60, το IR είχε γίνει πλέον ένα πολύ δημοφιλές θέμα καθώς πολλοί ερευνητές πίστευαν ότι μπορούν να αυτοματοποιήσουν μέχρι τότε χειροκίνητες διαδικασίες όπως η δεικτοδότηση και η αναζήτηση [2; 3].

Προκειμένου να πετύχει το στόχο της η κοινότητα IR όρισε δύο βασικές ενέργειες που έχουν γίνει αντικείμενα έρευνας για πολλά χρόνια και είναι: η δεικτοδότηση και η αναζήτηση. Η δεικτοδότηση αναφέρεται στον τρόπο με τον οποίο αναπαρίσταται η πληροφορία για τους σκοπούς της ανάκτησης. Η αναζήτηση αναφέρεται στον τρόπο με τον οποίο δομείται η πληροφορία όταν πραγματοποιείται ένα ερώτημα. Παρόλο που οι δύο αυτές διαδικασίες αποτελούν τον πυρήνα ενός συστήματος IR, άλλες διαδικασίες είναι αυτές που κερδίζουν έδαφος, όπως τεχνικές αναπαράστασης της πληροφορίας, με σκοπό να βελτιωθεί η αποτελεσματικότητα της ανάκτησης [4].

Στην παρούσα φάση το IR αντιμετωπίζει μία σειρά από θέματα. Αρχικά, εφαρμόστηκε σε ΒΔ βιβλιοθηκών, όπου σε ένα αρχείο αποθηκεύονταν γενικά χαρακτηριστικά κάθε εγγράφου, όπως ο τίτλος και ο συγγραφέας, και η αναζήτηση γινόταν βάσει αυτών των στοιχείων. Στη συνέχεια, και εξ αιτίας της αύξησης του μεγέθους των αποθηκευτικών μέσων, ολόκληρο το κείμενο αποθηκευόταν σε αρχείο και η αναζήτηση ήταν εφικτή σε ολόκληρες συλλογές από κείμενα. Έτσι μέχρι ενός σημείου το IR αντιπροσώπευε την ανάκτηση κειμένων. Αργότερα και έως σήμερα, δίνεται περισσότερη σημασία στον όρο πληροφορία (Information). Άλλωστε σήμερα δεν έχουμε μόνο έγγραφα πάνω στα οποία γίνεται η αναζήτηση αλλά και αρχεία πολυμέσων. Ωστόσο το βασικό κλειδί στην υπόθεση του IR είναι ανάκτηση κειμένων ή πληροφορίας που προσεγγίζουν περισσότερο τις ανάγκες του χρήστη που πραγματοποιεί την αναζήτηση.

Ένα από τα βασικά στοιχεία του IR είναι η μέτρηση του κατά πόσο τα ανακτημένα κείμενα είναι σχετικά με το ερώτημα που κάνουμε. [5]. Έτσι λοιπόν, ένα βασικό στοιχείο στο οποίο εστιάζουμε είναι η εύρεση μετρικών που θα μπορούν να αναπαραστήσουν αριθμητικά τη σχετικότητα των αποτελεσμάτων ενός συστήματος IR. Πολλές μετρικές έχουν αναπτυχθεί με τις δύο πιο γνωστές να είναι η ανάκληση και η ακρίβεια. Η ακρίβεια μας δίνει το ποσοστό (%) των σχετικών κειμένων εν συγκρίσει με αυτά που ανακτήθηκαν ενώ η ανάκληση μας δίνει το ποσοστό (%) των κειμένων που ανακτήθηκαν εν συγκρίσει με μία συλλογή που γνωρίζουμε ότι περιέχει όλα τα σχετικά.

Η συνηθισμένη απόκριση που έχει ένα σύστημα IR είναι αυτή που φαίνεται στο παρακάτω σχήμα στο οποίο φαίνεται ότι τα μεγέθη ακρίβεια και ανάκληση είναι αντιστρόφως ανάλογα. Αυτό σημαίνει πως για αν αυξήσουμε την ανάκληση θα μειωθεί η ακρίβεια. Φυσικά ισχύει και το αντίστροφο [6].



Εικόνα 1: Σχεδιάγραμμα ακρίβειας – ανάκλησης

Ένα σύστημα IR μπορεί να πετύχει κατά μέσο όρο περίπου 30% ανάκληση και 30% ακρίβεια. Οι τιμές αυτές δεν έχουν καμία σύγκριση με ένα σύστημα DBMS που τα ποσοστά αυτού προσεγγίζουν το 100%. Ωστόσο θα μπορούσε κανείς να πει πως και τα δύο συστήματα πραγματοποιούν την ίδια διαδικασία, δηλαδή ανάκτηση πληροφορίας. Αυτό βέβαια έχει να κάνει με τον τρόπο με τον οποίο δομείται ένα σύστημα DBMS και ο οποίος είναι τέτοιος ώστε να εξυπηρετεί απόλυτα τις ανάγκες ενός χρήστη.

Αυτή η δυσκολία που αντιμετωπίζουν τα συστήματα IR (μικρές τιμές ανάκλησης και ακρίβειας) γεννούν ένα άλλο επιστημονικό πεδίο το οποίο υπάρχει παράλληλα με το IR και είναι το IF (Information Filtering). Σε ένα κλασσικό άρθρο οι Belkin και Croft παρουσίασαν δύο διαφορετικούς ορισμούς για τα δύο παραπάνω θέματα οι οποίοι έχουν κοινές τεχνικές αλλά διαφέρουν σε τρία βασικά στοιχεία [7]. Πρώτον, στο IR όταν ο χρήστης κάνει ένα ερώτημα περιμένει άμεση απόκριση. Στο IF ο χρήστης μπορεί να περιμένει, εν γνώσει του, για μεγάλο χρονικό διάστημα μέχρι να του παρουσιαστεί μία απάντηση. Επιπρόσθετα το IF χειρίζεται και θέματα που από τη φύση του είναι δυναμικά και εντάσσει στο μηχανισμού του στοιχεία εκμάθησης σύμφωνα με τα κείμενα που προσθέτει στη συλλογή του. Τέλος, το βασικότερο είναι πως το IR αναζητά παραπλήσια κείμενα από μία μεγάλη συλλογή κειμένων σε αντίθεση με το IF το οποίο προσπαθεί να αφαιρέσει από μία συλλογή τα εισερχόμενα κείμενα που δεν είναι σχετικά.

Παρ' όλες τις διαφορές που έχουν τα δύο αυτά πεδία δεν πρέπει να αμελούμε πως έχουν παραπλήσιο σκοπό: να εξασφαλίσουν ότι τα κείμενα που θα παρουσιαστούν στο χρήστη είναι σχετικά με το ερώτημά του.

Τα διαγράμματα ακρίβειας/ανάκλησης είναι χρήσιμα εφόσον μελετούμε την απόδοση ανάκτησης διαφορετικών αλγορίθμων σε ένα σύνολο από πρότυπες πληροφοριακές ανάγκες. Ωστόσο υπάρχουν περιπτώσεις στις οποίες θα θέλαμε να συγκρίνουμε την απόδοση αλγορίθμων ανάκτησης για ατομικές πληροφοριακές ανάγκες. Οι λόγοι για να το κάνουμε αυτό είναι δύο:

1. η χρήση μέσων τιμών που προκύπτουν από την εκτέλεση διαφόρων ερωτημάτων μπορεί να αποκρύπτει σημαντικές ανωμαλίες στον αλγόριθμο ανάκτησης
2. όταν συγκρίνουμε δύο αλγορίθμους μπορεί να θέλουμε να μελετήσουμε κατά πόσο ο ένας είναι καλύτερος του άλλου για κάθε μία από τις πληροφοριακές ανάγκες που έχουμε και όχι συνολικά.

Σε τέτοιες περιπτώσεις υπολογίζουμε μία μόνο τιμή ακρίβειας για κάθε ερώτημα, η οποία θα μπορούσε να θεωρηθεί σαν σύνοψη του συνολικού διαγράμματος ακρίβειας/ανάκλησης. Συνήθως αυτή η τιμή είναι η ακρίβεια σε κάποιο συγκεκριμένο επίπεδο ανάκλησης. Φυσικά αυτές είναι λίγες από τις πολλές προσεγγίσεις που μπορούν να γίνουν.

3.2.1. Μοντέλα ανάκτησης πληροφορίας

Τα τρία κλασσικά μοντέλα στην Ανάκτηση Πληροφορίας είναι το Boolean, το Vector Space και το Πιθανοτικό. Στο μοντέλο Boolean, τόσο τα κείμενα όσο και τα ερωτήματα αντιμετωπίζονται ως ένα σύνολο από όρους δεικτοδότησης. Κατά συνέπεια το μοντέλο μπορεί να θεωρηθεί ως συνολοθεωρητικό. Στο Vector Space, τα κείμενα και τα ερωτήματα αναπαρίστανται ως διανύσματα σε έναν t-διάστατο χώρο. Έτσι λέμε ότι το μοντέλο είναι αλγεβρικό. Το Πιθανοτικό μοντέλο εισάγει έναν τρόπο αναπαράστασης, ο οποίος βασίζεται στην πιθανοθεωρία. Κατά συνέπεια το μοντέλο είναι πιθανοτικού χαρακτήρα. Το πιθανοτικό μοντέλο και Με τον καιρό προτάθηκαν διάφορες νέες προσεγγίσεις σε καθεμιά από τις κατηγορίες βασικών μοντέλων. Έτσι έχουμε στο συνολοθεωρητικό πεδίο τα μοντέλα, ασαφές (fuzzy) Boolean και επεκτεταμένο Boolean. Στα αλγεβρικά μοντέλα έχουμε το γενικευμένο vector space, την λανθάνουσας σημασιολογικής δεικτοδότησης (LSI) και το μοντέλο των νευρωνικών δικτύων. Στον πιθανοτικό τομέα εμφανίστηκαν τα δίκτυα εξαγωγής συμπεράσματος (inference networks) και τα δίκτυα πεποίθησης (belief networks). Εκτός από την χρήση του περιεχομένου των κειμένων, ορισμένα μοντέλα εκμεταλλεύονται και την εσωτερική δομή που φυσιολογικά υπάρχει στο γραπτό λόγο. Σε αυτή την περίπτωση λέμε ότι έχουμε ένα δομημένο μοντέλο. Για τη δομημένη ανάκτηση κειμένου, συναντούμε δύο μοντέλα, τις μη επικαλυπτόμενες λίστες (non-overlapping lists) και τους κοντινούς κόμβους (proximal nodes).

3.2.1.1. Τυπικός ορισμός των μοντέλων

Πριν προχωρήσουμε στην εξέταση των επί μέρους μοντέλων θα δώσουμε έναν τυπικό και ακριβή ορισμό για το τι είναι ένα μοντέλο ΑΠ. Ορισμός Ένα μοντέλο ανάκτησης πληροφορίας είναι η τετράδα [D, Q, F, R(qi, dj)] όπου:

1) D είναι ένα σύνολο από λογικές αναπαραστάσεις για τα κείμενα της συλλογής

2) Q είναι ένα σύνολο από λογικές αναπαραστάσεις για τις πληροφοριακές ανάγκες του χρήστη. Αυτές οι αναπαραστάσεις καλούνται ερωτήματα

3) F είναι ένα υπόβαθρο για την μοντελοποίηση της αναπαράστασης των κειμένων, των ερωτημάτων και των σχέσεων μεταξύ τους

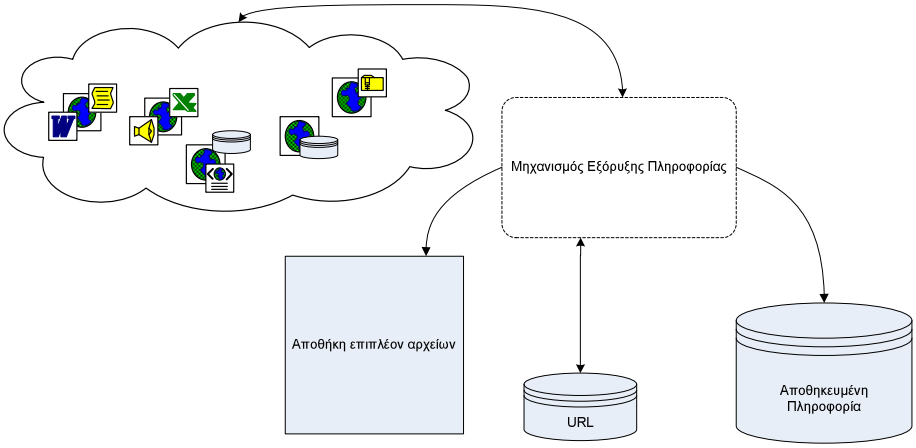
4) R(qi, dj) είναι μια συνάρτηση κατάταξης, η οποία συνδέει έναν πραγματικό αριθμό με ένα ερώτημα qi $\in$ Q και μια αναπαράσταση κειμένου dj $\in$ D. Μια τέτοια κατάταξη ορίζει μια διάταξη πάνω στα κείμενα πάντα με βάση το ερώτημα. qi.

Διαισθητικά ο παραπάνω ορισμός περιγράφει τη διαδικασία καθορισμού ενός μοντέλου ΑΠ. Η διαδικασία ορισμού ενός μοντέλου είναι η ακόλουθη. Αρχικά επινοείται ένας τρόπος αναπαράστασης για τα κείμενα και την πληροφοριακή ανάγκη του χρήστη. Έπειτα καθορίζεται ένα υπόβαθρο στο οποίο θα μπορούν αυτές οι αναπαραστάσεις να μοντελοποιηθούν. Το υπόβαθρο αυτό, θα πρέπει να μπορεί να παρέχει και τον μηχανισμό κατάταξης. Για παράδειγμα στο Boolean μοντέλο, το υπόβαθρο αυτό αποτελείται από τις αναπαραστάσεις των κειμένων και των ερωτήσεων ως σύνολα, και τις κλασσικές πράξεις πάνω στα σύνολα. Αντίστοιχα στο Vector space, το υπόβαθρο αποτελείται από τις διανυσματικές αναπαραστάσεις κειμένων στον t-διάστατο διανυσματικό χώρο και τις επιτρεπτές αλγεβρικές πράξεις πάνω σε διανύσματα.

3.2.2. Αρχιτεκτονική μηχανισμών εξόρυξης

Όλες οι μηχανές αναζήτησης πραγματοποιούν ανάκτηση πληροφορίας προκειμένου να μπορούν να εξυπηρετούν τους χρήστες τους. Έτσι, μέχρι σήμερα έχει κατασκευαστεί πληθώρα προγραμμάτων τα οποία είτε λειτουργώντας σαν αυτόνομες μονάδες είτε σε συνεργασία μεταξύ τους πραγματοποιούν εξόρυξη

πληροφορίας. Η γενική ιδέα ενός μηχανισμού εξόρυξης πληροφορίας είναι εξαιρετικά απλή και φαίνεται στο παρακάτω σχήμα.



Εικόνα 2: Μηχανισμός Εξόρυξης Πληροφορίας

Ένας τέτοιος μηχανισμός μπορεί να είναι ένας απλός υπολογιστής ή ακόμα και μερικές χιλιάδες υπολογιστές που λειτουργούν κάτω από την επίβλεψη ενός. Ο μηχανισμός ξεκινά να λειτουργεί περιδιαβαίνοντας σελίδες του Διαδικτύου. Οι HTML σελίδες αποθηκεύονται σε μία βάση δεδομένων μαζί με επιπρόσθετες πληροφορίες για αυτές οι οποίες μπορεί να περιλαμβάνουν: το URL, την ώρα που ανακτήθηκε η σελίδα, το μέγεθός της και άλλα. Σε μία ξεχωριστή (συνήθως) βάση δεδομένων αποθηκεύονται όλα τα URL που έχουν ανακτηθεί και τα οποία ανακτώνται ανά τακτά χρονικά διαστήματα. Παράλληλα κάθε σελίδα αναλύεται προκειμένου να εξαχθούν από αυτή όλα τα links που περιέχει (σύμβολο <a> στην HTML). Τα links που «διαβάζει» ο μηχανισμός συγκρίνονται με αυτά που υπάρχουν αποθηκευμένα στη βάση δεδομένων URL και γίνονται οι κατάλληλες προσθήκες. Τέλος, κάποια επιπλέον αρχεία (doc, css, xml, scripts, πολυμέσα) αποθηκεύονται συνήθως σε καταλόγους που ονομάζονται κατάλληλα από τον μηχανισμό, έτσι ώστε να είναι σε θέση να τα προσπελάσει ανά πάσα στιγμή.

Μερικοί από τους πιο γνωστούς μηχανισμούς που πραγματοποιούν εξόρυξη πληροφορίας είναι οι crawlers, τα bots, τα spiders κ.α. Η λειτουργία τους είναι ουσιαστικά ίδια και βασίζεται στην αρχιτεκτονική που φαίνεται στο παραπάνω σχήμα.

3.2.3. Τεχνολογίες ανάκτησης δεδομένων από το Διαδίκτυο

Η ανάκτηση πληροφορίας είναι μία έννοια η οποία αναφέρεται σε κάθε μηχανισμό ο οποίος μέσω ενός αλγορίθμου «επιστρέφει» αποτελέσματα από ένα σύνολο στοιχείων. Μιλώντας για ανάκτηση πληροφορίας από το διαδίκτυο θα πρέπει να αναλογιστούμε τη μοναδικότητα των στοιχείων που χαρακτηρίζουν το Διαδίκτυο και συνεπώς αλλάζουν τη διαδικασία ανάκτησης δεδομένων από αυτό. Τα κύρια χαρακτηριστικά του Διαδικτύου είναι:

- Εξαιρετικά μεγάλο μέγεθος
  - Σύμφωνα με πρόσφατους υπολογισμούς το μέγεθος του Διαδικτύου είναι περίπου 11 δισεκατομμύρια σελίδες
- Δυναμικός χαρακτήρας
  - Το Internet αλλάζει ώρα με τη ώρα ενώ στα κλασσικά συστήματα ανάκτησης δεδομένων υπάρχουν σταθερές βάσεις δεδομένων.
- Περιέχει ετερογενές υλικό
  - Υπάρχουν πολλοί διαφορετικοί τύποι αρχείων (κείμενα, εικόνες, βίντεο, ήχος, scripts) με αποτέλεσμα οι αλγόριθμοι ανάκτησης

δεδομένων να πρέπει να εφαρμοστούν τόσο σε απλό κείμενο όσο και πολυμεσικά δεδομένα.

- Υπάρχει μεγάλο εύρος γλωσσών
  - Οι γλώσσες που χρησιμοποιούνται στο Διαδίκτυο υπολογίζονται σε πάνω από 100.
- Διπλές εγγραφές
  - Η αντιγραφή είναι ένα βασικό χαρακτηριστικό ου Διαδικτύου. Δεν είναι τυχαίο πως 25-30% των σελίδων του Διαδικτύου αποτελούν αντίγραφα άλλων σελίδων.
- Πολλά links από μία σελίδα σε άλλη
  - Υπολογίζεται πως σε κάθε σελίδα περιέχονται κατά μέσο όρο 10 links προς άλλες σελίδες.
- Πολλοί και διαφορετικών ειδών χρήστες
  - Κάθε χρήστης έχει τα δικές του ανάγκες αλλά και τις δικές του γνώσεις και απαιτήσεις από το Διαδίκτυο.
- Διαφορετική συμπεριφορά από τους χρήστες
  - Έχει υπολογιστεί πως περίπου το 90% των χρηστών του Διαδικτύου παρατηρούν μόνο την πρώτη σελίδα από αυτές που του επιστρέφει μία μηχανή αναζήτησης. Παράλληλα, μόνο το 20% δοκιμάζει να αλλάξει το ερώτημα που έχει κάνει προκειμένου να βρει καλύτερα αποτελέσματα.

Στα κλασσικά συστήματα ανάκτησης πληροφορίας οι μετρικές που χρησιμοποιούνται για την αξιολόγηση είναι:

- Η ανάκληση
  - Το ποσοστό των σελίδων που έχουν επιστραφεί και είναι σχετικές
- Η ακρίβεια
  - Το ποσοστό των σχετικών σελίδων που έχουν επιστραφεί
- Η ακρίβεια στα πρώτα 10 αποτελέσματα

Σε ένα σύστημα όμως που έχει να κάνει με ανάκτηση πληροφορίας από το διαδίκτυο θα πρέπει:

«Τα αποτελέσματα που επιστρέφονται θα πρέπει να έχει υψηλή σχετικότητα με το ερώτημα και αλλά και υψηλή ποιότητα, δηλαδή, με λίγα λόγια, θα πρέπει τα αποτελέσματα να είναι μόνο τα αναγκαία και απαραίτητα».

Αυτό σημαίνει πως σε ένα τέτοιο σύστημα θα πρέπει να χρησιμοποιηθούν διαφορετικές μετρικές με τη βοήθεια των οποίων θα είναι σε θέση οι μηχανισμοί ανάκτησης πληροφορίας να μπορούν να αξιολογήσουν τα ερωτήματα των χρηστών και να επιστρέψουν τα πιο σωστά και πιο αντιπροσωπευτικά αποτελέσματα.

Η αρχιτεκτονική των μηχανισμών ανάκτησης πληροφορίας από το Διαδίκτυο διαφέρει από την αρχιτεκτονική των μηχανισμών ανάκτησης πληροφορίας γενικά. Τα στοιχεία που είναι απαραίτητα σε ένα μηχανισμό ανάκτησης πληροφορίας είναι

- Ο indexer
- Ο crawler και
- Ο query server.

Ο crawler χρησιμεύει στο να συλλέγονται σελίδες από το διαδίκτυο, ο indexer αναλαμβάνει να προβεί σε ανάλυση των ανακτημένων σελίδων και αναδόμηση αυτών προκειμένου να είναι εύκολη και εφικτή η αναζήτηση πάνω σε αυτές και τέλος ο query server είναι υπεύθυνος για την εξυπηρέτηση των ερωτημάτων από τους τελικούς χρήστες.

Αυτά τα τρία θεωρούνται τα βασικά δομικά στοιχεία ενός τέτοιου μηχανισμού ενώ δεν αποκλείεται σε σύνθετους μηχανισμούς ανάκτησης πληροφορίας από το διαδίκτυο να συναντήσουμε πολλά ακόμα υπο-συστήματα αλλά και αναβαθμίσεις και αλλαγές στα συστήματα που ήδη περιγράψαμε. Αυτού του είδους τα συστήματα δημιουργούν ένα off-line αντίγραφο του διαδικτύου και εφαρμόζουν αλγορίθμους αναζήτησης στο αντίγραφο που διατηρούν. Άλλωστε είναι σχεδόν αδύνατη η δυναμική αναζήτηση στις δισεκατομμύρια σελίδες του διαδικτύου. Φυσικά τίθενται μία σειρά από προβλήματα τα οποία έχουν να κάνουν με το πόσο επικαιροποιημένο είναι το off-line αντίγραφο. Όσο πιο επικαιροποιημένο είναι τόσο ακριβέστερα αποτελέσματα θα εμφανίζονται. Ένα παράδειγμα που δείχνει την αδυναμία των μηχανισμών ανάκτησης πληροφορίας του διαδικτύου όπου παρουσιάζεται έντονα το φαινόμενο της μη επικαιροποιημένης πληροφορίας είναι οι πρώτες σελίδες των μεγάλων ειδησεογραφικών πρακτορείων. Οι σελίδες αυτές είναι κατασκευασμένες με τέτοιον τρόπο ώστε μπορεί μέσα σε 12 ώρες να έχει αλλάξει εντελώς το περιεχόμενο (κειμενο και εικόνες) στη συγκεκριμένη σελίδα. Προκειμένου ο μηχανισμός ανάκτησης πληροφορίας από το διαδίκτυο να είναι ενημερωμένος για τις συγκεκριμένες αλλαγές θα πρέπει να προσπελαύνει συνέχεια τη συγκεκριμένη σελίδα και να εντοπίζει αλλαγές, κάτι το οποίο είναι αδύνατο για τα σημερινά δεδομένα του χαώδους διαδικτύου.

Για την ακριβέστερη ανάκτηση πληροφορίας από το διαδίκτυο, η αδόμητη πληροφορία που ανακτάται από τις σελίδες που περιδιαβαίνει ο crawler θα πρέπει να δομηθεί με κατάλληλο τρόπο και να αποθηκεύεται σε τέτοια μορφή ώστε να μη χάσει τη συσχέτισή τας από τα στοιχεία που την αποτελούν αλλά και από τις υπόλοιπες σελίδες που είναι όμοιές της. Τα στοιχεία που χρησιμοποιούνται για τη δόμηση των αποθηκευμένων σελίδων είναι συνήθως:

- Repository
  - Πρόκειται για το σημείο όπου αποθηκεύονται ολόκληρες οι σελίδες με τον HTML κώδικά τους.
- Document Index
  - Πρόκειται για πιο εξειδικευμένο χώρο αποθήκευσης πληροφορίας πια και όχι αρχείου όπου βέβαια υπάρχουν συσχετίσεις με τις σελίδες του repository καθώς και διάφορα στοιχεία checksum ή στατιστικά.
- Lexicon
  - Ένα λεξικό όπου είναι αποθηκευμένες περισσότερες από 20 εκατομμύρια λέξεις διαφόρων γλωσσών και χρησιμοποιούνται για ορθογραφικό έλεγχο των λέξεων των κειμένων
- Hit Lists
  - Πρόκειται για λίστες που περιέχουν στοιχεία που αφορούν μονοπάτια που οδηγούν από μία σελίδα του διαδικτύου σε άλλη. Αυτές οι λίστες χρησιμοποιούνται σε συνδυασμό με εξειδικευμένους αλγορίθμους προκειμένου να προκύψουν συσχετίσεις και δεσμοί μεταξύ των σελίδων
- Forward Index
  - Πρόκειται για λέξεις οι οποίες είναι ταξινομημένες βάσει ενός αύξοντα αριθμού που έχει ανατεθεί σε κάθε μία.
- Inverted Index
  - Είναι ακριβώς το ίδιο με το προηγούμενο μόνο που η ταξινόμηση γίνεται κατά φθίνουσα σειρά.

Οι περισσότεροι μηχανισμοί ανάκτησης πληροφορίας από το διαδίκτυο βασίζονται στον παραπάνω μηχανισμό που περιγράφηκε. Βασικός σκοπός τους είναι να λειτουργήσουν σαν μηχανές αναζήτησης και όχι για να προσφέρουν ένα ιστορικό του διαδικτύου. Επιπλέον, οι σελίδες που εμφανίζονται στον τελικό

χρήστη δεν ταξινομούνται βάσει συσχέτισης με το ερώτημα αλλά βάσει ενός αριθμού που έχουν οι μηχανές αναζήτησης για κάθε σελίδα και ο οποίος δείχνει πόσο «γνωστή» είναι η συγκεκριμένη σελίδα. Έτσι αν μία σελίδα ενός προσωπικού δικτυακού τόπου για δελφίνια περιέχει τη λέξη «δελφίνι» και την ίδια λέξη περιέχει κάποια σελίδα του CNN τότε οι μηχανές αναζήτησης στην αναζήτησή μας για τη λέξη δελφίνι θα βαθμολογήσουν περισσότερο τις σελίδες του πασίγνωστου CNN και λιγότερο τις σελίδες του προσωπικού δικτυακού τόπου.

3.2.4. Εξόρυξη γνώσης από αποθήκες δεδομένων

Η εξόρυξη γνώσης από μεγάλες αποθήκες δεδομένων που βρίσκονται στον παγκόσμιο ιστό, έχει εξελιχθεί σε ένα από τα βασικότερα ερευνητικά ζητήματα στον τομέα των βάσεων δεδομένων, των μηχανών γνώσης, της στατιστικής, καθώς επίσης και ως μία σημαντική ευκαιρία για καινοτομία στις επιχειρήσεις. Οι δικτυακές εφαρμογές που διαχειρίζονται μεγάλες αποθήκες δεδομένων, με σκοπό τη βελτίωση της ποιότητας των παρεχόμενων υπηρεσιών μέσω της μελέτης της συμπεριφοράς των πελατών και της εξαγωγής χρήσιμων συμπερασμάτων από αυτήν, αποτελούν αντικείμενο έρευνας.

Η τελευταία δεκαετία έχει επιφέρει μια αλματώδη αύξηση στην παραγωγή και συλλογή δεδομένων. Η πρόοδος στην τεχνολογία των βάσεων δεδομένων μας παρέχει νέες τεχνικές για την αποδοτική και αποτελεσματική συλλογή, αποθήκευση και διαχείριση των δεδομένων. Η δυνατότητα ανάλυσης και ερμηνείας των συνόλων δεδομένων και η εξαγωγή της «χρήσιμης» γνώσης από αυτά έχει ξεπεράσει κάθε όριο, και η ανάγκη για μια νέα γενιά εργαλείων και τεχνικών για ευφυή ανάλυση των δεδομένων έχει δημιουργηθεί. Αυτή η ανάγκη έχει προσελκύσει την προσοχή των ερευνητών από διάφορες περιοχές (τεχνητή νοημοσύνη, στατιστική, αποθήκες δεδομένων, διαδραστική ανάλυση και επεξεργασία, έμπειρα συστήματα και οπτικοποίηση δεδομένων) και ένας νέος ερευνητικός τομέας δημιουργείται, γνωστός ως εξόρυξη δεδομένων και γνώσης (Data and Knowledge Mining).

3.2.5. Εξόρυξη γνώσης και δεδομένων

Η ανακάλυψη γνώσης από βάσεις δεδομένων, αναφέρεται στη διεργασία εξόρυξης γνώσης από τις μεγάλες αποθήκες δεδομένων οι οποίες συλλέγουν τα δεδομένα μέσα από την τεράστια κίνηση του παγκοσμίου ιστού. Ο όρος εξόρυξη δεδομένων χρησιμοποιείται ως συνώνυμο της ανακάλυψης γνώσης από βάσεις δεδομένων, καθώς επίσης και για αναφορά στις πραγματικές τεχνικές που χρησιμοποιούνται για την ανάλυση και την εξαγωγή της από διάφορα σύνολα δεδομένων. Πολλοί ερευνητές θεωρούν τον όρο εξόρυξη δεδομένων μη αντιπροσωπευτικό της διαδικασίας που περιγράφει, υποστηρίζοντας ότι ο όρος εξόρυξη γνώσης θα ήταν μια πιο κατάλληλη περιγραφή. Ο όρος εξόρυξη δεδομένων (Data Mining) είναι αυτός που έχει επικρατήσει και χαρακτηρίζει τη διαδικασία της εύρεσης δομών γνώσης οι οποίες περιγράφουν με ακρίβεια μεγάλα σύνολα πρωτογενών δεδομένων. Οι δομές αυτές αναδεικνύουν γνώση (συσχετίσεις ή κανόνες) που είναι κρυμμένοι μέσα στα δεδομένα και δεν μπορούν να εξαχθούν με «γυμνό»' μάτι. Οι προκύπτουσες δομές είναι πλούσιες σε σημασιολογία και εκμεταλλεύονται πιθανές κοινές ιδιότητες των πρωτογενών δεδομένων.

Οι δύο βασικοί στόχοι της εξόρυξης δεδομένων (γνώσης) είναι η εφαρμογή τεχνικών περιγραφής και πρόβλεψης σε μεγάλα σύνολα δεδομένων. Η πρόβλεψη στοχεύει στον υπολογισμό της μελλοντικής αξίας ή στην πρόβλεψη της συμπεριφοράς κάποιων μεταβλητών που παρουσιάζουν ενδιαφέρον (π. χ. το ενδιαφέρον ενός αναγνώστη για διαφόρων κατηγοριών κείμενα) και οι οποίες βασίζονται στη συμπεριφορά άλλων μεταβλητών. Η περιγραφή επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης

δεδομένων με έναν κατανοητό και αξιοποιήσιμο τρόπο. Ως προς την εξόρυξη γνώσης, η περιγραφή τείνει να είναι περισσότερο σημαντική από την πρόβλεψη.

3.2.6. Ανακάλυψη γνώσης από βάσεις δεδομένων σε σχέση με την εξόρυξη γνώσης και δεδομένων

Η ανακάλυψη γνώσης από βάσεις δεδομένων αναφέρεται σε ολόκληρη τη διαδικασία ανακάλυψης χρήσιμης πληροφορίας από μεγάλα σύνολα δεδομένων. Ένας τυπικός ορισμός δόθηκε από τους Frawley, Piatesky-Shaphiro & Matheus:

Ανακάλυψη γνώσης από βάσεις δεδομένων είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα.

Για την κατανόηση του παραπάνω ορισμού, παρατίθενται οι βασικές έννοιες των όρων πάνω στους οποίους είναι βασισμένος.

- Τα δεδομένα περιγράφουν οντότητες ή συσχετίσεις του πραγματικού κόσμου. Παραδείγματος χάριν θα μπορούσε να είναι ένα σύνολο ακατέργαστων κειμένων προερχόμενα από μια πηγή νέων του διαδικτύου.
- Ένα πρότυπο είναι μια έκφραση σε μια γλώσσα η οποία περιγράφει ένα υποσύνολο δεδομένων εκμεταλλευόμενο κοινές ιδιότητες των δεδομένων του.
- Η διαδικασία ανακάλυψη γνώσης από βάσεις δεδομένων είναι μια διαδικασία πολλαπλών βημάτων, η οποία περιλαμβάνει την προ-επεξεργασία των δεδομένων, την αναζήτηση των προτύπων και την αξιολόγηση της εξαγόμενης γνώσης.
- Εγκυρότητα. Το εξαγόμενο πρότυπο (π. χ. περίληψη κειμένου) θα πρέπει να είναι συνεπές σε νέα δεδομένα με κάποιο βαθμό βεβαιότητας. Το ζήτημα της εγκυρότητας αποτελεί ένα από τα βασικά προβλήματα και αντικείμενο έρευνας στην εξόρυξη δεδομένων / πληροφορίας.
- Η εξαγωγή των προτύπων θα πρέπει να ακολουθείται από μερικές χρήσιμες διεργασίες όπως η αξιολόγησή τους από κάποιες συναρτήσεις χρησιμότητας. Για παράδειγμα η αυτόματη περίληψη ενός κειμένου θα πρέπει να μπορεί να αξιολογηθεί ως προς την χρησιμότητα / σαφήνειά και την πιστότητά του όσον αφορά το νόημα σε σχέση με το αρχικό κείμενο. Επίσης, θα ήταν χρήσιμο να εμπλουτιστεί η σημασιολογία των προτύπων, διατηρώντας όσο το δυνατόν περισσότερη γνώση από τα αρχικά δεδομένα η οποία μπορεί να φανεί χρήσιμη για τη λήψη αποφάσεων.
- Τελικά κατανοητό. Ο στόχος της εξόρυξης γνώσης είναι να προσδιοριστούν τα πρότυπα και να γίνουν κατανοητά, ώστε να μπορούν να οδηγήσουν ακόμη και τους μη ειδικούς σε χρήσιμα συμπεράσματα και αποφάσεις.

Η διαδικασία ανακάλυψη γνώσης είναι μια διαλογική και επαναληπτική διαδικασία που αποτελείται από μια σειρά από τα ακόλουθα βήματα:

- Την ανάπτυξη και κατανόηση της περιοχής της εφαρμογής, της σχετικά προγενέστερης γνώσης του προς εξέταση τομέα και τους στόχους του τελικού χρήστη.
- Την ολοκλήρωση των δεδομένων. Υπάρχουν διαφορετικά είδη αποθηκών πληροφοριών που μπορούν να χρησιμοποιηθούν στη διαδικασία εξόρυξης γνώσης. Κατά συνέπεια οι πολλαπλές πηγές δεδομένων μπορούν να συνδυαστούν καθορίζοντας το σύνολο στο οποίο τελικά η διαδικασία εξόρυξης πρόκειται να εφαρμοστεί.
- Τη δημιουργία του στόχου-συνόλου δεδομένων. Επιλογή του συνόλου δεδομένων (δηλαδή μεταβλητές, δείγματα δεδομένων) στο οποίο η διαδικασία εξόρυξης πρόκειται να εκτελεστεί.

- Τον καθαρισμό και την προ-επεξεργασία δεδομένων. Αυτό το βήμα περιλαμβάνει βασικές διαδικασίες όπως η αφαίρεση του θορύβου, η συλλογή των απαραίτητων πληροφοριών για τη διαμόρφωση ή τη μέτρηση του θορύβου, η απόφαση σχετικά με τις στρατηγικές διαχείρισης των ελλειπόντων πεδίων δεδομένων.
- Τον μετασχηματισμό των δεδομένων. Τα δεδομένα μετασχηματίζονται ή παγιώνονται σε μορφές κατάλληλες για εξόρυξη. Χρήση των μεθόδων μείωσης διαστάσεων ή μετασχηματισμού για τη μείωση του αριθμού των υπό εξέταση μεταβλητών ή την εύρεση κατάλληλης αντιπροσώπευσης των δεδομένων χωρίς μεταβλητές.
- Την επιλογή των στόχων και των αλγορίθμων εξόρυξης δεδομένων. Σε αυτό το βήμα αποφασίζουμε το στόχο της διαδικασίας εξόρυξης γνώσης, επιλέγοντας τους στόχους εξόρυξης δεδομένων που θέλουμε να επιτύχουμε. Επίσης, επιλέγονται οι μέθοδοι που θα χρησιμοποιηθούν. Αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου και παραμέτρων.
- Την εξόρυξη δεδομένων. Εφαρμόζοντας ευφυείς μεθόδους, ψάχνουμε για ενδιαφέροντα πρότυπα γνώσης. Τα πρότυπα θα μπορούσαν να είναι μιας συγκεκριμένης αντιπροσωπευτικής μορφής ή ενός συνόλου τέτοιων αντιπροσωπεύσεων, όπως κανόνες κατηγοριοποίησης, δέντρα, συσταδοποίηση, κλπ Η απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα προηγούμενα βήματα.
- Την αξιολόγηση των προτύπων. Τα εξαγόμενα πρότυπα αξιολογούνται με κάποια μέτρα, προκειμένου να προσδιοριστούν τα πρότυπα τα οποία αντιπροσωπεύουν τη γνώση, δηλαδή τα αληθινά ενδιαφέροντα πρότυπα.
- Τη σταθεροποίηση και παρουσίαση της γνώσης. Σε αυτό το βήμα, η εξορυγμένη γνώση ενσωματώνεται το σύστημα ή απλά την απεικόνισή μας και κάποιες τεχνικές αντιπροσώπευσης γνώσης χρησιμοποιούνται για να παρουσιάσουν την εξορυγμένη γνώση στο χρήστη. Επίσης, ελέγχουμε για επίλυση τυχών συγκρούσεων με προηγούμενη εξορυγμένη γνώση.

Η εξόρυξη δεδομένων ως βήμα της διαδικασίας εξόρυξης γνώσης ενδιαφέρεται κυρίως για τις μεθοδολογίες και τις τεχνικές εξαγωγής προτύπων δεδομένων ή τις περιγραφές δεδομένων από τις μεγάλες αποθήκες δεδομένων. Αφ' ετέρου, η διαδικασία εξόρυξης γνώσης περιλαμβάνει την αξιολόγηση και την ερμηνεία των προτύπων. Επίσης περιλαμβάνει την επιλογή της κωδικοποίησης των προτύπων, της προ-επεξεργασίας, της δειγματοληψίας και του μετασχηματισμού των δεδομένων πριν από το βήμα της εξόρυξης των δεδομένων.

3.2.7. Η διαδικασία εξόρυξης δεδομένων

Η εξόρυξη δεδομένων περιλαμβάνει τα μοντέλα συναρμολογήσεων των υπό εξέταση δεδομένων, ή εναλλακτικά την εξαγωγή των προτύπων από αυτά. Ουσιαστικά, οι παράμετροι του μοντέλου είναι γνωστές από τα δεδομένα ή τα πρότυπα που προσδιορίζονται, αντιπροσωπεύουν τη γνώση που έχει εξαχθεί από ένα σύνολο δεδομένων.

Υπάρχει μια μεγάλη συλλογή αλγορίθμων εξόρυξης δεδομένων, πολλοί από τους οποίους χρησιμοποιούν έννοιες και τεχνικές από διαφορετικούς τομείς όπως η στατιστική, η αναγνώριση προτύπων, η μηχανική μάθηση, οι αλγόριθμοι και οι βάσεις δεδομένων. Μια θεμελιώδης ιδιότητα των αλγορίθμων εξόρυξης δεδομένων, και αυτή που διαφοροποιεί τους περισσότερους από αυτούς από άλλες παρόμοιες τεχνικές που υιοθετούνται στη μηχανική μάθηση και τη στατιστική, είναι ότι οι αλγόριθμοι εξόρυξης δεδομένων έχουν σχεδιαστεί με έμφαση στην εξελιξιμότητα όσον αφορά το μέγεθος του συνόλου δεδομένων εισαγωγής. Η πλειοψηφία των αλγορίθμων εξόρυξης δεδομένων θα μπορούσε να περιγραφεί σε

υψηλό επίπεδο με τον όρο ενός απλού πλαισίου. Συγκεκριμένα μπορούν να αντιμετωπισθούν ως σύνθεση των τριών ακολούθων συστατικών:

- Την περιγραφή του μοντέλου. Υπάρχουν δύο παράγοντες σχετικοί με το μοντέλο:
- Η λειτουργία του μοντέλου. Καθορίζει τους βασικούς στόχους κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων.
- Η παραστατική μορφή του μοντέλου. Η απεικόνιση του μοντέλου καθορίζει και το ταίριασμά του με την απεικόνιση των δεδομένων και τη δυνατότητα να ερμηνευθεί το μοντέλο με κατανοητούς όρους. Χαρακτηριστικά, πιο περίπλοκα μοντέλα ταιριάζουν καλύτερα στα δεδομένα αλλά μπορεί να είναι δυσκολότερο να γίνουν κατανοητά και να ανταποκριθούν σε πραγματικές συνθήκες.
- Την αξιολόγηση του μοντέλου. Με βάση κάποια κριτήρια αξιολόγησης (π. χ. μέγιστη πιθανότητα) θα μπορούσαμε να καθορίσουμε πόσο καλά ένα συγκεκριμένο μοντέλο ταιριάζει με τα κριτήρια της διαδικασίας εξόρυξης γνώσης. Γενικά, η αξιολόγηση του μοντέλου αναφέρεται και στην εγκυρότητα των προτύπων και στην αξιολόγηση της ακρίβειας, της χρησιμότητας και της δυνατότητας κατανόησης του μοντέλου.
- Τους αλγορίθμους αναζήτησης. Αναφέρεται στην προδιαγραφή ενός αλγορίθμου να βρίσκει συγκεκριμένα μοντέλα και παραμέτρους, δοσμένου ενός συνόλου δεδομένων, μιας οικογένειας μοντέλων και ενός κριτηρίου αξιολόγησης. Υπάρχουν δύο τύποι αλγορίθμων αναζήτησης:
- Αυτοί που αναζητούν παραμέτρους. Αυτός ο τύπος αλγορίθμων ψάχνει για παραμέτρους, οι οποίες βελτιστοποιούν ένα κριτήριο αξιολόγησης για το μοντέλο. Οι αλγόριθμοι εκτελούν το στόχο αναζήτησης παίρνοντας ως είσοδο ένα σύνολο δεδομένων και μια απεικόνιση μοντέλου.
- Αυτοί που αναζητούν μοντέλα. Εκτελούν μια επαναληπτική διαδικασία αναζήτησης για την αντιπροσώπευση των δεδομένων. Για κάποια συγκεκριμένη απεικόνιση του μοντέλου, εφαρμόζεται η μέθοδος αναζήτησης παραμέτρων και η ποιότητα των αποτελεσμάτων αξιολογείται.

3.2.8. Κατηγορίες μεθόδων εξόρυξης πληροφορίας

Τα τελευταία χρόνια διάφορες τεχνικές και μέθοδοι εξόρυξης δεδομένων έχουν αναπτυχθεί. Διαφορετικά κριτήρια κατηγοριοποίησης μπορούν να χρησιμοποιηθούν για να κατηγοριοποιήσουν τις μεθόδους και τα συστήματα εξόρυξης δεδομένων, βασισμένες στους τύπους των βάσεων δεδομένων που θα χρησιμοποιηθούν, τους τύπους γνώσης που θα εξαχθούν και τις τεχνικές που θα εφαρμοστούν. Η κατηγοριοποίηση των μεθόδων εξόρυξης πληροφορίας βασίζεται στα ακόλουθα κριτήρια:

- Είδος πηγής δεδομένων που χρησιμοποιείται. Π. χ. ένα σύστημα εξόρυξης πληροφορίας που χρησιμοποιεί δεδομένα μια σχεσιακής βάσης δεδομένων μπορεί να ονομαστεί σχεσιακό.
- Είδος γνώσης που εξάγεται. Από ένα σύστημα εξόρυξης δεδομένων θα μπορούσαν να εξαχθούν διάφορα είδη γνώσης, όπως κανόνες συσχέτισης, συσταδοποίηση, κανόνες κατηγοριοποίησης, χαρακτηριστικοί κανόνες. Ένα σύστημα εξόρυξης δεδομένων θα μπορούσε να ταξινομηθεί σύμφωνα με το επίπεδο γενίκευσης της εξαγόμενης γνώσης, η οποία θα μπορούσε να είναι γενική, πρώτου επιπέδου ή πολυεπίπεδη γνώση.
- Είδος χρησιμοποιούμενων τεχνικών. Τα συστήματα εξόρυξης δεδομένων θα μπορούσαν να ταξινομηθούν σύμφωνα με τις χρησιμοποιούμενες τεχνικές εξόρυξης δεδομένων. Για παράδειγμα, θα μπορούσαν να ταξινομηθούν σε

αυτόνομα συστήματα, συστήματα προσανατολισμένα στα δεδομένα, συστήματα οδηγούμενα από ερωταποκρίσεις καθώς και διαλογικά συστήματα. Επίσης, σύμφωνα με την προσέγγιση που χρησιμοποιείται θα μπορούσαν να ταξινομηθούν σε συστήματα γενικής εξόρυξης, εξόρυξης βασισμένης στα πρότυπα, εξόρυξης βασισμένης στη στατιστική ή στα μαθηματικά κλπ.

3.2.9. Εύρεση προτύπων συσχέτισης

Η ανακάλυψη χρήσιμης πληροφορίας, μέσα σε συγκεκριμένα έγγραφα, αποτελεί το πεδίο δράσης της διαδικασίας της εύρεσης προτύπων συσχέτισης (Association Patterns) . Οι Arimura Hiroki, Wataki Atsushi, Fujino Ryoichi και Arikawa Setsuo [62], μελέτησαν την ανακάλυψη πολύ απλών προτύπων, που τα ονόμασαν πρότυπα συσχέτισης ζευγών λέξεων -εγγύτητας (k-proximity two-words association patterns). Σε μία δεδομένη συλλογή κειμένων και με τη χρήση μιας αντικειμενικής συνθήκης, ορίζεται το πρότυπο συσχέτισης. Το πρότυπο αυτό, εκφράζει ένα κανόνα που αναφέρει ότι αν βρεθεί η υπολέξη που περιέχεται στο πρότυπο, ακολουθούμενη από μία άλλη δεδομένη υπολέξη, σε συγκεκριμένη απόσταση γραμμάτων, τότε η αντικειμενική συνθήκη θα διατηρηθεί με μεγάλη πιθανότητα.

Οι κανόνες αυτοί είναι πολύ ευέλικτοι για την περιγραφή των τοπικών ομοιοτήτων που περιέχονται στα δεδομένα του κειμένου. Το είδος των κανόνων αυτών, χρησιμοποιείται για παράδειγμα στην βιοπληροφορική, στην βιβλιογραφική έρευνα και στην έρευνα στο διαδίκτυο. Ως γενικό πλαίσιο εργασίας, ο αλγόριθμος ανακάλυψης προτύπων λαμβάνει ένα σύνολο δειγμάτων με μία συγκεκριμένη συνθήκη και βρίσκει όλα ή μερικά από τα πρότυπα, τα οποία μεγιστοποιούν ένα συγκεκριμένο κριτήριο.

Διακρίνουμε το πρόβλημα του προτύπου βέλτιστης εμπιστοσύνης όπου, δεδομένου ενός συνόλου από έγγραφα και με μία αντικειμενική συνθήκη για αυτό το σύνολο, υπολογίζεται το πρότυπο που μεγιστοποιεί την τιμή των κριτηρίων που έχουν τεθεί για τα συγκεκριμένα έγγραφα. Ένα δεύτερο πρόβλημα, αναφέρεται στην ελαχιστοποίηση του εμπειρικού λάθους, όπου αναζητείται ένα πρότυπο που θα ελαχιστοποιεί τον αριθμό των εγγράφων που έχουν επεξεργαστεί με λάθος τρόπο.

Χαρακτηριστικές εφαρμογές που χρησιμοποιούν την εύρεση προτύπων συσχέτισης, είναι αυτές που αναλύουν απλά έγγραφα κειμένου, όπως προτείνουν και οι Montes-y-Gomez M., Gelbukh A. και Lopez-Lopez A. [63]. Προσπαθούν να ανακαλύψουν τις σχέσεις που υπάρχουν ανάμεσα στα διάφορα θέματα που παρουσιάζονται σε αφημερίδες. Επιχειρούν να ανακαλύψουν τον τρόπο που τα θέματα της λεγόμενης πρώτης σελίδας, επηρεάζουν και όλα τα υπόλοιπα θέματα της ειδησεογραφίας. Οι συσχετίσεις που υπάρχουν ανάμεσα στα διάφορα ειδησεογραφικά θέματα, καλούνται εφήμερες (Ephemeral Associations). Άλλη χαρακτηριστική εφαρμογή, αποτελεί η ανακάλυψη προτύπων σε σύνολα ακολουθιών DNA, που προτείνουν οι Kiem Hoang και Phuc Do [64]. Μελετούν υποακολουθείες που εμφανίζονται πολύ συχνά στο σύνολο των ακολουθιών DNA, για την ανακάλυψη εκείνων των κανόνων συσχέτισης, που βασίζονται στην επανάληψη.

3.2.10. Ανάκτηση γνώσης από βάσεις δεδομένων

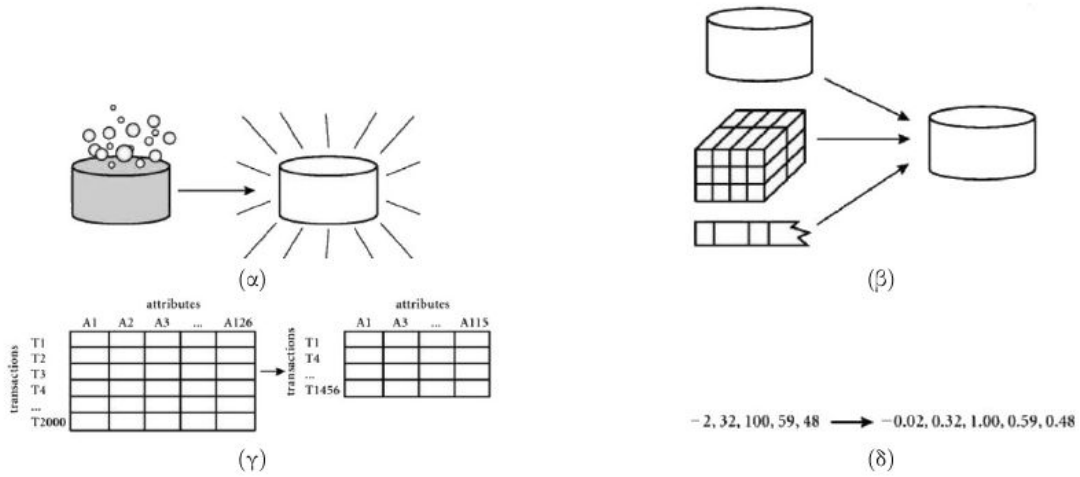
Η ανάκτηση γνώσης από βάσεις δεδομένων (Knowledge Discovery in Databases - KDD) είναι η μη τετριμμένη διαδικασία της αναγνώρισης έγκυρων, καινότυπων, ενδεχόμενα χρήσιμων και τελικά κατανοητών προτύπων δεδομένων. Τα ακατέργαστα δεδομένα είναι πάντοτε «ακάθαρτα» με την έννοια ότι πάντα θα υπάρχουν διπλοεγγραφές, ελλιπή πεδία και μη ακριβές τιμές δεδομένων. Είναι

επιθυμητό επομένως, τα αποτελέσματα των αναζητήσεων να πρέπει να περάσουν από κάποιο στάδιο εκκαθάρισης πριν παρουσιαστούν στον χρήστη. Η εκκαθάριση δεδομένων στην KDD διαδικασία είναι ένα βασικό βήμα για την αφαίρεση του θορύβου και των outliers[1], την συγκέντρωση των σχετικών πληροφοριών για μοντελοποίηση του θορύβου και την λήψη αποφάσεων για τα ελλιπή δεδομένα.

Τα καθαρά δεδομένα υπονοούν και σχετικά δεδομένα παρότι η σχετικότητα των δεδομένων είναι συνήθως υποκειμενική. Είναι όμως γεγονός ότι μια ακριβής περίληψη ενός κειμένου μπορεί να χρησιμοποιηθεί για να εκτιμηθεί η σχετικότητα ή μη του αρχικού κειμένου με τα ενδιαφέροντα του χρήστη. Παράλληλα, μια προηγούμενη αντιστοίχηση των εξαγομένων κειμένων με ορισμένα πεδία ενδιαφέροντος μπορεί να βοηθήσει στον εντοπισμό των outliers. Αυτό σημαίνει ότι εκείνα τα έγγραφα που δεν εμπίπτουν στις κατηγορίες ενδιαφέροντος του χρήστη, μπορούν να αγνοηθούν.

3.3. Προεπεξεργασία Δεδομένων

Τα δεδομένα που κατακλύζουν τις σύγχρονες βάσεις δεδομένων και τον παγκόσμιο ιστό σήμερα, είναι πολύ επιρρεπή σε θόρυβο, σε ανεπάρκεια ή συνοχή λόγω κυρίως του τεράστιου όγκου και της ετερογένειας τον πηγών τους. Δεδομένα χαμηλής ποιότητας οδηγούν σε χαμηλής ποιότητας εξόρυξη πληροφορίας. Το θεμελιώδες ερώτημα που τίθεται είναι: πώς μπορούν να προεπεξεργαστούν τα δεδομένα, ώστε να βελτιωθεί η ποιότητά τους και επομένως τα αποτελέσματα της εξόρυξης πληροφορίας.



Εικόνα 3: Τεχνικές προεπεξεργασίας δεδομένων (α)Καθαρισμός δεδομένων (β)Ολοκλήρωση δεδομένων (γ)Αφαίρεση δεδομένων (δ)Μετασχηματισμός δεδομένων

Υπάρχει ένα πλήθος μεθόδων που χρησιμοποιούνται για την προεπεξεργασία δεδομένων. Το καθάρισμα δεδομένων μπορεί να έχει εφαρμογή στην αφαίρεση του θορύβου από τα δεδομένα και στην διόρθωση των ασυνεπειών σε αυτά. Η ολοκλήρωση των δεδομένων συνενώνει δεδομένα από διάφορες πηγές σε συναφή αποθήκη δεδομένων, όπως π. χ. μια βάση δεδομένων. Ο μετασχηματισμός των δεδομένων, όπως η κανονικοποίηση μπορεί να χρησιμοποιηθεί από τη διαδικασία προεπεξεργασίας δεδομένων. Για παράδειγμα, η κανονικοποίηση μπορεί να βελτιώσει την ακρίβεια και την αποτελεσματικότητα των αλγορίθμων εξόρυξης δεδομένων ενσωματώνοντας μετρικές απόστασης. Η αφαίρεση δεδομένων, μπορεί να μειώσει το μέγεθος των δεδομένων, συναθροίζοντας, απαλείφοντας τα

[1] δεδομένα που βρίσκονται εκτός του διαστήματος τυπικής απόκλισης των υπολοίπων δεδομένων και ως εκ' τούτου αποτυγχάνουν να αναπαραστήσουν σωστά την πληροφορία

πλεονάζοντα χαρακτηριστικά, ή ομαδοποιώντας τα δεδομένα. Αυτές οι τεχνικές δεν είναι αμοιβαία αποκλειόμενες· μπορούν να δουλέψουν μαζί. Για παράδειγμα, το καθάρισμα δεδομένων μπορεί να περιλαμβάνει μετασχηματισμούς για την διόρθωση λανθασμένων δεδομένων. Οι τεχνικές προεπεξεργασίας δεδομένων, όταν εφαρμόζονται πριν την εξόρυξη πληροφορίας, μπορούν να βελτιώσουν σημαντικά την ποιότητα της πληροφορίας που εξορύσσεται ή τον χρόνο που απαιτείται γι' αυτή τη διαδικασία.

3.3.1. Αφαίρεση σημείων στίξης

Τα σημεία στίξης (punctuation) ενός κειμένου δεν προσδίδουν σημασιολογική πληροφορία σε αυτό και άρα δεν δεικτοδοτούνται. Είναι επομένως αναγκαίο, ένα σύστημα ανάκτησης πληροφορίας να αφαιρεί κάθε σημείο στίξης από το αρχικό κείμενο σε πρώιμα στάδια της προεπεξεργασίας. Ιδιαίτερη μέριμνα πρέπει να λαμβάνεται ώστε να συγκρατείται το τέλος της κάθε πρότασης (π. χ. με κάποιο άλλο διαχωριστικό πέραν της τελείας) ώστε να είναι δυνατός ο μετέπειτα διαχωρισμός των προτάσεων. Η διαδικασία θα πρέπει να λαμβάνει όσο το δυνατός καλύτερα υπ' όψιν τις γλωσσολογικές ιδιομορφίες της εκάστοτε γλώσσας ώστε να μην προκύπτουν λάθη κατά τη διαδικασία της αφαίρεσης των σημείων στίξης. Ορισμένα παραδείγματα:

- Ne’er: χρήση language-specific πηγών για τον κατάλληλο μετασχηματισμό
- State-of-the-art: διαχωρισμός λέξεων με παύλες σε ξεχωριστά tokens
- U.S.A. vs. USA: απομάκρυνση ενδιάμεσων τελειών σε ακρωνύμια

3.3.2. Αφαίρεση αριθμών

Γενικά, οι αριθμοί ενός κειμένου δεν δεικτοδοτούνται (τουλάχιστον όχι όπως το υπόλοιπο κείμενο) για λόγους παρόμοιους με αυτών των σημείων στίξης. Η αντιμετώπισή τους μπορεί να ποικίλει από IR σε IR σύστημα και εξαρτάται κυρίως από τις απαιτήσεις που θέτονται. Σπάνια χρειάζεται να ανακτηθεί μια ημερομηνία π. χ. από ένα μεγάλο κείμενο αλλά η πληροφορία αυτή μπορεί να αποθηκευτεί ως meta-δεδομένο για το κείμενο.

3.3.3. Κεφαλαία γράμματα

Η διάκριση μεταξύ κεφαλαίων και μικρών γραμμάτων, αμελητέα μόνο σημασιολογική πληροφορία μπορεί να δώσει για το κείμενο. Για το λόγο αυτό, και για ομοιομορφία των προς επεξεργασία λέξεων, όλα τα κεφαλαία γράμματα συνήθως μετασχηματίζονται σε μικρά.

3.4. Περίληψη Πληροφορίας

Η διαδικασία της περίληψης κειμένου (Text Summarization) , αποσκοπεί στην παρουσίαση των κύριων σημείων ενός εγγράφου, σε μία περιεκτική μορφή. Μία πραγματική περίληψη, θα πρέπει να εκφράζει την ουσία του εγγράφου, αποκαλύπτοντας το βαθύτερο νόημα του περιεχομένου του. Σκοπός της είναι, η ανακάλυψη ενδιαφέρουσας και απροσδόκητης πληροφορίας. Σύμφωνα με τον Crangle Colleen [65], υπάρχουν δύο κύριες αντιλήψεις για την εξαγόμενη περίληψη του αρχικού κειμένου. Η πρώτη αναφέρει ότι η περίληψη θα περιέχει προτάσεις οι οποίες περιέχονται μόνο στο αρχικό κείμενο. Η δεύτερη είναι πιο σύνθετη και αναφέρει ότι εκτός των αρχικών προτάσεων του κειμένου, είναι δυνατόν να υπάρχουν και άλλες, κατασκευασμένες από τον μηχανισμό περίληψης. Οι προτάσεις αυτές, είτε θα δημιουργούνται με τη χρήση τμημάτων των αρχικών προτάσεων, είτε με την επεξεργασία των αρχικών και την παραγωγή νέων, που δεν θα περιέχουν τμήματα, που υπάρχουν στις αρχικές προτάσεις. Μπορούμε να αναφερθούμε στις δύο αυτές διαφορετικές κλάσεις τεχνικών περίληψης κειμένου χρησιμοποιώντας τις έννοιες αφαίρεση και εξαγωγή.

Σε αντίθεση με τις τεχνικές της αφαίρεσης, οι οποίες απαιτούν τεχνικές Natural Language Processing - NLP, συμπεριλαμβανομένων γραμματικών και λεξικών για την ανάλυση του κειμένου, η εξαγωγή μπορεί να θεωρηθεί ως μια διεργασία επιλογής σημαντικών αποσπασμάτων (προτάσεων, παραγράφων, κ.λπ.) από το αρχικό κείμενο και συνένωσής του σε μια νέα πιο σύντομη έκδοση.

Οι περιλήψεις κειμένων μπορεί να είναι είτε συσχετιζόμενες με κάποιο ερώτημα χρήστη (προτιμήσεις του χρήστη), είτε γενικές. Το πρώτο είδος επιστρέφει περιεχόμενο του κειμένου που ανταποκρίνεται στις προτιμήσεις του χρήστη, μια διαδικασία που περιέχει πολλά κοινά με την διαδικασία ανάκτησης κειμένων και ως εκ' τούτου, οι αλγόριθμοι που χρησιμοποιούνται συνήθως πηγάζουν από αυτή. Από την άλλη μεριά, μια γενική περίληψη παρέχει μια συνολική άποψη για τα περιεχόμενα του κειμένου. Μια καλή γενική περίληψη πρέπει να περιέχει τα βασικά σημεία του κειμένου διατηρώντας παράλληλα τον πλεονασμό στο ελάχιστο. Σε αυτή την εργασία αξιοποιούνται τεχνικές που αφορούν και τα δύο είδη περίληψης: α)γενική και β)προσωποποιημένη στο χρήστη.

3.4.1. Αλγόριθμοι για αυτόματη εξαγωγή περίληψης

Ένα από τα σημαντικότερα συστήματα του μηχανισμού που σχεδιάζουμε και βρίσκεται στον πυρήνα του μηχανισμού είναι οι συναρτήσεις και οι αλγόριθμοι για αυτόματη εξαγωγή περίληψης. Οι προσπάθειες για αυτοματοποίηση της διαδικασίας εξαγωγής περίληψης ξεκινούν από τη δεκαετία του 50 όταν ο H.P. Luhn [19] προσπαθούσε να βρει έναν αλγόριθμο για την παραγωγή περίληψης κειμένου. Η δουλειά του θεωρείται από τις πλέον κλασσικές και ολοκληρωμένες και πάνω σε αυτή βασίζονται ακόμα και πολύ πρόσφατες θεωρίες [20]. Σε αυτές τις τεχνικές η ανάλυση γίνεται σε επίπεδο λέξης μέσα στην πρόταση. Ουσιαστικά η τεχνική βασίζεται στο γεγονός πως θα πρέπει να υπάρχουν κάποια στοιχεία σε όλο το κείμενο που αφορούν τις λέξεις και αποδεικνύουν πως λέξεις ή και ολόκληρες φράσεις δε θα πρέπει να λείπουν από την περίληψη του κειμένου [21], [22], [23], [24]. Πιο σύνθετες τεχνικές ελέγχουν το μέγεθος της πρότασης ή και την επανάληψη λέξεων με συγκεκριμένη σειρά. Με αυτό τον τρόπο δομούνται ιεραρχικά οι προτάσεις που περικλείουν το «νόημα» το κειμένου ενώ παράλληλα είναι εφικτή η «νοηματική» συσχέτιση προτάσεων, λέξεων και συστοιχιών λέξεων. Σε αυτές τις τεχνικές χρησιμοποιούνται στατιστικά από το ίδιο το κείμενο που αναλύεται ενώ παράλληλα δύνανται να χρησιμοποιηθούν στοιχεία από πρότυπες περιλήψεις προκειμένου να «γνωρίζει» ποιος είναι ο τρόπος με τον οποίο δομείται μία περίληψη[23], [25], [26].

Πολλές τεχνικές για αυτόματη εξαγωγή περίληψης βασίζονται σε NLP (Natural Language Processing) και σε πληροφορίας ομιλίας. Μερικές προορίζονται αποκλειστικά σε τεχνικές που έχουν αναπτυχθεί στο πλαίσιο της ανάκτησης πληροφορίας (IR). Άλλες πάλι προσπαθούν να ισορροπήσουν μεταξύ του NLP και του IR. [32], [33]

Φυσικά οι τεχνικές για την αυτόματη εξαγωγή περίληψης δε βασίζονται στην εξαγωγή των σημαντικότερων στοιχείων από ένα κείμενο. Πολλές τεχνικές υπάρχουν που βασίζονται στη δημιουργία από το μηδέν μίας περίληψης που αντιπροσωπεύει σε μεγάλο βαθμό το νόημα του κειμένου. Κάποιες από τις τεχνικές αυτές βασίζονται σε μοντέλα γνώσης και πιο συγκεκριμένα προσπαθούν να μοντελοποιήσουν γνωστικά το νόημα ενός κειμένου έχοντας σαν βάση στατιστικά στοιχεία που εξάγονται από το κείμενο.[29], [30]

Πολλές από τις προσπάθειες για αυτόματη εξαγωγή περίληψης έχουν δοκιμαστεί στο πεδίο της έρευνας που εντοπίζουμε και στην παρούσα εργασία. Πιο συγκεκριμένα, πολλές ερευνητικές εργασίες έχουν γίνει πάνω στο θέμα της εξαγωγής περίληψης από νέα, άρθρα, ειδήσεις [31]. Μάλιστα, δεδομένου ότι οι ειδήσεις και τα άρθρα αναφέρονται συχνά σε γεγονότα πολλές προσπάθειες έχουν

εντοπιστεί στην εξαγωγή των γεγονότων με διάφορους τρόπος και εν συνεχεία στη δόμηση της περίληψης γύρω από το συγκεκριμένο γεγονός. Μάλιστα οι συγκεκριμένες τεχνικές έχουν την ευκολία και τη δυνατότητα για παρακολούθηση των αλλαγών που πραγματοποιούνται σε ένα συγκεκριμένο θέμα. Πολλές βέβαια τεχνικές βασίζονται απλά στο γεγονός ότι από πολλά όμοια ή παραπλήσια άρθρα που ασχολούνται με το ίδιο θέμα μπορεί να εξαχθεί μία και μόνον περίληψη έχοντας στοιχεία από όλα τα άρθρα.[34], [35], [36]

Η εξαγωγή περίληψης που αφορά τις ειδήσεις και τα άρθρα που είναι συνήθως πολύ επίκαιρα είναι ένα θέμα που έχει απασχολήσει πολλούς ερευνητές. Μάλιστα, έχουν οριστεί και μετρικές οι οποίες αφορούν την ισορροπία που μπορεί να υπάρχει ανάμεσα στη χρησιμότητα και στην ποιότητα μίας περίληψης αλλά και στο κατά πόσο είναι σύμφωνο με το χρόνο (up-to-date). Σε αυτή την περίπτωση, το ενδιαφέρον δε στρέφεται αποκλειστικά στη σωστή απάντηση στις ερωτήσεις του χρήστη, αλλά στη σωστή δόμηση της πληροφορίας που παρουσιάζεται στο χρήστη.

3.4.2. Χρησιμότητα της περίληψης κειμένου

Στο επίκαιρο σενάριο της συνδυαστικής έκρηξης της πληροφορίας που εμφανίζεται στις μέρες μας, η αναζήτηση για καλύτερες τεχνικές εξαγωγής πληροφορίας (Information Retrieval - IR) συνεχίζει να γοητεύει τους επιστήμονες της πληροφορικής. Παρότι όμως τα σύγχρονα συστήματα για αναζήτηση και ανάκτηση πληροφορίας είναι ικανά να ανακτούν χιλιάδες εγγράφων στην επιφάνεια εργασίας των χρηστών και μάλιστα σε πολύ σύντομο χρονικό διάστημα, απέχουν πολύ από την ιδανική λύση. Ο χρήστης πρέπει να κάνει πολλές κρίσεις που έχουν να κάνουν με τη σχετικότητα των εγγράφων με τα ενδιαφέροντά του «ξαφρίζοντας» μέσα από πολλαπλά έγγραφα, τα περισσότερα εκ' των οποίων είναι άσχετα. Η διαδικασία αυτή είναι ιδιαίτερα επίπονη και χρονοβόρα για τον χρήστη που επιθυμεί να εντοπίσει γρήγορα και εύκολα το κείμενο που επιθυμεί.

Είναι λοιπόν προφανές ότι η κοινότητα των χρηστών θα ωφεληθεί σημαντικά εάν τα ανακτημένα έγγραφα «συμπυκνωθούν» με κάποιον τρόπο και παρουσιαστούν πίσω στον τελικό χρήστη με τη μορφή αναγνώσιμης και εύκολα διαχειρίσιμης περίληψης. Δυστυχώς, οι απαιτήσεις για ακρίβεια και ανάκληση επιβάλουν αντικρουόμενες απαιτήσεις στο σύστημα. Σε αυτό το ζήτημα είναι εύλογο να θεωρηθεί ότι μια αναζήτηση με υψηλή ακρίβεια με τα ενδιαφέροντα του χρήστη είναι πιο πιθανό να ικανοποιήσει τον μέσο χρήστη σε σχέση με μια εξαντλητική αναζήτηση ενός μεγάλου πλήθους κειμένων. Αυτά τα θέματα, μαζί με την αυξανόμενη ποικιλία των συλλογών κειμένων, αναδεικνύουν τον τομέα της αυτοματοποιημένης περίληψης κειμένων ως έναν από τους βασικότερους της ανάκτησης πληροφορίας.

Οι περιλήψεις κειμένων, μπορούν να χρησιμοποιηθούν από αναλυτές πληροφοριών, έτσι ώστε να είναι σε θέση να γνωρίζουν αν θα πρέπει να μελετήσουν κάποια κείμενα στο σύνολο τους, και κάποια άλλα με διαφορετικό και πιο περιεκτικό τρόπο. Οι περιλήψεις μπορούν να αποκαλύψουν ομοιότητες στο περιεχόμενο των κειμένων, οι οποίες μπορούν να χρησιμοποιηθούν για την μετέπειτα ομαδοποίηση ή κατηγοριοποίηση των εγγράφων. Η διαδικασία της κατηγοριοποίησης ή ομαδοποίησης των περιλήψεων περισσοτέρων του ενός εγγράφου, μέσα σε μία συλλογή, μπορεί να αποκαλύψει αναπάντεχες σχέσεις μεταξύ των εγγράφων. Επιπλέον, η περίληψη μιας συλλογής από σχετιζόμενα έγγραφα, που έχουν επεξεργαστεί μαζί, μπορεί να αποκαλύψει αθροιστική πληροφορία, που υπάρχει μόνο στο επίπεδο της συλλογής των εγγράφων.

3.4.3. Η διαδικασία της περίληψης

Μια αποτελεσματική περίληψη κειμένου εντοπίζει την σημαντική πληροφορία από μια ή περισσότερες πηγές και παράγει μια συντομευμένη έκδοση της αρχικής

πληροφορίας. Η διαδικασία της αυτοματοποιημένης περίληψης περικλείει τουλάχιστον τέσσερα διακριτά στάδια επεξεργασίας:

- Ανάλυση του κειμένου
- Αναγνώριση / Εντοπισμός των σημαντικών τμημάτων του κειμένου
- Συμπύκνωση πληροφορίας και
- Παραγωγή της αναπαράστασης της περίληψης που προκύπτει.

Εικόνα 4: Διαδικασία Εξαγωγής Περίληψης

3.4.4. Αξιολόγηση της εξαγόμενης περίληψης

Η αξιολόγηση της περίληψης που προκύπτει από ένα σύστημα αυτόματης εξαγωγής περίληψης, είναι μια εργασία εξίσου σημαντική με την ίδια τη διαδικασία εξαγωγής. Η αξιολόγηση όμως πρέπει να είναι «φθηνή», από άποψη υπολογιστικού κόστους και συνάμα εφαρμόσιμη και αποτελεσματική για ένα ευρύ φάσμα κειμένων που εισέρχονται στο σύστημα. Στη συνέχεια περιγράφονται οι πλέον συνηθισμένοι τρόποι αξιολόγησης μιας περίληψης.

3.4.5. Αξιολόγηση με συσχέτιση προτάσεων

Η συσχέτιση των εξαγόμενων προτάσεων με το αρχικό κείμενο περιλαμβάνει μετρικές ακρίβειας και ανάκλησης. Αυτές οι μέθοδοι, προϋποθέτουν την ύπαρξη μιας διαθέσιμης «απόλυτα σωστής» περίληψης (στην οποία μπορούμε να υπολογίσουμε την ακρίβεια και την ανάκληση). Μπορούμε να λάβουμε μια τέτοια περίληψη με αρκετούς τρόπους. Πιο συνηθέστερα, λαμβάνεται με τη βοήθεια διαφόρων ανθρώπων που παράγουν περιλήψεις, και στη συνέχεια βρίσκοντας ένα «μέσο όρων» αυτών. Αυτή η μέθοδος όμως είναι συνήθως προβληματική.

3.4.6. Μέθοδοι βασιζόμενοι σε περιεχόμενο

Αυτές οι μέθοδοι υπολογίζουν την ομοιότητα ανάμεσα σε δύο κείμενα σε ένα πιο λεπτομερές επίπεδο από αυτό των απλών προτάσεων. Η βασική μέθοδος συνίσταται από τον υπολογισμό της ομοιότητας μεταξύ του αρχικού κειμένου και της περίληψής του με χρήση της μετρικής ομοιότητας συνημιτόνου:

$$\cos(X,Y) = \frac{\sum x_i * y_i}{\sqrt{\sum (x_i)^2} * \sqrt{\sum (y_i)^2}},$$

όπου τα και βασίζονται στο μοντέλο διανυσματικού χώρου.

3.4.7. Συσχέτιση ομοιότητας

Αφορά τον υπολογισμό της σχετικής μείωσης στο πληροφοριακό περιεχόμενο όταν γίνεται χρήση της περίληψης αντί του αρχικού κειμένου.

3.4.8. Αξιολόγηση βασισμένη σε εργασίες

Αυτές οι τεχνικές μετρούν την ανθρώπινη απόδοση χρησιμοποιώντας τις περιλήψεις για μια συγκεκριμένη εργασία (αφού έχουν παραχθεί οι περιλήψεις). Μπορούμε για παράδειγμα να μετρήσουμε την αποτελεσματικότητα της χρήσης

περιλήψεων αντί των κειμένων για κατηγοριοποίηση αυτών. Αυτού του είδους η αξιολόγηση απαιτεί μια προ-κατηγοριοποιημένη συλλογή κειμένων (corpus).

3.5. Κατηγοριοποίηση Πληροφορίας

Η κατηγοριοποίηση της πληροφορίας είναι ένα θέμα που απασχολεί ολοένα και περισσότερο τα τελευταία χρόνια την ακαδημαϊκή κοινότητα. Βασική αρχή στην κατηγοριοποίηση αποτελούν τα μοντέλα μάθησης. Είναι ουσιαστικά η βάση ενός μηχανισμού κατηγοριοποίησης. Ένας μηχανισμός κατηγοριοποίησης υλοποιεί μία διαδικασία μέσω της οποίας προβάλλεται το διάνυσμα του κειμένου εισόδου στο χώρο και μέσω συγκρίσεων προκύπτει η κλάση στη οποία πιθανώς ανήκει το κείμενο εισόδου. Στην περίπτωση της κατηγοριοποίησης κειμένου τα χαρακτηριστικά είναι λέξεις του κειμένου και οι κλάσεις είναι κατηγορίες κειμένου (π.χ. πολιτικά, αθλητικά, πολιτισμός κλπ). Συχνά, οι μηχανισμοί κατηγοριοποίησης είναι πιθανοτικοί όσον αφορά τη διαδικασία με την οποία κατηγοριοποιούν, η οποία είναι πιθανοτική κατανομή.

Ο κυρίαρχος στόχος της κατηγοριοποίησης πληροφορίας είναι να πραγματοποιήσει διαδικασία μάθησης στους μηχανισμούς κατηγοριοποίησης χρησιμοποιώντας επαγωγικές διαδικασίες. Προκειμένου να γίνει αντιληπτό αυτό θα αναλύσουμε στην πορεία μια σειρά από διαφορετικούς αλγορίθμους κατηγοριοποίησης. Όλοι οι αλγόριθμοι απαιτούν μόνο ένα μικρό σύνολο από «πληροφορία εκπαίδευσης» σαν είσοδο. Η «πληροφορία εκπαίδευσης» χρησιμοποιείται για να αρχικοποιήσει τις παραμέτρους του μοντέλου κατηγοριοποίησης. Στη διαδικασία δοκιμών και αποτίμησης, μπορούμε να προσδιορίσουμε την αποδοτικότητα κάθε αλγορίθμου.

Ένα κοινό χαρακτηριστικό στις διαφορετικές εκδόσεις των αλγορίθμων είναι η αναπαράσταση των κειμένων με ένα διάνυσμα από λέξεις, το οποίο είναι δημοφιλέστατο και στα συστήματα IR. Οι τιμές της συχνότητας των λέξεων και η ανάστροφη συχνότητα κειμένων υπολογίζονται και ανάλογα με την τεχνική εκμάθησης που χρησιμοποιείται, μερικά ή όλα από τα στοιχεία εισόδου εισάγονται στην πληροφορία εκπαίδευσης.

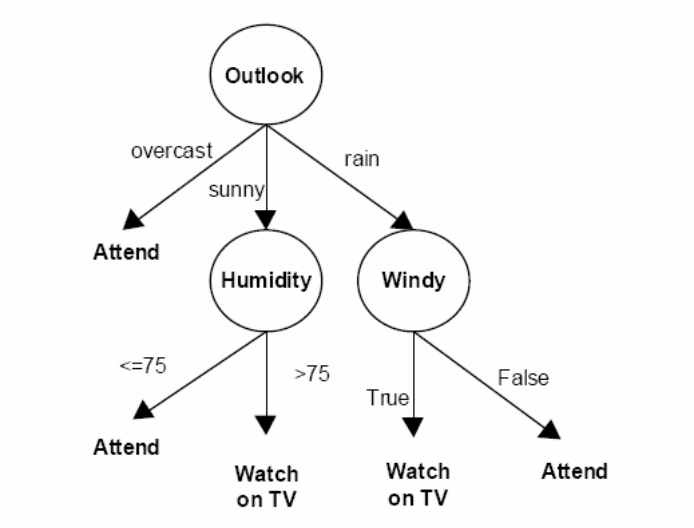
3.5.1. Αλγόριθμοι για κατηγοριοποίηση πληροφορίας

Υπάρχει πληθώρα αλγορίθμων που πραγματοποιούν αυτόματη κατηγοριοποίηση κειμένων βασισμένοι στο περιεχόμενο του κειμένου. Παρακάτω παρουσιάζονται οι πιο σημαντικοί από αυτούς.

3.5.1.1. Δέντρα απόφασης (Decision Trees)

Σε αυτό τον αλγόριθμο αρχικά έχουμε μια σειρά εγγραφών. Κάθε εγγραφή έχει την ίδια δομή, ένα ζευγάρι με χαρακτηριστικό/τιμή. Ένα από αυτά τα ζευγάρια αντιπροσωπεύει την κατηγορία της εγγραφής. Ο στόχος είναι να προσδιοριστεί ένα δέντρο το οποίο με βάση απαντήσεις σε ερωτήματα σε ότι αφορά χαρακτηριστικά που δεν αφορούν κάποια κατηγορία να προβλεφθεί η κατηγορία στην οποία θα ενταχθεί το χαρακτηριστικό. Ένας μηχανισμός προσδιορισμού κατηγορίας μέσω δέντρου δημιουργείται για κάθε ξεχωριστή κατηγορία χρησιμοποιώντας την προσέγγιση του Quinlan [8]. Αλγόριθμοι όπως ο ID3, C4.5 ή ο C5 είναι απλώς παραδείγματα που προκύπτουν από πρότυπα δέντρα απόφασης. Συνήθως τα χαρακτηριστικά κατηγοριών έχουν δυαδικές τιμές (0 ή 1).

Στο παρακάτω γράφημα βλέπουμε ένα παράδειγμα δέντρου απόφασης που δύναται να αποφασίσει αν κάποιος πρέπει να πάει να δει έναν ποδοσφαιρικό αγώνα ή να τον παρακολουθήσει από την τηλεόρασή του, βασισμένο στις καιρικές συνθήκες.



Εικόνα 5: Δέντρο Απόφασης

Εκτός από δυαδικές τιμές για την κατηγοριοποίηση, μπορεί να χρησιμοποιηθεί μέθοδος που χρησιμοποιεί κλάση πιθανοτήτων όπου η έξοδος είναι η πιθανότητα να ανήκει ένα αντικείμενο σε μια συγκεκριμένη κατηγορία. Ένα πιο αναλυτικό άρθρο για τη συγκεκριμένη τεχνική μπορεί να βρεθεί στο [9].

3.5.1.2. Naïve Bayes

Ένας μηχανισμός κατηγοριοποίησης βασισμένος στην τεχνική Naïve Bayes δημιουργείται χρησιμοποιώντας πληροφορία εκπαίδευσης για να ευρεθεί η πιθανότητα κάθε κατηγορίας δεδομένου ενός κειμένου προς κατηγοριοποίηση. Το θεώρημα του Bayes μπορεί να χρησιμοποιηθεί για να υπολογιστεί η πιθανότητα:

$$P(C = c_k \mid \vec{x}) = \frac{P(\vec{x} \mid C = c_k)P(C = c_k)}{P(\vec{x})} \quad (1)$$

Ο πρώτος όρος του αριθμητή είναι συνήθως δύσκολο να υπολογιστεί χωρίς να απλοποιηθεί η παράσταση. Για το συγκεκριμένο μηχανισμό κατηγοριοποίησης, υποθέτουμε πως τα χαρακτηριστικά X1 … Xn είναι ανεξάρτητα υπό όρους, δεδομένης μίας μεταβλητής κατηγορίας C. Αυτή η υπόθεση απλοποιεί την παραπάνω παράσταση στην:

$$P(\vec{x} \mid C = c_k) = \prod_i P(x_i \mid C = C_k) \quad (2)$$

Παρά το γεγονός ότι η θεώρηση της ανεξαρτησίας είναι γενικά αναληθής όσον αφορά την εμφάνιση κειμένων μέσα σε ένα έγγραφο, ο παραπάνω αλγόριθμος είναι αποτελεσματικός.

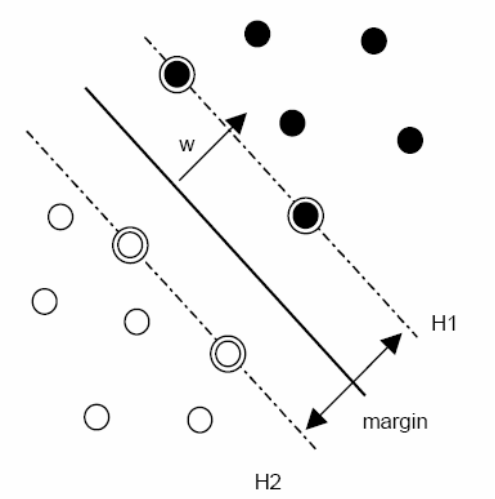
3.5.1.3. k-Nearest Neighbor (κοντινότερος γείτονας)

Ο αλγόριθμος kNN, είναι μία ακόμα στατιστική προσέγγιση στην αναγνώριση μοτίβου και κατηγοριοποίηση πληροφορίας [10]. Ο συγκεκριμένος αλγόριθμος, για ένα δοκιμαστικό κείμενο βρίσκει του κ κοντινότερους γείτονες ανάμεσα στα κείμενα εκπαίδευσης με την προσέγγιση να υπολογίζεται σαν μια ομοιότητα, και χρησιμοποιεί τις κατηγορίες των κ αυτών γειτόνων για να υπολογίσει τα βάρη με τα οποία θα συμμετέχει το κείμενο στην προσπάθεια ένταξης σε μία κατηγορία. Το αποτέλεσμα που εξάγεται υπολογίζοντας όλα τα βάρη, δίνει ένα αποτέλεσμα για την κατηγοριοποίηση του κειμένου.

3.5.1.4. Support Vector Machine

Το SVM είναι μια καινούρια μέθοδος κατηγοριοποίησης η οποία προτάθηκε από τον Vapnik [11; 12], και έχει ήδη αποκτήσει μεγάλη δημοσιότητα.

Στην πιο απλή του μορφή, ένα SVM ορίζεται σαν έναν υπερεπίπεδο που δύναται να διαχωρίσει ένα σύνολο θετικών από ένα σύνολο αρνητικών στοιχεία που αφορούν μια συγκεκριμένη κατηγορία. Αυτό φαίνεται και στο παρακάτω σχήμα όπου υποθέτοντας ότι οι μαύρες κουκίδες αφορούν τα θετικά στοιχεία και οι άσπρες τα αρνητικά στοιχεία ορίζεται με τη βοήθεια του SVM ένα μέγιστο υπερεπίπεδο που αποτελεί το διαχωριστικό ανάμεσα στα στοιχεία.



Εικόνα 6: Γραμμικά χωρισμένα υπερεπίπεδα

Στη γραμμική μορφή του αλγορίθμου, το περιθώριο μεταξύ των στοιχείων μπορεί να οριστεί σαν η απόσταση του υπερεπιπέδου από τα κοντινότερα θετικά και αρνητικά στοιχεία. Η μεγιστοποίηση αυτού του περιθωρίου μπορεί να αποτελέσει ένα πρόβλημα βελτιστοποίησης. Φυσικά τα περισσότερα παραδείγματα δε μπορούν να διαχωριστούν με τη χρήση της γραμμικής μορφής του αλγορίθμου γι' αυτό χρησιμοποιούνται πίνακες προκειμένου να υπολογιστούν τα περιθώρια και οι αποστάσεις.

Οι αλγόριθμοι για SVM έχουν αποδειχθεί ότι έχουν καλή γενικά απόδοση ακόμα και σε δύσκολα προβλήματα κατηγοριοποίησης μερικά από τα οποία είναι η αναγνώριση γραφικού χαρακτήρα, η αναγνώριση προσώπου, η κατηγοριοποίηση κειμένων. Η απλή γραμμική μορφή έχει πολύ καλή απόδοση, υφίσταται γρήγορη εκμάθηση και παράλληλα μπορεί να κατηγοριοποιεί εξαιρετικά γρήγορα.

Περισσότερα στοιχεία για το SVM μπορούν να βρεθούν στο [13].

3.6. Αξιοποίηση Πληροφορίας

Η πληροφορία που ανακτάται τόσο από το μηχανισμό εξόρυξης όσο και από το μηχανισμό κατηγοριοποίησης είναι υπέρογκη. Αρκεί να φανταστεί κάποιος ότι από 100 τυχαίες ηλεκτρονικές διευθύνσεις εξάγονται 90-95 κείμενα, από τα οποία λαμβάνουμε 2000 διακριτές λέξεις (πριν τη διαδικασία του stemming) και από τις οποίες προκύπτουν 8000-10000 συσχετίσεις κείμενο-λέξη-βάρος. Για το λόγο αυτό θα πρέπει να υπάρχει ένας ισχυρός μηχανισμός που να είναι σε θέση να αξιοποιήσει τη συγκεκριμένη πληροφορία και να μπορεί να βελτιώσει τους τρόπους που γίνονται ερωτήματα στη βάση και προσθήκες νέων εγγραφών.

Αυτό που θα πρέπει να μας απασχολήσει περισσότερο για το συγκεκριμένο σύστημα είναι να δημιουργηθεί ένας μηχανισμός διαχείρισης της πληροφορίας. Η πληροφορία δε θα πρέπει να είναι στάσιμη. Συνεχώς θα ανανεώνεται, και θα

πρέπει ανελλιπώς να διαγράφονται ή να τροποποιούνται τα στοιχεία τα οποία δε συγκεντρώνουν το ενδιαφέρον των χρηστών του συστήματος.

Προκειμένου να αξιοποιηθεί η πληροφορία θα πρέπει να δημιουργηθούν περιβάλλοντα διαχείρισης και μηχανισμοί ανάλυσης των ερωτημάτων και εύρεσης απάντησης. Παράλληλα θα πρέπει να υπάρχει τρόπος με τον οποίο να είναι εφικτή η ανάλυσης πληροφορίας από τις κινήσεις του χρήστη. Ωστόσο αυτό είναι ένα θέμα που θα καλυφθεί στην επόμενη ενότητα.

Τα συστήματα διαχείρισης πληροφορίας και ανάλυσης ερωτημάτων χρήστη, θα βασίζονται σε web interface προκειμένου να είναι άμεση η διασύνδεση με τον «πραγματικό» κόσμο. Είναι ουσιαστικά μία προσπάθεια να προσεγγίσουμε περισσότερα πραγματικά δεδομένα ξεφεύγοντας από την πληροφορία εκπαίδευσης. Επίσης, με αυτό τον τρόπο θα κάνουμε το μηχανισμό μας πιο διάφανο προς το χρήστη καθώς και πιο φιλικό.

Τα εργαλεία διαχείρισης δε θα περιέχουν πολύπλοκες συναρτήσεις, μα ούτε και πολύπλοκο περιβάλλον. Ο όγκος της πληροφορίας κάνει απαγορευτική την άμεση προσέγγισή της, συνεπώς ο διαχειριστής του συστήματος θα πρέπει να είναι σε θέση να έχει μια γενική εποπτεία του συστήματος διατηρώντας παράλληλα ανεκτά τα επίπεδα πρόσβασης σε εξειδικευμένα στοιχεία του συστήματος.

3.7. Προφίλ Χρήστη σε Δυναμικά Περιβάλλοντα

Ένα πολύ σημαντικό στοιχείο της εργασία είναι το προφίλ χρήστη σε δυναμικό περιβάλλον. Είναι το στοιχείο που χαρακτηρίζει την πύλη ποιοτικού περιεχομένου και είναι ένα από τα βασικά στοιχεία που δίνουν νόημα στη λέξη ποιότητα της πύλης.

Το δυναμικό περιβάλλον της πύλης θα δίνει τη δυνατότητα πρόσβασης σε πληροφορία η οποία ενδιαφέρει το χρήστη, καταργώντας τα περιθώρια εμφάνισης ανεπιθύμητων αποτελεσμάτων. Προκειμένου να γίνει κατανοητό θα πρέπει να προσδιοριστεί ο όρος προφίλ χρήστη.

Στο άκουσμα του όρου προφίλ χρήστη θα περίμενε κανείς να έρθει αντιμέτωπος με προσωπικά στοιχεία του χρήστη [όνομα, επώνυμο κλπ.]. Όσο κι αν ακούγεται παράξενο, σε ένα δυναμικό περιβάλλον ίσως δεν έχει και τόσο μεγάλη σημασία ο προσδιορισμός του χρήστη σαν φυσικό πρόσωπο αλλά περισσότερο σαν χρήστης του διαδικτύου. Βασικός στόχος της δημιουργίας του προφίλ ενός χρήστη είναι να προσδιοριστεί με όσο μεγαλύτερη ακρίβεια η δράση του φυσικού προσώπου όταν έρχεται αντιμέτωπος με το διαδίκτυο. Είναι μεγάλο επίτευγμα να μπορεί κανείς να προσδιορίσει την επόμενη κίνηση που θα πραγματοποιήσει ο χρήστης [πχ ποιο σύνδεσμο θα ακολουθήσει στην επόμενη κίνηση]. Ακούγεται σαν παιχνίδι πρόβλεψης και ίσως θα μπορούσε να παρομοιαστεί με κάτι τέτοιο. Ωστόσο είναι κάτι πιο σύνθετο και βασίζεται σε μία πληθώρα στοιχείων. Τι ερωτήματα πραγματοποιεί ο χρήστης, ποιες σελίδες επισκέπτεται πιο συχνά από τα αποτελέσματα που του εμφανίζονται, τι έχει δηλώσει σαν «αγαπημένες κατηγορίες» αποτελούν μερικά από τα βασικά στοιχεία πάνω στα οποία βασίζεται η δημιουργία του προφίλ ενός χρήστη.

Στο συγκεκριμένο σύστημα, το ενδιαφέρον μας επικεντρώνεται στην αξιολόγηση που κάνει ο χρήστης όταν του παρουσιάζονται τα αποτελέσματα της αναζήτησής του. Ένα παράδειγμα θα ήταν αρκετό για να κατανοήσει κανείς το νόημα που έχει το «δυναμικό προφίλ» στη συγκεκριμένη δικτυακή πύλη. Έστω ένας χρήστης του διαδικτύου που χρησιμοποιεί τη συγκεκριμένη δικτυακή πύλη και επιθυμεί να βλέπει καθημερινά τα περιεχόμενα της κατηγορίας business. Το προφίλ έχει ήδη δημιουργηθεί και περιλαμβάνει την πολύ γενική κατηγορία business. Όταν παρουσιάζονται στο χρήστη αποτελέσματα (τίτλος άρθρου, μικρό απόσπασμα άρθρου), τότε ο χρήστης επιλέγει κάποιο ή κάποια αποτελέσματα για να τα εξετάσει περαιτέρω. Το κάθε κείμενο όμως αποτελείται, συν τοις άλλοις, και από κάποιες λέξεις-κλειδιά. Μόλις κάποιος χρήστης επιλέξει κάποιο κείμενο, οι

λέξεις-κλειδιά που υπάρχουν στο συγκεκριμένο, αυτομάτως αποκτούν αξία για το συγκεκριμένο χρήστη και εισάγονται αυτόματα στο προφίλ του. Αυτή η πληροφορία είναι πολύ σημαντική προκειμένου το σύστημα να είναι σε θέση να κάνει μεγαλύτερη αξιολόγηση των κειμένων που θα παρουσιάσει στο χρήστη. Έτσι, την επόμενη φορά που ο χρήστης θα δει τα αποτελέσματα για την κατηγορία που επιθυμεί τα κείμενα θα είναι ταξινομημένα (και) βάσει των λέξεων-κλειδιών που έχουν τη μεγαλύτερη βαθμολογία για κάθε χρήστη. Με αυτό τον τρόπο αποκτά μεγαλύτερη αξία το κείμενο που περιέχει πολλές λέξεις-κλειδιά για ένα συγκεκριμένο χρήστη. Η συγκέντρωση των αποτελεσμάτων συνολικά για τους χρήστες μίας κατηγορίας μπορεί να οδηγήσει σε μεγαλύτερη διαβάθμιση κάθε κατηγορίας και δημιουργία εικονικών υποκατηγοριών που θα είναι χωρισμένες βάση της απόκρισης των χρηστών. Θεωρητικά ένα τέτοιο μοντέλο, εικονικής ουσιαστικά, κατηγοριοποίησης είναι πιο αποτελεσματικό από κάθε αλγοριθμικό μοντέλο καθώς η κατηγοριοποίηση δε γίνεται από τη μηχανή αλλά από τον άνθρωπο.

4

ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

Στο κεφάλαιο αυτό περιγράφονται σχετικές εργασίες με κάθε υποσύστημα της συγκεκριμένης εργασίας. Οι σχετικές εργασίες περιλαμβάνουν:

- Συλλογή Δεδομένων
- Φιλτράρισμα Δεδομένων
- Προεπεξεργασία πληροφορίας
- Κατηγοριοποίηση πληροφορίας
- Αυτόματη εξαγωγή περίληψης
- Προσωποποίηση στο χρήστη

4. ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

Η αναζήτηση για σχετικές εργασίες μας φέρνει αντιμέτωπους με μία σειρά από συστήματα που έχουν αναπτυχθεί προκειμένου να διευκολύνουν τους χρήστες κατά την προσπάθεια αναζήτησης για πληροφορίες που αφορούν άρθρα. Τα συστήματα αυτά έχουν, το καθένα, ένα διαφορετικό τρόπο προσέγγισης του θέματος. Σε άλλα σημεία συγκλίνουν και σε άλλα αποκλίνουν ενώ η «γενική ιδέα» εντοπίζεται κυρίως στους μηχανισμούς κατηγοριοποίησης και προσωποποίησης στο χρήστη.

Ελπιδοφόρα είναι τα μηνύματα που έρχονται από το δικτυακό τόπο Google [119] όπου μία νέα υπηρεσία ειδήσεων έχει ήδη αρχίσει να προβάλλεται από τη Google και να χρησιμοποιείται από ολοένα και περισσότερους χρήστες. Σε αυτή την περίπτωση οι χρήστες βλέπουν σε βασικές κατηγορίες τα νέα και έχουν τη δυνατότητα προσωποποίησης των νέων που τους παρουσιάζονται. Σημαντικό είναι το γεγονός πως η υπηρεσία παρέχεται στη γλώσσα που επιθυμεί ο χρήστης με αποτέλεσμα να είναι σε θέση να αναγνώσει όλες τις τελευταίες ειδήσεις που έχουν συλλεχθεί από μεγάλα ειδησεογραφικά πρακτορεία.

4.1. Συλλογή δεδομένων

Για τη συλλογή δεδομένων από το διαδίκτυο χρησιμοποιούνται οι ευρέως γνωστοί crawlers. Το πλήθος τους είναι αμέτρητο ενώ, αν εξαιρέσουμε τους εξειδικευμένους crawlers (focused crawlers) παρατηρούμε πως οι περισσότεροι έχουν σαν σκοπό να συλλέξουν όλες τις HTML σελίδες από τις οποίες απαρτίζεται ένας δικτυακός τόπος μαζί με τα βοηθητικά αρχεία (pdf, εικόνες, video, css, javascript) και ουσιαστικά να δημιουργήσουν ένα offline-instance του δικτυακού τόπου τον οποίο προσπελαύνουν.

Οι crawlers που έχουν κατασκευαστεί για το διαδίκτυο αγγίζουν σε αριθμό τις μερικές χιλιάδες καθώς η κατασκευή τους είναι σχεδόν τετριμμένη. Στη συνέχεια θα παρουσιάσουμε συγκεκριμένου crawlers που αξίζουν προσοχής για τα ιδιαίτερα χαρακτηριστικά που παρουσιάζουν.

4.1.1. WebCrawler

Πρόκειται για έναν από τους πρώτους crawlers που κατασκευάστηκαν από τον Pinkerton το 1994 [37]. Βασίστηκε στη βιβλιοθήκη WWW προκειμένου να είναι σε θέση να κατεβάζει σελίδες από το διαδίκτυο ενώ χρησιμοποιούσε ένα δεύτερο πρόγραμμα προκειμένου να διαβάζει τα URL τα οποία πρέπει να προσπελάσει. Ο αλγόριθμος προσπέλασης ήταν κατά πλάτος αναζήτηση του γραφήματος μίας ιστοσελίδας σε συνδυασμό με αποφυγή των σελίδων που έχει ήδη επισκεφθεί. Ένα αξιοσημείωτο στοιχεία ήταν η δυνατότητα να ακολουθεί συγκεκριμένα μόνο links σε ένα δικτυακό τόπο – και όχι όλα – βάση του ερωτήματος που έθετε ο χρήστης. Ήταν κάτι σαν ένας crawler πραγματικού χρόνου που φυσικά μπορούσε να ανταποκριθεί πλήρως λόγω του μικρού μεγέθους που είχε το διαδίκτυο.

4.1.2. Google Crawler

Ένας από τους πιο σημαντικούς crawlers που κατασκευάστηκαν και διατηρούνται ακόμα και σήμερα, με σημαντικές βέβαια βελτιώσεις είναι ο Google Crawler των Brin και Page, 1998 [37]. Βασίζεται στις γλώσσες προγραμματισμού C++ και Python και παρουσιάζει εξαιρετικά μεγάλη πολυπλοκότητα. Επειδή η χρήση των σελίδων που κατέβαζε ο crawler προοριζόταν για εκτενή αναζήτηση μέσα σε σειρές από κείμενα, ο συγκεκριμένος crawler βασίστηκε στη διαδικασία indexing. Στο μηχανισμό υπάρχει ένας URL εξυπηρετητής που αποστέλλει λίστες

με URL προς τους crawlers του συστήματος οι οποίοι λειτουργούν παράλληλα. Οι crawlers εξάγουν από τις σελίδες το κείμενο αλλά και όσα URLs εντοπίζουν. Αυτά στέλνονται πίσω στον URL εξυπηρετητή για έλεγχο και σε περίπτωση που δεν τα έχει επισκεφθεί ποτέ ο crawler προστίθενται στη λίστα του εξυπηρετητή.

4.1.3. Mercator

Ο Mercator [39], [41] είναι ένας κατανεμημένος τμηματοποιημένος web crawler γραμμένος εξ' ολοκλήρου σε γλώσσα προγραμματισμού Java. Η τμηματοποίηση του προκύπτει από τη χρήση δύο διαφορετικών πρωτοκόλλων.

- Protocol modules
  - Τα τμήματα πρωτοκόλλων είναι υπεύθυνα για την ομαλή σύνδεση του μηχανισμού στις σελίδες και για την εξασφάλιση πως ο μηχανισμός θα είναι σε θέση να «κατεβάσει» τη σελίδα.
- Processing modules
  - Από την άλλη μεριά τα τμήματα επεξεργασίας είναι αυτά που αφορούν την ανάλυση της σελίδας και την εξαγωγή του κειμένου και συνδέσμων από αυτή. Η απλή διαδικασία επεξεργασίας περιλαμβάνει ανάλυση της σελίδας και εξαγωγή των συνδέσμων που αυτή περιέχει ενώ σε μία πιο σύνθετη μορφή της περιλαμβάνει αλγορίθμους για την αποτελεσματική εξαγωγή του κειμένου.

4.1.4. WebFountain

Πρόκειται για έναν κατανεμημένο τμηματικό crawler παραπλήσιο του mercator, με τη διαφορά ότι είναι γραμμένος σε C++ [40], [41]. Περιλαμβάνει έναν κεντρικό μηχανισμό και μία σειρά από "ant" (μερμύγκι) μηχανισμούς. Πρόκειται δηλαδή για το ρυθμιστή της κατάστασης και τους εργάτες. Ο μηχανισμός αυτός περιέχει στοιχεία που τον κάνουν πολύ φιλικό προς τις σελίδες που επισκέπτεται. Σκοπός του είναι η διατήρηση ενός off-line instance του διαδικτύου. Αυτό έχει σαν αποτέλεσμα, μία από τις μετρικές τις οποίες προσμετρά ο συγκεκριμένος μηχανισμός να είναι το κατά πόσο η σελίδες που διαθέτει ανταποκρίνονται στις πραγματικές σελίδες που βρίσκονται on-line στους δικτυακούς τόπος και όχι απλά μία παλαιότερη έκφανσή τους. Για να πετύχει μεγαλύτερο freshness όπως ονομάζεται η συγκεκριμένη μετρική χρησιμοποιεί διαφορετική συχνότητα επίσκεψης στις σελίδες που έχει αποθηκευμένες στη βάση δεδομένων του.

4.1.5. WebRACE

Πρόκειται για έναν crawler ο οποίος είναι γραμμένος σε Java και αποτελεί ένα κομμάτι ενός γενικότερου συστήματος που ονομάζεται eRACE [42]. Το συγκεκριμένο σύστημα λαμβάνει εντολές από τους τελικούς χρήστες για να ξεκινήσει να κατεβάσει σελίδες και συμπεριφέρεται σαν proxy server. Το σύστημα μπορεί να εξυπηρετήσει και αιτήσεις για αλλαγές στοιχείων σε σελίδες: μόλις μία σελίδα αλλάξει, τότε ο crawler την ξανακατεβάζει και ειδοποιεί τον τελικό χρήστη που ενδιαφέρεται πως η σελίδα έχει αλλάξει και πως πλέον στον proxy είναι αποθηκευμένη μία νέα σελίδα. Το πιο σημαντικό στοιχείο του συγκεκριμένου crawler είναι η χαρακτηριστική διαφορά που παρουσιάζει συγκριτικά με όσους crawlers έχουμε δει. Στο συγκεκριμένο crawler δεν υπάρχει ένα feed URL από το οποίο θα ξεκινήσει να αναζητά σελίδες. Το URL feed είναι δυναμικό και διαμορφώνεται από τα ερωτήματα των χρηστών. Μετά τη χρήση του καταστρέφεται και ο μηχανισμός βρίσκεται σε αναμονή μέχρι να του δοθεί κάποιο νεότερο ερώτημα.

4.1.6. Ubicrawler

Ο Ubicrawler είναι ένας κατανεμημένος crawler γραμμένος σε Java και δε διαθέτει κεντρικοποιημένη διαδικασία [43]. Είναι κατασκευασμένος από έναν αριθμό από όμοιους "agents" και μία συνάρτηση ανάθεση που αναθέτει σε κάθε agent κάποια εργασία. Οι agents δεν επικοινωνούν μεταξύ τους άμεσα αλλά όλες οι διαδικασίες διευθετούνται από την κεντρική συνάρτηση ανάθεσης. Καμία σελίδα δεν προσπελαύνεται διπλή φορά καθώς κάθε agent φροντίζει να ενημερώσει για τις σελίδες που έχει επισκευθεί εκτός και αν κάποιος από τους agents καταστραφεί. Πρόκειται για έναν πολύ σταθερό crawler, σχεδιασμένο με τέτοιον τρόπο ώστε να πετυχαίνει μέγιστη κλιμάκωση και μικρή ευαισθησία σε σφάλματα.

4.1.7. Crawlers Ανοιχτού Κώδικα

Μία σειρά από crawlers ανοιχτού κώδικα διανέμονται ελεύθερα στο διαδίκτυο. Κυρίως είναι προϊόντα κάποιου ιδιώτη που κατασκευάζονται για να καλύψουν συγκεκριμένες ανάγκες που έχουν οι τελικοί χρήστες, ανάγκες που συχνά δεν καλύπτονται από τους εμπορικούς crawlers. Η χρήση τους είναι συνήθως ως εξής. Κάποιος χρήστης που δεν καλύπτεται από έναν εμπορικό crawler λαμβάνει τον κώδικα ενός open source συστήματος και το αλλάζει με σκοπό να το φέρει στα μέτρα του. Συνήθως οι open source crawlers δεν έχουν εξειδικευμένες λειτουργικότητες ωστόσο προσφέρονται στους τελικούς χρήστες οι οποίοι μπορούν να τους τροποποιήσουν ελεύθερα.

Μερικά παραδείγματα από crawlers ανοιχτού κώδικα ακολουθούν

- GNU Wget [133]
- Heritrix [134]
- ht://Dig [135]
- HTTrack [136]
- Larbin [137]
- Methabot [138]
- Nutch [139]
- WebSPHINX (Miller and Bharat, 1998) [140]
- WIRE - Web Information Retrieval Environment (Baeza-Yates and Castillo, 2002) [141]

4.2. Φιλτράρισμα δεδομένων – Εξαγωγή κειμένου από HTML σελίδες

Η διαδικασία της εξαγωγής κειμένου για τον σκοπό για τον οποίο χρησιμοποιείται στη συγκεκριμένη εργασία ξεφεύγει από το σκοπό που έχουν οι ελάχιστες εμπορικές εφαρμογές. Έτσι η εξαγωγή χρήσιμου κειμένου από HTML σελίδες αποτελεί αντικείμενο έρευνας ενώ η εξαγωγή όλου του κειμένου μίας HTML σελίδας αποτελεί μία τετριμμένη διαδικασία

Η εξαγωγή κειμένου από HTML σελίδες είναι μία απλοϊκή διαδικασία η οποία βασίζεται στην αφαίρεση των HMTL tags και στη διατήρηση του υπόλοιπου κειμένου μέσα από μία HTML σελίδα. Στην περίπτωσή μας όμως, αυτός ο μηχανισμός δεν είναι αρκετός. Το σύστημά μας θα πρέπει να υλοποιεί έναν έξυπνο αλγόριθμο ο οποίος θα είναι σε θέση να ξεχωρίσει το επιθυμητό κείμενο από κείμενο που μπορεί να αφορά το navigation menu ή κάποιες διαφημίσεις. Με απλά λόγια, ο μηχανισμός μας θα πρέπει να είναι φτιαγμένος με τέτοιον τρόπο ώστε να ανακτάται μόνον ο τίτλος και το κείμενο του άρθρου που αφορά κάποια είδηση. Κάθε άλλο κείμενο στη σελίδα είναι μη επιθυμητό και άρα ο μηχανισμός θα πρέπει να το απορρίπτει.

Τέτοιοι μηχανισμοί κατασκευάζονται σε πειραματικό επίπεδο και κυρίως για ερευνητικούς σκοπούς. Απλοϊκά προγράμματα που να μπορούν να απομονώσουν

κομμάτι μίας HTML σελίδας και να ανακτήσουν την πληροφορία που βρίσκεται σε ένα συγκεκριμένο κομμάτι υπάρχουν, αλλά θα πρέπει να προσαρμοστούν σε κάθε διαφορετική ιστοσελίδα. Δεν είναι εφικτό να υπάρχει ένα γενικό σύστημα το οποίο να έχει τη δυνατότητα να αναλύσει τα σημεία που εντοπίζεται χρήσιμο κείμενο. Για το λόγο αυτό στηριζόμαστε στη θεωρία του web clipping σύμφωνα με την οποία είναι εφικτός ο διαχωρισμός περιοχών σε μία σελίδα και μάλιστα είναι εφικτό να δημιουργηθεί αλγόριθμος ο οποίος να εξάγει αυτόματα το χρήσιμο κείμενο από μία HTML σελίδα. Σε γενικές γραμμές οι μηχανισμοί αυτοί βασίζονται στο γεγονός πως η HTML σελίδα μπορεί να αναλυθεί σε δενδρική μορφή. Τα φύλλα του δένδρου αναπαριστούν το κείμενο που υπάρχει στη σελίδα με αποτέλεσμα να είναι εφικτό να εντοπιστούν άμεσα τα σημεία μέσα στο κείμενο που περιέχουν κείμενο. Σε επόμενη φάση θα πρέπει να βρεθούν τα φύλλα τα οποία περιέχουν χρήσιμο κείμενο. Στην πιο απλή περίπτωση υπολογίζεται ο λόγος bytes κειμένου / bytes κώδικα + bytes κειμένου για κάθε κόμβο που έχει φύλλα. Με αυτό τον τρόπο επιτυγχάνεται το αυτονόητο. Σημεία που έχουν πολύ περισσότερο κείμενο απ’ ότι κώδικα προφανώς και έχουν χρήσιμο κείμενο. Θέτοντας ένα αυστηρό όριο για το συγκεκριμένο λόγο έχουμε σαν αποτέλεσμα το να εντοπίσουμε τις θέσεις που έχουν αποκλειστικά και μόνο κείμενο. Ο αλγόριθμος που περιγράφηκε είναι απλός και αποτελεσματικός και συχνά χρησιμοποιείται ατόφιος σε όλα τα συστήματα εξαγωγής χρήσιμου κειμένου.

4.3. Προεπεξεργασία δεδομένων

Στη θεωρία, τα βασισμένα σε κείμενο χαρακτηριστικά ενός εγγράφου μπορούν να περιλαμβάνουν κάθε λέξη / φράση η οποία μπορεί να εμφανίζεται σε ένα δεδομένο σύνολο κειμένων. Όμως, επειδή κάτι τέτοιο είναι υπολογιστικά μη-ρεαλιστικό, χρειαζόμαστε κάποια μέθοδο προεπεξεργασίας κειμένων για την αναγνώριση των λέξεων - κλειδιών (κωδικολέξεων ή αλλιώς keywords) και φράσεων οι οποίες μπορεί να μας είναι χρήσιμες. Διάφορες τεχνικές έχουν προταθεί για την αναγνώριση των keywords ενός κειμένου όπως τα Hidden Markov Models [66], η Naive Bayes [68] και τα Support Vector Machines [67] όμως όλες αυτές οι μέθοδοι τείνουν να κάνουν χρήση συγκεκριμένης γνώσης μετα-πληροφορίας για τη γλώσσα του κειμένου. Άλλες μέθοδοι χρησιμοποιούν στατιστικές πληροφορίες, όπως η συχνότητα μιας λέξης. Μια ευρέως γνωστή τεχνική είναι η TF-IDF (Term Frequency - Inverse Document Frequency), όπου TF είναι το πλήθος των εμφανίσεων ενός όρου σε ένα δεδομένο σύνολο κείμενων συγκρινόμενο με το πλήθος των κειμένων που περιέχουν το συγκεκριμένο όρο, και IDF είναι ένα μέτρο των συνολικών κειμένων σε μια συλλογή κειμένων, συγκρινόμενο με το συνολικό αριθμό κειμένων που περιέχουν μια δεδομένη λέξη [69]. Σχετικές τεχνικές, οι οποίες περιλαμβάνουν άλλες στατιστικές που πηγάζουν από το σύνολο των κειμένων, έχουν επίσης προταθεί τα πρόσφατα χρόνια· π. χ. κέρδος πληροφορίας [70], odds ratio [71], CORI [72], κλπ. Οι τεχνικές αυτές προσφέρουν μια βελτιωμένη προσέγγιση.

4.3.1. Ανάλυση

Στην ανάκτηση πληροφορίας, η σχέση μεταξύ ενός ερωτήματος χρήστη και ενός κειμένου καθορίζεται κυρίως από το πλήθος των όρων που έχουν κοινούς. Δυστυχώς, οι λέξεις έχουν πολλές μορφολογικές παραλλαγές οι οποίες δεν αναγνωρίζονται από αλγόριθμους που βασίζονται στο ταίριασμα όρων χωρίς να προηγηθεί κάποιας μορφής επεξεργασία φυσικής γλώσσας (Natural Language Processing). Στις περισσότερες των περιπτώσεων, αυτές οι παραλλαγές έχουν παρόμοιες εννοιολογικές ερμηνείες και μπορούν να αντιμετωπισθούν ως ισοδύναμες στα πλαίσια εφαρμογών ανάκτησης πληροφορίας (σε αντίθεση με τις γλωσσολογικές). Ως εκ' τούτου, ένα πλήθος αλγορίθμων κατάλληλων για τη διαδικασία του stemming έχουν αναπτυχθεί ώστε να περιορίσουν τις μορφολογικές παραλλαγές στην αρχική τους ρίζα.

Το πρόβλημα του stemming έχει προσεγγιστεί από μια μεγάλη ποικιλία μεθόδων που περιγράφονται στο [73] και περιλαμβάνουν αφαίρεση της κατάληξης, τμηματοποίηση λέξης και λεξιλογική μορφοποίηση. Δύο από τους διασημότερους αλγορίθμους, ο Lovins [74] και ο Porter [75], βασίζονται στην αφαίρεση της κατάληξης. Ο αλγόριθμος Lovins βρίσκει το μακρύτερο ταίριασμα από μια μεγάλη λίστα καταλήξεων, ενώ ο Porter [75] χρησιμοποιεί έναν επαναληπτικό αλγόριθμο με μικρότερο αριθμό καταλήξεων και μερικούς κανόνες. Ένας ακόμη αλγόριθμος, ο Paice/Husk [76], χρησιμοποιεί αποκλειστικά ένα σύνολο κανόνων ενώ ακολουθεί επαναληπτική προσέγγιση.

Στο [77] περιγράφονται τα προβλήματα που σχετίζονται με αυτές τις προσεγγίσεις. Οι περισσότεροι stemmers λειτουργούν χωρίς λεξικό και επομένως αγνοούν το νόημα των λέξεων, κάτι που οδηγεί σε ορισμένα λάθη κατά τη διαδικασία του stemming. Λέξεις διφορούμενες μειώνονται στην ίδια ρίζα και λέξεις με παρόμοιο νόημα δεν μειώνονται στην ίδια ρίζα. Για παράδειγμα, ο Porter stemmer μειώνει τις λέξεις general, generous, generation, generic στην ίδια ρίζα.

Παράλληλα, η έξοδος (stems) που παράγεται από τους αλγορίθμους, συνήθως δεν περιέχει πραγματικές λέξεις, κάτι που την κάνει δύσχρηστη για εργασίες που έχουν να κάνουν με ανάκτηση πληροφορίας. Διαδραστικές τεχνικές οι οποίες απαιτούν είσοδο από τον χρήστη απαιτούν από αυτόν την εργασία με stems και όχι πραγματικών λέξεων. Προβλήματα αυτού του τύπου αντιμετωπίζονται προσεγγίζοντας τη διαδικασία με μορφολογική ανάλυση.

Υπάρχει ένας μεγάλος αριθμός εργασιών που έχουν εξετάσει τον αντίκτυπο των stemming αλγορίθμων στην απόδοση της ανάκτησης πληροφορίας. Στο [78] δίνεται μια καλή περίληψη, αναφέροντας ότι τα συνδυασμένα αποτελέσματα των προηγούμενων μελετών καθιστούν ασαφές εάν η διαδικασία του stemming είναι χρήσιμη. Στις περιπτώσεις όπου το stemming είναι χρήσιμο τείνει να ασκήσει μόνο μικρή επίδραση στην απόδοση, και η επιλογή του stemmer μεταξύ των πιο κοινών παραλλαγών δεν είναι σημαντική. Εντούτοις, δεν υπάρχει κανένα στοιχείο ότι ένα λογικός stemmer μπορεί να βλάψει την απόδοση της ανάκτησης πληροφορίας.

Αντίθετα, μια πρόσφατη μελέτη [79] εντοπίζει μια αύξηση 15-35% στην απόδοση ανάκτησης όταν το stemming χρησιμοποιείται σε μερικές συλλογές (CACM και npl). Αναφέρεται ότι αυτές οι συλλογές έχουν και ερωτήματα και έγγραφα τα οποία είναι εξαιρετικά σύντομα. Για συλλογές με μεγαλύτερα κείμενα, οι stemming αλγόριθμοι χαρακτηρίζονται από μια σχετική αύξηση στην απόδοση της διαδικασίας ανάκτησης πληροφορίας.

4.4. Κατηγοριοποίηση πληροφορίας

Η αυτόματη κατηγοριοποίηση κειμένων είναι η διαδικασία ανάθεσης ετικετών κατηγορίας (προκαθορισμένων) σε νέα κείμενα που καταφθάνουν, στηριζόμενη στην πιθανότητα η οποία προτείνεται από τη βάση γνώσης που προϋπάρχει. Η διαδικασία έχει εγείρει ορισμένες προκλήσεις για τις στατιστικές μεθόδους που συνήθως χρησιμοποιούνται, και την αποτελεσματικότητά τους στην επίλυση πραγματικών προβλημάτων, τα οποία συχνά είναι πολλών διαστάσεων και έχουν μη σαφώς καθορισμένη κατανομή μεταξύ των κειμένων προς κατηγοριοποίηση. Η ανίχνευση του θέματος ενός κειμένου, για παράδειγμα, είναι η πιο κοινή εφαρμογή της κατηγοριοποίησης κειμένων. Ένας ολοένα και αυξανόμενος αριθμός μεθόδων αντιμετώπισης του προβλήματος προτείνονται, μεταξύ των οποίων μοντέλα παλινδρόμησης, κατηγοριοποίηση κοντινότερων γειτόνων [80], [81], πιθανοτικές προσεγγίσεις με μεθόδους Bayes [82], [83], επαγωγική εκμάθηση κανόνων [84], [85], νευρωνικά δίκτυα [86], on-line εκμάθηση [87] και Support Vector Machines [88]. Παρότι η πλούσια βιβλιογραφία που υπάρχει πάνω στον τομέα της

κατηγοριοποίησης κειμένων, ασφαλείς εκτιμήσεις και συγκρίσεις μεταξύ των μεθόδων είναι συνήθως δύσκολες.

Για να είναι δυνατή η παραγωγή μιας κατηγοριοποιημένης περίληψης, που θα ανταποκρίνεται στα ενδιαφέροντα του τελικού χρήστη, πρέπει να εντοπιστεί η κατηγορία του κειμένου. Λέξεις κλειδιά, οι οποίες είναι μοναδικές για κάποιο πεδίο (κατηγορία) αποτελούν πολύ καλές ενδείξεις για την κατηγορία του κειμένου [89]. Άλλες εναλλακτικές επιλογές, όπως συντακτικές και στατιστικές εκφράσεις έχουν επίσης χρησιμοποιηθεί [90], [91], [92] Το βασικό θέμα της αναγνώρισης του θέματος με χρήση NLP έχει αναλυθεί διεξοδικά στο [93].

Άλλες επαναστατικές τεχνικές, όπως η χρήση κωδικών ελέγχου [94], η χρήση αιτιολογικών δικτύων έχουν προταθεί και αποτελούν ουσιαστικά μια τροποποιημένη έκδοση του Bayes αλγορίθμου του [95] που αποδίδουν καλά σε εργασίες κατηγοριοποίησης κειμένων. Καμία από τις προηγούμενες τεχνικές δεν αντιμετωπίζει τα σημασιολογικά θέματα.

4.5. Αυτόματη εξαγωγή περίληψης

Παρουσιάζει ενδιαφέρον το γεγονός ότι πολλές διεργασίες ανάκτησης πληροφορίας, όπως η κατηγοριοποίηση κειμένου και η εξόρυξη πληροφορίας, μοιράζονται τους ίδιους στόχους και προβλήματα με την εξαγωγή περίληψης. Τα προβλήματα των συστημάτων ανάκτησης, λόγω του διλήμματος ακρίβειας - ανάκτησης, μπορούν να μειωθούν κάνοντας χρήση μιας αυτόματα εξαγόμενης περίληψης στοχευμένη στο προσωποποιημένο προφίλ (ενδιαφέροντα) του χρήστη.

Η έρευνα στον τομέα της αυτόματης περίληψης, θεωρούμενη ως εξαγωγή, αφαίρεση ή περίληψη χρήσιμου κειμένου, έχει μεγάλη ιστορία με αρχικό ``ξέσπασμα' τις προσπάθειες στη δεκαετία του 60 της πρωτοποριακής εργασίας του Luhn, ακολουθείται από τις δύο επόμενες δεκαετίες με σχετικά μικρή έρευνα στο θέμα, και κορυφώνεται τη δεκαετία του 90 και ως της μέρες μας με πολλές ερευνητικές προσπάθειες [97], [98], [99]. Σε κάθε περίπτωση, η δουλειά που έχει γίνει και που ουσιαστικά αφορά προτάσεις υλοποίησης κατατάσσονται σε δύο υποομάδες: εξαγωγή κειμένου και εξαγωγή γεγονότων. Στην εξαγωγή κειμένου, όπου «αυτό που βλέπεις είναι αυτό που παίρνεις», μερικά τμήματα που υπάρχουν στο αρχικό κείμενο μεταφέρονται αυτούσια στην περίληψη του. Η εξαγωγή κειμένου είναι μια «ανοιχτή» προσέγγιση στο πρόβλημα της περίληψης εφόσον δεν υπάρχει κάποια προηγούμενη υπόθεση για το τι είδους πληροφορία περιεχομένου είναι χρήσιμη. Το τι είναι σημαντικό για το πηγαίο κείμενο θεωρείται ως αξιοπρόσεκτο σε σχέση με κάποια γενικά, γλωσσολογικά, σημαντικά κριτήρια τα οποία εφαρμόζονται κατά τη διαδικασία εξαγωγής. Με την εξαγωγή γεγονότων αυτό που συμβαίνει είναι το αντίθετο: «αυτό που ξέρεις είναι αυτό που παίρνεις», δηλαδή αυτό που έχεις ήδη αποφασίσει πως είναι το θέμα του περιεχομένου που αναζητάς στο πηγαίο κείμενο, αυτό είναι που τελικά παίρνεις στην περίληψη του. Αυτή είναι μια «κλειστή» προσέγγιση, εννοώντας ότι το πηγαίο κείμενο δεν κάνει κάτι παραπάνω από το να παρέχει ένα στιγμιότυπο από κάποιες ήδη προκαθορισμένες απαιτήσεις. Η μέθοδος εξαγωγής κειμένου στοχεύει στο να κάνει το σημαντικό περιεχόμενο να «αναδυθεί» μόνο του από κάθε κείμενο. Αντίθετα η μέθοδος εξαγωγής γεγονότων στοχεύει να βρει εμφανή στοιχεία σημαντικών ιδεών (γνωμών), ανεξαρτήτως της κατάστασης του κειμένου.

Οι τεχνικές προεπεξεργασίας που χαρακτηρίζουν τις δύο προαναφερόμενες μεθόδους εξαγωγής είναι πολύ διαφορετικές. Στην εξαγωγή κειμένου, η προεπεξεργασία στη ουσία συνενώνει τα στάδια ερμηνείας και μετασχηματισμού. Σημεία «κλειδιά» του κειμένου, συνήθως ολόκληρες προτάσεις, αναγνωρίζονται από ένα μείγμα από στατιστικά, τοπικά και άλλα κριτήρια και επιλέγονται. Στη συνέχεια η παραγωγή της περίληψης είναι ουσιαστικά μια διαδικασία εξομάλυνσης των επιλεγμένων τμημάτων. Για παράδειγμα, διόρθωση αναφορών που

περιέχονται σε επιλεγμένες προτάσεις και δεν αναφέρονται στην περίληψη. Θα μπορούσαμε να δούμε αυτή την στρατηγική εξαγωγής ως εξής: το πηγαίο κείμενο αντιμετωπίζεται χωρίς καμία ερμηνεία και η αναπαράστασή του τίθεται σε ένα στάδιο μετασχηματισμού το οποίο είναι στην ουσία εξαγωγικό. Η εξαγόμενη περίληψη είναι επομένως γλωσσολογικά «κοντά» στο αρχικό κείμενο όσον αφορά την δομή της. Γενικά, με τις περιλήψεις που παράγονται με αυτόν τον τρόπο είναι σαν να έχουμε μια «θολή εικόνα» για το αρχικό κείμενο. Οι επιλεγμένες προτάσεις συνήθως έχουν κάποια συσχέτιση μεταξύ τους αλλά και με το τμήμα του κειμένου που θα εκτιμούσαμε ως σημαντικό - το νόημά του. Όμως αυτή η μη εντελώς σαφής αναπαράσταση του αρχικού κειμένου γίνεται ακόμη πιο θολή δεδομένου ότι το εξαγόμενο κείμενο της περίληψης, παρότι εξομαλυμένο, δεν είναι συνήθως εντελώς κατανοητό. Αυτό αποτελεί και το σημαντικότερο πρόβλημα της μεθόδου αυτής.

Με την εξαγωγή γεγονότων, τα στάδια ερμηνείας και μετασχηματισμού επίσης ενώνονται. Η αρχική προεπεξεργασία κειμένου σχεδιάζεται ώστε να εντοπίζει και να επεξεργάζεται τα τμήματα του αρχικού κειμένου που σχετίζονται σε γενικές και προκαθορισμένες αρχές ή συσχετίσεις. Δεν υπάρχει ανεξάρτητη αναπαράσταση του πηγαίου κειμένου, μόνο άμεση εισαγωγή πηγαίου υλικού, αλλαγμένο λίγο έως πολύ σε σχέση με την αρχική του αναπαράσταση σύμφωνα με τις απαιτήσεις της κάθε ανεξάρτητης εφαρμογής.

Πιθανοτικά μοντέλα [100],[101] κατανομής των όρων στα κείμενα έχουν βρει χρησιμότητα στον τομέα της αυτόματης εξαγωγής περίληψης, το ίδιο και οι κλασικές TF-IDF (term frequency inverse document frequency) μέθοδοι [102] οι οποίες χρησιμοποιούνται στις περισσότερες εργασίες αυτόματης περίληψης κειμένων και παράγουν ένα ad-hoc σχήμα ζυγίσματος των λέξεων διότι δεν εξάγονται απ' ευθείας από κάποιο μαθηματικό μοντέλο κατανομής όρων ή σχετικότητας. Επιπλέον, κάποιες ερευνητικές εργασίες [103] προσεγγίζουν το πρόβλημα με Poisson και αρνητικές διωνυμικές κατανομές ή με χρήση του k-mixture μοντέλου [104] το οποίο πλησιάζει το μοντέλο του αρνητικού διωνύμου αλλά είναι υπολογιστικά σημαντικά απλούστερο.

Στην πράξη παρατηρούνται σημαντικές παραλλαγές στις προαναφερόμενες μεθόδους προσέγγισης του προβλήματος που συχνά συσχετίζονται με τον επιθυμητό βαθμό μείωσης του μήκους εισόδου. Έτσι, για μικρές πηγές, η εξαγωγή μιας μοναδικής πρότασης μπορεί να φαντάζει σωστή (αν και επικίνδυνη) και αποφεύγει το πρόβλημα της συνοχής νοήματος των προτάσεων εξόδου (μιας και αυτή είναι μόνο μία). Παρόμοια, για τύπου μικρής εισόδου, μπορεί να είναι καταλληλότερη η επεξεργασία όλου του πραγματικού μήκους του κειμένου (Young and Hayes). Από την άλλη μεριά, όπου η εξαγωγή περίληψης βασίζεται στην εξαγωγή γεγονότων από πολλές πηγές, μπορεί να απαιτούνται περισσότεροι μετασχηματισμοί των συνδυασμένων τους αναπαραστάσεων, όπως στο σύστημα POETIC [105], όπου η διαδικασία περίληψης είναι δυναμικά εξαρτώμενη από τα συμφραζόμενα. Είναι φανερό ότι χρειαζόμαστε α) περισσότερη αποτελεσματικότητα στην αυτοματοποιημένη περίληψη από ότι η εξαγωγή κειμένου μας προσφέρει και β) περισσότερη ευελιξία από ότι η εξαγωγή γεγονότων μας παρέχει.

Πέρα από τη διαδικασία εξαγωγής, είναι σημαντικός ο ρόλος της δομής του κειμένου αλλά και των συμφραζομένων στην εξαγωγή αποτελεσματικής περίληψης. Βελτιώσεις επομένως στη διαδικασία περίληψης θα περιλαμβάνουν μεθόδους σύλληψης της δομής αυτής στο αρχικό κείμενο και χρήση της κατά τη διαδικασία εξαγωγής των χρήσιμων τμημάτων του κειμένου. Παραδείγμα της προσπάθειας αυτής αποτελεί η Rhetorical Structure Theory [106]. Οι προσεγγίσεις που εφαρμόζονται συνήθως έχουν να κάνουν με το είδος της πληροφορίας,

γλωσσολογικά, επικοινωνιακά πεδία ενδιαφέροντος που καθορίζουν τη δομή, με το είδος της δομής και τις συσχετίσεις μεταξύ δομών διαφόρων ή του ίδιου κειμένου.

Συνοπτικά θα λέγαμε ότι διακρίνουμε δύο κύριους τρόπους εξαγωγής της περίληψης του αρχικού κειμένου. Ο πρώτος είναι οι ευρετικές μέθοδοι, που βασίζονται κυρίως στον τρόπο σκέψης και εργασίας του ανθρώπου. Πολλές από αυτές, αξιοποιούν την όποια οργάνωση του εγγράφου. Έτσι, προτάσεις που βρίσκονται στις αρχικές και τις τελικές παραγράφους του κειμένου είναι πολύ πιθανό να περιέχονται στην τελική περίληψη. Ο δεύτερος τρόπος, αποτελείται από μεθόδους που βασίζονται στην αναγνώριση λέξεων κλειδιών, φράσεων και ομάδων λέξεων. Το έγγραφο αναλύεται με την χρήση στατιστικών ή/και γλωσσολογικών τεχνικών, για να βρεθούν τα στοιχεία εκείνα που αναπαριστούν το περιεχόμενο του εγγράφου. Αφού ολοκληρωθεί η διαδικασία της περίληψης, ορισμένοι περιλήπτες επιτελούν κάποια περιορισμένη μετα-επεξεργασία ομαλοποίησης των προτάσεων της περίληψης. Δημιουργούν μία λίστα προτάσεων, σε μία προσπάθεια να δοθεί συνέπεια και ευφράδεια στην περίληψη. Γενικά, απομακρύνουν τα ακατάλληλα συνδετικά λέξεων και φράσεων, και εξακριβώνουν σε ποιόν αναφέρονται οι αντωνυμίες του κειμένου ώστε η τελική περίληψη να έχει μια συνοχή.

4.5.1. Συστήματα περίληψης βασισμένα στη γνώση

Από την γέννηση τους, η ανάπτυξη των συστημάτων αντίληψης κειμένων ήταν άρρηκτα συνδεδεμένη με το πεδίο της αναπαράστασης γνώσης και των μεθόδων λογικής [107]. Αυτή η στενή σχέση αιτιολογήθηκε από την παρατήρηση ότι για να έχουμε μια επαρκή κατανόηση του κειμένου απαιτείται γραμματική γνώση σχετικά με τη συγκεκριμένη γλώσσα του κειμένου, αλλά και ενσωμάτωση προηγούμενης γνώσης με την οποία πραγματεύεται το κείμενο. Έτσι, οι συμπερασματικές δυνατότητες των γλωσσών αναπαράστασης γνώσης θεωρούνται πολύ σημαντικές για συστήματα που θα κατανοούν κείμενα. Βασισμένα σε αυτού του είδους την αντίληψη, μια σειρά από συστήματα εξαγωγής περίληψης, βασισμένα στην αναπαράσταση γνώσης, αναπτύχθηκαν (Schankian-type Conceptual Dependency reprecentations). Τα συστήματα αυτά αποτέλεσαν την πρώτη γενιά συστημάτων δημιουργίας αυτοματοποιημένης περίληψης βασιζόμενα στη γνώση.

Ακολούθησε μια δεύτερη γενιά συστημάτων η οποία υιοθέτησε μια πιο «ώριμη» προσέγγιση αναπαράστασης γνώσης, βασιζόμενη στην ήδη υπάρχουσα μεθοδολογία υβριδικών, βασισμένων σε κατηγοριοποίηση, γλωσσών αναπαράστασης [108]. Αυτές οι αρχές χρησιμοποιήθηκαν σε συστήματα περίληψης όπως τα: SUSY [109], SCISOR [110] και TOPIC [111] Αλλά ακόμη και αυτού του είδους τα συστήματα αδυνατούσαν να εξάγουν αποτελεσματικά αξιόλογες μεταφράσεις.

4.5.2. Αναγνώριση Θεμάτων

Το θέμα της αναγνώρισης θεμάτων (Topic Identification), αναφέρεται στην διαδικασία της έρευνας σε έγγραφα κειμένου, για την ανακάλυψη συγκεκριμένων δομών. Σύμφωνα με τους Mather A. Laura και Note Jarrod [112], μία ολοκληρωμένη εφαρμογή, που θα αφορά το θέμα της εύρεσης θεμάτων, θα πρέπει να έχει τη δυνατότητα επεξεργασίας εγγράφων κειμένου, με σκοπό την ανακάλυψη κανόνων και αλγορίθμων, που θα αναγνωρίζουν εγκυκλοπαιδική δομή και εγκυκλοπαιδικά θέματα. Αν αναγνωριστούν συγκεκριμένα θέματα σε έγγραφα κειμένου, τότε αυτά μπορούν να αξιοποιηθούν κατάλληλα και να ενσωματωθούν σε κάποια εγκυκλοπαίδεια. Με αυτό τον τρόπο η εγκυκλοπαίδεια θα είναι ενημερωμένη και η εταιρεία που διαχειρίζεται μία τέτοια εφαρμογή, θα έχει σίγουρα ένα ανταγωνιστικό πλεονέκτημα έναντι των υπολοίπων. Για την υλοποίηση αυτή, απαιτείται η χρησιμοποίηση της επεξεργασίας φυσικής γλώσσας

(Natural Language Processing), η ανάκτηση πληροφορίας (Information Retrieval) και η υπολογιστική γλωσσολογία (Computational Linguistics). Αρχικά απαιτείται η αναγνώριση περιοχών δευτερεύουσας σημασίας (Subtopic Regions) μέσα στο κείμενο, και στην συνέχεια η εύρεση των θεμάτων που σχετίζονται με τις περιοχές αυτές. Για τους σκοπούς αυτούς, αναγνωρίζονται οι φράσεις των ουσιαστικών, τα όρια των προτάσεων και των παραγράφων του κειμένου (Tokenization) . Στην συνέχεια, απομακρύνονται όλες οι συχνές λέξεις (stopwords), μετατρέπεται κάθε λέξη στον ενικό αριθμό και υπολογίζεται η ρίζα της κάθε λέξης. Ακολούθως, ανακαλύπτονται οι περιοχές δευτερεύουσας σημασίας (Subtopic Regions) και προστίθενται ετικέτες στο κείμενο που έχει επεξεργαστεί μέχρι τώρα, για την αναγνώριση των ορίων του κάθε επιθέματος. Τέλος, αναγνωρίζονται τα προεξέχοντα και τα δευτερεύουσας σημασίας θέματα του εγγράφου (Topics, Subtopics). Αφού βρεθούν οι περιοχές δευτερεύουσας σημασίας, υπολογίζεται η βαθμολογία του κάθε θέματος, η οποία θα υποδείξει την υπεροχή του αντίστοιχου θέματος στην αντίστοιχη περιοχή.

4.5.3. Περίληψη κειμένου βασισμένη στο χρόνο

Παρότι είναι λίγη σχετικά η έρευνα στο συγκεκριμένο τομέα, ορισμένοι ερευνητές έχουν ασχοληθεί με το πως είναι δυνατή η εξαγωγή προσωρινών εκφράσεων από ένα κείμενο, αναζητώντας και κανονικοποιώντας αναφορές σε ημερομηνίες, χρόνο και παρερχόμενο χρόνο. Η δουλειά αυτή είναι σημαντική για την ανάλυση του περιεχομένου του κειμένου αλλά όχι για αυτή καθ' αυτή την περίληψή του. Το 1999, το Novelty Detection workshop στο Πανεπιστήμιο του Johns Hopkins εισήγαγε το New Information Detection - NID, έργο του οποίου ήταν η καταγραφή της «νέας» πληροφορίας σε ένα θέμα επισημαίνοντας την πρώτη πρόταση που την περιείχε. Προβλήματα σχετικά με τον επιτυχή καθορισμό της έννοιας «νέο» εμπόδισαν το σύστημα αυτό ώστε να επιτύχει. Η έρευνα αυτή σχετίζεται και με τον τομέα του automatic timeline construction που επικεντρώνεται στην εξαγωγή ασυνήθιστων λέξεων και φράσεων από μία συνεχή ροή νέων και στην περαιτέρω ομαδοποίηση των συστατικών αυτών ώστε να απομονωθούν θέματα μέσα σε ένα νέο.

4.5.4. Αξιολόγηση της περίληψης κειμένου

Μια περίληψη κειμένου είναι γενικά δύσκολο να αξιολογηθεί, κυρίως λόγω των υποκειμενικών κριτηρίων που τίθενται. Ανακατανομή τμημάτων του κειμένου, προτάσεων, παράληψη προφανώς ασήμαντων φράσεων, κ.ο.κ. όλα αυτά καταλήγουν σε μια μεγάλη ποικιλία «καλών» περιλήψεων. Πώς καταλήγουμε όμως στην καλύτερη περίληψη και πως μπορούμε να πούμε πως αυτή που παράγει ο μηχανισμός μας προσεγγίζει τη βέλτιστη?

Υπάρχουν γενικότερα οι εξής μέθοδοι που χρησιμοποιούνται για την αξιολόγηση μια εξαγόμενης περίληψης:

- Χρήση αρκετών πρωτοτύπων παραδειγμάτων από τεχνικές περίληψης κειμένου για τις οποίες γνωρίζουμε την απόδοσή τους
- Συμμετοχή ανθρώπων με την ανάγνωση των περιλήψεων και την βαθμολόγησή τους με κριτήριο το πόσο αντιπροσωπευτική θεωρείται σε σχέση με το αρχικό κείμενο
- Θεωρούμε ότι η περίληψη του κειμένου είναι ένα υποσύνολο του κειμένου και ελέγχουμε εάν μπορεί να αντιπροσωπεύσει επαρκώς το αρχικό κείμενο σε θέματα όπως: είναι δυνατό να κατηγοριοποιηθεί το κείμενο με βάση την περίληψή του ή να εντοπιστεί εάν ανταποκρίνεται στις προτιμήσεις του χρήστη χωρίς να εξεταστεί το αρχικό κείμενο. Μπορεί ένας χρήστης να εμπεδώσει σωστά το κείμενο έχοντας διαβάσει μόνο την περίληψή του και απαντώντας σε tests Μπορεί ο χρήστης να αντιστοιχίσει σωστές λέξεις - κλειδιά σε μια περίληψη.

- Συγκρίνουμε την ομοιότητα μεταξύ προτάσεων επιλεγμένων από ανθρώπους, ως αντιπροσωπευτικές για το κείμενο, και των προτάσεων που προέκυψαν από την αυτοματοποιημένη περίληψη, ή συγκρίνουμε το βαθμό αντιπροσωπευτικότητας που δίνουν οι χρήστες σε μια πρόταση σε σχέση με αυτόν που δίνει ο μηχανισμός. Οι τεχνικές αυτού του είδους αναφέρονται συνήθως και ως corpus-based.

4.5.5. Copernic Summarizer

Πρόκειται για ένα εμπορικό προϊόν το οποίο πραγματοποιεί αυτόματη εξαγωγή περίληψης στα Αγγλικά, Γαλλικά και Γερμανικά. Χρησιμοποιείται για να παράγει περιλήψεις κειμένων και δικτυακών τόπων προσφέροντας με αυτό τον τρόπο μία γενική εικόνα των εγγράφων προτού ο χρήστης τα διαβάσει ολόκληρα.

Χρησιμοποιώντας πολύπλοκους στατιστικούς αλγορίθμους και γλωσσική ανάλυση, εντοπίζει τις πιο καίριες εκφράσεις του κειμένου και εξάγει τις πιο σημαντικές προτάσεις τόσο σε ένα δικτυακό τόπο όσο και σε ένα κείμενο. Ενώνοντας αυτές τις προτάσεις παράγεται η περίληψη του κειμένου.

Ως εμπορικό πρόγραμμα, δεν είναι εφικτή η αναλυτική προσέγγιση των τρόπων με τους οποίους πραγματοποιείται η εξαγωγή περίληψης.

4.5.6. MS Word Summarizer

Το πρόγραμμα MS Word στις πιο πρόσφατες εκδόσεις του περιέχει ένα μηχανισμό αυτόματης εξαγωγής περίληψης κειμένων το οποίο απαρτίζεται από προτάσεις του κειμένου που απομονώνονται. Αναλυτικές πληροφορίες για τις μεθόδους που χρησιμοποιούνται για την εξαγωγή περίληψης δεν υπάρχουν ωστόσο τα αποτελέσματα του μηχανισμού δεν είναι καθόλου ικανοποιητικά συγκριτικά με αλγορίθμους και μηχανισμούς που υπάρχουν.

4.5.7. MEAD Summarizer

.Πρόκειται ίσως για τον πιο ολοκληρωμένο μηχανισμό αυτόματης εξαγωγής περίληψης που υπάρχει. Οι πληροφορίες είναι περιορισμένες ωστόσο υπάρχουν πολλές δημοσιεύσεις που αφορούν το μηχανισμό όπου και φαίνονται οι δυνατότητές του. Βασικός σκοπός του είναι η εξαγωγή περίληψης από πολλαπλά έγγραφα και έχει τη δυνατότητα να ξεχωρίσει νοηματικά ίδιες προτάσεις και να μην πραγματοποιεί διπλοεγγραφές κατά τη διαδικασία εξαγωγή περίληψης. Περισσότερες πληροφορίες για το μηχανισμό υπάρχουν στα [59] και [60]

4.5.8. SUMMARIST

Ο Summarist είναι ένας μηχανισμός ο οποίος πραγματοποιεί αυτόματη εξαγωγή περίληψης κειμένων. Πρόκειται για ένα σύστημα το οποίο βασίζεται σε οντολογίες προκειμένου να αποκτήσει γνώση επί των λέξεων και χρησιμοποιεί αμιγώς NLP (Natural Language Processing). Η βασική συνάρτηση στην οποία στηρίζεται είναι:

Κατηγοριοποίηση = Εντοπισμός τίτλου + μετάφραση + παραγωγή

Για κάθε βήμα από τα παραπάνω το σύστημα εφαρμόζει τις ακόλουθες τεχνικές:

4.5.8.1. Εντοπισμός Τίτλου

Με γενίκευση των τεχνικών ανάκτησης πληροφορίας και προσθέτοντας τεχνικές εντοπισμού τίτλου, χρησιμοποιείται ο μηχανισμός SENSUS αλλά και λεξικά, ο μηχανισμός πραγματοποιεί εντοπισμό σεναρίων μέσα στο κείμενο. Επιτρέπει πολυγλωσσική ανάλυση και πιο συγκεκριμένα οι γλώσσες στις οποίες

πραγματοποιείται ο εντοπισμός είναι: Αγγλικά, Ισπανικά, Ιαπωνικά, Ινδονησιανά και Αραβικά.

4.5.8.2. Μετάφραση

Το κομμάτι αυτό του μηχανισμού δεν κάνει τη μετάφραση των κειμένων αλλά χρησιμοποιεί τεχνικές στατιστικής ανάλυσης από την Ανάκτηση Πληροφορίας αλλά και LSA (Latent Semantic Analysis) όπως και λεξικά για να πραγματοποιήσει διασύνδεση των τίτλων και των σεναρίων που έχουν εντοπιστεί σε ένα κείμενο προκειμένου να εντοπιστεί το «νόημα» του κειμένου.

4.5.8.3. Δημιουργία

Ο μηχανισμός χρησιμοποιεί τρια διαφορετικά συστήματα για τη δημιουργία της αυτόματης περίληψης: μία λίστα λέξεων-κλειδιών, ένα μηχανισμό δημιουργίας φράσεων και ένα μηχανισμό δημιουργίας προτάσεων από λέξεις κλειδιά και φράσεις. Οι τρεις μηχανισμού λειτουργού σειριακά με τον τρόπο που αναφέρονται προκειμένου να δημιουργήσουν το επιθυμητό αποτέλεσμα.

Αναλυτικές πληροφορίες για το μηχανισμό υπάρχουν στο [61]

4.6. Προσωποποίηση στο χρήστη

Σύμφωνα με τον Mobasher [44], «η προσωποποίηση στο διαδίκτυο μπορεί να περιγραφή σαν κάθε ενέργεια που σαν σκοπό έχει να κάνει τη Διαδικτυακή εμπειρία ενός χρήστη να είναι βάσει των αναγκών που έχει κάθε χρήστης». Σε γενικές γραμμές αυτό σημαίνει αλλαγή της παρουσίασης των δεδομένων ενός Δικτυακού τόπου προς το χρήστη σύμφωνα με τις εκάστοτε ρητές και εννοούμενες επιλογές του χρήστη. Αυτό είναι σχετικά εύκολο όταν αναφέρομαστε σε ένα και μόνον δικτυακό τόπο. Ο χρήστης καλείται να δηλώσει ρητά τις προτιμήσεις του ενώ παράλληλα το σύστημα «μαθαίνει» τις προτιμήσεις του χρήστη. Αυτό συναντάται σε πολλούς δικτυακούς τόπους.

Ο έλεγχος της δραστηριότητας του χρήστη σε πολλαπλούς δικτυακούς τόπους και ο εντοπισμός των πραγματικών αναγκών του και επιλογών είναι μία μεγάλη πρόκληση. Αυτό συνεπάγεται πως τη στιγμή που ένας χρήστης επισκέπτεται ένα δικτυακό τόπο, υπάρχει ήδη ένα προφίλ του και το σύστημα είναι άμεσα σε θέση να προσαρμοστεί στις ανάγκες του συγκεκριμένου χρήστη. Πολλές προσεγγίσεις πάνω στο συγκεκριμένο θέμα έχουν δοκιμαστεί: Single Sign On συστήματα [142] [46], προσωποποίηση στη μεριά του χρήστη [45] και βέβαια όλα τα συστήματα spyware και ad trackers. Πολλά από αυτά τα συστήματα παρουσιάζουν προβλήματα με τη νομοθεσία καθώς προσβάλλουν την ιδιωτικότητα του χρήστη ενώ τα συστήματα που εφαρμόζουν την προσωποποίηση στη μεριά του χρήστη έχουν χαμηλή αποδοτικότητα.

Μία σειρά από πρωτοβουλίες στην W3C έχουν σαν σκοπό την καθολική προσωποποίηση. Το OPS (Open Profiling Standard) [47] είναι ένα προτεινόμενο W3C standard το οποίο έχει υποβληθεί από τις εταιρίες Netscape, Verisign και Firefly από το 1997. Παρουσιάζει ένα σχήμα τυποποίησης και ένα πρωτόκολλο ανταλλαγής δεδομένων που αφορούν το προφίλ ενός χρήστη, όπως για παράδειγμα το όνομα, τη διεύθυνση και τον ταχυδρομικό κώδικα. Ωστόσο, δεν τέθηκε ποτέ σε χρήση. Η ιδέα ανταλλαγής πληροφορίας είναι πολύ χρήσιμη, όμως πολλοί χρήστες δε θα επιθυμούσαν τη δημοσιοποίηση τέτοιων στοιχείων. Για την προσωποποίηση θα ήταν χρησιμότερο να διαμοιράζονται πληροφορίες που αφορούν την περιαγωγή ενός χρήστη στους δικτυακούς τόπους.

Το PIDL (Personalized Information Description Language) [48] είναι ένα πρωτόκολλο που υποβλήθηκε στην W3C από την εταιρία NEC το 1999. Πρόκειται για έναν τρόπο δόμησης εγγράφου που περιέχει στοιχεία για τις προτιμήσεις ενός χρήστη κατά τη διάρκεια που βρίσκεται σε διάφορους δικτυακούς τόπους. Είναι

προφανές πως κάτι τέτοιο έρχεται ενάντια στα στοιχεία ιδιωτικότητας του χρήστη που έχουμε ήδη αναφέρει. Είχε προταθεί αρχικά για χρήση σε multicast, μία τεχνολογία που τελικά δεν αναπτύχθηκε όσο αναμενόταν.

Το CC/PP (Composite Capabilities/Preference Profiles) [49] είναι ένα W3C στάνταρ που προτάθηκε το 1999 και βρίσκεται μέχρι και σήμερα σε χρήση. Επιτρέπει σε κινητούς χρήστες να εκφράσουν τις προτιμήσεις ενός χρήστη σε έναν κεντρικοποιημένο εξυπηρετητή. Παρά το γεγονός ότι οι κινητές τεχνολογίες έχουν πολλούς περιορισμούς στην ανταλλαγή δεδομένων, αυτή η αρχιτεκτονική θα μπορούσε να αποτελέσει τη βάση για ένα σύστημα διαμοιρασμού των προτιμήσεων ενός χρήστη.

Το P3P (Platform for Privacy Preferences) [50] έρχεται σε αντίθεση με κάθε σύστημα προσωποποίησης που βασίζεται στο διαμοιρασμό των στοιχείων ενός χρήστη μεταξύ δικτυακών τόπων. Αυτή η σύσταση της W3C που έγινε το 2002 έχει σχεδιαστεί ώστε να επιτρέπει στους χρήστες να ελέγχουν τα προσωπικά τους δεδομένα που θα παρουσιάζονται στους διάφορους δικτυακούς τόπους που επισκέπτεται.

Κανένα από τα παραπάνω δεν επιτρέπει την προσωποποίηση σε πολλαπλούς δικτυακούς τόπους. Αν αναλογιστούμε τα εμπορικά συστήματα θα δούμε πως πρόκειται για ένα σημαντικό κομμάτι τους, κυρίως όσον αφορά θέματα μάρκετινγκ. Οι εταιρίες επιθυμούν να γνωρίζουν τις ανάγκες των «πελατών» τους πρωτού αυτοί επισκευθούν το «κατάστημά» τους. Έτσι, πολλοί δικτυακοί τόποι, όπως για παράδειγμα η προσωποποίηση και οι συστάσεις που παρουσιάζονται στο δικτυακό τόπο του Amazon.com [143] το εφαρμόζουν σε ατομικό επίπεδο. Από τις πρώτες κιόλας σελίδες που επισκέπτεται ο χρήστης διαμορφώνεται ένα προφίλ του προκειμένου ο δικτυακός τόπος να προσαρμόζεται σιγά - σιγά στις ανάγκες του.

Η μελέτη του θέματος που αφορά τις επιλογές ενός χρήστη καθώς και τη συμπεριφοράς αυτού κατά την επίσκεψη πολλών διαφορετικών δικτυακών τόπων έχει πραγματοποιηθεί από πολλές εταιρίες και έχουν γίνει πολλές προτάσεις. Αν εξαιρέσουμε τις προσπάθειες στις οποίες ανακύπτουν ηθικά αλλά και νομικά ζητήματα παραβίασης της ιδιωτικότητας καταλήγουμε αποκλειστικά στα συστήματα SSO (Single Sign On) όπως είναι το Microsoft Passport [142] και το Liberty Alliance [51]. Αυτά παρέχουν μία ενιαία βάση δεδομένων που περιέχει τα προσωπικά στοιχεία και τις επιλογές του. Οι χρήστες προσθέτουν από μόνοι τους στοιχεία στη βάση δεδομένων στα οποία έχουν ελεύθερη πρόσβαση εταιρίες που είναι συμβεβλημένες με τα εκάστοτε SSO συστήματα.

Βασικό πρόβλημα αυτής της προσέγγισης είναι η εξασφάλιση της ασφάλειας του συστήματος καθότι ο χρήστης μπορεί να αποθηκεύει ευαίσθητα δεδομένα. Το συγκεκριμένο θέμα τονίζεται ακόμα και στα προϊόντα των εταιριών (για παράδειγμα η Sun το τονίζει ιδιαίτερα στο πρόγραμμα Liberty [52]). Πως θα εμπιστευτεί ένας χρήστης το πρόγραμμα το οποίο του τονίζει ιδιαίτερα πως δεν είναι ασφαλές; Τα νεότερα SSO συστήματα όπως το Liberty Alliance [51] και το SXIP [53] έχουν δώσει ιδιαίτερη προσοχή στο συγκεκριμένα θέμα προκειμένου να βελτιωθούν. Μάλιστα το SIXP επιτρέπει σε ένα χρήστη να διαθέτει πολλαπλά προφίλ ανάλογα με το μέγεθος των δεδομένων που επιθυμεί να είναι ορατά σε διάφορους δικτυακούς τόπους ορίζοντας με αυτό τον τρόπο αυτόνομα το επίπεδο ασφάλειας. Παράλληλα είναι ένα σύστημα ανοιχτού κώδικα προκειμένου οι χρήστες να μπορούν να δουν επακριβώς τι στοιχεία τους διαμοιράζονται και με ποιον τρόπο. Αυτό βέβαια δεν ξεπερνά τα προβλήματα που παρουσιάζονται. Οι χρήστες πρέπει να αποφασίσουν αν οι εταιρίες στις οποίες θα εμπιστευτούν τα προσωπικά τους δεδομένα είναι έμπιστες ή όχι. Αυτό συνεπάγεται και την αποτυχία τετοιων συστημάτων με χαρακτηριστικό παράδειγμα το σύστημα Passport σαν τεχνολογία καθότι οι χρήστες δεν έχουν κάποια ιδιαίτερη προτίμηση στα SSO συστήματα. Παράλληλα, όπως αναφέρει και ο Gartner [144], «όσο οι χρήστες δε

δείχνουν να αποδέχονται τέτοια συστήματα οι εταιρίες δεν πρόκειται να κάνουν απολύτως καμία επένδυση».

Υπάρχουν βέβαια και συστήματα τα οποία δεν απαιτούν την εισαγωγή στοιχείων από το χρήστη αλλά χρησιμοποιούν μεταδεδομένα που υπάρχουν από τα ίχνη που αφήνει ένας χρήστης καθώς πραγματοποιεί περιαγωγή σε σελίδες του διαδικτύου. Το WAWA (Wisconsin Adaptive Web Assistant) [54] είναι ένα σύστημα το οποίο προσπαθεί να εντοπίσει τις σελίδες που μπορεί να αφορούν κάποιο χρήστη ανάλογα με το history που εντοπίζει στο φυλλομετρητή. Αντίστοιχα το Syskill and Webert [57] είναι ένα πρόγραμμα το οποίο μαθαίνει να βαθμολογεί τις σελίδες που επισκέπτεται ο χρήστης και αποφασίζει ποιες είναι οι σελίδες που πιθανόν ενδιαφέρουν το χρήστη. Το σύστημα αυτό χρησιμοποιεί το προφίλ χρήστη που το ίδιο κατασκευάζει και προτείνει στο χρήστη συνδέσμους που ενδεχόμενα τον ενδιαφέρουν το χρήστη ή πραγματοποιεί ερωτήματα σε μηχανές αναζήτησης με λέξεις κλειδιά από το διαμορφωμένο προφίλ χρήστη. Ο Chan [56] περιγράφει ένα παραπλήσιο σύστημα το οποίο περιέχει δύο στοιχεία: το Web Access Graph (WAG) και τον Page Interest Estimator (PIE). Το WAG εντοπίζει ίχνη σε ιστοσελίδες που μπορεί να αφορούν το χρήστη και το PIE «μαθαίνει» τον τρόπο με τον οποίο επισκέπτεται ένας χρήστης μία σελίδα βάσει των επιλογών που κάνει.

Οι Widyantoro, Ioerger και Yen [55] ανέπτυξαν ένα σύστημα το οποίο βασίζεται σε έναν τριπλό περιγραφέα προκειμένου να καταγράφουν τη δυναμική ενός χρήστη απέναντι στο διαδίκτυο. Το μοντέλο αυτό διατηρεί μία μία περιγραφή για κάθε ίχνος που αφήνει ο χρήστης στο διαδίκτυο σε ένα μεγάλο βάθος χρόνου και το συνδυάζει με δεδομένα που αποθηκεύονται προσωρινά προκειμένου να κάνει προβλέψεις για τις ιστοσελίδες που μπορεί να αφορούν το χρήστη.

Οι Goecks και Shavlik [58] προτείνουν ένα σύστημα που «μαθαίνει» τα ενδιαφέροντα του χρήστη ελέγχοντας περισσότερα στοιχεία που αφορούν τις σελίδες που επισκέπτεται. Παρατηρούν για παράδειγμα τις κινήσεις που κάνει ο χρήστης με το ποντίκι εκτός από την απλή διαδικασία ελέγχου των σελίδων που επισκέπτεται ο χρήστης.

5

ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

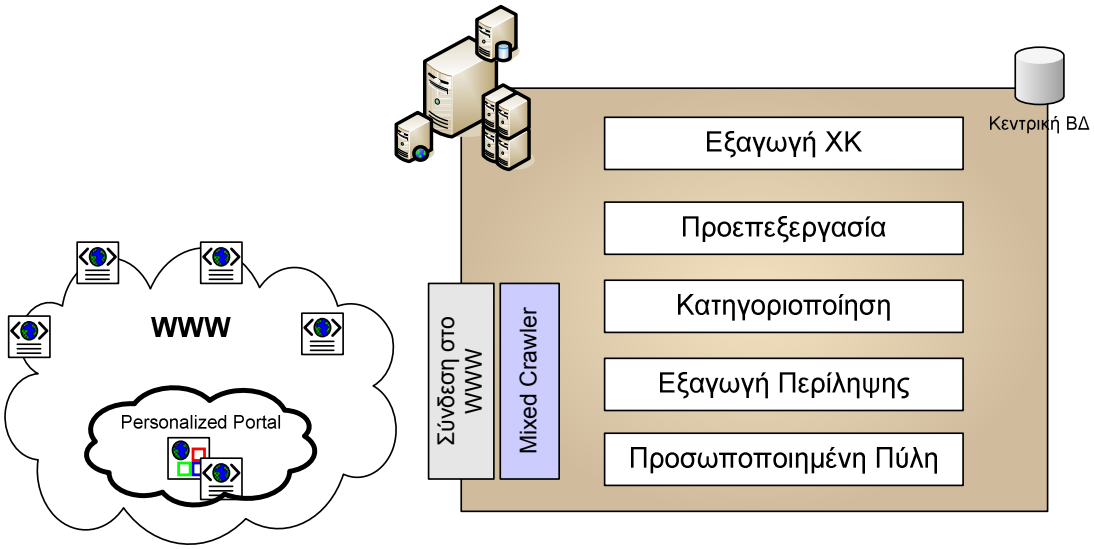
Στο κεφάλαιο αυτό περιγράφεται η αρχιτεκτονική του συστήματος που αναπτύχθηκε. Πιο συγκεκριμένα υπάρχουν στοιχεία που αφορούν:

- Τη γενική αρχιτεκτονική
- Τα υποσυστήματα συλλογής πληροφορίας, εξαγωγής κειμένου, προεπεξεργασίας, κατηγοριοποίησης, εξαγωγής περίληψης, παρουσίασης πληροφορίας

5. ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

5.1. Γενική Αρχιτεκτονική

Το σύστημα που σχεδιάστηκε αποτελείται από μικρότερα υποσυστήματα προκειμένου να είναι εύκολη η αυτόνομη σχεδίαση, κατασκευή και χρήση τους καθότι οι μηχανισμοί που οδηγούν στο επιθυμητό αποτέλεσμα είναι πολλοί. Το παρακάτω γενικό σχήμα μας δίνει τη γενική αρχιτεκτονική του συστήματος



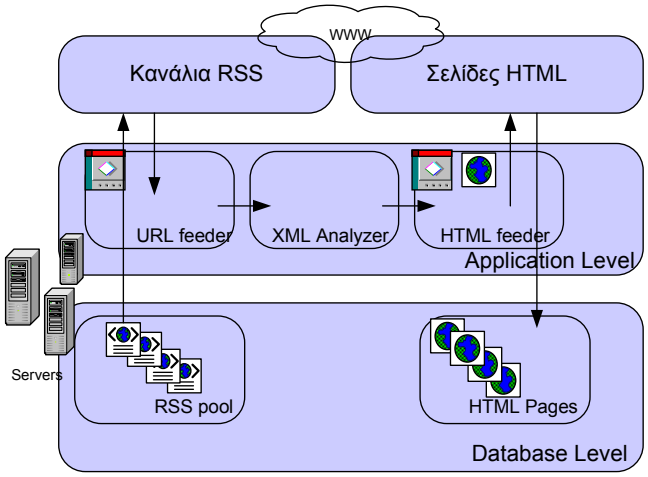
Εικόνα 7: Γενική Αρχιτεκτονική του Συστήματος

5.2. Υποσυστήματα

Εν συνεχεία θα παρουσιαστούν τα υποσυστήματα του μηχανισμού προκειμένου να γίνει κατανοητή η λειτουργία του μηχανισμού σε κάθε διαφορετικό επίπεδο υλοποίησης.

5.2.1. Συλλογή πληροφορίας

Για τη συλλογή πληροφορίας για το σύστημά μας και πιο συγκεκριμένα για την συνεχή και αδιάκοπτη συλλογή άρθρων από το Διαδίκτυο εκμεταλλευόμαστε την τάση που επικρατεί σε όλους τους δικτυακούς τόπους να προσφέρουν κανάλια άμεσης επικοινωνίας με τους χρήστες και δε μιλούμε για κάτι διαφορετικό από τα RSS.

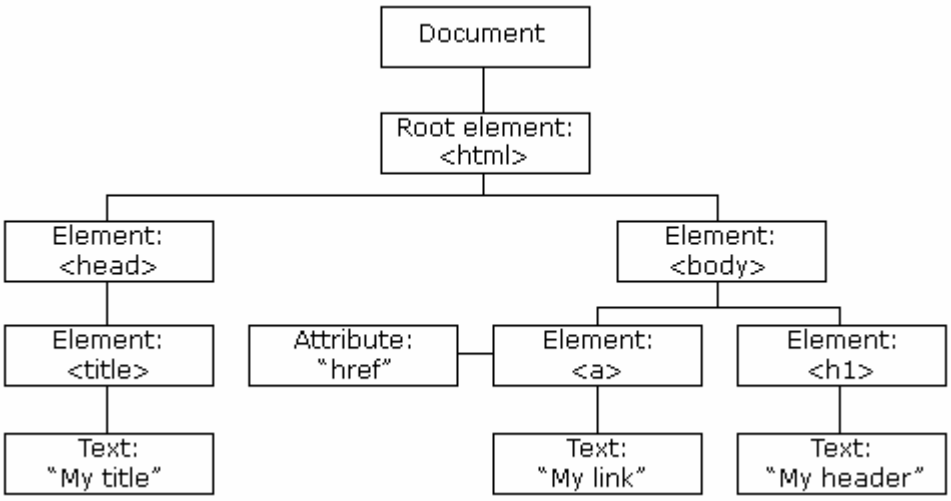


Εικόνα 8: Μηχανισμός Συλλογής Πληροφορίας

Η αρχιτεκτονική του μηχανισμού είναι απλή και εκμεταλλεύεται πλήρως το γεγονός ότι οι μεγαλύτερες δικτυακές πύλες ενημέρωσης προσφέρουν στους χρήστες RSS feeds. Όπως φαίνεται από το σχήμα, ένας απλοϊκός mixed selective crawler χρησιμοποιείται προκειμένου να λαμβάνει το σύστημά μας HTML σελίδες. Πρόκειται για έναν mixed crawler διότι συνδυάζει τη χρήση wrapper και crawler. Ο wrapper είναι ένας μηχανισμός αναγνώρισης προτύπων που συνήθως ακολουθείται από επεξεργασία αυτών. Στην περίπτωσή μας ο wrapper στο μηχανισμό συλλογής πληροφορίας εντοπίζει μέσα στα XML αρχεία εκείνα τα σημεία τα οποία περιέχουν πληροφορίες για τα άρθρα που θέλουμε να εξάγουμε. Μέσα από αυτά τα αρχεία προκύπτουν τα URL seeds τα οποία επανατροφοδοτούν το ίδιο μηχανισμό για να προχωρήσει στο «κατέβασμα» των σελίδων HTML, που περιέχουν άρθρα, από τη φυσική τους θέση χωρίς να χρειαστεί καμία απολύτως αναζήτηση. Ο wrapper συνεπώς χρησιμοποιείται για να μπορέσουμε να εξάγουμε τον τίτλο του άρθρου και τη διεύθυνση στην οποία βρίσκεται με τη βοήθεια των RSS feeds και εν συνεχεία το πρόγραμμα αλλάζει μορφή και μετατρέπεται σε crawler ο οποίος «επισκέπτεται» τα URLs που έχει εξάγει ο wrapper και από αυτά λαμβάνει τον HTML κώδικα. Η βάση δεδομένων δε χρειάζεται τις ενδιάμεσες πληροφορίες και έτσι οι πληροφορίες που έχει είναι η λίστα με τα RSS. Τις πληροφορίες που αποθηκεύονται για κάθε άρθρο θα τις δούμε στη συνέχεια του κεφαλαίου.

5.2.2. Εξαγωγή Χρήσιμου κειμένου (φιλτράρισμα)

Για την εξαγωγή του χρήσιμου κειμένου χρησιμοποιείται η ιδιότητα της HTML να μπορεί να αναπαρασταθεί σε δενδρική μορφή σύμφωνα με το DOM (Document Object Model) μοντέλο, όπως φαίνεται και στο παρακάτω σχήμα.



Εικόνα 9: HTML Document Object Model (DOM)

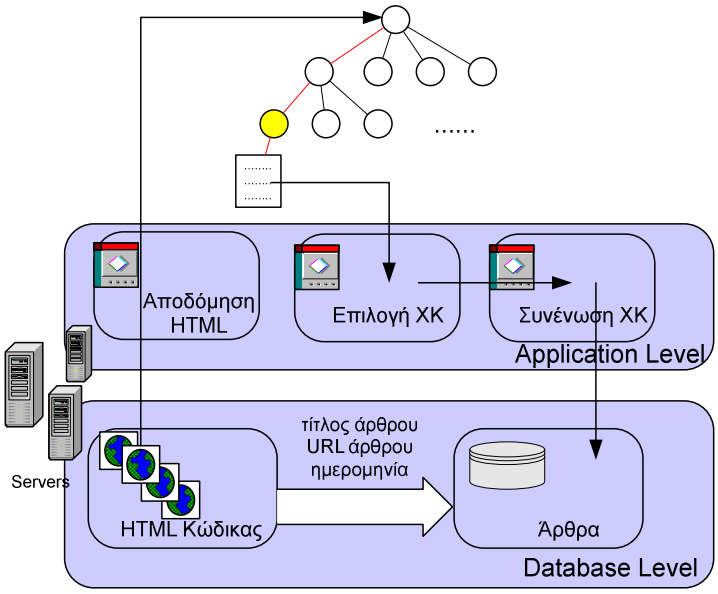
Το παραπάνω σχήμα είναι η DOM αναπαράσταση του παρακάτω HTML κώδικα.

Κώδικας 1: HTML κώδικας όπως προκύπτει από το DOM μοντέλο

```
<html>
  <head>
    <title>My Title</title>
  </head>
  <body>
    <a href="#">My Link</a>
    <h1>My Header</h1>
  </body>
</html>
```

Βασιζόμενοι , λοιπόν, στο γεγονός ότι κάθε HTML κώδικας μπορεί να αποδομηθεί στα βασικά του στοιχεία σε δενδρική μορφή χρησιμοποιούμε ένα

μηχανισμό όπως αυτός που φαίνεται στο παρακάτω σχήμα προκειμένου να εξάγουμε το χρήσιμο κείμενο από τις HTML σελίδες.



Εικόνα 10: Προεπεξεργασία κειμένου

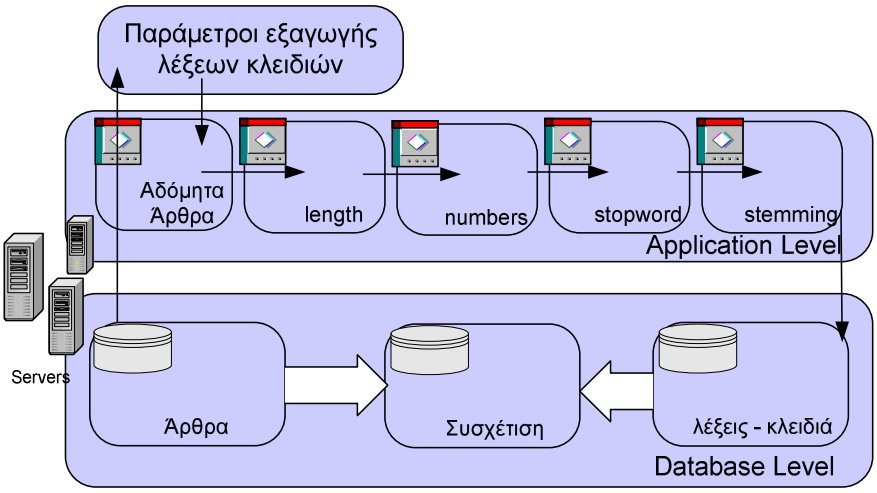
Και πάλι σε αυτή την περίπτωση εργαζόμαστε σε δύο επίπεδα αυτό της εφαρμογής και αυτό της Βάσης Δεδομένων. Από τη ΒΔ λαμβάνουμε τον HTML κώδικα καθώς και πληροφορίες για το άρθρο που έχουν συλλεχθεί από το προηγούμενο στάδιο και προχωρούμε σε αποδόμηση της HTML σελίδας προκειμένου να εντοπίσουμε τα φύλλα του δένδρου που ενδεχόμενα περιέχουν χρήσιμες πληροφορίες για το μηχανισμό.

5.2.3. Προεπεξεργασία κειμένου

Η προεπεξεργασία κειμένου βασίζεται σε συγκεκριμένη τεχνική η οποία αφορά την προεπεξεργασία κάθε είδους κειμένου προκειμένου να εξαχθούν οι λέξεις κλειδιά. Ο μηχανισμός προεπεξεργασία ή εξαγωγής των λέξεων κλειδιών μπορεί να διαβάσει πληροφορίες, είτε από αρχείο είτε από τη βάση δεδομένων. Παράλληλα θα πρέπει να δοθούν συγκεκριμένες μεταβλητές οι οποίες αφορούν την επεξεργασία που θα πραγματοποιηθεί. Αυτές οι παράμετροι αφορούν:

- Το ελάχιστο μήκος λέξης,
- Το αν οι αριθμοί θα αποθηκευτούν ή όχι,
- Το είδος της stopword λίστας που θα χρησιμοποιηθεί,
- Τον αλγόριθμο stemming που θα χρησιμοποιηθεί [75],[76].

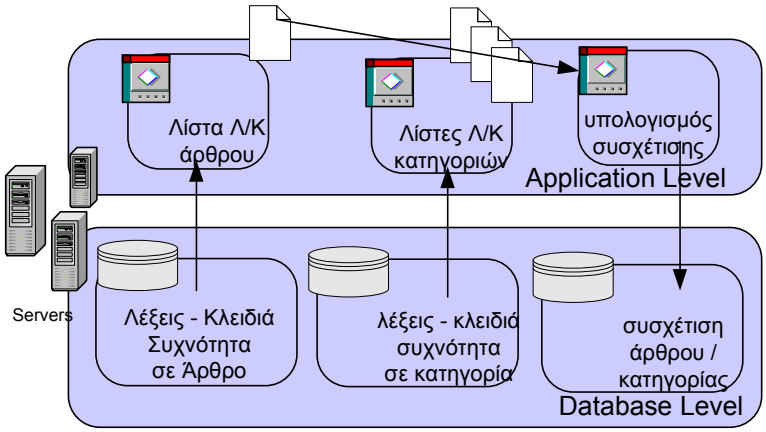
Η διαδικασία για την προεπεξεργασία κειμένου και την εξαγωγή των λέξεων κλειδιών φαίνεται στο παρακάτω σχεδιάγραμμα. Όπως είναι εμφανές ο μηχανισμός διαβάζει τα νέα άρθρα από τη βάση δεδομένων καθώς και τις παραμέτρους που τίθενται για κάθε άρθρο και σύμφωνα με τη διαδικασία που περιγράψαμε παραπάνω εξάγει τις λέξεις κλειδιά για κάθε κείμενο και τις συσχετίζει με κάθε άρθρο.



Εικόνα 11: Προεπεξεργασία κειμένου και εξαγωγή λέξεων-κλειδιών

5.2.4. Κατηγοριοποίηση Κειμένου

Το υποσύστημα της κατηγοριοποίησης κειμένου αποτελεί ένα κεντρικό συστατικό του μηχανισμού που αναπτύχθηκε και σε συνδυασμό με εκείνο της εξαγωγής περίληψης, βρίσκονται στο δεύτερο επίπεδο ανάλυσης του συστήματος αποτελώντας τον πυρήνα του μηχανισμού. Το υποσύστημα περιγράφεται από το παρακάτω σχήμα.



Εικόνα 12: Κατηγοριοποίηση Κειμένου

Η είσοδος του υποσύστηματος κατηγοριοποίησης κειμένου είναι XML αρχεία τα οποία περιέχουν την έξοδο του υποσυστήματος εξαγωγής κωδικολέξεων και πιο συγκεκριμένα: τα keywords του κειμένου και τις συχνότητες εμφάνισής τους στο κείμενο. Ο βασικός στόχος του υποσυστήματος αυτού είναι η εφαρμογή αλγορίθμων κατηγοριοποίησης στο κείμενο και επομένως η αντιστοίχηση του κειμένου με κάποια από τις ήδη υπάρχουσες κατηγορίες. Βασικό ρόλο σε αυτή τη διαδικασία παίζει η ύπαρξη μιας σωστής, πλήρης και αποτελεσματικής βάσης γνώσης πάνω στην οποία θα στηρίζεται η κατηγοριοποίηση. Πιο αναλυτικά, χρειαζόμαστε κάποιες βασικές κατηγορίες άρθρων, στις οποίες θα εμπίπτουν τα περισσότερα των νέων άρθρων που έρχονται στο σύστημα, καθώς και ένα πλήθος αντιπροσωπευτικών της κάθε κατηγορίας κειμένων, τα οποία έχουν περάσει από το μηχανισμό εξαγωγής keywords και στην ουσία «ταΐζουν» το σύστημα με την αναγκαία γνώση, ώστε να μπορεί με χρήση απλών μετρικών να κατηγοριοποιεί νεοαφιχθέντα άρθρα.

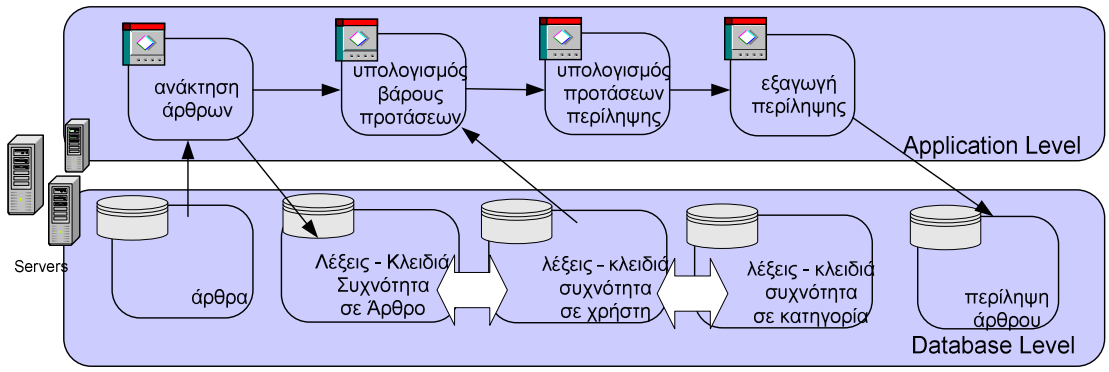
Το υποσύστημα κατηγοριοποίησης βασίζεται στην μετρική ομοιότητας συνημίτονου, σε εσωτερικά γινόμενα καθώς και σε υπολογισμούς ζυγίσματος όρων. Η χρήση αυτών των μετρικών γίνεται ύστερα από την αρχικοποίηση του

training set της βάσης γνώσης και μέσω μιας διαδικασίας η οποία on the fly ελέγχει τη συσχέτιση του κάθε keyword του προς κατηγοριοποίηση κειμένου με τις υπάρχουσες κατηγορίες. Οι συσχετίσεις που θα βρεθούν αθροίζονται και κανονικοποιούνται με αποτέλεσμα να προκύπτει για κάθε κείμενο ένα ποσοστό ομοιότητας (relativity) με κάθε μια από τις υπάρχουσες κατηγορίες. Εάν το training set είναι αποτελεσματικό και αξιόπιστο, τα άρθρα που περιέχουν πολλά keywords σχετικά με κάποια από τις κατηγορίες, θα πρέπει να έχουν κατηγοριοποιηθεί με συσχέτιση μεγαλύτερη ως προς αυτή. Στην πράξη βέβαια, ακόμη και αν το κείμενο είναι εντελώς αντιπροσωπευτικό κάποιας κατηγορίας, δεν αποκτά συσχέτιση 100% με μία και μόνο κατηγορία, αφού είναι φυσικό να περιέχει ορισμένα keywords τα οποία συσχετίζονται και με τις υπόλοιπες κατηγορίες (το άθροισμα των συσχετίσεων ενός κειμένου με όλες τις κατηγορίες, προφανώς σε κάθε περίπτωση είναι 1).

Η έξοδος του υποσυστήματος κατηγοριοποίησης, οι συσχετίσεις δηλαδή του κειμένου με κάθε κατηγορία, αποθηκεύονται στη βάση δεδομένων του συστήματος.

5.2.5. Εξαγωγή Περίληψης Κειμένου

Για τη διαδικασία εξαγωγής της περίληψης του κειμένου χρησιμοποιείται ενιαία αρχιτεκτονική παρά το γεγονός ότι το σύστημα δημιουργεί τόσο γενικευμένες περιλήψεις για όλους τους χρήστες αλλά και προσωποποιημένες για κάθε εγγεγραμμένο χρήστη του συστήματος. Η διαδικασία εξαγωγής αυτόματης περίληψης βασίζεται σε ευρεστικές μεθόδους και πιο συγκεκριμένα σε αξιολόγηση και βαθμοδότηση των προτάσεων του κειμένου προκειμένου να εξαχθούν οι πιο σημαντικές από αυτές και να αποτελέσουν την περίληψη του κειμένου.



Εικόνα 13: Εξαγωγή Περίληψης Άρθρου

Το υποσύστημα εξαγωγής περίληψης κειμένου του μηχανισμού αποτελεί ένα ανεξάρτητο υποσύστημα το οποίο δέχεται ως είσοδο τα αποτελέσματα του keyword extraction (αποθηκευμένα στη βάση ή σε μορφή XML) που περιέχουν: τα keywords που κρατήθηκαν, τη συχνότητα εμφάνισής τους στο κείμενο, τις θέσεις τους (σε ποιες προτάσεις εμφανίζονται, π. χ. 1η, 3η, κ.ο.κ.) και το πόσες προτάσεις πρέπει να κρατηθούν για την τελική περίληψη. Τα στοιχεία αυτά, μαζί με την πληροφορία για τον τίτλο του κειμένου, είναι αρκετά ώστε να μπορεί το υποσύστημα αυτό να επιχειρεί μια βαθμολόγηση των προτάσεων του κειμένου. Θα πρέπει να πούμε σε αυτό το σημείο ότι, ο μηχανισμός αυτόματης εξαγωγής περίληψης δεν χρειάζεται απαραίτητα αυτό καθ' αυτό το κείμενο αν και για να παραχθεί η τελική περίληψη ενός κειμένου αυτό είναι αναγκαίο. Με το προηγούμενο εννοούμε ότι, το υποσύστημα αυτό μπορεί να παράγει μια τελική κατάταξη των προτάσεων του κειμένου απλά και μόνο με τις εισόδους που περιγράφηκαν νωρίτερα και ενώ το αρχικό κείμενο βρίσκεται αποθηκευμένο μία φορά μόνο στην βάση δεδομένων. Το τελευταίο δεδομένο εισόδου του υποσυστήματος περιγράφει πόσες προτάσεις επιθυμούμε να έχουμε ως έξοδο για περίληψη του αρχικού κειμένου. Το πλήθος των προτάσεων μπορεί να καθοριστεί είτε ως ποσοστό % των προτάσεων του αρχικού κειμένου είτε ως συνολικό πλήθος

χαρακτήρων. Για παράδειγμα, αν το αρχικό κείμενο είχε 20 προτάσεις και κρατάμε ένα ποσοστό 30% επί των προτάσεων, στην περίληψη θα κρατηθούν οι 6 σημαντικότερες προτάσεις του κειμένου, αντίθετα, εάν επιθυμούμε η περίληψη του κειμένου να περιέχει περίπου ένα συγκεκριμένο πλήθος χαρακτήρων, θα επιλεχθούν τόσες προτάσεις από τις σημαντικότερες ώστε και να καλύπτεται το πλήθος χαρακτήρων που τέθηκε και να μην ξεπερνιέται κατά πολύ αυτό. Στην ουσία επιλέγεται η βέλτιστη επιλογή μήκους χαρακτήρων στο όριο να επιλεχθεί μια παραπάνω πρόταση ή μια λιγότερη.

Η έξοδος επομένως του υποσυστήματος αυτόματης εξαγωγής περίληψης κειμένου είναι μια φθίνουσα σειρά προτάσεων με βάση το σκορ που αξιολογεί ο μηχανισμός πως πρέπει να έχουν όσον αφορά την σημαντικότητά τους για να αναπαραστήσουν το κείμενο. Η βαθμολόγηση των προτάσεων του κειμένου γίνεται βάσει των keywords όπου αυτές περιέχουν και αφορά στις παρακάτω σημαντικές παραμέτρους:

- υπάρχει το keyword και στον τίτλο του κείμενου
- υπάρχει πληροφορία για την κατηγορία που ανήκει το κείμενο
- υπάρχει πληροφορία για τις προτιμήσεις του χρήστη σε κατηγορία ή keywords

Το ζύγισμα των παραπάνω παραμέτρων είναι κεφαλαιώδους σημασίας για τον μηχανισμό αυτόματης εξαγωγής περίληψης καθώς η εύρεση των βέλτιστων παραγόντων που θα χρησιμοποιηθούν θα κρίνει και το σκορ που θα λάβουν οι προτάσεις, επομένως και την περίληψη του κειμένου.

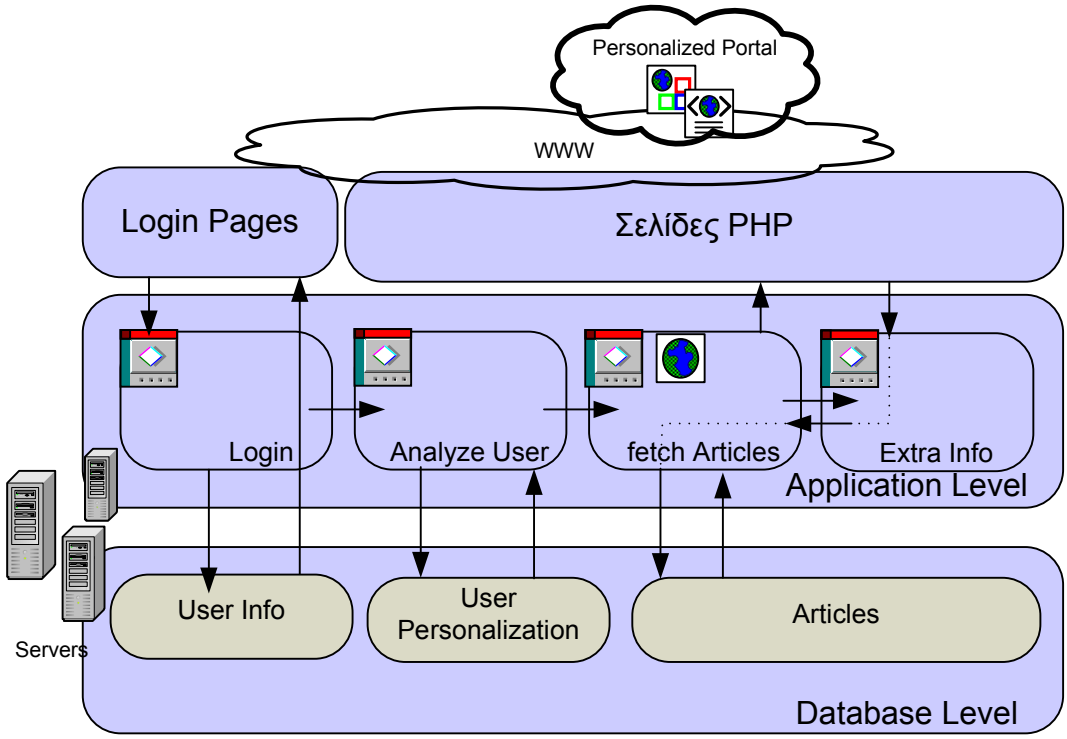
Ένα άλλο σημαντικό θέμα είναι η σειρά εμφάνισης των προτάσεων στην τελική περίληψη που προκύπτει. Είναι πιθανό, προτάσεις που βρίσκονται όχι στην αρχή του κειμένου να είναι πιο αντιπροσωπευτικές του νοήματος του κειμένου και επομένως να λαμβάνουν υψηλότερο σκορ από το μηχανισμό σε σχέση με άλλες οι οποίες βρίσκονται νωρίτερα στο κείμενο. Η παρουσίαση όμως τυχαίων προτάσεων στον τελικό χρήστη, κάθε άλλο παρά κατανοητή περίληψη είναι. Είναι σωστότερο επομένως, αφού έχει επιλεχθεί το πλήθος των προτάσεων που θα απαρτίζουν μια περίληψη, να γίνει μια ταξινόμησή τους σε σχέση με τη σειρά εμφάνισής τους στο κείμενο, διατηρώντας έτσι τη νοηματική συνοχή του κειμένου πριν παρουσιαστούν στον τελικό χρήστη.

5.2.6. Παρουσίαση Πληροφορίας και Προσωποποίηση στο χρήστη

Το τελικό στάδιο της αρχιτεκτονικής του συστήματος είναι η παρουσίαση της πληροφορίας στον τελικό χρήστη. Πρόκειται ίσως για το πιο σημαντικό στάδιο του συστήματος καθώς αποτελεί το περιβάλλον διεπαφής με τους χρήστες. Απώτερος σκοπός είναι ο χρήστης να μην αντιλαμβάνεται όλες τις διεργασίες που λαμβάνουν χώρα και να απολαμβάνει ποιοτικά και γρήγορα αποτελέσματα βάση των προσωπικών του επιλογών. Για την προσωποποίηση στο χρήστη μπορούν να χρησιμοποιηθούν δύο μέθοδοι:

- Ο χρήστης να δώσει κάποια πληροφορία στο σύστημα και το σύστημα να ξεκινήσει παρουσιάζοντας εξ αρχής προσωποποιημένα αποτελέσματα και να συγκλίνει γρήγορα στις ανάγκες του χρήστη.
- Ο χρήστης να μη δώσει καθόλου πληροφορία στο σύστημα και το σύστημα να ξεκινήσει παρουσιάζοντας γενικές πληροφορίες και να αργήσει να συγκλίνει στις προσωπικές επιλογές του χρήστη.

Σε κάθε περίπτωση το επιθυμητό επιτυγχάνεται και πρόκειται για τη σύγκλιση των πληροφοριών που παρουσιάζονται στις ανάγκες του χρήστη. Το παρακάτω σχήμα απεικονίζει την αρχιτεκτονική του συγκεκριμένου μηχανισμού.



Εικόνα 14: Αρχιτεκτονική της Προσωποποιημένης Πύλης

Όπως φαίνεται και από το παραπάνω σχήμα η αρχιτεκτονική του μηχανισμού βασίζεται στην συνεχή ανάδραση και δεν είναι στατικός ο τρόπος παρουσίασης των δεδομένων στον τελικό χρήστη. Αφού ο χρήστης κάνει log-in στο σύστημα εμφανίζεται κάποια πληροφορία η οποία είναι προσωποποιημένη στο χρήστη ανάλογα με τις επιλογές που έχει κάνει. Έτσι του εμφανίζονται αρχικά άρθρα από τις κατηγορίες που έχει επιλέξει ενώ άρθρα από κατηγορίες που δεν έχει επιλέξει έχουν πολύ χαμηλή βαθμολογία και δε συμπεριλαμβάνονται καθόλου στα πρώτα αποτελέσματα. Κάθε ενέργεια που πραγματοποιεί ο χρήστης στο δικτυακό τόπο καταγράφεται σε ένα log του οποίου η ανάλυση οδηγεί στα συμπεράσματα για τις επιλογές και μη του χρήστη προκειμένου το σύστημα να μπορεί να διαμορφώσει αυτόματα το προφίλ του χρήστη.

6

ΤΕΧΝΟΛΟΓΙΕΣ ΥΛΟΠΟΙΗΣΗΣ

Στο κεφάλαιο αυτό οι τεχνολογίες που χρησιμοποιήθηκαν για την υλοποίηση κάθε υποσυστήματος του μηχανισμού.

6. ΤΕΧΝΟΛΟΓΙΕΣ ΥΛΟΠΟΙΗΣΗΣ

Οι τεχνολογίες που χρησιμοποίηθηκαν σε κάθε επίπεδο του μηχανισμού είναι διαφορετικές προκειμένου να επιτευχθεί η μέγιστη απόδοση συνολικά του συστήματος με τη χρήση κάθε μίας από αυτές.

Η επιλογή της τεχνολογίας που θα ακολουθηθεί κατά την κατασκευή ενός σύνθετου συστήματος είναι εξαιρετικά σημαντική προκειμένου να δημιουργηθεί ένα καθολικό σύστημα το οποίο να είναι ευέλικτο, να υποστηρίζει εύκολα αλλαγές και αναβαθμίσεις, να αποτελείται από υποσυστήματα και τέλος να βασίζεται σε ανοιχτά πρότυπα. Το σύστημα που υλοποιήθηκε είναι σύνθετο καθότι έχει βάση το διαδίκτυο αλλά ένα σημαντικό κομμάτι του, ίσως ο πυρήνας, κρύβεται στο μηχανισμό που πραγματοποιεί κατηγοριοποίηση κειμένου και γενικότερα διαχείριση πληροφορίας. Ο τελευταίος μηχανισμός ουσιαστικά δεν έχει καμία επαφή με το διαδίκτυο και φυσικά δεν είναι και απαραίτητο να έχει. Βέβαια, τα δεδομένα που δέχεται προέρχονται από εξόρυξη πληροφορίας στο διαδίκτυο (HTML σελίδες) ενώ τα δεδομένα που εξάγει χρησιμοποιούνται προκειμένου να τροφοδοτήσουν το portal με περιεχόμενο.

6.1. Βάση Δεδομένων

Οι πιθανές επιλογές που έχουμε όσον αφορά τη βάση δεδομένων του συστήματος προέρχονται από την επιλογή των τεχνολογιών για τους μηχανισμούς κατηγοριοποίησης και κατασκευής του Portal. Συνεπώς θα πρέπει να επιλεγεί μία βάση δεδομένων η οποία να είναι πλήρως συμβατή με το μηχανισμό που θα κατηγοριοποιεί καθώς επίσης και με τη γλώσσα προγραμματισμού που θα χρησιμοποιηθεί για την κατασκευή του portal. Θεωρούμε πως ο μηχανισμός δημιουργίας του δυναμικού προφίλ δύναται να ενταχθεί, είτε στο μηχανισμό κατηγοριοποίησης είτε στο μηχανισμό κατασκευής του portal.

6.1.1. Γιατί MySQL

Η MySQL είναι η δημοφιλέστερη Βάση Δεδομένων ανοιχτού κώδικα που προσφέρεται από το Δίκτυο MySQL. Η αρχιτεκτονική της την κάνουν να είναι εξαιρετικά γρήγορη και πολύ εύκολη σε αλλαγές και αναβαθμίσεις. Επιτρέπει επαναχρησιμοποίηση κώδικα όπου αυτό είναι αναγκαίο και παρέχει ένα μινιμαλιστικό τρόπο δημιουργίας στοιχείων διαχείρισης βάσης δεδομένων τέτοια ώστε να κάνουν τη MySQL ασύγκριτη σε ταχύτητα, σε κατάληψη χώρου, σταθερότητα και ευκολία. Ο μοναδικός στο είδος του διαχωρισμός του κεντρικού πυρήνα του server από το μηχανισμό αποθήκευσης κάνει δυνατή την ύπαρξη αυστηρού ελέγχου σε συναλλαγές και μείωση ταχύτητας ή ύπαρξη θεαματικά μεγάλης ταχύτητας με απευθείας προσπέλαση των δεδομένων στοιχεία που μπορεί να χρησιμοποιηθούν ανάλογα με τις ανάγκες των χρηστών.

Η MySQL περιλαμβάνει αποθήκευση σε μηχανή InnoDB, η οποία υποστηρίζει ασφάλεια στις συναλλαγές και ACID-συμβατή μηχανή αποθήκευσης με commit, rollback, crash recovery και low-level locking δυνατότητες.

Η έκδοση της MySQL που βρίσκεται αυτή τη στιγμή σε σταθερή κατάσταση είναι η 4.1.12 και υποστηρίζει πολλά στοιχεία που αφορούν την απόδοση, τη διεθνοποίηση και τη δυνατότητα ένταξης του MySQL server σε άλλα στοιχεία υλικού και λογισμικού. Τα πιο βασικά στοιχεία που χαρακτηρίζουν τη MySQL είναι:

- Υποερωτήματα, που επιτρέπουν στους χρήστες να κάνουν σύνθετα ερωτήματα με μεγάλη ευκολία και αποδοτικά.
- Γρήγορη επικοινωνία μεταξύ server και client μέσα από ένα καινούριο πρωτόκολλο

- Μικρότερη κατανάλωση πόρων από το server μέσα από βελτιστοποίηση στις βιβλιοθήκες
- Υποστήριξη Unicode, διεθνείς χαρακτήρες και υποστήριξη αποθήκευσης στην πλειοψηφία των συνόλων χαρακτήρων
- Υποστήριξη τύπων GIS για ερωτήματα που αφορούν χάρτες και γεωγραφικά δεδομένα

Τα παραπάνω στοιχεία κάνουν τη MySQL ένα υπερπολύτιμο εργαλείο στα χέρια κάποιου χρήστη και τη θέτουν στην 1η θέση για επιλογή ως βάση δεδομένων του συστήματός μας. [21]

6.1.2. Γιατί PostgreSQL

Η PostgreSQL είναι μια σχεσιακή βάση δεδομένων βασισμένη στα αντικείμενα. Ουσιαστικά προέρχεται από την POSTGRES, V 4.2, που έχει δημιουργηθεί στο πανεπιστήμιο της Καλιφόρνια στο τμήμα Επιστήμης των Υπολογιστών του Μπέρκλεϋ. Μάλιστα το συγκεκριμένα σύστημα υλοποίησε πολλές λειτουργικότητες πολλά χρόνια πριν εφαρμοστούν στα πιο γνωστά από τα σημερινά συστήματα βάσεων δεδομένων.

Η PostgreSQL είναι ένας ανοιχτού κώδικα απόγονος του αρχικού κώδικα που γράφηκε στο Μπέρκλεϋ. Υποστηρίζει SQL92 και SQL99 και προσφέρει πολλά στοιχεία που υποστηρίζουν οι περισσότερες βάσεις δεδομένων τελευταίας τεχνολογίας όπως:

- Σύνθετα ερωτήματα
- Foreign Keys
- Triggers
- Διαφορετικές όψεις
- Ακεραιότητα στις συναλλαγές
- Συνεργασία ταυτόχρονων πολλαπλών εκδόσεων

Επιπρόσθετα, η PostgreSQL μπορεί να εμπλουτιστεί σε στοιχεία από κάποιον έμπειρο χρήστη με πολλούς τρόπους ώστε να υποστηρίζει νέα:

- Τύπους δεδομένων
- Συναρτήσεις
- Διαχειριστές
- Συναθροιστικές συναρτήσεις
- Μεθόδους ευρετηρίου
- Διαδικασιακές γλώσσες

Τέλος, αξίζει να τονιστεί η γενναιοδωρία της άδειας κάτω από την οποία βρίσκεται η PostgreSQL σύμφωνα με την οποία μπορεί να χρησιμοποιηθεί, αλλαχθεί και διακινηθεί από τον καθένα χωρίς κανένα κόστος. [22]

6.1.3. Επιλέγοντας τη Βάση Δεδομένων

Σύμφωνα με τα παραπάνω αλλά και λαμβάνοντας υπόψη μας τους σκοπούς που έχει το σύστημά μας καταλήξαμε στην επιλογή της MySQL σαν τη βάση δεδομένων που θα χρησιμοποιηθεί στο σύστημα. Συγκρίνοντας τις δύο βάσεις δεδομένων μπορούμε να καταλήξουμε στο ότι διαθέτουν πολλά κοινά στοιχεία, ωστόσο η MySQL φαίνεται να είναι πιο διαδεδομένη, ένας λόγος ο οποίος την κάνει πιο ισχυρή. Επιπρόσθετα τα στοιχεία διεθνοποίησης που διαθέτει φαίνονται πολύ χρήσιμα για ένα σύστημα το οποίο μελλοντικά μπορεί να επεκταθεί ώστε να υποστηρίζει πολλές γλώσσες. Ένα άλλο στοιχείο που μας οδηγεί στην επιλογή της MySQL είναι και το γεγονός πως οι βοηθητικοί crawlers που τροφοδοτούν το σύστημά μας με σελίδες HTML υποστηρίζουν βάση δεδομένων MySQL. Τέλος θα πρέπει να λάβουμε υπόψη μας το γεγονός πως δημιουργούμε ένα σύστημα

πολυεπίπεδο με τη βάση δεδομένων να είναι ο ουσιαστικός σύνδεσμος μεταξύ των περισσοτέρων κομματιών και συνεπώς μία βάση δεδομένων με μεγάλη σταθερότητα και αξιοπιστία θα προσέδιδε κύρος στο συνολικό σύστημα.

Καταλήγουμε λοιπόν στη χρήση Mysql Server έκδοση 4.12. [21]

6.2. Τεχνολογία Μηχανισμού Κατηγοριοποίησης

Ο μηχανισμός κατηγοριοποίησης είναι ένα σύστημα το οποίο αναλαμβάνει μια πολύ μεγάλη και επίπονη διαδικασία. Προκειμένου να καταλάβουμε τι τεχνολογία πρέπει να χρησιμοποιηθεί θα συνοψίσουμε της εργασίες του μηχανισμού σε μία παράγραφο.

Ο μηχανισμός κατηγοριοποίησης διαβάζει την έξοδο ενός crawler που είναι απλές HTML σελίδες. Εν συνεχεία εξάγει το κείμενο από αυτές και το φιλτράρει προκειμένου να βρει τις λέξεις-κλειδιά. Αυτές τις αποθηκεύει στη βάση δεδομένων μαζί με τα βάρη που έχει κάθε μία για κάθε κείμενο. Τέλος υπολογίζει την κατηγορία που ανήκει ένα κείμενο από τη συσχέτιση που έχει ο πίνακας με τις λέξεις ενός κειμένου με τα βάρη τους με τις αντίστοιχες λέξεις και τα βάρη τους για την κατηγορία.

Συνεπώς ένας τέτοιος μηχανισμός θα πρέπει να μπορεί να επικοινωνήσει άμεσα και γρήγορα με τη βάση καθώς και να κάνει γρήγορους υπολογισμούς όπου αυτοί είναι απαραίτητοι. Το ερώτημα που τίθεται εδώ είναι αν θα χρησιμοποιηθεί κάποια αντικειμενοστραφής γλώσσα ή μία γλώσσα διαδικασιακή.

6.2.1. Γιατί C

Η επιλογή της C μπορεί να γίνει για ένα σύνολο από λόγους μεταξύ των οποίων είναι οι εξής: Η C μπορεί να χρησιμοποιηθεί σαν χαμηλού επιπέδου γλώσσα προγραμματισμού επιτρέποντας άμεση πρόσβαση στους πόρους του υπολογιστή και άρα στην αποτελεσματική και χωρίς overhead αξιοποίησή τους. Εξάλλου, είναι η καθιερωμένη γλώσσα για χαμηλού επιπέδου προγραμματισμό που ένας μηχανικός θα απαιτηθεί να κάνει για την καλύτερη αξιοποίηση του υλικού που σχεδιάζει και αναπτύσσει. Ταυτόχρονα, μπορεί να χρησιμοποιηθεί και σαν γλώσσα υψηλού επιπέδου καθώς η πληθώρα των διαθέσιμων βιβλιοθηκών υπερκαλύπτουν τις απαιτήσεις ανάπτυξης λογισμικού επιπέδου εφαρμογής (Application Layer Software). Επίσης είναι σχετικά μικρή και εύκολη στην εκμάθηση, υποστηρίζει top-down και modular σχεδιασμό, υποστηρίζει δομημένο (structured) προγραμματισμό και είναι αποτελεσματική (efficient) αφού παράγει συμπαγή και γρήγορα στην εκτέλεση προγράμματα. Ακόμα είναι φορητή (portable), ευέλικτη (flexible), ισχυρή (powerful), δε βάζει περιορισμούς, γεγονός που συχνά αποβαίνει σε βάρος της και αποτελεί με τη C++ την ευρύτερα χρησιμοποιούμενη γλώσσα σε ερευνητικά και αναπτυξιακά προγράμματα. Να αναφέρουμε ακόμα ότι υπάρχει μία πολλή μεγάλη εγκατεστημένη βάση εφαρμογών που αναπτύχθηκαν με τη γλώσσα αυτή και πρέπει να συντηρούνται και να εξελίσσονται και τέλος η γνώση της C αποτελεί ένα πολύ καλό εφόδιο για την εκμάθηση της Java καθώς αυτή υιοθετεί το μεγαλύτερο ποσοστό των δομικών στοιχείων της C [23].

6.2.2. Γιατί C++

Πρόκειται μία γλώσσα προγραμματισμού που δημιουργήθηκε ως κύριος αντίπαλος της Java και προφανώς υποστηρίζει αντικειμενοστραφή προγραμματισμό. Από το 1998 το C++ Standard αποτελείται από δύο κομμάτια: ο πυρήνας και οι βασικές βιβλιοθήκες. Η τελευταία έκδοση περιέχει βασικές βιβλιοθήκες της C++ και ένα μεγάλο κομμάτι από τις βασικές βιβλιοθήκες της C. Παράλληλα υπάρχουν πολλές βιβλιοθήκες που έχουν συγκεκριμένους σκοπούς και επικεντρώνονται σε συγκεκριμένα στοιχεία και δεν περιλαμβάνονται στις Standard βιβλιοθήκες. Αξιοσημείωτο είναι και το γεγονός ότι είναι σχετικά απλό να ενταχθούν βιβλιοθήκες της C μέσα σε προγράμματα γραμμένα σε C++.

Είναι πολύ σημαντικό να γίνει κατανοητό, πως δεν υπάρχει πλέον μία μοναδική γλώσσα που να ονομάζεται C++. Ο όρος αντιπροσωπεύει μία οικογένεια παρόμοιων γλωσσών οι οποίες είναι συχνά υπό- ή υπέρ- σύνολα μεταξύ τους.

Βασικά στοιχεία της C++ περιλαμβάνουν δηλώσεις, function-like casts, inline functions, function overloading, classes, exception handling κ.α. Η C++ συνήθως πραγματοποιεί μεγαλύτερο έλεγχο τύπων σε μεταβλητές απ' ότι η C. Πολλά στοιχεία της C++ τα υιοθέτησε και η C ωστόσο η C99 παρουσίασε πολλά στοιχεία που δεν υιοθετήθηκαν ούτε και υπάρχουν στην C++. Μία πολύ συνηθισμένη πηγή σύγχυσης είναι το ζήτημα ορολογίας: εξαιτίας της παραγωγής από τη C, στη C++ ο όρος αντικείμενο σημαίνει περιοχή μνήμης, όπως και στη C, και όχι ένα class instance, κάτι το οποίο συμβαίνει στις περισσότερες γλώσσες προγραμματισμού. [24]

6.2.3. Γιατί Java

Αντίστοιχα, η επιλογή της Java μπορεί να γίνει για ένα σύνολο από λόγους μεταξύ των οποίων είναι οι εξής: Αναπτύχθηκε κατ' αρχήν ως γλώσσα για ανάπτυξη ενσωματωμένου λογισμικού (embedded software) και καλύπτει τις αντίστοιχες ανάγκες ενός Μηχανικού συστημάτων. Είναι φορητή, γεγονός που διασφαλίζει τη δυνατότητα εκτέλεσης των Java προγραμμάτων ανεξάρτητα πλατφόρμας υλικού και λογισμικού. Επίσης διαθέτει πολύ μεγάλη βιβλιοθήκη έτοιμων κλάσεων, οι οποίες διευκολύνουν σε μεγάλο βαθμό τη γρήγορη ανάπτυξη αξιόπιστων εφαρμογών και γνωρίζει ραγδαία εξάπλωση σε ερευνητικά και αναπτυξιακά προγράμματα. Ακόμα μπορεί να χρησιμοποιηθεί για προγραμματισμό στο διαδίκτυο και όσον αφορά την υποστήριξη της Αντικειμενοστραφούς Προσέγγισης είναι πολύ πιο καθαρή από τη C++ και έτσι θα μπορούσε να θεωρηθεί σαν λογική συνέχεια της C. Τέλος υιοθετεί μεγάλο μέρος της C.

Η Java παρουσιάστηκε σαν μία γλώσσα που είχε αφαιρέσει τα «βρώμικα» στοιχεία της C++ και είχε εισάγει ένα σύνολο από καλά στοιχεία άλλων γλωσσών όπως η Smalltalk. Η ιστορία της γλώσσας ξεκίνησε όταν μία ομάδα ερευνητών στην προσπάθειά της να αναπτύξει ενσωματωμένο λογισμικό (embedded software) για έξυπνες καταναλωτικές συσκευές στα πλαίσια του project Green, αποφάσισε να αναπτύξει μία νέα γλώσσα μετά τη διαπίστωσή της ότι η C και η C++ δεν ανταποκρίνονται στις απαιτήσεις της. Έτσι τον Αύγουστο του 1991 εμφανίστηκε μία νέα αντικειμενοστραφής γλώσσα με το όνομα OAK, που είναι το ακρωνύμιο του Object Application Kernel. Η γλώσσα απλά προστέθηκε στον κατάλογο των καλών γλωσσών προγραμματισμού με ουσιαστική υποστήριξη σε εφαρμογές τύπου πελάτη-εξυπηρετητή (client-server) και τίποτα παραπάνω.

Μόλις τον Απρίλιο του 1993 έκανε την εμφάνισή του το NCSA MOSAIC 1.0 ως πρώτο γραφικό πρόγραμμα πλοήγησης στο διαδίκτυο (Web browser) και έτσι η γλώσσα άρχισε να κάνει τα πρώτα της βήματα στο χώρο του διαδικτύου με πολύ θετικά αποτελέσματα. Το στοιχείο αυτό ώθησε τη Sun, μετά από μία αποτυχημένη προσπάθειά της να πουλήσει τη γλώσσα (Αύγουστος 93), να χρηματοδοτήσει την ανάπτυξή της για το 1994, αν και το προηγούμενο έτος είχε διακόψει ως μη επιτυχημένο το αντίστοιχο project. Στα μέσα του 1994, αναπτύχθηκε το πρώτο πειραματικό πρόγραμμα πλοήγησης με Java κάτω από το όνομα του WebRunner. Το φθινόπωρο του ίδιου έτους, ο Van Hoff υλοποιεί με Java τον πρώτο Java διερμηνευτή.

Μόλις τον Ιανουάριο του 1995, η γλώσσα πήρε τη σημερινή της ονομασία και εμφανίστηκε η πρώτη επίσημη τεκμηρίωσή της με τη μορφή ενός "white paper". Το Μάιο του ίδιου έτους, η Sun παρουσιάζει επίσημα τη Java και το HotJava. Ταυτόχρονα, η Netscape αγόρασε άδεια χρήσης της Java και ενσωμάτωσε τη γλώσσα στη δεύτερη έκδοση του Netscape, του γνωστού προγράμματος πλοήγησης. Στη συνέχεια, ο ένας μετά τον άλλο, οι μεγάλοι κατασκευαστές λογισμικού ανακοίνωσαν την απόφασή τους να χρησιμοποιήσουν τη Java, με αποκορύφωμα την απόφαση της Microsoft το Δεκέμβρη του 1995. Η Java

καθιερώθηκε πια ως η γλώσσα που θα πρωτοστατήσει στην ερχόμενη δεκαετία. Μία αναλυτική αναφορά στο χρονικό της εξέλιξης της γλώσσας μπορείτε να βρεθεί στο [24]

6.2.4. Γιατί Perl

Η Perl είναι μια γενικού σκοπού γλώσσα προγραμματισμού που αρχικά δημιουργήθηκε για την επεξεργασία κειμένου και τώρα χρησιμοποιείται σε μια πλειάδα συστημάτων, συμπεριλαμβανομένων των συστημάτων διαχείριση, ανάπτυξη συστημάτων δικτύου, δικτυακός προγραμματισμούς, ανάπτυξη GUI και άλλα.

Η γλώσσα αυτή σκοπεύει να είναι απλή, αποδοτική και τέλεια παρά «όμορφη». Τα κύρια στοιχεία της είναι η ευκολία στη χρήση, η υποστήριξη διαδικασιακού και αντικειμενοστραφή προγραμματισμού και παράλληλα υποστηρίζει πολύ ισχυρούς μηχανισμούς επεξεργασίας κειμένου.

Η γενικότερη δομή της προέρχεται κυρίως από τη γλώσσα προγραμματισμού C. Είναι μια διαδικασιακή γλώσσα προγραμματισμού που χρησιμοποιεί μεταβλητές, παραστάσεις, αποδόσεις, μπλοκ κώδικα, συναρτήσεις ελέγχου και υπορουτίνες.

Λαμβάνει υπόψη της τον προγραμματισμό σε shell και τα προγράμματα σε perl είναι μεταφραζόμενα. Όλες οι μεταβλητές διαχωρίζονται με ένα συγκεκριμένο χαρακτηριστικό που προηγείται αυτών, επιτρέποντας έτσι καλύτερη σύνταξη. Όπως και το shell του UNIX, η Perl έχει πολλές έτοιμες συναρτήσεις οργανωμένες σε βιβλιοθήκες που αναλαμβάνουν τις περισσότερες απλές εργασίες όπως ταξινόμηση ή διασύνδεση με λειτουργίες του συστήματος.

Η Perl χρησιμοποιεί συσχετιζόμενους πίνακες από το awk και «κανονικές εκφράσεις» από το sed. Αυτά τα στοιχεία απλοποιούν την ανάλυση λέξεων, τη διαχείριση κειμένου και τη διαχείριση δεδομένων.

Στην έκδοση 5 της perl, προστέθηκαν στοιχεία για να υποστηρίζουν σύνθετους τύπους δεδομένων και δομές δεδομένων καθώς επίσης και μοντέλα αντικειμενοστραφούς προγραμματισμού.

Σε όλες τις εκδόσεις της perl ο τύπος δεδομένων μίας μεταβλητής βρίσκεται αυτόματα, ενώ αυτόματη είναι και η διαχείριση της μνήμης. Ο μεταφραστής γνωρίζει τον τύπο και τις απαιτήσεις σε αποθηκευτικό χώρο για κάθε τύπο του προγράμματος. Καθορίζει το χώρο που θα καταλαμβάνει κάθε πρόγραμμα και απελευθερώνει πόρους όποτε αυτό είναι εφικτό. Επιτρεπόμενες μετατροπές μεταξύ τύπων γίνονται αυτόματα.

Τα παραπάνω βέβαια σημαίνουν ότι δεν επιτρέπονται διαρροές στη μνήμη, σταμάτημα του μεταφραστή ή να διακοπεί η αναπαράσταση των εσωτερικών δεδομένων [25].

6.2.5. Επιλογή της τεχνολογίας υλοποίησης

Δεδομένων των παραπάνω και βάση των απαιτήσεων που έχει το σύστημά μας δεν καταλήγουμε σε μία γλώσσα υλοποίησης για τα υποσυστήματα που σχεδιάζουμε αλλά σε δύο. Έτσι, λοιπόν, δεδομένου ότι η γλώσσα C++ είναι πιο κοντά στις διαδικασίες πυρήνα ενός συστήματος τη χρησιμοποιούμε για να υλοποιήσουμε τις διαδικασίες προεπεξεργασίας, κατηγοριοποίησης και αυτόματης εξαγωγής περίληψης προκειμένου να επιτύχουμε μέγιστη χρήση των πόρων του συστήματος αλλά και των προτερημάτων της C++. Οι διαδικασίες αυτές απαιτούν μεγάλη υπολογιστική ισχύ και γρήγορη εκτέλεση σύνθετων αλγορίθμων σε πραγματικό χρόνο γεγονός που μας οδηγεί στη χρήση C++. Για διαδικασίες του συστήματος οι οποίες εκτελούνται σε επίπεδο εφαρμογής και δεν απαιτούν άμεση αντίδραση από το μηχανισμό όπως είναι η συλλογή δεδομένων από το Διαδίκτυο επιλέγουμε τη γλώσσα προγραμματισμού Java η οποία προσφέρει περισσότερη ευελιξία σε τέτοιου είδους εφαρμογές. Ο mixed crawler, λοιπόν, είναι υλοποιημένος με τη χρήση της γλώσσας προγραμματισμού Java. Για να επιτύχουμε

επικοινωνία του μηχανισμού που χρησιμοποιεί Java με τους μηχανισμούς που υλοποιήθηκαν σε C++ χρησιμοποιήσαμε κεντρικοποιημένη βάση δεδομένων ενώ θέσαμε συγκεκριμένο τρόπο σειριακής εκτέλεσης των διαδικασιών προκειμένου να διατηρηθεί η ακεραιότητα του συστήματος.

6.3. Τεχνολογία Δημιουργίας Portal

Όσον αφορά την τεχνολογία που θα χρησιμοποιηθεί για τη δημιουργία του portal θα πρέπει να επισημανθεί ότι θα χρησιμοποιηθεί κάποια τεχνολογία δημιουργίας δυναμικών σελίδων. Οι σελίδες θα πρέπει να έχουν απλή δομή και κατανοητή προκειμένου να μην αποπροσανατολίζεται ο χρήστης. Για το σκοπό αυτό η δυνατότητα που μας δίνεται είναι να χρησιμοποιήσουμε μία εκ των PHP ή JSP. Η τεχνολογία ASP.NET αποκλείεται γιατί αίρει το χαρακτήρα ανοικτού κώδικα που βασίζεται σε ανοικτά στάνταρ.

6.3.1. Γιατί PHP

Η ευκολία στη χρήση αλλά και η ομοιότητα με της πιο κοινές γλώσσες δομημένου προγραμματισμού κάνουν την PHP μία γλώσσα η οποία ελκύει τους προγραμματιστές και οι πιο έμπειροι από αυτούς βρίσκουν εύκολη τη δημιουργία σύνθετων εφαρμογών από την πρώτη στιγμή που θα έρθουν σε επαφή με την PHP. Επίσης επιτρέπει στους έμπειρους χρήστες να δημιουργήσουν εφαρμογές Διαδικτύου με δυναμικό περιεχόμενο χωρίς να χρειάζεται να αναλωθούν σε πρακτικές ή να χρειαστεί να αποστηθίσουν σειρές από συναρτήσεις.

Ένα από τα πιο ελκυστικά κομμάτια της PHP είναι το γεγονός ότι είναι κάτι περισσότερο από μια προγραμματιστική γλώσσα. Εξαιτίας της κλιμακωτής σχεδίασής της, μπορεί να χρησιμοποιηθεί και για τη δημιουργία γραφικών περιβαλλόντων απεικόνισης, και για την εκτέλεση προγραμμάτων μέσω της γραμμής εντολών

Η PHP επιτρέπει την αλληλεπίδραση με ένα μεγάλο αριθμό σχεσιακών βάσεων δεδομένων όπως είναι οι Mysql, Oracle, IBM DB2, Microsoft SQL Server, PostgreSQL και SQLite ενώ η σύνταξη που χρησιμοποιείται είναι απλή και κατανοητή. Τρέχει στα περισσότερα λειτουργικά συστήματα όπως UNIX, Linux, Windows και Mac OS X και μπορεί να υποστηριχθεί σχεδόν από όλους τους γνωστούς εξυπηρετητές εφαρμογών Διαδικτύου.

Η PHP είναι αποτέλεσμα μίας σειράς προσπαθειών από πολλούς συμμετέχοντες. Τα δικαιώματα παρέχονται με ένα SD-style license. Τέλος, μετά την έκδοση 4 η PHP υποστηρίζεται από τη μηχανή Zend [26].

6.3.2. Γιατί JSP

Η JSP έρχεται σαν απάντηση της Java στις τεχνολογίες εφαρμογών διαδικτύου. Χρησιμοποιεί τεχνολογία που βασίζεται είτε σε Java Servlets ή σε Java Beans και προσφέρει δυνατότητα ανάλογα με την επιλογή της τεχνολογίας να δημιουργηθούν από πολύ απλές Διαδικτυακές εφαρμογές μέχρι πολύ σύνθετες.

Όσον αφορά την αρχιτεκτονική, η jsp μπορεί να θεωρηθεί σαν servlet με πολύ υψηλού επιπέδου αφαίρεση η οποία υλοποιείται σαν επέκταση του API 2.1 των Servlet.

Όσον αφορά τη σύνταξη, μία σελίδα γραμμένη σε JSP μπορεί να χωριστεί στα εξής κομμάτια

Στατικό περιεχόμενο (π.χ. HTML)

- JSP directives
- JSP μεταβλητές και στοιχεία κώδικα
- JSP action
- Tags γραμμένα από το χρήστη

Πρόκειται για τη γλώσσα προγραμματισμού που χρησιμοποιείται στις περισσότερες σύνθετες εφαρμογές που δημιουργούνται στο Διαδίκτυο γιατί προσφέρει τη δυνατότητα με τη χρήση συνδυασμού καθαρής Java, μέσω των Beans και μίας C-like γλώσσας προγραμματισμού για τη δημιουργία απλού δυναμικού περιεχομένου. Ωστόσο προορίζεται κυρίως για έμπειρους χρήστες που μπορούν να καταλάβουν τη διαφορά αντικειμενοστραφούς και συναρτησιακού προγραμματισμού και να τα συνδυάσουν κατάλληλα προκειμένου να επιτευχθεί το επιθυμητό αποτέλεσμα [27].

6.4. Τελική επιλογή τεχνολογιών

Η τελική επιλογή τεχνολογιών όπως αναφέρθηκε και στην αρχή του κεφαλαίου βασίζεται στο γεγονός ότι θα γίνει συνδυασμός τεχνολογιών που θα συνδυάζουν καθαρό αντικειμενοστραφή κώδικα με σελίδες του διαδικτύου. Θα μπορούσε κανείς να πει πως η επιλογή Java, JSP και Oracle θα ήταν ιδανικός για ένα τέτοιο σύστημα καθότι είναι εκ των πραγμάτων τεχνολογίες που η δυνατότητα διασύνδεσής τους είναι εύκολη και οι δυνατότητες που προσφέρει ο συγκεκριμένος συνδυασμός είναι πολλές.

Ωστόσο, επειδή ακριβώς τα υποσυστήματα που απαρτίζουν το μηχανισμό που δημιουργήσαμε μπορούν να λειτουργήσουν ανεξάρτητα και αυτόνομα, η επιλογή των τεχνολογιών έγινε περισσότερο βάση γενικών αρχών και προτύπων προκειμένου να καταλήξουμε σε ένα τελικό σύστημα ανοιχτό, και ευέλικτο το οποίο θα μπορεί να επιδέχεται βελτιώσεις σε κάθε κομμάτι του ξεχωριστά. Έγινε, δηλαδή, προσπάθεια να μη δημιουργηθούν επικαλύψεις στον κώδικα αλλά η διασύνδεση των υποσυστημάτων να γίνει σε επίπεδο βάσης δεδομένων. Αυτό βέβαια δε μας απαγορεύει να χρησιμοποιούμε ένα κεντρικό μηχανισμό που θα κάνει διαχείριση όλων των υποσυστημάτων. Συνεπώς καταλήγουμε σε γλώσσα διαδικτύου PHP με υποστήριξη βάσης δεδομένων MySQL γιατί επιθυμούμε απλότητα σε επίπεδο web site, και σε Java και C++ με υποστήριξη βάσης δεδομένων MySQL προκειμένου να γίνονται όλες οι διαδικασίες που χρειάζονται εκτενείς αναλύσεις και υπολογισμούς.

6.5. Μηχανισμός συλλογής ειδήσεων

Όπως ήδη αναφέρθηκε για το μηχανισμό συλλογής ειδήσεων χρησιμοποιήθηκε η γλώσσα προγραμματισμού Java. Η επιλογή αυτής της γλώσσας για το συγκεκριμένο μηχανισμό είναι γιατί προσφέρει μεγάλη ευελιξία στη διαχείριση πόρων του διαδικτύου αλλά και γιατί διαθέτει APIs ανάλυσης βάση του DOM μοντέλου τόσο σελίδων HTML

6.6. Μηχανισμός εξαγωγής χρήσιμου κειμένου

Ο μηχανισμός εξαγωγής του χρήσιμου κειμένου είναι ένα επίπεδο πιο κάτω από το μηχανισμό συλλογής ειδήσεων. Πρόκειται για ένα σύστημα το οποίο δεν έχει καμία αλληλεπίδραση με το επίπεδο δικτύου, ούτε και με το επίπεδο χρήστη. Αυτό έχει σαν αποτέλεσμα να πρόκειται για μία διαδικασία που ανήκει σε αυτές χαμηλότερου επιπέδου. Η υλοποίησή της γίνεται αποκλειστικά με C++ καθότι περιέχει πληθώρα διαδικασιών γλωσσολογικής ανάλυσης, ανάλυσης κειμένου, εκτενή χρήση regular expressions και υλοποίηση αλγορίθμων για stemming.

6.7. Μηχανισμός κατηγοριοποίησης και εξαγωγής περίληψης

Ο μηχανισμός περίληψης είναι ένα σύστημα το οποίο αναλαμβάνει μια πολύ μεγάλη και επίπονη διαδικασία. Προκειμένου να καταλάβουμε τι τεχνολογία πρέπει να χρησιμοποιηθεί θα συνοψίσουμε της εργασίες του μηχανισμού σε μία παράγραφο. Ο μηχανισμός περίληψης δέχεται ως είσοδο αρχεία, ή καλύτερα, δομημένη μορφή XML με στοιχεία για το κείμενο και προχωράει σε μια διαδικασία εξαγωγής λέξεων - κλειδιά για αυτό. Ακολουθεί η διαδικασία αντιστοίχησης λέξεων

σε προτάσεις και ακολούθως η βαθμολόγηση των προτάσεων για την εξαγωγή των σημαντικότερων αυτών ώστε να προκύψει η περίληψη του κειμένου. Μια αντίστοιχη διαδικασία ακολουθείται και στην περίπτωση κατηγοριοποίησης ενός κειμένου. Ο μηχανισμός συνεχίζει με ένα επίπεδο προσωποποίησης όπου παράγεται μια προσωποποιημένη περίληψη του κειμένου και αποστέλλεται στο χρήστη (που διαθέτει κινητή συσκευή μικρού μήκους) σε κατάλληλη μορφή (π. χ. RSS Feed). Η επικοινωνία με τη βάση δεδομένων είναι διαρκής σε κάθε φάση του μηχανισμού (αποθήκευση/ανάκτηση keywords και συχνοτήτων, κειμένων, κατηγοριών, στοιχείων προσωποποίησης κ.λπ.).

Είναι φυσική συνέπεια ότι ένας τέτοιος μηχανισμός θα πρέπει να μπορεί να επικοινωνήσει άμεσα και γρήγορα με τη βάση καθώς και να κάνει γρήγορους υπολογισμούς (εσωτερικά γινόμενα, υπολογισμός μέτρων, πράξεις σε πίνακες, κ.λπ.) όπου αυτοί είναι απαραίτητοι. Το ερώτημα που τίθεται εδώ είναι αν θα χρησιμοποιηθεί κάποια αντικειμενοστραφής γλώσσα ή μία γλώσσα διαδικαστική και ποια θα μπορούσε να είναι αυτή.

6.8. Μηχανισμός παρουσίασης πληροφορίας και προσωποποίησης

Ο μηχανισμός παρουσίασης πληροφορίας και προσωποποίησης ανήκει στο κομμάτι που αφορά το δικτυακό τόπο και ως εκ τούτου υλοποιείται αποκλειστικά σε PHP που είναι και η γλώσσα κατασκευής του δικτυακού. Παράλληλα, για την καλύτερη παρουσίαση των δεδομένων στον τελικό χρήστη γίνεται εκτενής χρήση τεχνολογίας AJAX. Σε αυτό το κομμάτι έχει προκύψει αρκετές φορές το ζήτημα επιλογής τεχνολογίας καθότι μία πιο ολοκληρωμένη πρόταση θα ήταν υλοποίηση όλων των μηχανισμών σε Java και επιλογή JSP με Enterprise Java beans για το δικτυακό τόπο. Ωστόσο, η απόκριση της γλώσσας προγραμματισμού Java στις διαδικασίες πυρήνα του συστήματος μας είναι πολύ πιο αργή από τη C++. Αυτό συμβαίνει κυρίως, όπως έχει ήδη αναφερθεί, στην καλύτερη αντιμετώπιση που έχει η C++ όταν εκτελεί διαδικασίες χαμηλού επιπέδου.

6.9. Διασύνδεση μηχανισμών

Η διασύνδεση των μηχανισμών βασίζεται αποκλειστικά στο επίπεδο βάσης δεδομένων αλλά και στη σειριακή εκτέλεση των διαδικασιών που προσφέρει το λειτουργικό σύστημα. Το γεγονός ότι χρησιμοποιούνται πολλαπλά επίπεδα στην υλοποίηση είναι σωτήριο για ένα τέτοιο σύστημα καθότι υπάρχει ένα επίπεδο το οποίο είναι κοινό για όλα τα υποσυστήματα και συνεπώς είναι εφικτή η ανταλλαγή δεδομένων. Παράλληλα, όλοι οι μηχανισμοί του συστήματος έχουν σχεδιαστεί με τέτοιο τρόπο ώστε να δέχονται δεδομένα από δύο διαφορετικά κανάλια και αντίστοιχα να εξάγουν τα δεδομένα σε δύο διαφορετικά κανάλια, το ένα αυτό της βάσης δεδομένων και το άλλο σε μορφή XML. Μιλούμε για το κλασσικό πρότυπο μίας n-tier αρχιτεκτονικής η οποία επιτυγχάνει διασύνδεση των αυτόνομων μηχανισμών που την αποτελούν στο επίπεδο καναλιού επικοινωνίας. Με αυτό τον τρόπο έχουν μηχανισμούς που αποδεσμεύονται όσο αφορά το κομμάτι της υλοποίησης και δεν έχουν κανένα περιορισμό αρκεί να μπορούν να «διαβάσουν» δεδομένα από βάση δεδομένων ή από XML αρχεία και αντίστοιχα να είναι σε θέση να «γράψουν» σε βάση δεδομένων ή σε XML αρχεία.

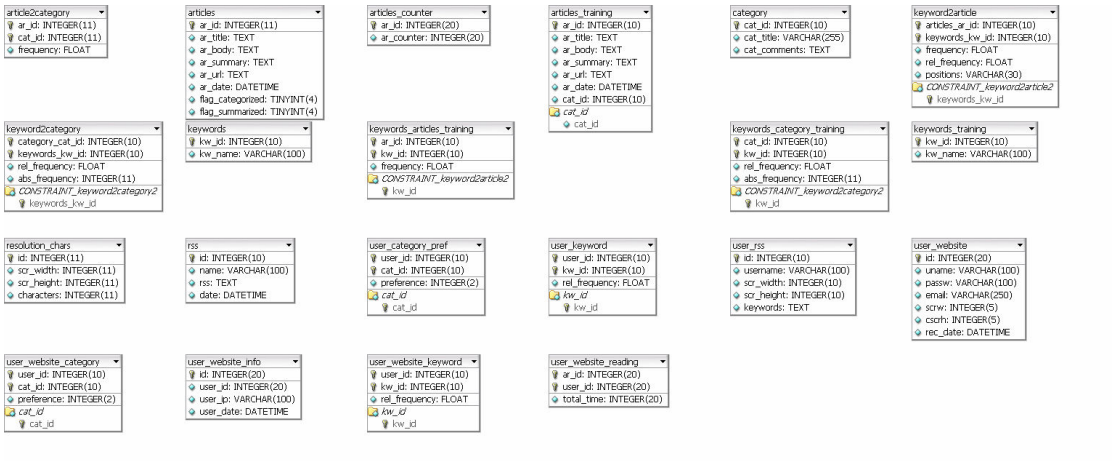
7

ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

Στο κεφάλαιο αυτό περιγράφεται η Βάση Δεδομένων του συστήματος. Πιο αναλυτικά παρουσιάζονται εκτενώς οι πίνακες που χρησιμοποιεί συνολικά το σύστημα σε όλα τα στάδια λειτουργίας του.

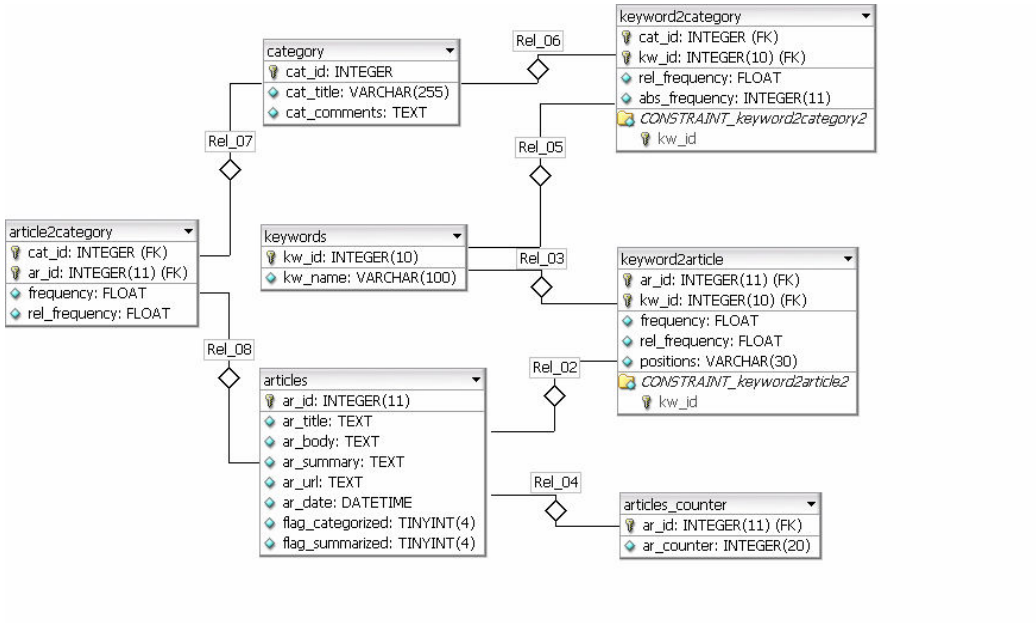
7. ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

Η βάση δεδομένων που χρησιμοποιούμε στο σύστημά μας είναι η MySQL 5.0.44 και η οποία αποτελεί και το ουσιαστικό επίπεδο διασύνδεσης μεταξύ των διαφορετικών υποσυστημάτων που έχουν υλοποιηθεί. Μία γενική εικόνα της βάσης δεδομένων φαίνεται στο παρακάτω σχήμα.

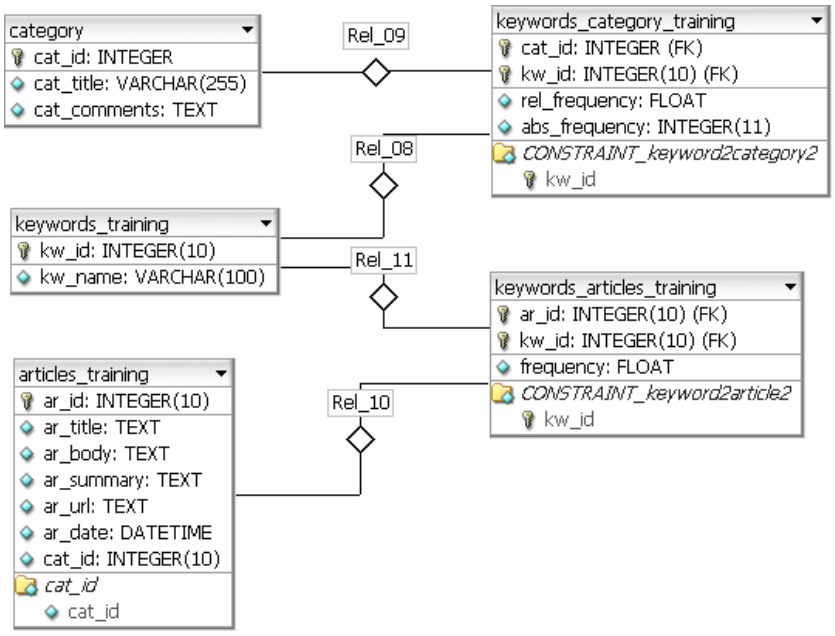


Εικόνα 15: Οι πίνακες της βάσης δεδομένων

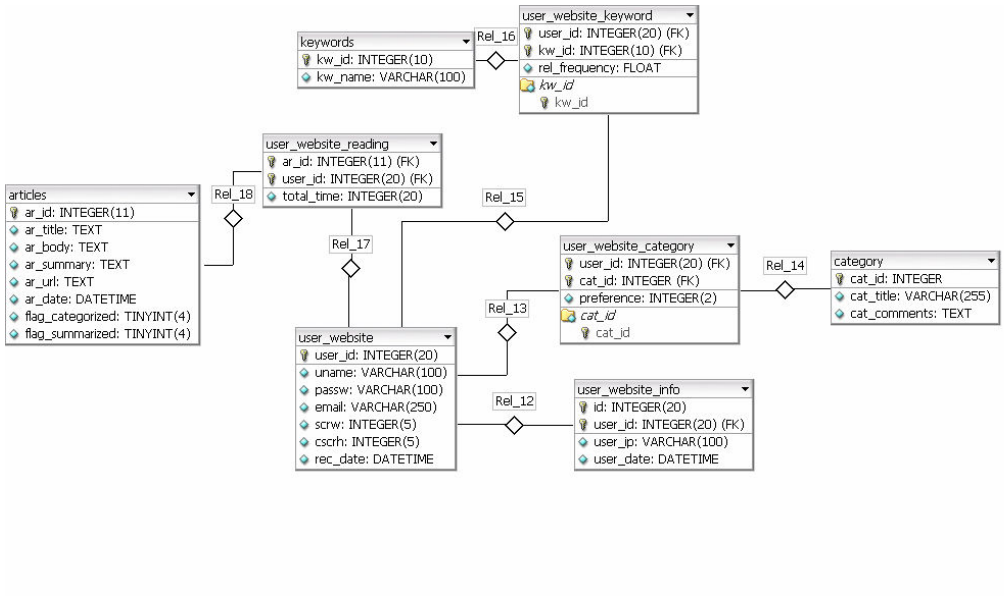
Η εικόνα της βάσης δεδομένων είναι πολύ γενική και οι πίνακές της μπορούν να ομαδοποιηθούν προκειμένου να παρουσιαστεί ο ακριβής τρόπος με τον οποίο γίνεται η αλληλεπίδραση μεταξύ των πινάκων της.



Εικόνα 16: Πίνακες που αφορούν τα άρθρα που εισέρχονται στο σύστημα



Εικόνα 17: Πίνακες που αφορούν τη βάση γνώσης του συστήματος



Εικόνα 18: Πίνακες που αφορούν τους χρήστες του συστήματος

7.1. Ανάλυση γενικών πινάκων

Στο επόμενο κομμάτι ακολουθεί η ανάλυση των πινάκων που υπάρχουν στη βάση δεδομένων και λεπτομερής παρουσίασή τους σε κάθε σημείο χρήσης τους στις διαδικασίες του συστήματος.

7.1.1. rss

Ο πίνακας rss χρησιμεύει προκειμένου να παρέχει στον mixed crawler πληροφορίες για το ποιες ιστοσελίδες θα πρέπει να προσπελάσει.

id

Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

name

Το όνομα του συγκεκριμένου rss. Καθότι είναι ένα στοιχείο που θα εμφανίζεται στις σελίδες του δικτυακού τόπου θα πρέπει να είναι μικρό και περιγραφικό

rss

Το rss feed από το οποίο ο mixed crawler θα «διαβάζει» για να εντοπίσει τις καινούριες ειδήσεις που υπάρχουν στους ειδησεογραφικούς δικτυακούς τόπους του συστήματος.

date

Η ημερομηνία κατά την οποία προστέθηκε το rss feed

7.1.2. articles

Ο πίνακας articles περιέχει όλα τα στοιχεία που αφορούν τα άρθρα που προστίθενται στο σύστημα.

ar_id

Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

ar_title

Ο τίτλος του άρθρου όπως αυτός αναγνωρίστηκε μέσα από τις σελίδες των rss feeds και όχι από την ανάλυση των σελίδων.

ar_body

Το κύριο σώμα του κειμένου ή όπως έχει ήδη αναφερθεί το Χρήσιμο Κείμενο

ar_summary

Η γενική περίληψη του άρθρου όπως προκύπτει από το μηχανισμό αυτόματης εξαγωγής περίληψης. Στην περίπτωση της δυναμικής δημιουργίας περίληψης το σύστημα τη συνθέτει σε πραγματικό χρόνο και δεν την ανακτά από τη ΒΔ.

ar_url

Το URL το οποίο οδηγεί στη σελίδα του άρθρου όπως αυτό ανακτήθηκε μέσα από το rss feed.

ar_date

Η ημερομηνία (timestamp) κατά την οποία ανακτήθηκε ένα άρθρο.

flag_categorized

Πρόκειται για μία μεταβλητή αναγνώρισης για να εντοπίσουμε ποια άρθρα έχουν κατηγοριοποιηθεί και ποια όχι προκειμένου ο μηχανισμός κατηγοριοποίησης να είναι σε θέση να αναγνωρίσει ποια άρθρα θα πρέπει να προβούν σε κατηγοριοποίηση

flag_summarized

Πρόκειται για μία μεταβλητή αναγνώρισης για να εντοπίσουμε ποια άρθρα έχουν περάσει από το μηχανισμό αυτόματης εξαγωγής περίληψης και ποια όχι προκειμένου ο μηχανισμός αυτόματης εξαγωγής περίληψης να είναι σε θέση να αναγνωρίσει ποια άρθρα θα πρέπει να προβούν σε διαδικασία αυτόματης εξαγωγής περίληψης.

7.1.3. keywords

Πρόκειται για έναν πίνακα που περιέχει όλες τις λέξεις κλειδιά που έχουν καταγραφεί στο μηχανισμό.

kw_id

Μοναδικό αναγνωριστικό κλειδί για τις εγγραφές του συγκεκριμένου πίνακα.

kw_name

Η λέξη κλειδί (stemmed).

7.1.4. category

Ο πίνακας αυτός περιέχει τα στοιχεία των κατηγοριών που υπάρχουν στο σύστημα. Οι κατηγορίες προκύπτουν από τα στοιχεία που διαθέτει η βάση γνώσης του συστήματος.

cat_id

Το μοναδικό αναγνωριστικό κλειδί για τις εγγραφές του συγκεκριμένου πίνακα

cat_name

Το όνομα της συγκεκριμένης κατηγορίας.

cat_comments

Μικρή περιγραφή για τα στοιχεία κάθε κατηγορίας. Πρόκειται για ένα προαιρετικό πεδίο της ΒΔ που χρησιμεύει για να υπάρχουν μεταδεδομένα εφόσον χρειαστούν μελλοντικά από το σύστημα.

7.1.5. keyword2article

Ο πίνακας αυτός χρησιμεύει για να γίνει η συσχέτιση των λέξεων κλειδιών με τα άρθρα του συστήματος. Συγκεκριμένα μας παρουσιάζει ποιες λέξεις κλειδιά υπάρχουν σε κάθε κείμενο του συστήματος.

articles_ar_id

Πρόκειται για ένα ξένο κλειδί που αναφέρεται στο μοναδικό αναγνωριστικό κλειδί του άρθρου

keywords_kw_id

Πρόκειται για ένα ξένο κλειδί που αναφέρεται στο μοναδικό αναγνωριστικό κλειδί της λέξης κλειδί

frequency

Η απόλυτη συχνότητα με την οποία εμφανίζεται μία λέξη κλειδί σε ένα κείμενο.

rel_frequency

Η σχετική συχνότητα με την οποία εμφανίζεται μία λέξη κλειδί σε ένα κείμενο. Η σχετική συχνότητα υπολογίζεται ως:

$$rel\_fr_m = \frac{abs\_fr_m}{\sum_{k=1}^{1} abs\_fr_k}$$

positions

Πρόκειται για τις θέσεις μέσα στο κείμενο όπου εντοπίζονται οι λέξεις κλειδιά. Οι θέσεις αφορούν ουσιαστικά τις προτάσεις του κειμένου όπου ευρίσκονται οι λέξεις κλειδιά.

7.1.6. article2category

Πρόκειται για έναν πίνακα ο οποίος περιέχει στοιχεία που συσχετίζουν τα άρθρα του συστήματος με κατηγορίες. Κάθε άρθρο δεν αντιστοιχίζεται σε μία μόνο κατηγορία, αλλά το σύστημα μας υπολογίζει τη συσχέτιση με κάθε κατηγορία που υπάρχει στο σύστημα.

ar_id

Πρόκειται για ένα ξένο κλειδί που αναφέρεται στο μοναδικό αναγνωριστικό κλειδί του άρθρου.

cat_id

Πρόκειται για ένα ξένο κλειδί που αναφέρεται στο μοναδικό αναγνωριστικό κλειδί της κατηγορίας.

frequency

Πρόκειται για τη συχνότητα που εκφράζει τη συσχέτιση μεταξύ άρθρου και κατηγορίας. Υπολογίζεται σαν η συσχέτιση του άρθρου με την κατηγορία.

7.1.7. articles_counter

Σε αυτό τον πίνακα καταγράφονται τα hits που έχει δεχθεί κάθε άρθρο. Χρησιμοποιείται σαν μετρική που μπορεί να «δείξει» ποια είναι τα άρθρα για τα οποία δείχνουν ενδιαφέρον οι χρήστες του συστήματος.

ar_id

Πρόκειται για ένα ξένο κλειδί που αναφέρεται στο μοναδικό αναγνωριστικό κλειδί του άρθρου.

ar_counter

Πρόκειται για έναν ακέραιο αριθμό που καταγράφει τα hits που έχει δεχθεί ένα άρθρο.

7.1.8. user_website

Πρόκειται για τον πίνακα που αποθηκεύει τις προσωπικές πληροφορίες κάθε χρήστη

id

Το μοναδικό αναγνωριστικό πρωτεύον κλειδί για τις εγγραφές του συγκεκριμένου πίνακα

uname

Το ψευδώνυμο του χρήστη (username).

passw

Ο κωδικός του χρήστη. Για τη βελτιστοποίηση της ασφάλειας του συστήματος ο κωδικός είναι κωδικοποιημένος με md5 κωδικοποίηση

email

Το e-mail του χρήστη. Εφόσον ο χρήστη χρησιμοποιεί τη δυνατότητα RSS του δικτυακού τόπου, τότε το e-mail του χρήστη δεν είναι αναγκαία πληροφορία. Εφόσον ο χρήστης χρησιμοποιεί τις υπηρεσίες του δικτυακού τόπου τότε το e-mail μπορεί να βοηθήσει σε κάποιες υπηρεσίες όπως είναι υλοποιημένες στο δικτυακό τόπο.

scrw

Πρόκειται για το μήκος της οθόνης του χρήστη σε pixels (screen width). Χρησιμεύει στο να αποσταλεί το σωστό μέγεθος κειμένου στον τελικό χρήστη.

scrh

Πρόκειται για το ύψος της οθόνης του χρήστη σε pixels (screen height). Χρησιμεύει στο να αποσταλεί το σωστό μέγεθος κειμένου στον τελικό χρήστη. Δεδομένου ότι η σύνηθης κύλιση σελίδες είναι προς τον κάθετο άξονα (scrolling) το ύψος της οθόνης είναι ενδεικτικό.

rec_date

Πρόκειται για την ημερομηνία εγγραφής του χρήστη στο σύστημα (timestamp).

7.1.9. user_website_category

Ο πίνακας αυτός αποθηκεύει τις πρωταρχικές επιλογές του χρήστη που αφορούν τις κατηγορίες προτίμησης των χρηστών.

user_id

Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τους χρήστες

cat_id

Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τις κατηγορίες

preference

Πρόκειται για την επιλογή του χρήστη user_id όσον αφορά την κατηγορία cat_id. Το preference μπορεί να πάρει τιμές από -5 έως 5 με το -5 να αντιπροσωπεύει δυσαρέσκεια προς την κατηγορία και το 5 να αντιπροσωπεύει πλήρη προτίμηση προς την κατηγορία.

7.1.10. user_website_info

Ο πίνακας αυτό χρησιμοποιείται σαν log για τις ενέργειες του χρήστη. Καταγράφει τις ημερομηνίες και την IP από την οποία έχουν πραγματοποιήσει σύνδεση οι χρήστες και βοηθά στην καλύτερη παρουσίαση των νέων άρθρων στους χρήστες αφού το σύστημα είναι σε θέση να γνωρίζει ποια άρθρα έχουν προστεθεί στο σύστημα από την τελευταία φορά που το επισκεφθήκαν οι χρήστες του συστήματος.

id

Το μοναδικό αναγνωριστικό κλειδί που αφορά τις εγγραφές που γίνονται στο συγκεκριμένο πίνακα. Χρησιμοποιείται γιατί το ξένο κλειδί user_id δε μπορεί να είναι κλειδί στον συγκεκριμένο πίνακα λόγω της πληθώρας των εγγραφών χρήστη που υπάρχουν στο συγκεκριμένο πίνακα και αφορούν μεμονωμένους χρήστες.

user_id

Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τους χρήστες

user_ip

Η IP από την οποία έχει συνδεθεί ο χρήστης

user_date

Η ημερομηνία που συνδέθηκε ο χρήστης

7.1.11. user_website_keyword

user_id

Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τους χρήστες

kw_id

Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τις λέξεις κλειδιά

rel_frequency

Η σχετική συχνότητα που αντιπροσωπεύει κατά πόσο ο χρήστης ενδιαφέρεται για τη συγκεκριμένη λέξη κλειδί. Οι τιμές είναι θετικές και αρνητικές ενώ συνήθεις τιμές για το συγκεκριμένο πεδίο είναι -2,00 έως 2,00. Το φαινόμενο μία λέξη να ξεφεύγει από αυτά τα όρια είναι (α) ο χρήστης να μην ενδιαφέρεται καθόλου για μία λέξη κλειδί και (β) ο χρήστης να ενδιαφέρεται πολύ για μία λέξη κλειδί όταν οι τιμές είναι μικρότερες του -2 και μεγαλύτερες του +2 αντίστοιχα.

7.1.12. user_website_reading

Ο πίνακας χρησιμεύει για να καταμετρήσουμε το χρόνο που κάθε χρήστης «σπαταλά» για να διαβάσει ένα άρθρο. Το χρησιμοποιούμε σαν μετρική προκειμένου να καταμετρήσουμε το ενδιαφέρον του χρήστη για συγκεκριμένα κείμενα

ar_id

Το ξένο κλειδί που αντιπροσωπεύει το άρθρο.

user_id

Το ξένο κλειδί που αντιπροσωπεύει το χρήστη.

total_time

Ο συνολικός χρόνος που έχει σπαταλήσει ο χρήστης στο συγκεκριμένο άρθρο.

7.2. Πίνακες της βάσης γνώσης

Οι προηγούμενοι πίνακες αντιπροσωπεύουν όλους τους πίνακες που χρησιμοποιούνται άμεσα από το δικτυακό τόπο και από το δυναμικό RSS προκειμένου να παρέχουν και να αποθηκεύουν όλες τις απαραίτητες πληροφορίες. Στην πορεία θα παρουσιάσουμε τους πίνακες που χρησιμοποιούνται για να συνθέσουν τη βάση γνώσης πάνω στην οποία στηρίζονται οι βασικότεροι αλγόριθμοι του συστήματός μας.

7.2.1. articles_training

Πίνακας που χρησιμοποιείται για να αποθηκεύσει τη βάση γνώσης πάνω στην οποία στηρίζεται το σύστημα για τους αλγορίθμους κατηγοριοποίησης, αυτόματης εξαγωγής περίληψης και προσωποποίησης στο χρήστη.

ar_id

Το μοναδικό αναγνωριστικό κλειδί που αφορά τις εγγραφές που γίνονται στο συγκεκριμένο πίνακα και αφορούν τα άρθρα

ar_title

Ο τίτλος του άρθρου

ar_body

Το σώμα του άρθρου

ar_summary

Η περίληψη του συγκεκριμένου άρθρου

ar_url

Ο σύνδεσμος όπου βρίσκεται το άρθρο. Πρόκειται για το σύνδεσμο από τον οποίο «κατέβηκε» ο HTML κώδικας της σελίδας.

ar_date

Η ημερομηνία κατά την οποία συλλέχθηκε το άρθρο από τον αυτοματοποιημένο μηχανισμό

cat_id

Η κατηγορία στην οποία ανήκει το άρθρο. Ας μην ξεχνάμε πως πρόκειται για τη βάση γνώσης και συνεπώς τα άρθρα είναι προκατηγοριοποιημένα.

7.2.2. keywords_articles_training

Ο πίνακας αυτός χρησιμοποιείται προκειμένου να αποθηκευτούν οι λέξεις κλειδιά που έχουν εξαχθεί από τα άρθρα. Για τη βάση γνώσης δεν είναι ένας πίνακας ο οποίος έχει άμεση χρησιμότητα για τους αλγορίθμους του μηχανισμού. Ωστόσο, είναι ένας βοηθητικός πίνακας γιατί χρησιμοποιείται προκειμένου να ελεγχθούν άρθρα τα οποία βρίσκονται στη βάση γνώσης και είναι προβληματικά. Ως προβληματικά αναφέρονται τα άρθρα των οποίων οι λέξεις κλειδιά δεν ανταποκρίνονται στην ενότητα την οποία αντιπροσωπεύει μία κατηγορία.

ar_id

Το μοναδικό ξένο κλειδί που αναφέρεται στο άρθρο από το οποίο εξάγουμε τις λέξεις κλειδιά.

kw_id

Το μοναδικό ξένο κλειδί που αναφέρεται στις λέξεις κλειδιά που εξάγονται από το άρθρο με αναγνωριστικό κλειδί ar_id.

Τα δύο παραπάνω χρησιμοποιούνται ως κλειδιά για το συγκεκριμένο πίνακα και συνεπώς δε μπορεί να υπάρχουν διπλές εγγραφές με τα ίδια χαρακτηριστικά ar_id και kw_id.

frequency

Η απόλυτη συχνότητα που αναφέρεται στη συχνότητα με την οποία εμφανίζεται μία λέξη κλειδί σε ένα άρθρο. Είναι μία μετρική η οποία μπορεί να δείξει τη σημαντικότητα μίας λέξης κλειδιού για ένα άρθρο.

7.2.3. keywords_category_training

Πρόκειται ίσως για τον πιο σημαντικό πίνακα της βάσης γνώσης. Σε αυτό τον πίνακα αποθηκεύονται πληροφορίες που αφορούν τις λέξεις κλειδιά που αντιπροσωπεύουν μία κατηγορία, ενώ παράλληλα αποθηκεύεται πληροφορία που αφορά το πόσο σημαντική είναι μία λέξη για μία κατηγορία.

cat_id

Το μοναδικό ξένο κλειδί που αφορά την κατηγορία στην οποία ανήκει μία λέξη κλειδί.

kw_id

Το μοναδικό ξένο κλειδί που αφορά τη λέξη κλειδί που ανήκει σε μία κατηγορία.

rel_frequency

Πρόκειται για τη σχετική συχνότητα με την οποία εμφανίζεται μία λέξη κλειδί σε μία κατηγορία.

abs_frequency

Η απόλυτη συχνότητα με την οποία εμφανίζεται μία λέξη κλειδί σε μία κατηγορία.

7.2.4. keywords_training

Ο πίνακας αυτός χρησιμοποιείται προκειμένου να αποθηκευτούν όλες οι λέξεις κλειδιά πού εξάγονται από τα άρθρα της βάσης γνώσης.

kw_id

Το μοναδικό αναγνωριστικό κλειδί που αντιπροσωπεύει μία λέξη κλειδί.

kw_name

Η λέξη.

7.2.5. resolution_chars

Αυτός ο πίνακας χρησιμοποιείται προκειμένου να υπάρχει αποθηκευμένη πληροφορία στη σύστημα για να γνωρίζουμε ανά πάσα στιγμή πόσοι χαρακτήρες πρέπει να εμφανίζονται στις διαφορετικές αναλύσεις που μπορεί να έχει η οθόνη του χρήστη.

id

Το μοναδικό αναγνωριστικό κλειδί που αφορά το συγκεκριμένο πίνακα.

scr_width

Το μήκος της οθόνης σε pixels (screen width).

scr_height

Το ύψος της οθόνης σε pixels (screen height).

characters

Οι χαρακτήρες που μπορούν να εμφανίζονται σε οθόνες με το συγκεκριμένο scr_width και scr_height.

8

ΑΝΑΠΤΥΞΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

Στο κεφάλαιο αυτό περιγράφεται ο τρόπος ανάπτυξης του συστήματος συνολικά αλλά και κάθε μηχανισμού ξεχωριστά. Γίνεται ανάλυση των αλγορίθμων που χρησιμοποιούνται αλλά και ο τρόπος λειτουργίας κάθε υποσυστήματος.

8. ΑΝΑΠΤΥΞΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

Μέχρι αυτή τη στιγμή έχουμε αναφερθεί στα βασικά συστήματα που αποτελούν τη βάση για το σύστημά μας αλλά και τους αλγορίθμους που χρησιμοποιούμε προκειμένου να υλοποιήσουμε το κάθε υποσύστημα. Σε αυτό το κεφάλαιο θα εστιάσουμε την προσοχή μας στην υλοποίηση κάθε συστήματος ξεχωριστά. Δεδομένου ότι οι γραμμές κώδικα που γράφηκαν συνολικά για την κατασκευή του συστήματος είναι υπερβολικά πολλές για να παρουσιαστούν θα εστιάσουμε την προσοχή μας στα πιο σημαντικά σημεία καθώς και στις τεχνικές με τις οποίες υλοποιήθηκε κάθε αλγόριθμος σε κάθε σημείο.

8.1. Αλγοριθμικά θέματα

Για να αναλύσουμε πως κάθε αλγόριθμος εφαρμόζεται πάνω στα κείμενα, παρουσιάζουμε μια σύνοψη της διαδικασίας εκτέλεσης.

Αλγόριθμος 1: Αλγόριθμος διαδικασίας εκτέλεσης συστήματος

```
String Text = fetch next text();
List kwfr(text) = create keyword frequency list(text);
List * kwfr cat(category) = create keyword frequency list(text, category);
Categorize (kwfr(text), *kwfr cat(category));
if !Categorize then
String Stext = Summarize(text,kwfr(text));
List kwfr(Stext) = create keyword frequency list(Stext);
List * kwfr cat(category) = create keyword frequency list(Stext, category);
Categorize (kwfr(Stext), *kwfr cat(category));
if !Categorize then
Category = "generic";
end if
else
Summarize(Category,Personal Data);
end if
```

Παρά το γεγονός ότι η αλγοριθμική διαδικασία δείχνει μόνο την κατηγοριοποίηση των άρθρων, τελικά μέσω αυτής επιτυγχάνουμε τους τρεις βασικούς στόχους: κατηγοριοποίηση, περίληψη και αλληλεπίδραση μεταξύ των μηχανισμών. Ξεκινάμε προσπαθώντας να κατηγοριοποιήσουμε το νέο άρθρο βάσει του συνόλου εκμάθησης που προυπάρχει στη βάση δεδομένων, δημιουργώντας μια λίστα από αντιπροσωπευτικές κωδικολέξεις (οι οποίες είναι stemmed από την διαδικασία προεπεξεργασίας) μαζί με την συχνότητα εμφάνισής τους. Έπειτα κατασκευάζουμε όμοιες λίστες για όλες τις κατηγορίες που υπάρχουν στη βάση δεδομένων. Αυτές οι λίστες αποτελούνται από τις ίδιες κωδικολέξεις ακολουθούμενες από την συχνότητά τους στην εκάστοτε κατηγορία. Εξετάζουμε την ομοιότητα συνημιτόνου αυτών των λιστών με σκοπό να καθορίσουμε την κατηγορία του κειμένου.

Πίνακας 1: Ομοιότητα μεταξύ κειμένου και κατηγορίας

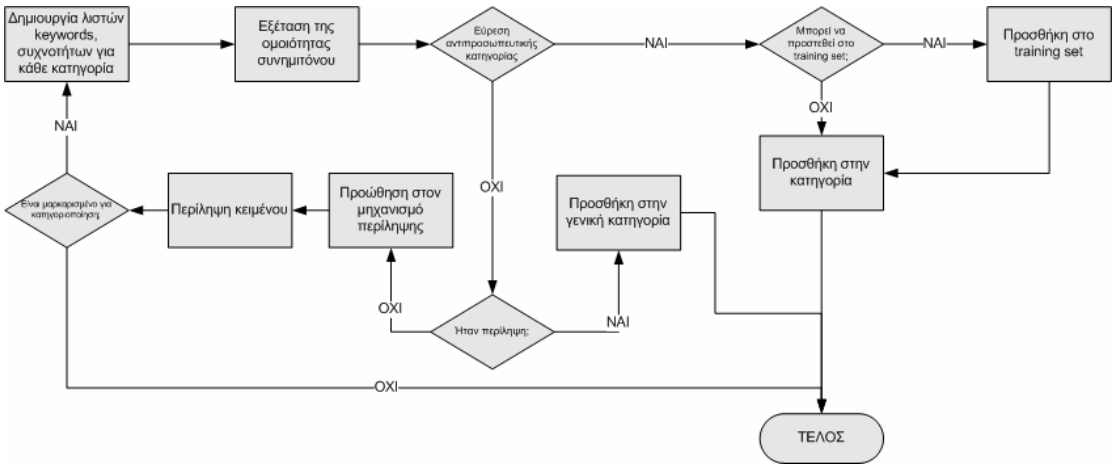
Κατηγορία	Συχνότητα
business	0,742862
entertainment	0,449297
health	0,532352
politics	0,418447
science	0,526925
sports	0,642862
education	0,596509

Εάν το κείμενο δεν μπορεί να ταξινομηθεί σε κάποια από τις υπάρχουσες κατηγορίες, τότε προωθείται στον μηχανισμό περίληψης όπου εξάγεται μια γενική (generic) περίληψη η οποία στη συνέχεια εξετάζεται αν μπορεί να κατηγοριοποιηθεί. Ένα κείμενο μπορεί να κατηγοριοποιηθεί επιτυχώς όταν:

- η ομοιότητα συνημιτόνου με κάποια κατηγορία είναι πάνω από ένα όριο, και
- η διαφορά των ομοιοτήτων συνημιτόνου μεταξύ της ισχυρότερης και των υπολοίπων κατηγοριών είναι πάνω από ένα όριο.

Τα όρια αυτά εξηγούνται αναλυτικά στη συνέχεια.

Τελικά, εάν η ομοιότητα συνημιτόνου μεταξύ του κειμένου και της αντιπροσωπευτικής του κατηγορίας είναι πολύ μεγάλη, και παρόμοια η διαφορά των ομοιοτήτων συνημιτόνου μεταξύ της ισχυρότερης και των υπολοίπων κατηγοριών είναι επίσης πολύ μεγάλη, τότε το κείμενο προστίθεται στο δυναμικό σύνολο κειμένων εκπαίδευσης που χρησιμοποιεί ο μηχανισμός. Η προηγούμενη διαδικασία αποτυπώνεται και στο διάγραμμα ροής του παρακάτω σχήματος.



Εικόνα 19: Το διάγραμμα ροής των διεργασιών του συστήματος.

8.1.1. Προεπεξεργασία κειμένου

Το υποσύστημα προεπεξεργασίας κειμένου και εξαγωγής κωδικολέξεων αποτελεί ένα ανεξάρτητο μηχανισμό που φέρνει εις πέρας μια σημαντική διεργασία του όλου συστήματος, καθώς τροφοδοτεί τους μηχανισμούς που ακολουθούν με την απαραίτητη είσοδο. Πρόκειται για μια αλγοριθμική, ακολουθιακή διαδικασία η οποία περιγράφεται από τον παρακάτω αλγόριθμο.

Αλγόριθμος 2: Δημιουργία πίνακα λέξεων κλειδιών(XML, options)

```
String Text = fetch next text(XML);
String Title = fetch next title(XML);
parseTitle(Title);
Text = removePunctuation(Text);
Text = removeStopwords(Text);
list Keywords = keepKeywordsPercentage(Text);
Keywords = stemming(Keywords);
list keyword frequency list = measure keyword frequencies(Text,Keywords);
list keyword positions = get keywords positions(Text,Keywords);
return keyword frequency list;
```

Η βασική συνάρτηση measure_keyword_frequencies(Text,Keywords) περιγράφεται από τον παρακάτων αλγόριθμο:

Αλγόριθμος 3: measure keyword frequencies(Text, Keywords)

```
for all kw in Keywords do
if kw is found in Text then
```

```
keyword frequency list[kw][appearances]++;
end if
end for
return keyword frequency list;
```

και η get_keywords_positions(Text,Keywords) παρουσιάζεται στον παρακάτω αλγόριθμο:

Αλγόριθμος 4: get_keywords_positions(Text, Keywords)

```
for all kw in Keywords do
for all sentence in Text do
if kw is found in sentence then
keyword positions list[kw][positions].push back(position);
end if
end for
end for
return keyword positions list;
```

Οι προηγούμενοι αλγόριθμοι επιτυγχάνουν το ζητούμενο της διαδικασίας του keyword extraction: την εξαγωγή των keywords από το κείμενο, την καταγραφή των συχνοτήτων εμφάνισής τους στο κείμενο και την καταγραφή των προτάσεων στις οποίες εμφανίζονται (θέσεις στο κείμενο). Για να συμβεί αυτό, το κείμενο περνάει από ορισμένα στάδια προεπεξεργασίας, όπως η αφαίρεση των σημείων στίξης, των stopwords καθώς και το stemming του κειμένου. Για την εφαρμοσιμότητα του αλγορίθμου σε πραγματικές συνθήκες λειτουργίας του μηχανισμού, όπου τα άρθρα καταφθάνουν με γοργούς ρυθμούς και η προεπεξεργασία δεν θα πρέπει να διαρκεί πολύ, είναι σημαντικό τα διάφορα μέρη του αλγορίθμου να υλοποιούνται με τρόπο βέλτιστο. Η χρήση επομένως τεχνικών που βασίζονται σε κανονικές εκφράσεις (regular expressions) για την εκτεταμένη διαχείριση συμβολοσειρών την οποία κάνει το υποσύστημα προεπεξεργασίας κειμένου είναι επιβεβλημένη.

8.1.2. Αυτόματη Περίληψη Κειμένου

Η διαδικασία παραγωγής περίληψης βασίζεται σε ευρετικές μεθόδους. Αυτό σημαίνει ότι η περίληψη δεν παράγεται «από την αρχή», αλλά αποτελείται από τις πιο αντιπροσωπευτικές προτάσεις του κειμένου. Με αυτό εννοούμε ότι σε κάθε πρόταση δίνεται ένα «σκορ» το οποίο μας οδηγεί στην κατασκευή της περίληψης.

Για την παραγωγή της περίληψης ενός άρθρου, 6 ξεχωριστοί παράγοντες χρησιμοποιούνται για την δημιουργία της αλλά και για την αλληλεπίδραση με τον μηχανισμό κατηγοριοποίησης:

(a) η συχνότητα του keyword στο κείμενο (πόσες φορές εμφανίζεται το keyword στο κείμενο

(b) η συχνότητα εμφάνισης του keyword στον τίτλο του κειμένου

(c) το ποσοστό των keywords μέσα στην πρόταση

(d) το ποσοστό των keywords στο κείμενο

(e) η ικανότητα του κάθε keyword να αναπαραστήσει μια κατηγορία, και

(f) η ικανότητα του κάθε keyword να αναπαραστήσει τις επιλογές και τις επιθυμίες του κάθε ξεχωριστού χρήστη ή μιας κατηγορίας χρηστών με ίδιο προφίλ.

Σύμφωνα με τους δύο πρώτους παράγοντες [(a) και (b)], παράγουμε την πρώτη και αρχική εξίσωση για μια γενική βαθμολόγηση των προτάσεων:

$S_i = \sum w_{k,i}(k_1 + k_2)$	(1)

Όπου $w_{k,i}$, είναι η συχνότητα του k-οστού keyword της πρότασης i, k_1 είναι μια σταθερά που αναπαριστά την επίδραση του παράγοντα (α), και k_2 είναι μια σταθερά που αναπαριστά την επίδραση του παράγοντα (β') στην διαδικασία περίληψης.

Μέσα από εκτενή πειραματική διαδικασία, καταλήξαμε σε τιμές για τα k_1 και k_2. Το ορίζεται από την ακόλουθη σχέση:

$k_1 = 1 + 0.1x$	(2)

Όπου x οι φορές που ένα keyword εμφανίζεται στον τίτλο του κειμένου. Παρόμοια, το k_2 ορίζεται από την ακόλουθη σχέση:

$K_2 = 1 + 1.2y$	(3)

Όπου y είναι η πιθανότητα το keyword να βρίσκεται n φορές σε μια πρόταση. Θεωρώντας μια πρόταση με μήκος m (m keywords) και το κείμενο με μήκος t, η παράμετρος y βγαίνει από την ακόλουθη σχέση:

$y = n/t * m/t$	(4)

Για να κανονικοποιήσουμε τις τιμές που προκύπτουν από την εξίσωση (1), προτείνουμε την χρήση των παραγόντων (c) και (d). Η κανονικοποίηση χρειάζεται διότι, οι μεγάλες σε μήκος προτάσεις του κειμένου, τείνουν να βαθμολογούνται υψηλότερα σε σχέση με τις μικρές σε μήκος. Ο παράγοντας (c) αναπαριστά το ποσοστό των keywords στο κείμενο. Πιο συγκεκριμένα, εάν για παράδειγμα τρία keywords έχουν εξαχθεί από μια πρόταση η οποία αποτελείται από πέντε keywords και ο αριθμός των συνολικά εξαχθέντων keywords από το κείμενο είναι είκοσι πέντε, τότε ο παράγοντας (c) ισούται με τρία πέμπτα (3/5) και ο παράγοντας (d) με τρία εικοστά πέμπτα (3/25).

Η κανονικοποίηση που αναφέρθηκε χρησιμοποιείται για να επιλυθούν κάποια προβλήματα που εγείρονται, όπως στο παράδειγμα που ακολουθεί. Υποθέτουμε ότι ένα κείμενο έχει πολλές μικρές προτάσεις και μία η οποία είναι πολύ μεγάλη. Η μεγάλη πρόταση αποτελείται από 20 keywords και τα keywords που εξήχθησαν (χρήσιμα) είναι 5. Μια μικρή πρόταση, η οποία είναι πολύ αντιπροσωπευτική για το κείμενο αποτελείται από 4 keywords, όλα από τα οποία είναι χρήσιμα. Έστω επίσης ότι ο συνολικός αριθμός των εξαχθέντων keywords για το κείμενο είναι 30. Η μεγάλη πρόταση είναι πολύ πιθανό να βαθμολογηθεί υψηλότερα σύμφωνα με την εξίσωση (1), αφού το μήκος της την «βοηθά» να έχει περισσότερα keywords. Οι δύο παράγοντες που προτείνονται, κανονικοποιούν αυτή την πιθανή «αδικία». Η μεγάλη πρόταση θα έχει 5/20 και 5/30 αντίστοιχα, ενώ η μικρή πρόταση θα έχει 4/4 και 4/30 για τους παράγοντες (c) και (d) αντίστοιχα. Με αυτό τον τρόπο, η μικρή σε μήκος πρόταση θα αντιμετωπιστεί ως πιο σημαντική σε σχέση με την μεγάλη, κάτι που ισχύει για το συγκεκριμένο κείμενο. Η κανονικοποίηση εφαρμόζεται απ' ευθείας στην εξίσωση (1) και το , όπου το είναι ο παράγοντας κανονικοποίησης που ισούται με το γινόμενο των (c) και (d).

Οι παράγοντες (e), η ικανότητα του keyword να αντιπροσωπεύει την κατηγορία, και (f), η ικανότητα του keyword να ανταποκρίνεται στις επιλογές του μοναδικού χρήστη, παρουσιάζονται αναλυτικά στις ενότητες που ακολουθούν αφού η επίδρασή τους στην διαδικασία είναι σημαντική και μετατρέπουν το σύστημα εξαγωγής περίληψης σε ένα πλήρως προσωποποιημένο μηχανισμό.

8.1.3. Μηχανισμός Κατηγοριοποίησης

Το υποσύστημα κατηγοριοποίησης βασίζεται στην μετρική ομοιότητας συνημιτόνου, σε εσωτερικά γινόμενα πινάκων και σε υπολογισμούς ζυγίσματος βαρών. Πιο συγκεκριμένα, το σύστημα αρχικοποιείται με ένα σύνολο κειμένων (άρθρα ειδήσεων) εκμάθησης τα οποία συλλέγονται από σημαντικές ειδησεογραφικές ιστοσελίδες (major news portals). Τα κείμενα αυτά είναι προ-κατηγοριοποιημένα από ανθρώπους και παρουσιάζονται ως ήδη κατηγοριοποιημένα στα news portals. Το σύνολο κειμένων εκπαίδευσης αποτελείται από αυτά τα

προκατηγοριοποιημένα κείμενα και από κείμενα που προσθέτονται δυναμικά από τον μηχανισμό όταν εντοπίζονται κείμενα με μεγάλη σχετικότητα με κάποια από τις υπάρχουσες κατηγορίες. Το σύστημα κατηγοριοποίησης δέχεται ως είσοδο την εξαγωγή του μηχανισμού προεπεξεργασίας. Αυτή είναι (α) ένα XML αρχείο (ή δομή) που περιέχει stemmed keywords, την απόλυτη και σχετική συχνότητα εμφάνισής τους αλλά και την θέση τους στο κείμενο και (β) ένα XML αρχείο που περιέχει το ίδιο το κείμενο. Η πληροφορία που αποθηκεύεται στο δεύτερο αρχείο XML αφορά στο id στον τύπο, στον τίτλο και στο σώμα του κειμένου.

Ύστερα από την αρχικοποίηση του συνόλου κειμένων εκπαίδευσης, ο μηχανισμός της κατηγοριοποίησης δημιουργεί λίστες από keywords τα οποία είναι αντιπροσωπευτικά της κάθε μία κατηγορίας, αποτελούμενες από keywords με υψηλή συχνότητα εμφάνισης σε μια συγκεκριμένη κατηγορία και μικρή ή μηδενική εμφάνιση για τις άλλες κατηγορίες. Η δημιουργία των λιστών είναι βοηθητική για την κατηγοριοποίηση των νεοεισερχομένων άρθρων αλλά αποδεικνύεται βοηθητική και για την διαδικασία της εξαγωγής περίληψης.

Αφού η διαδικασία περίληψης κειμένου του συστήματος βασίζεται στην επιλογή των πιο αντιπροσωπευτικών προτάσεων οι οποίες επιλέγονται ζυγίζοντάς τες κατάλληλα, τα αποτελέσματα της κατηγοριοποίησης μπορούν να βοηθήσουν στην επιλογή πιο αποτελεσματικού ζυγίσματος για τις προτάσεις. Η κοινή λογική λέει ότι ένα keyword που έχει πολύ υψηλή συχνότητα εμφάνισης για μια συγκεκριμένη κατηγορία, πρέπει να δίνει περισσότερο βάθος σε μια πρόταση που εμφανίζεται, ενώ ένα keyword που έχει μικρή ή μηδενική συχνότητα εμφάνισης για μια κατηγορία μπορεί να προσθέτει λιγότερο στο συνολικό σκορ της πρότασης. Ακόμα παραπέρα, ένα keyword που συμπεριλαμβάνεται στα εξαγόμενα keywords ενός άρθρου που είναι αντιπροσωπευτικό για μια κατηγορία διαφορετική από αυτή στην οποία ανήκει το άρθρο, μπορεί να δώσει αρνητικό βάρος σε μια πρόταση. Η επόμενη εξίσωση χρησιμοποιείται για τον υπολογισμό της επίδρασης της διαδικασίας της κατηγοριοποίησης σε αυτήν της περίληψης.

$$k_3 = \begin{cases} A \cdot cw_i & \text{όπου } A > 1 \text{ και } cw \text{ το } \underline{\text{θετικό}} \text{ βάρος κατηγορίας} \\ -A \cdot cw_i & \text{όπου } A > 1 \text{ και } cw \text{ το } \underline{\text{αρνητικό}} \text{ βάρος κατηγορίας} \\ 1 & \text{για ουδέτερα ή μη βαθμολογημένα από το σύστημα keywords ή εάν } A = 0 \end{cases} \quad (5)$$

Η παράμετρος A πρέπει να είναι μεγαλύτερη από το 1 και χρησιμοποιείται για να προσθέσει βάρος για την παράμετρο k_3. Εάν θέλουμε η διαδικασία περίληψης να βασίζεται κυρίως στο k_3, τότε οι τιμές ζυγίσματος για το A χρησιμοποιούνται, αντίθετα, αν η διαδικασία περίληψης πρέπει να βασίζεται ισοδύναμα σε όλες τις "k" μεταβλητές, τότε το δεν πρέπει να είναι μεγαλύτερο από τις τιμές που έχουν ανατεθεί στα k_1 και k_2. Η παράμετρος cw αποτυπώνει την σχετική συχνότητα ενός keyword στην κατηγορία. Η ποσότητα αυτή μπορεί να μας παρέχει πληροφορία για το πόσο σημαντικό (αντιπροσωπευτικό) είναι ένα keyword για την κατηγορία.

Με τη χρήση της τελευταίας εξίσωσης η εξίσωση (1) μετατρέπεται ως εξής:

$$S_i = \sum w_{k,i}(k_1 + k_2)k_3 \quad (6)$$

8.1.4. Μηχανισμός προσωποποίησης

Ο μηχανισμός προσωποποίησης του συστήματος, που υποστηρίζεται ως ένα μέσο επικοινωνίας μεταξύ όλων των διαδικασιών και των χρηστών, μπορεί να χρησιμοποιηθεί για να προσωποποιηθεί η περίληψη σε κάθε χρήστη. Σε ένα σύγχρονο, αποτελεσματικό και χρήσιμο σύστημα, ο χρήστης θα πρέπει να βλέπει προσωποποιημένο περιεχόμενο ανάλογα με τα κριτήρια που έχει θέσει και τις προτιμήσεις του. Στην περίπτωσή μας, θα πρέπει να λαμβάνει προσωποποιημένη περίληψη των άρθρων μόνο που τον ενδιαφέρουν και όχι απλά μιας γενικής μορφής περίληψη που προκύπτει από μια απλή αλγοριθμική διαδικασία.

Σύμφωνα με τις αλγοριθμικές διαδικασίες που ακολουθεί το σύστημα που αναπτύχθηκε, δημιουργούνται λίστες από keywords για κάθε χρήστη οι οποίες

αντιπροσωπεύουν τις προτιμήσεις του. Πιο συγκεκριμένα, τα keywords σχηματίζουν δύο ειδών λίστες: μια λίστα «θετικών» keywords που φαίνεται να ταιριάζουν στις επιλογές του χρήστη (ή της ομάδας χρηστών), και μια λίστα «αρνητικών» keywords τα οποία δεν ενδιαφέρουν τον χρήστη. Αυτές οι λίστες συνεπάγονται από τις επιλογές των χρηστών για τις κατηγορίες και τα keywords που τον ενδιαφέρουν. Η πρόθεση μας είναι να βαθμολογήσουμε υψηλότερα τις προτάσεις κειμένων που περιέχουν «θετικά» keywords και χαμηλότερα τις προτάσεις που περιέχουν «αρνητικά» keywords. Με αυτή την προοπτική, χρησιμοποιείται μια ακόμη παράμετρος, η k_4, η οποία δρα ως παράγοντας προσωποποίησης.

Η μεταβλητή για την προσωποποίηση χρησιμοποιείται όπως και αυτή για την κατηγοριοποίηση και δίνεται από την ακόλουθη εξίσωση:

$$k_4 = \begin{cases} B \cdot uw_i & \text{όπου } B > 1 \text{ και } uw \text{ το } \underline{\text{θετικό}} \text{ βάρος χρήστη} \\ -B \cdot uw_i & \text{όπου } B > 1 \text{ και } uw \text{ το } \underline{\text{αρνητικό}} \text{ βάρος χρήστη} \\ 1 & \text{για ουδέτερα ή μη βαθμολογημένα από τον χρήστη keywords ή εάν } B = 0 \end{cases} \quad (7)$$

Η παράμετρος uw αποτυπώνει τη σχετική συχνότητα ενός keyword για τον χρήστη. Αυτή μπορεί να μας παρέχει πληροφορία για το πόσο σημαντικό (ισχυρό) είναι ένα keyword για τον χρήστη. Αυτή η παράμετρος προστίθενται στην εξίσωση (6) η οποία γίνεται:

$$S_i' = \sum w_{k,i}(k_1 + k_2)k_3k_4 \quad (8)$$

Οι παράμετροι A και B στις εξισώσεις (5) και (7) αντίστοιχα, χρησιμοποιούνται σε συνδυασμό μεταξύ τους. Εάν δεν σκοπεύουμε να χρησιμοποιήσουμε κάποιον από τον παράγοντα κατηγοριοποίησης ή προσωποποίησης, μπορούμε να θέσουμε την τιμή 0 για την αντίστοιχη παράμετρο. Εάν θέλουμε να εστιάσουμε την προσοχή μας κυρίως στον παράγοντα προσωποποίησης και λιγότερο στην κατηγοριοποίησης, τότε μπορούμε να θέσουμε B=2 και A=1. Αυτό σημαίνει ότι ο παράγοντας k_4 θα έχει διπλάσια επίδραση από τον k_3. Ο επόμενος πίνακας δείχνει την επίδραση των παραμέτρων (e) και (f) σύμφωνα με τις τιμές των A και B.

Πίνακας 2: Επίδραση των παραγόντων A και B στο ζύγισμα των προτάσεων

A	B	Αποτέλεσμα
0	0	Οι παράγοντες κατηγοριοποίησης και προσωποποίησης δεν υπολογίζονται στο αποτέλεσμα
0	1	Μόνο ο παράγοντας προσωποποίησης έχει επίδραση στο αποτέλεσμα
1	0	Μόνο ο παράγοντας κατηγοριοποίησης έχει επίδραση στο αποτέλεσμα
1	2	Ο παράγοντας προσωποποίησης έχει διπλάσια ισχύ από τον παράγοντα κατηγοριοποίησης
1	10	Ο παράγοντας προσωποποίησης έχει τόσο μεγαλύτερη ισχύ από τον παράγοντα κατηγοριοποίησης που η συμμετοχή του παράγοντα κατηγοριοποίησης είναι αμελητέα
1	1	Ίδια επίδραση από τους δύο παράγοντες
1.2	1.8	Οι τιμές που πραγματικά χρησιμοποιούνται στο μηχανισμό

Όπως παρατηρείται από την εξίσωση (8), μερικές «ειδικές» περιπτώσεις μπορούν να λάβουν χώρα από τους μηδενισμούς που εισάγουν οι παράμετροι και

k_3 και k_4. Ο επόμενος πίνακας δείχνει την αντίδραση του αλγορίθμου στις τέσσερις διαφορετικές καταστάσεις.

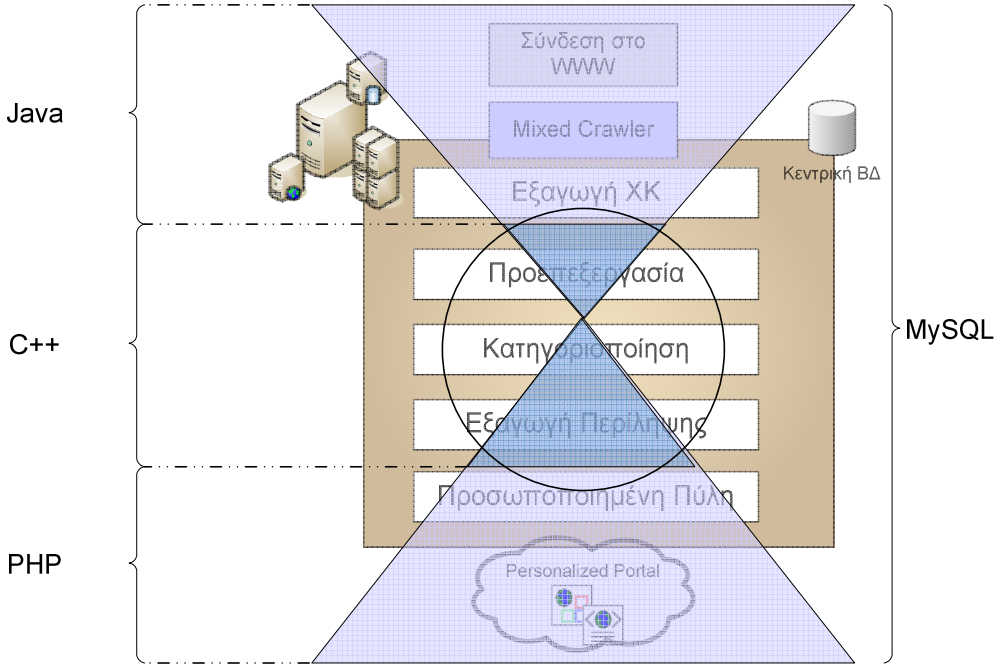
Πίνακας 3: Αντίδραση του αλγορίθμου σε διάφορες τιμές του k_3 και k_4

Μεταβλητή k_3	Μεταβλητή k_4	Αποτέλεσμα
Θετικό	Θετικό	Θετικό
Θετικό	Αρνητικό	Αρνητικό
Αρνητικό	Θετικό	Θετικό (το k_3 δεν υπολογίζεται στο αποτέλεσμα)
Αρνητικό	Αρνητικό	Αρνητικό

Μια ειδική περίπτωση συμβαίνει όταν η μεταβλητή κατηγοριοποίησης είναι αρνητική και η μεταβλητή προσωποποίησης είναι θετική. Σε αυτή την περίπτωση θεωρούμε ότι, η επιλογή του χρήστη για το συγκεκριμένο keyword ως αντιπροσωπευτικό των ενδιαφερόντων του, υπερισχύει της μη αντιπροσωπευτικότητας του keyword για συγκεκριμένη κατηγορία. Επιπρόσθετα, όταν και οι δύο μεταβλητές είναι αρνητικές, το αποτέλεσμα παραμένει αρνητικό αφού οι αρνήσεις σε αυτή την περίπτωση σημαίνουν ακόμα πιο αρνητικό σκορ για την πρόταση.

8.2. Υλοποίηση του συστήματος

Για την υλοποίηση του συστήματος όπως έχει ήδη αναφερθεί χρησιμοποιήθηκε συνδυασμός τεχνολογιών. Για τα συστήματα που επικοινωνούν σε υψηλό επίπεδο με τον έξω κόσμο (εκτός μηχανισμού) χρησιμοποιήθηκαν αντίστοιχα γλώσσες οι οποίες λειτουργούν σε υψηλό επίπεδο και είναι η Java και η PHP ενώ για τα συστήματα που λειτουργούν στον πυρήνα του συστήματος χρησιμοποιούμε τη γλώσσα προγραμματισμού C++. Για τη διασύνδεση όλων των συστημάτων χρησιμοποιούμε τη βάση δεδομένων. Αποτελεί το κανάλι επικοινωνίας όλων των συστημάτων που κατασκευάζουμε καθότι κάθε σύστημα αντλεί πληροφορίες από αυτή και σε αυτή τις αποθηκεύει.



Εικόνα 20: Τεχνολογίες Υλοποίηση του Μηχανισμού

8.3. Ιστορικό

Η σύλληψη της ιδέας για την κατασκευή του συστήματος ξεκινά από το 2004 οπότε και κατασκευάστηκε ένας μηχανισμός συλλογής άρθρων ο οποίος βασιζόταν σε αυτόματη κατηγοριοποίηση και προσωποποίηση στον τελικό χρήστη. Η εργασία ολοκληρώθηκε με την Προπτυχιακή Διπλωματική εργασία:

Βασισμένοι στα συστήματα που είχαν υλοποιηθεί αλλά και στους αλγορίθμους που είχαν χρησιμοποιηθεί ξεκινά στο τέλος του 2005 η ανάπτυξη ενός συνολικού συστήματος το οποίο θα περιλαμβάνει κάθε μηχανισμό που είναι απαραίτητος και θα υλοποιεί βέλτιστους αλγόριθμους με βέλτιστο τρόπο. Δεδομένου ότι οι μηχανισμοί τελικής παρουσίασης στο χρήστη αλλά και ο μηχανισμός συλλογής πληροφορίας είναι τετριμμένοι και εύκολη στο σχεδιασμό και στην υλοποίηση ελήφθη η απόφαση να ξεκινήσουν υλοποιήσεις που αφορούν στην προεπεξεργασία κειμένου και εξαγωγή λέξεων κλειδιών, στην κατηγοριοποίηση κειμένων αλλά και στην αυτόματη περίληψη. Εφόσον το σημαντικότερο κομμάτι του συστήματός μας που αποτελεί τον πυρήνα είναι τα συστήματα που προαναφέρθηκαν προχωρήσαμε σε υλοποίηση αυτών των συστημάτων και σε πειράματα για να εξασφαλίσουμε την καλή λειτουργία αυτών των συστημάτων. Μέχρι το τέλος του 2006 η υλοποίηση του πυρήνα του συστήματος ήταν έτοιμη με αποτέλεσμα να είναι εφικτή η εκκίνηση της υλοποίηση των άλλων κομματιών που ουσιαστικά είχαν την ανάγκη των μηχανισμών πυρήνα.

Οι περιφερειακοί μηχανισμοί χωρίστηκαν σε δύο κατηγορίες καθότι είναι αυτοί που έχουν άμεση επικοινωνία με τον «έξω» κόσμο αλλά έχουν και διαφορετικά επίπεδα πολυπλοκότητας υλοποίησης. Από τη μία μεριά υπάρχει ο μηχανισμός συγκέντρωσης πληροφορίας από το διαδίκτυο και εξαγωγής χρήσιμου κειμένου και από την άλλη μεριά βρίσκουμε το μηχανισμό προσωποποίησης και παρουσίασης της πληροφορίας στον τελικό χρήστη. Ο μεν πρώτος βασίζεται σε ένα απλοϊκό αλγόριθμο ενώ, περισσότερη πολυπλοκότητα εμφανίζεται κατά τη διάρκεια εξαγωγής χρήσιμου κειμένου ενώ ο δεύτερος μηχανισμούς βασίζεται σε σύνθετους αλγορίθμους για προσωποποίηση των δεδομένων στο χρήστη ενώ παράλληλα όλοι πρέπει να υλοποιηθούν σε PHP.

Την ίδια στιγμή παρουσιάζεται έντονα η ανάγκη για περισσότερα κείμενα εισαγωγής στο μηχανισμό προκειμένου να βελτιωθούν οι αλγόριθμοι του πυρήνα του μηχανισμού. Έτσι, η υλοποίηση των πρώτων περιφερειακών μηχανισμών κρίνεται αναγκαία.

Τέλος, προκειμένου να υπάρχει περιβάλλον διεπαφής με το χρήστη και για να εκτελεσθούν τα πειράματα που αφορούν το τελικό αποτέλεσμα που εμφανίζεται στους χρήστες τους συστήματος κατασκευάστηκε το Portal.

8.4. Ανάλυση των υποσυστημάτων

Στο μηχανισμό που δημιουργούμε καθένα από τα υποσυστήματα υλοποιεί κάποιον σύνθετο αλγόριθμο. Γι' αυτό το λόγο θα εξετάσουμε κάθε σύστημα ξεχωριστά.

8.4.1. Συλλογή Άρθρων από το Διαδίκτυο

Για τη συλλογή των πιο πρόσφατων άρθρων από το διαδίκτυο υλοποιήθηκε ένας μηχανισμός που θα «τρέχει» ανά τακτά χρονικά διαστήματα και θα συγκεντρώνει άρθρα από το διαδίκτυο. Η ιδέα για την υλοποίηση του συγκεκριμένου μηχανισμού στηρίζεται στο γεγονός πως οι μεγαλύτεροι ειδησεογραφικοί δικτυακοί τόποι διαθέτουν RSS feeds τα οποία ανανεώνονται συνέχεια με τις νέες ειδήσεις που προκύπτουν. Η ιδέα είναι να ελέγχονται ανά μία ώρα όλα τα RSS feeds των ειδησεογραφικών πρακτορείων και αν υπάρχουν νέες καταχωρήσεις τότε ο μηχανισμός να διαβάζει όλα τα νέα άρθρα και να τα προσθέτει στη βάση δεδομένων.

Δεδομένης της μορφής που έχουν τα RSS feeds ο μηχανισμός αυτός μπορεί να συλλέξει την εξής πληροφορία:

- Τίτλος Άρθρου
- URL Άρθρου

Κώδικας 2: Στοιχεία που εξάγονται από RSS feed

```
<item>
<title>Taliban extends hostage deadline</title>
<link>http://edition.cnn.com/2007/WORLD/asiapcf/07/22/afghan.hostages.reut/index.
html?eref=edition
</link>

<description>Read full story for latest details.
<a href=http://rss.cnn.com/~a/rss/edition?a=CfB4PD><img
src="http://rss.cnn.com/~a/rss/edition?i=CfB4PD" border="0"></img></a>
</description>
<pubDate>Sun, 22 Jul 2007 11:27:25 EDT</pubDate>
</item>
```

Για να εξαχθούν τα συγκεκριμένα στοιχεία χρησιμοποιείται ένας απλός wrapper ο οποίος προσπαθεί να εντοπίσει όλα τα στοιχεία <title> και όλα τα στοιχεία <link>. Οι υπόλοιπες πληροφορίες (<description> και <pubDate>)δεν είναι σημαντικές για το μηχανισμό αλλά αποθηκεύονται ως μεταδεδομένα.

Η διαδικασία για την εξαγωγή όλων των νέων άρθρων είναι απλή:

- Διάβασμα από τη ΒΔ (πίνακας rss) όλων των RSS που έχουν εισαχθεί στο σύστημα για συλλογή άρθρων
- Ανάλυση όλων των στοιχείων των RSS και προσωρινή αποθήκευσή τους σε μεταβλητή Vector (Java).
- Έλεγχος βάσει URL ή/και τίτλου προκειμένου να εξασφαλιστεί πως το άρθρο δε βρίσκεται ήδη στη βάση δεδομένων
- Εισαγωγή στη ΒΔ (πίνακας articles) όλων των τίτλων URL που δεν υπήρχαν στο σύστημα.

Αυτό είναι το ένα κομμάτι του μηχανισμού που συλλέγει όλα τα νέα άρθρα που έχουν εισαχθεί στα ειδησεογραφικά πρακτορεία την τελευταία μία ώρα, δεδομένου ότι ο μηχανισμός εκτελείται κάθε μία ώρα. Από αυτή τη διαδικασία συλλέγουμε όλα τα νέα άρθρα αλλά συγκεκριμένα οι πληροφορίες που έχουμε είναι ο τίτλος του και το URL του. Στο άλλο κομμάτι του μηχανισμού συλλέγουμε όλες τις HTML σελίδες από τα URL που έχει εντοπίσει ο wrapper. Πιο συγκεκριμένα τα βήματα του μηχανισμού είναι τα εξής:

- Ανάγνωση όλων των URL που συνέλλεξε ο wrapper
- Σύνδεση με τα URL ένα προς ένα και «κατέβασμα» της HTML σελίδας
- Αποθήκευση σε μεταβλητή vector όλων των HTML σελίδων που συγκεντρώθηκαν

Αυτό που μας ενδιαφέρει είναι να μπορέσουμε να τροφοδοτήσουμε το μηχανισμό εξαγωγής χρήσιμου κειμένου με τον HTML κώδικα. Φτάνοντας σε αυτό το σημείο του μηχανισμού έχουμε επιτύχει να διαθέτουμε για κάθε νέο άρθρο που εντοπίστηκε τα εξής στοιχεία: Τίτλος, URL, ημερομηνία, μεταδεδομένα, HTML κώδικας.

8.4.2. Εξαγωγή Χρήσιμου Κειμένου

Η εξαγωγή χρήσιμου κειμένου είναι μία διαδικασία η οποία περιλαμβάνει την απομόνωση των χρήσιμων κομματιών μίας ιστοσελίδας τα οποία στη συγκεκριμένη περίπτωση είναι τα άρθρα – ειδήσεις. Η ανάλυση και εξαγωγή του κειμένου βασίζεται στον τρόπο με τον οποίο είναι δομημένες οι σελίδες που περιέχουν

άρθρα – ειδήσεις αλλά και στο DOM μοντέλο στο οποίο μπορεί να αποδομηθεί μία HTML σελίδα.

Ο μηχανισμός εξαγωγής χρήσιμου κειμένου ακολουθεί μετά τη διαδικασία συλλογής άρθρων από το Διαδίκτυο ενώ για μεγαλύτερη ταχύτητα μπορεί να εκτελείται παράλληλα από τη στιγμή που έστω και μία νέα σελίδα συλλέγεται από τους ειδησεογραφικούς δικτυακούς τόπους.

Η εξαγωγή χρήσιμου κειμένου υλοποιείται με τη χρήση της γλώσσας προγραμματισμού Java ενώ παράλληλα έχει ξεκινήσει προσπάθεια μετατροπής του συγκεκριμένου μηχανισμού ούτως ώστε η ανάλυση να γίνεται με C++. Άλλωστε πρόκειται για μία ανάλυση χαμηλού επιπέδου με χρήση πολύπλοκων αλγορίθμων και ως εκ τούτου είναι αναμενόμενη η χρήση της C++ να οδηγήσει σε ακόμα μεγαλύτερες ταχύτητες εκτέλεσης.

Ας περάσουμε όμως στην υλοποίηση του συγκεκριμένου μηχανισμού. Όπως έχουμε ήδη δει στο κεφάλαιο 5.2.2 ο HTML κώδικας μπορεί να αναπτυχθεί σε δενδρική μορφή σύμφωνα με το DOM μοντέλο. Αυτό συνεπάγεται πως θα υπάρχουν κόμβοι αλλά και φύλλα. Στη συγκεκριμένη περίπτωση οι κόμβοι αποτελούν τα HTML tags ενώ τα φύλλα περιέχουν το κείμενο που βρίσκεται μέσα στα tags. Τα φύλλα του συγκεκριμένου δέντρου περιέχουν όλο το κείμενο όλης της ιστοσελίδας. Ωστόσο εμείς ενδιαφερόμαστε μόνο για το κομμάτι που περιέχει το άρθρο και όχι για οποιαδήποτε άλλη πληροφορίας η οποία μπορεί να είναι κάποιο άλλο κείμενο της σελίδας ή μενού πλοήγησης. Προκειμένου να πετύχουμε τη σωστή εξαγωγή πληροφορίας κάνουμε μία απλή διαπίστωση. Ο κόμβος πατέρας των φύλλων με χρήσιμο κείμενο έχει τις εξής ιδιότητες:

- Τα φύλλα του παρουσιάζουν μεγάλο ποσοστό σε κείμενο συγκριτικά με όλο το κείμενο που έχει η HTML σελίδα.
- Οι γειτονικοί του κόμβοι έχουν και αυτοί φύλλα με μεγάλο ποσοστό κειμένου συγκριτικά με όλο το κείμενο που έχει η HTML σελίδα.
- Έχουν πολύ περισσότερο κείμενο μέσα σε tags που αφορούν διαμόρφωση κειμένου (<b>, <i>, <h1>, <h2>, κλπ) παρά σε tags που αφορούν links (<a>)

Όπως φαίνεται και από τις ιδιότητες που έχουν τα φύλλα θα πρέπει να ορίσουμε συγκεκριμένες μεταβλητές για να μπορέσουμε να εξάγουμε το χρήσιμο κείμενο. Η μία μεταβλητή που χρειαζόμαστε αφορά το συνολικό κείμενο της σελίδας (μέγεθος κειμένου σε bytes). Η δεύτερη μεταβλητή αφορά το μέγεθος κειμένου κάθε φύλλου (μέγεθος κειμένου σε bytes). Η τρίτη μεταβλητή αφορά το μέγεθος κειμένου φύλλων που αφορά links. Τέλος θα πρέπει να χρησιμοποιηθούν μεταβλητές που θα εκφράζουν τη γειτονικότητα των φύλλων και συνεπώς να χρησιμοποιηθεί ένας αλγόριθμος για την αρίθμηση των κόμβων του δέντρου προκειμένου η αρίθμηση των φύλλων να είναι σειριακή. Έτσι παρά το γεγονός ότι τα φύλλα δεν είναι στο ίδιο βάθος θα πρέπει να ορίσουμε μία μεταβλητή που να αποθηκεύει την αρίθμηση των φύλλων. Επειδή ο αλγόριθμος κατασκευής του δένδρου από την ανάλυση της HTML σελίδας είναι depth first χρησιμοποιούμε έναν επιπλέον μετρητή ο οποίος σηματοδοτεί το κάθε φύλλο και αυξάνεται με την εύρεση νέου φύλλου.

Από τα προαναφερθέντα καταλήγουμε στους παρακάτω παράγοντες:

- S_H = το συνολικό μέγεθος του κειμένου σε bytes. Υπολογίζεται προσθέτωντας όλα τα S_{Lx}.
- S_{Lx} = το μέγεθος κειμένου σε bytes για το φύλλο X. Υπολογίζεται μετρώντας τα bytes αλφαριθμητικών χαρακτήρων σε ένα φύλλο
- S_{Ax} = το μέγεθος κειμένου του φύλλου X που περιέχεται σε tag <a> (link). Υπολογίζεται μετρώντας τα bytes αλφαριθμητικών μέσα σε tags <a> ενός φύλλου.
- I_X = το αναγνωριστικό κάθε φύλλου σύμφωνα με το μετρητή φύλλων

Για την αναγνώριση ενός φύλλου σαν φύλλο που περιέχει χρήσιμο κείμενο θα πρέπει να ισχύουν συγκεκριμένες προϋποθέσεις που αφορούν τα ποσοστά κειμένου μέσα σε αυτό συγκριτικά με το συνολικό κείμενο της σελίδας και συγκριτικά με το κείμενο που αφορά συνδέσμους. Έτσι για κάθε φύλλο ελέγχουμε τις ποσότητες:

$LP = S_{Ax} / S_{Lx}$. Πρόκειται για το Link Percentage το οποίο είναι μία ποσότητα που μας δείχνει πόσο από το κείμενο ενός φύλλου είναι κείμενο που βρίσκεται σε link. Αν αυτή η ποσότητα είναι μεγάλη αυτό σημαίνει πως ο συγκεκριμένος κόμβος είναι ένα navigation menu που η πλειονότητα του κειμένου του βρίσκεται μέσα σε links συνεπώς δε μπορεί να είναι το κείμενο ενός άρθρου το οποίο συνήθως δεν περιέχει πολλά links.

$TP = S_{Lx} / S_H$. Πρόκειται για το Text Percentage το οποίο είναι μία ποσότητα που μας δείχνει πόσο κείμενο περιέχει ένα φύλλο συγκριτικά με το κείμενο ολόκληρης της σελίδας. Αν αυτή η ποσότητα είναι μεγάλη τότε συνεπάγεται πως το κείμενο αυτού του φύλλου ενδέχεται να είναι «χρήσιμο κείμενο».

Αφού απορρίψουμε όλα τα φύλλα με μεγάλο LP και κρατήσουμε όλα τα φύλλα με μεγάλο TP υπολογίζουμε πόσο κοντά (distance) είναι οι κόμβοι με μεγάλο TP. Ο αλγόριθμος είναι απλός και συνίσταται στον υπολογισμό της διαφοράς των τιμών I_X κάθε φύλλου. $D_{X,Y} = I_Y - I_X$.

Τα νούμερα που ορίζουν τα όρια για τα LP, TP και D εξήχθησαν μετά από πειραματικές διαδικασίες σε διάφορους δικτυακούς τόπους που περιείχαν άρθρα και ειδήσεις. Χαρακτηριστικό είναι το παράδειγμα που φαίνεται από το παρακάτω σχήμα για τη λειτουργία του μηχανισμού.

Εικόνα 21: Χαρακτηρισμός περιοχών ιστοσελίδας από το ηχανισμός εξαγωγής χρήσιμου κειμένου

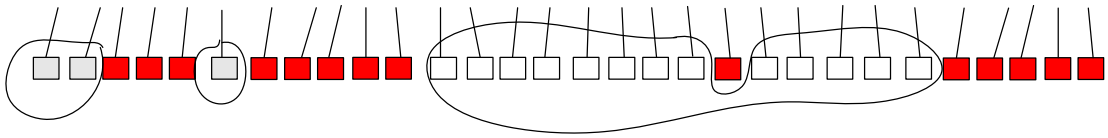
Όπως φαίνεται και από το παραπάνω σχήμα υπάρχουν περιοχές στο δικτυακό τόπο οι οποίες περιέχουν το κείμενο του άρθρου ενώ άλλες έχουν κείμενο το οποίο δεν αφορά το άρθρο. Οι περιοχές που είναι με κόκκινο χρώμα έχουν αποκλειστεί από χρήσιμο κείμενου λόγω πολύ υψηλού LP. Οι περιοχές με μπλε χρώμα είναι περιοχές που έχουν αποκλειστεί είτε λόγω πολύ χαμηλού TP ή λόγω πολύ ψηλού D. Οι περιοχές με πράσινο χρώμα είναι αυτές που επιλέγονται από το σύστημα σαν το κύριο σώμα του άρθρου.

Ο αλγόριθμος για το σωστό υπολογισμό των παραπάνω περιλαμβάνει τα παρακάτω βήματα:

- Αποδόμηση της HTML σελίδας
- Δημιουργία του DOM μοντέλου με τα tags να αποτελούν κόμβους και τα φύλλα να περιλαμβάνουν μόνο κείμενο.
- Μαρκάρισμα κάθε φύλλου του δένδρου με ένα μοναδικό αναγνωριστικό για το σωστό υπολογισμό της απόστασης.
- Υπολογισμούς των bytes αλφαριθμητικών κάθε φύλλου
- Μαρκάρισμα του κειμένου που βρίσκεται μέσα σε σύνδεσμο (<a> tag)
- Για κάθε φύλλο
    - Υπολογισμός του LP

- Αν το LP είναι μεγαλύτερο από 0,42 τότε το κείμενο του φύλλου απορρίπτεται
- Αν το TP είναι μικρότερο από 0,18 τότε το κείμενο του φύλλου απορρίπτεται
- Υπολογισμός των D για τα φύλλα που έχουν απομείνει και αν D>3 τότε απόρριψη του κειμένου του φύλλου.

Η επιλογή βάσει γειτνίασης των φύλλων δεν είναι τόσο απλή όσο περιγράφεται παραπάνω. Ουσιαστικά περιλαμβάνει ένα σύνθετο αλγόριθμο που δημιουργεί ομάδες από γειτονικά φύλλα όπως φαίνεται στο παρακάτω σχήμα.



Εικόνα 22: Ομάδες γειτονικών φύλλων

Όπως μπορούμε να δούμε υπάρχουν αρχικά δύο φύλλα τα οποία περιέχουν αρκετό κείμενο ώστε να χαρακτηριστεί χρήσιμο κείμενο αλλά είναι πολύ μακριά από άλλα τέτοια φύλλα. Στη συνέχεια παρουσιάζεται ένα μεμονωμένο και έπειτα μία συστάδα από φύλλα τα οποία έχουν χαρακτηριστεί σαν φύλλα με χρήσιμο κείμενο και τα αποδέχεται ο μηχανισμός. Το συγκεκριμένο παράδειγμα θα μπορούσε να είναι της σελίδες που είδαμε στο παραπάνω σχήμα. Τα πρώτα φύλλα είναι αυτά που περιέχουν τον τίτλο της σελίδας (όχι του άρθρου) ή γενικά στοιχεία που υπάρχουν στη σελίδα ενώ στο σημείο που είναι πολλά φύλλα μαζί βλέπουμε το κυρίως σώμα. Το κόκκινο φύλλο ενδιάμεσα θα μπορούσε να είναι το φύλλο που περιέχει το κείμενο της εικόνας του άρθρου που προφανώς και θέλουμε να απορρίψουμε.

Με αυτό τον τρόπο ο μηχανισμός εξαγωγής χρήσιμου κειμένου είναι σε θέση να μας παρέχει αποκλειστικά και μόνο με χρήσιμο κείμενο που εξάγει από τις σελίδες που έχει ανακτήσει το σύστημα με το μηχανισμό συλλογής άρθρων από το διαδίκτυο.

8.4.3. Προεπεξεργασία

Η προεπεξεργασία των κειμένων που δέχεται ο μηχανισμός ως είσοδο, αποτελεί μια βασική και σημαντική διαδικασία του όλου συστήματος, καθώς είναι αυτή που τροφοδοτεί τα συστήματα ανάκτησης πληροφορίας που ακολουθούν με την κατάλληλη είσοδο, η οποία θα πρέπει να είναι σε τέτοια μορφή, ώστε ο μηχανισμός να μπορεί να παράγει ικανοποιητικά αποτελέσματα σαν σύνολο. Αφορά και τη διαδικασία της εξαγωγής κωδικολέξεων (keyword extraction) και πρόκειται ουσιαστικά για μια ακολουθιακή διαδικασία, η οποία μπορεί να θεωρηθεί ως ένα module του όλου συστήματος (και επομένως να αντιμετωπιστεί ξεχωριστά από αυτό).

Το υποσύστημα προεπεξεργασίας δέχεται ως είσοδο ένα πλήθος παραμέτρων:

- Το όνομα του XML αρχείου που περιέχει τα απαραίτητα στοιχεία του κειμένου (τίτλος, σώμα, ID και ενδεχόμενα την κατηγορία του)
- Το ελάχιστο μήκος λέξεων που πρέπει να κρατηθούν
- Ένα σύνολο από λέξεις τερματισμού (stopwords), οι οποίες αφαιρούνται από το κείμενο
- Πληροφορία σχετικά με το ελάχιστο μήκος λέξεων που πρέπει να κρατηθούν και για το αν θα κρατηθούν τα ψηφία (αριθμοί) του κειμένου

Η διαδικασία που ακολουθείται στη συνέχεια περιγράφεται από τα παρακάτω βήματα:

- Parsing του XML αρχείου ώστε να εξαχθούν τα στοιχεία που περιέχει (τίτλος, σώμα κειμένου, είδος (κατηγορία) και αναγνωριστικό (ID).
- Αφαίρεση των σημείων στίξης (punctuation removal) από τον τίτλο του κειμένου και πέρασμα από τον stemmer
- Διαχωρισμός των προτάσεων του κειμένου
- Αφαίρεση των σημείων στίξης του κειμένου
- Αφαίρεση των μεγάλων κενών που υπάρχουν στις προτάσεις του κειμένου. Πλέον κάθε λέξη έχει απόσταση ενός κενού από την επόμενη
- Διαγραφή των stopwords με σύγκριση των λέξεων των προτάσεων με αυτές που έχουν δοθεί ως είσοδος
- Εξαγωγή μεμονωμένων λέξεων από τις προτάσεις (keywords)
- Πέρασμα των keywords του κειμένου από τη διαδικασία του stemming.
- Αντιστοίχηση των keywords με τις αρχικές προτάσεις του κειμένου και εύρεση απόλυτης συχνότητας εμφάνισης του κάθε keyword μέσα στο κείμενο
- Κράτημα του ποσοστού των keywords που μας ενδιαφέρει (εξαρτάται από τις διαδικασίες που ακολουθούν το k/w extraction και είναι συνήθως 30-50% των συνολικών keywords

Η έξοδος που προκύπτει από τη διαδικασία προεπεξεργασίας κειμένου και εξαγωγής keywords που περιγράφηκε είναι:

- Μια λίστα από keywords διατεταγμένη κατά φθίνουσα σειρά συχνότητας εμφάνισης
- Οι σχετικές και απόλυτες συχνότητες εμφάνισης του κάθε keyword μέσα στο κείμενο
- Οι προτάσεις στις οποίες εμφανίζεται το κάθε keyword (π.χ. 1η, 3η, κ.ο.κ)

Οι παραπάνω έξοδοι του μηχανισμού keyword extraction, κωδικοποιούνται κατάλληλα σε αρχείο XML και παρέχονται ως είσοδος στο μηχανισμό που ακολουθεί. Επίσης, είναι δυνατό με κατάλληλο switch στη συνάρτηση που υλοποιεί τη διαδικασία, η έξοδος να αποθηκευθεί απ' ευθείας στη βάση δεδομένων του συστήματος απ' όπου ο μηχανισμός ανάκτησης πληροφορίας που ακολουθεί (περίληψη ή κατηγοριοποίηση κειμένου) να λάβει τις απαραίτητες εισόδους ασύγχρονα.

8.4.4. Κατηγοριοποίηση

Η κατηγοριοποίηση κειμένου επιτελεί μια ουσιαστική διαδικασία για το μηχανισμό καθώς μπορεί, δεδομένης μιας βάσης γνώσης κειμένων και ενός συνόλου κατηγοριών, να χαρακτηρίσει ένα νέο κείμενο ταξινομώντας το κατάλληλα με κάποια συσχέτιση στις κατηγορίες. Συνήθως, και εφόσον το κείμενο ανήκει σε κάποια από τις κατηγορίες, η συσχέτιση με αυτή την κατηγορία θα είναι σχετικά μεγάλη, ενώ με τις υπόλοιπες θα είναι πολύ μικρότερη.

Το υποσύστημα κατηγοριοποίησης μπορεί να λειτουργήσει με δύο τρόπους: είτε κατηγοριοποιώντας το κείμενο που δίνεται στην είσοδο, είτε προσθέτοντας το κείμενο της εισόδου στην δυναμική βάση γνώσης (training set) του συστήματος. Φυσικά για την ύστερη περίπτωση, προηγούμενη γνώση για την κατηγορία του κειμένου είναι απαραίτητη. Οι είσοδοι του υποσυστήματος κατηγοριοποίησης κειμένου περιλαμβάνουν τα εξής:

- Το XML αρχείο του keyword extraction που περιέχει τα keywords του κειμένου, τις συχνότητες εμφάνισής τους και τις θέσεις τους στο κείμενο (το τελευταίο στοιχείο δεν χρειάζεται για την κατηγοριοποίηση).

- Ένα training set flag που αντιπροσωπεύει τον τρόπο λειτουργίας από τους δύο που περιγράφηκαν προηγουμένως.
- Το ποσοστό των keywords που θα πρέπει να κρατηθούν από το XML αρχείο των keywords. Αυτό δίνεται ξεχωριστά και ανεξάρτητα από τη διαδικασία του keyword extraction γιατί κρατάμε διαφορετικά μεγέθη του συνόλου των keywords σε κάθε περίπτωση. Οι λόγοι και οι επιλογές οι οποίες γίνονται περιγράφονται αναλυτικά στη συνέχεια.

8.4.4.1. Ποσοστό των keywords για training set

Για την περίπτωση που έχουμε να κάνουμε προσθήκη στο training set της βάσης γνώσης μας, κρατούμε ένα ποσοστό 50% των αρχικών keywords λόγω του ότι θέλουμε το κείμενο να προσθέσει τη δικιά του συσχέτιση στην κατηγορία όπου εισάγεται (να μεταβάλει επομένως ελαφρώς την κατηγορία) και επομένως νέα κείμενα που εισέρχονται στο σύστημα και μοιάζουν με αυτό που προστέθηκε στο training set, να κατηγοριοποιούνται στην ίδια κατηγορία. Επίσης η επιλογή του 50% αποκλείει την συμπερίληψη keywords στην κατηγορία τα οποία έχουν πολύ μικρή συχνότητα εμφάνισης στο κείμενο και επομένως δεν είναι τόσο αντιπροσωπευτικά αυτού, άρα και της κατηγορίας.

8.4.4.2. Ποσοστό των keywords για κατηγοριοποίηση

Το ποσοστό των keywords που κρατούνται για την κατηγοριοποίηση ενός νέου άρθρου από το σύστημα είναι 30%. Το αποτέλεσμα αυτό προέκυψε ύστερα από πειραματική διαδικασία η οποία περιγράφεται σε προηγούμενη ενότητα και μας επιτρέπει να έχουμε πολύ καλή συσχέτιση των keywords με το αρχικό κείμενο, χωρίς παράλληλα να χρειάζεται να υπερφορτώνουμε τη βάση δεδομένων με μη χρήσιμα στοιχεία. Η επιλογή αυτή μας δίνει παράλληλα και πλεονέκτημα χρόνου στην εκτέλεση του αλγορίθμου κατηγοριοποίησης, μιας και υπάρχουν λιγότερα keywords τα οποία θα πρέπει να συγκριθούν με σχετικότητες αποθηκευμένες στη βάση δεδομένων.

8.4.4.3. Διαδικασία κατηγοριοποίησης

Η διαδικασία της κατηγοριοποίησης ενός νέου άρθρου προχωρά ως εξής:

1. Ανακτώνται οι συχνότητες των keyword του κειμένου με κάθε κατηγορία από τη ΒΔ (τα keyword που εμφανίζονται στο κείμενο και υπάρχουν και στην κατηγορία). Έχουμε επομένως, εκτός από τον αρχικό πίνακα συχνοτήτων των keywords του κειμένου, και ένα πίνακα για κάθε κατηγορία που περιέχει συχνότητα εμφάνισης για την κατηγορία του κάθε keyword του κειμένου που εμφανίζεται και στην κατηγορία. Προφανώς, αν κάποιο από τα keywords του κειμένου δεν εμφανίζεται στην εκάστοτε κατηγορία, η αντίστοιχη θέση στον πίνακα συχνοτήτων της κατηγορίας θα έχει την τιμή 0 (μη εμφάνιση).
2. Υπολογίζονται τα μέτρα των προηγούμενων πινάκων, το εσωτερικό τους γινόμενο και από αυτά, η ομοιότητα συνημίτονου μεταξύ τους. Έχουμε επομένως για κάθε κατηγορία, μια συσχέτιση του κειμένου, κάτι που είναι και το ζητούμενο.
3. Οι συσχετίσεις κειμένου-κατηγορίας ταξινομούνται κατά φθίνουσα σειρά.
4. Πλέον οι συσχετίσεις μπορούν να αποθηκευτούν στη βάση και η κατηγοριοποίηση του κειμένου έχει ολοκληρωθεί.

8.4.4.4. Διαδικασία προσθήκης στο training set

Όπως ήδη αναφέρθηκε, το module της κατηγοριοποίησης είναι εκείνο που διαχειρίζεται τη βάση γνώσης του συστήματος και επομένως έχει τη δυνατότητα

προσθήκης νέων άρθρων, αντιπροσωπευτικών των κατηγοριών, σε αυτή. Η διαδικασία που ακολουθείται είναι απλή: έχοντας ως δεδομένα τα keywords του κειμένου (50% των συνολικών) και την κατηγορία την οποία αντιπροσωπεύει το κείμενο, εισάγονται (εάν δεν υπάρχουν ήδη) τα keywords του κειμένου και ανανεώνονται (η εισάγονται) οι συσχετίσεις των keywords με την κατηγορία, στους κατάλληλους training πίνακες. Τέλος εισάγεται και η συσχέτιση των keywords που εισήχθησαν (η τροποποιήθηκαν) με το κείμενο που μόλις μπήκε στο training set. Φυσικά σε κάθε περίπτωση εξασφαλίζεται η μη ύπαρξη διπλοεγγραφών στη βάση και η γενικότερη συνέπεια των δεδομένων. Η διαδικασία ανανέωσης του training set είναι πολύ ευκολότερη αφού δεν εμπεριέχει υπολογισμούς εσωτερικών γινομένων ή ομοιοτήτων συνημίτονου όπως η διαδικασία της κατηγοριοποίησης, παρά μόνο συναλλαγές με τη ΒΔ.

8.4.5. Αυτόματη Εξαγωγή Περίληψης

Η διαδικασία της εξαγωγής περίληψης κειμένου, δέχεται για είσοδο την έξοδο του keyword extraction του μηχανισμού, δηλαδή τις εξαγμένες κωδικολέξεις μαζί με τις συχνότητες εμφάνισής τους στο κείμενο καθώς και τις θέσεις τους στις προτάσεις. Επίσης ως είσοδος δίνεται το μέγεθος της απάντησης που επιθυμούμε και αν υπάρχει, πληροφορία για την κατηγορία του κειμένου.

Έχει ήδη αναφερθεί, ότι η διαδικασία αυτόματης εξαγωγής περίληψης δίνει ένα βαθμό, ή αλλιώς ένα σκορ, σε κάθε πρόταση του κειμένου ανάλογα με τη σχετικότητα που εκτιμά πως έχει. Το σκορ της κάθε πρότασης σχηματίζεται με τη βοήθεια ενός πλήθους παραμέτρων που αφορούν στη συχνότητα εμφάνισης του keyword στο κείμενο, στην πιθανότητα εμφάνισης του keyword στον τίτλο του κειμένου, στην κατηγορία στην οποία ανήκει το κείμενο και τέλος στις ιδιαίτερες προτιμήσεις του χρήστη για τις κατηγορίες και επομένως για ορισμένα keywords. Το σύνολο των παραμέτρων συνοψίζεται στη σχέση. Για την υλοποίησή μας, και ύστερα από πειράματα, καταλήξαμε στο να θέσουμε τον μεν παράγοντα , ενώ τον παράγοντα . Ο πρώτος αφορά στην περίπτωση εμφάνισης του keyword στον τίτλο, ενώ ο δεύτερος στη συχνότητα εμφάνισης του keyword στο κείμενο. Η διαδικασία έχοντας αυτές τις δύο παραμέτρους μπορεί να δώσει μια βασική βαθμολόγηση για τις προτάσεις του κειμένου και επομένως μια περίληψη.

Η ποιότητα της περίληψης αυξάνεται δραματικά με την χρήση των επόμενων δύο ευρετικών.

- Έχοντας την πληροφορία για την κατηγορία του κειμένου, η διαδικασία αναζητεί για κάθε keyword κάθε πρότασης την σχετικότητα που έχει το keyword με την κατηγορία. Σημειώνοντας το πόσο σχετικό είναι το κάθε keyword ή όχι με την κατηγορία, οι προτάσεις βαθμολογούνται με θετικό βάρος για κάθε keyword σχετικό που περιέχουν και με αρνητικό βάρος για κάθε keyword μη σχετικό που έχουν. Το τελικό σκορ των προτάσεων προκύπτει ύστερα από την προσθαφαίρεση όλων των βαρών. Με αυτό τον τρόπο, οι προτάσεις που περιέχουν keywords αντιπροσωπευτικά της κατηγορίας επιτυγχάνουν υψηλότερο σκορ σε σχέση με άλλες που δεν περιέχουν πολλά αντιπροσωπευτικά keywords και επιτυγχάνουν ουδέτερο σκορ (κοντά στο 0), ή με άλλες που έχουν πολλά αντιπροσωπευτικά άλλων κατηγοριών keywords και επιτυγχάνουν αρνητικό σκορ.
- Θέλοντας να παράγουμε μια προσωποποιημένη περίληψη για κάποιον χρήστη με δεδομένο προφίλ στο σύστημα, βαθμολογούμε υψηλότερα προτάσεις που περιέχουν keywords αντιπροσωπευτικά των προτιμήσεων του χρήστη (keywords που ανήκουν με υψηλή θετική βαρύτητα στο προφίλ του χρήστη), ή βαθμολογούμε χαμηλότερα ή αρνητικά προτάσεις που περιέχουν keywords μη αντιπροσωπευτικά των προτιμήσεων του χρήστη. Η βαθμολόγηση γίνεται όπως και στην περίπτωση της κατηγοριοποιημένης περίληψης αναζητώντας για κάθε keyword, κάθε πρότασης, τη σημαντικότητα που έχει για τον χρήστη.

8.4.6. Προσωποποίηση στο χρήστη

Η προσωποποίηση στο χρήστη είναι ένα από τα πιο σημαντικά κομμάτια του συστήματος καθώς σε αυτό το στάδιο διαμορφώνεται το δυναμικό προφίλ και προβάλλονται πίσω στο χρήστη όλα τα αποτελέσματα των προηγούμενων μηχανισμών.

Η προσωποποίηση στο χρήστη γίνεται σε επίπεδο διαδικτύου με τη συνεργασία PHP, AJAX και βάσης δεδομένων. Η προσωποποίηση βασίζεται σε συγκεκριμένες παραμέτρους προκειμένου να είναι πληρέστερη και να είναι εφικτή η καλύτερη δημιουργία προφίλ χρήστη. Οι παράμετροι που θέσαμε στο σύστημα για την προσωποποίηση είναι:

- Οι επιλογές του χρήστη που αφορούν τις κατηγορίες που έχει το σύστημα (μόλις κάνει εγγραφή)
  - Βαθμολόγηση των κατηγοριών ανάλογα με το πόσο ενδιαφέρουν το χρήστη
- Οι επιλογές του χρήστη μόλις του εμφανίζονται άρθρα
  - Επιλογή του χρήστη να διαβάσει ένα άρθρο
  - Επιλογή του χρήστη να μη διαβάσει ένα άρθρο
- Οι επιλογές του χρήστη την ώρα που διαβάζει ένα άρθρο
  - Πόσο χρόνο καταναλώνει σε ένα άρθρο
  - Πόσα άρθρα του ίδιου θέματος διαβάζει κατά τη διάρκεια μίας συνεδρίας
  - Η επιλογή να στείλει το άρθρο σε ένα φίλο
  - Η επιλογή να τυπώσει το άρθρο
  - Η επιλογή να αναζητήσει παραπλήσια άρθρα σε όλη τη συλλογή.

Όλα τα παραπάνω αποτελούν παραμέτρους που διαμορφώνουν το προφίλ ενός χρήστη. Όμως ας δούμε τι εννοούμε όταν αναφερόμαστε στο προφίλ ενός χρήστη. Δεδομένων των διαδικασιών με τις οποίες εξάγονται τα αποτελέσματα τόσο για την κατηγοριοποίηση (επιλογή σε ποια κατηγορία ανήκει ένα άρθρο που μόλις μπήκε στο σύστημα) όσο και για τις περιλήψεις έχουμε δει πως αυτό που έχει τη μεγαλύτερη σημασία είναι να εντοπίσουμε τις λέξεις κλειδιά. Έτσι, λοιπόν, και για το προφίλ του χρήστη αυτό που πραγματοποιούμε είναι να δημιουργήσουμε λίστες με λέξεις κλειδιά που έχουν κάποια βάρη. Σε αυτή την περίπτωση τα βάρη είναι θετικά και αρνητικά και προδίδουν το κατά πόσο ο χρήστης ενδιαφέρεται για κάποια λέξη κλειδί ή όχι καθώς και το μέγεθος ενδιαφέροντος.

8.4.6.1. Αλγόριθμος διαμόρφωσης αρχικού προφίλ

Ως δεδομένα έχουμε στο σύστημά μας έχουμε 7 κατηγορίες τις οποίες τις χαρακτηρίζουν λέξεις κλειδιά με συγκεκριμένα βάρη. Ο παρακάτω πίνακας δείχνει ένα τέτοιο παράδειγμα για μία από τις κατηγορίες του συστήματός μας.

Πίνακας 4: Πίνακας συσχέτισης λέξεων κλειδιών με κατηγορία

Cat_id	Kw_id	Rel_frequency	Abs_frequency
1	42	0.00105974	298
1	43	0.000927275	201
1	44	0.00172208	201
1	41	0.0103325	188
1	37	0.00516625	150
1	228	0.0149689	148

1	45	0.00251689	141

Το σκεπτικό είναι πως κάθε χρήστης αντιπροσωπεύει μία υποκατηγορία ή πιο σωστά, μία σειρά από υποκατηγορίες. Αυτό σημαίνει πως εφόσον οι λίστες με τις λέξεις κλειδιά δύνανται να χαρακτηρίσουν μία κατηγορία αυτό συνεπάγεται και πως λίστες με λέξεις κλειδιά δύνανται να χαρακτηρίσουν τις επιλογές και τις προτιμήσεις ενός χρήστης. Αυτό που μας ενδιαφέρει συνεπώς είναι να μπορέσουμε από τις διαδικασίες που περιγράψαμε παραπάνω να καταλήγουμε σε λέξεις κλειδιά και συγκεκριμένα βάρη σε κάθε μία προκειμένου να χαρακτηρίσουμε το χρήστη. Σε πρώτη φάση αυτό που κάνουμε είναι να διαμορφώσουμε κάποιο αρχικό προφίλ για το χρήστη κατά τη διάρκεια που πραγματοποιεί εγγραφή στο σύστημα. Δεδομένου ότι θέλουμε να κρατήσουμε τις διαδικασίες όσο το δυνατόν πιο διαφανείς προς τους χρήστες είναι ίσως το μόνο σημείο που μπορούμε ανώδυνα να βάλουμε το χρήστη στη διαδικασία του να συμπληρώσει κάποια στοιχεία για το προφίλ του.

Η διαδικασία εγγραφής και γενικά το περιβάλλον διεπαφής αποτελούν την βασική μονάδα επικοινωνίας του χρήστη με το σύστημα. Ένας χρήστης εγγράφεται στο σύστημα δίνοντας πληροφορίες για το μέγεθος της συσκευής που χρησιμοποιεί και δίνοντας πληροφορίες για τις κατηγορίες που θέλει να παρακολουθεί. Ένας χρήστης είναι δυνατόν να αλλάξει τα στοιχεία του μελλοντικά, κάτι που βέβαια δεν επηρεάζει άμεσα τα στοιχεία που έχουν ήδη συλλεγεί για το προφίλ του, εκτός κι αν ο ίδιος επιθυμεί δημιουργία από την αρχή του προφίλ που ήδη έχει. Οι πληροφορίες αποθηκεύονται στην κεντρικοποιημένη βάση δεδομένων και ανανεώνονται συνεχώς με το δυναμικό προφίλ του όπως θα δούμε στην επόμενη ενότητα. Όταν ο χρήστης βρίσκεται στη διαδικασία εγγραφής στο σύστημα του παρουσιάζονται όλες οι κατηγορίες του συστήματος και του ζητήται να δηλώσει την προτίμησή του για κάθε κατηγορία. Ο χρήστης καλείται να επιλέξει μία βαθμολογία για κάθε κατηγορία από -5 έως 5. Το -5 μεταφράζεται σαν η κατηγορία δε με αντιπροσωπεύει καθόλου ενώ το +5 σημαίνει πως η κατηγορία αντιπροσωπεύει απόλυτα το χρήστη. Η επιλογή του 0 σαν προτίμηση κατηγορίας μεταφράζεται σαν ουδέτερη στάση απέναντι στην κατηγορία. Εκμεταλλευόμενοι τις απαντήσεις των χρηστών μπορούμε να διαμορφώσουμε ένα αρχικό προφίλ για το χρήστη. Αυτό γίνεται ως εξής. Αρχικά δημιουργούμε εγγραφές για τις κατηγορίες που αρέσουν στο χρήστη και γι αυτές που ο χρήστης δεν προτιμά. Αυτό θα μας βοηθήσει να κάνουμε ένα πρώτο ξεκαθάρισμα των άρθρων ανάμεσα σε αυτά που ο χρήστης θέλει να δει και σε αυτά που δεν τον ενδιαφέρουν, ανάλογα με τις γενικές κατηγορίες που έχει επιλέξει. Ο χρήστης όμως δεν επιλέγει απλώς τι θέλει να βλέπει και τι δε θέλει. Έχει δώσει και κάποια βαθμολογία για κάθε κατηγορία. Χρησιμοποιώντας αυτά τα δεδομένα μπορούμε να δημιουργήσουμε μία πιο αναλυτική περιγραφή του προφίλ. Το αναλυτικό προφίλ όπως έχει ήδη αναφερθεί περιλαμβάνει λίστες με λέξεις κλειδιά όπως αυτές που υπάρχουν για τις κατηγορίες που δείχνουν ποιες λέξεις κλειδιά ενδιαφέρουν το χρήστη και ποιες δεν τον αφορούν. Σε αυτή την περίπτωση επιτρέπονται τόσο θετικά βάρη όσο και αρνητικά. Ο υπολογισμός των βαρών για τις λέξεις κλειδιά του χρήστη υπολογίζονται από τον παρακάτω αλγόριθμο.

Κώδικας 3: Αλγόριθμος εξαγωγής των λέξεων κλειδιών του χρήστη

```
For each (selection s) {
 If (s!=0) {
  Keyword_name_usr = select 20*s keywords from category keywords
  // the keywords used for categorization, summarization etc
  Keyword_weight_usr = select (2*s*relative frequency) from category keywords
  // the same list as above
}
else {
  Keyword_name_usr = select 10 keywords from category keywords
  Keyword_weight_usr = select relative_frequency from category. keywords
}
```

```
Insert into user profile keyword_name_usr, keyword_weight_usr
  If exists
    Update user profile set keyword_weight += keyword_weight_usr where
keyword_name = keyword_name_usr
}
```

Υποθέτουμε ότι ο χρήστης κάνει κάποιες επιλογές για τις κατηγορίες και επιλέγει από -5 έως 5. Από αυτές τις επιλογές επιλέγουμε 20s λέξεις κλειδιά, όπου s είναι η επιλογή του χρήστη (s Є [-5..5]) από τη λίστα με τις λέξεις κλειδιά που αφορούν την κατηγορία, όπως ο πίνακας που είδαμε παραπάνω. Εν συνεχεία, επιλέγουμε τη σχετική συχνότητα κάθε λέξης και την πολλαπλασιάζουμε με 2s. Αν για παράδειγμα ο χρήστης έχει επιλέξει για μία κατηγορία την επιλογή -3 και μία συγκεκριμένη λέξη κλειδί για την κατηγορία έχει σχετική συχνότητα 0,12 τότε στον πίνακα του χρήστη η συγκεκριμένη λέξη θα πάρει σχετική συχνότητα -0,12. αυτός ο αριθμός μας δείχνει και το πόσο ο χρήστης ενδιαφέρεται για τη συγκεκριμένη λέξη κλειδί. Στο παράδειγμα που δείξαμε ο χρήστης δεν ενδιαφέρεται για τη συγκεκριμένη λέξη. Πραγματοποιώντας αυτή τη διαδικασία καταλήγουμε σε μία αρχική λίστα με λέξεις κλειδιά και σχετικές συχνότητες για το χρήστη οι οποίες μας δίνουν τα παρακάτω στοιχεία:

- Πολλές λέξεις κλειδιά από τις κατηγορίες που έχει επιλέξει ο χρήστης με μεγάλο σκόρ, είτε θετικό είτε αρνητικό και παράλληλα πολύ λίγες λέξεις από τις κατηγορίες που έχει δηλώσει ο χρήστης με χαμηλό σκορ. Πρόκειται για κατηγορίες που είναι αδιάφορες στο χρήστη και άρα, λέξεις κλειδιά από αυτές τις κατηγορίε δεν είναι απαραίτητες για το προφιλ του χρήστη.
- Μεγάλη θετική τιμή για τις σχετικές συχνότητες των λέξεων κλειδιών που ανήκουν στις κατηγορίες που έχει επιλέξει ο χρήστης με μεγάλο σκορ και μεγάλη απόλυτα αρνητική τιμή για τις σχετικές συχνότητες των λέξεων κλειδιών που ανήκουν σε κατηγορίες που έχει επιλέξει ο χρήστης με πολύ μικρό σκορ.

Αυτά τα στοιχεία μπορούν να μας δώσουν πληροφορίες για να εξάγουμε τα παρακάτω στοιχεία:

- Επιλογή κειμένων από τις κατηγορίες που ενδιαφέρουν το χρήστη
- Αποφυγή επιλογής κειμένων από κατηγορίες που δεν ενδιαφέρουν το χρήστη
- Επιλογή κειμένων από κατηγορίες που ενδιαφέρουν το χρήστη ενώ παράλληλα δεν ανήκουν σε κατηγορίες που δεν ενδιαφέρουν το χρήστη (να θυμίσουμε πως ένα κείμενο ανήκει σε πολλές κατηγορίες)
- Ξεκαθάρισμα των αποτελεσμάτων του μηχανισμού αυτόματης εξαγωγής περίληψης προσθέτοντας τον παράγοντα προσωποποίησης.

Η προαναφερθείσα διαδικασία, συμπεριλαμβανομένης και της κατασκευής της λίστας με τις λέξεις κλειδιά πραγματοποιήθηκε προκειμένου να έχουμε κάποια πρώτα στοιχεία για το αρχικό προφίλ του χρήστη. Στη συνέχεια θα περάσουμε στην κατασκευή του δυναμικού προφίλ χρήστη, το οποίο μεταβάλλεται με τη χρήση του δικτυακού τόπου. Είναι σημαντικό το γεγονός πως όσο περισσότερο χρησιμοποιεί ο χρήστης το δικτυακό τόπο, τόσο καλύτερα διαμορφώνεται το προφίλ του.

8.4.6.2. Δυναμική διαμόρφωση προφίλ χρήστη

Όσο ο χρήστης χρησιμοποιεί το δικτυακό τόπο, τόσο καλύτερα διαμορφώνεται το προφίλ του από τα στοιχεία που συλλέγονται από τις επιλογές του. Όπως έχουμε ήδη αναφέρει τα στοιχεία που ελέγχονται για τη δυναμική διαμόρφωση του προφίλ του χρήστη είναι:

- Οι επιλογές του χρήστη μόλις του εμφανίζονται άρθρα

- Επιλογή του χρήστη να διαβάσει ένα άρθρο
- Επιλογή του χρήστη να μη διαβάσει ένα άρθρο

- Οι επιλογές του χρήστη την ώρα που διαβάζει ένα άρθρο
  - Πόσο χρόνο καταναλώνει σε ένα άρθρο
  - Πόσα άρθρα του ίδιου θέματος διαβάζει κατά τη διάρκεια μίας συνεδρίας
  - Η επιλογή να στείλει το άρθρο σε ένα φίλο
  - Η επιλογή να τυπώσει το άρθρο
  - Η επιλογή να αναζητήσει παραπλήσια άρθρα σε όλη τη συλλογή.

Ας δούμε αναλυτικά πως συμπεριφέρεται ο μηχανισμός ανάλογα με τις επιλογές που κάνει ένας χρήστης.

8.4.6.3. Επιλογές του χρήστη μόλις εμφανίζονται σε αυτόν άρθρα

Από το χρήστη του συστήματός μας περιμένουμε όταν του εμφανιστούν τα τελευταία 20 άρθρα, κάποια από αυτά να τα διαβάσει και άλλα να μην τα δει καθόλου. Και οι δύο αυτές αντιδράσεις κάτι μπορεί να σημαίνουν όμως και γι αυτό κάθε τέτοιο στοιχεία είναι αντικείμενο μελέτης για το μηχανισμό μας. Αυτό που μπορούμε να καταλάβουν δημιουργώντας εικονικά προφίλ στο μηχανισμό μας είναι πως ο χρήστης θα επιλέξει να διαβάσει τα άρθρα που τον ενδιαφέρουν ενώ στα υπόλοιπα δε θα δώσει σημασία. Αυτή τη συμπεριφορά χρήστη την καταγράφουμε και την εκμεταλλευόμαστε προκειμένου να διαμορφώσουμε το προφίλ του. Από τα άρθρα που παρουσιάζουμε στο χρήστη επιλέγουμε τις λέξεις κλειδιά. Για κάθε άρθρο που επιλέγει ο χρήστης να διαβάσει προσθέτουμε τις συγκεκριμένες λέξεις κλειδιά στο προφίλ του βάση της σχετικής συχνότητας που παρουσιάζουν στο συγκεκριμένο άρθρο. Πρόκειται για μία πολύ μεγάλη σχετική συχνότητα κάτι που είναι επιθυμητό καθότι πρόκειται για λέξεις κλειδιά σε ένα άρθρο που ενδιαφέρει το χρήστη. Όσον αφορά τα άρθρα που δεν επέλεξε ο χρήστης. Σε αυτή την περίπτωση συγκεντρώνουμε όλες τις λέξεις κλειδιά από αυτά τα άρθρα και ανανεώνουμε τις λέξεις κλειδιά του προφίλ χρήστη με αρνητική σχετική συχνότητα. Σε αυτή την περίπτωση και προκειμένου να διατηρηθεί η ακεραιότητα του μηχανισμού δεν αφαιρούμε με την πολύ μεγάλη σχετική συχνότητα που έχουν οι λέξεις κλειδιά αλλά με το ¼ αυτής. Έτσι σε περίπτωση που ένας χρήστης δεν ανάγνωσε ένα άρθρο που τον ενδιέφερε επειδή του διέφυγε δεν υπάρχει μεγάλη διαφορά στο προφίλ του. Αντίθετα, για τα άρθρα που επιλέγει ο χρήστης παρατηρείται μεγάλη αλλαγή στο προφίλ του.

8.4.6.4. Επιλογές του χρήστη κατά τη διάρκεια ανάγνωσης ενός άρθρου

Την ώρα που ο χρήστης επιλέγει να διαβάσει ένα άρθρο, όπως ήδη είπαμε οι λέξεις κλειδιά αυτού του άρθρου προστίθενται στο προφίλ του. Αυτό που δεν είπαμε είναι πως υπάρχει μία δικλείδα ασφαλείας για την περίπτωση που ο χρήστης κάνει εσφαλμένη επιλογή. Έτσι, αν ο χρήστης ανοίξει ένα άρθρο για να το διαβάσει και το κλείσει μέσα σε 7 δευτερόλεπτα τότε αυτό δεν προσμετράται σε αυτά που έχει διαβάσει. Αντίθετα θεωρείται σε αυτά που δεν έχει διαβάσει. Για τον υπολογισμό του χρόνου αυτού, χρησιμοποιείται τεχνολογία AJAX. Επιπλέον υπάρχει μία δικλείδα ασφαλείας για την περίπτωση που ο χρήστης «ξεχάσει» για οποιονδήποτε λόγο ανοιχτό το παράθυρο που περιέχει το σώμα του άρθρου. Ο χρόνος αυτός έχει οριστεί στα 2 λεπτά, κάτι το οποίο σημαίνει πως μετά την πάροδο δύο λεπτών που ο χρήστης έχει ανοιχτό ένα άρθρο, θεωρείται πως το έχει ξεχάσει ανοιχτό και έτσι αυτός ο χρόνος δεν είναι αντιπροσωπευτικός για το χρόνο που δαπάνησε ο χρήστης στο συγκεκριμένο άρθρο.

Ο χρόνος που καταναλώνει ο χρήστης σε ένα άρθρο είναι φυσικά ευθέως ανάλογος με το μέγεθος του άρθρου. Ας υπενθυμίσουμε σε αυτό το σημείο πως στο χρήστη προβάλλεται η εξής πληροφορία μόλις διαβάζει ένα άρθρο:

- Ο τίτλος του άρθρου
- Η ημερομηνία που καταγράφηκε το άρθρο στο σύστημα
- Οι πιθανές κατηγορίες στις οποίες ανήκει
- Η περίληψη του άρθρου
- Το σώμα του άρθρου, όπως αυτό έχει εξαχθεί από το μηχανισμό εξαγωγής χρήσιμου κειμένου.

Ο χρήστης βλέπει τα 4 πρώτα στοιχεία άμεσα ενώ το σώμα του άρθρου μπορεί να το δει επιλέγοντας ένα σύνδεσμο και αποτελεί και αυτό στοιχείο που καταγράφεται για την προσωποποίηση. Τα στοιχεία που μπορεί να επιλέξει ο χρήστης μόλις βλέπει ένα άρθρο και μπορεί να αναγνωρίσει ο μηχανισμός είναι:

- Να διαβάσει το κύριο σώμα του άρθρου
- Να τυπώσει το άρθρο (περίληψη ή κύριο σώμα)
- Να ακολουθήσει το σύνδεσμο προς το δικτυακό τόπο που φιλοξενεί το άρθρο
- Να στείλει το άρθρο σε ένα φίλο
- Να διαβάσει ταυτόσημα άρθρα που υπάρχουν σε άλλους δικτυακούς τόπους
- Να διαβάσει παρόμοια άρθρα των τελευταίων 24 ωρών
- Να διαβάσει σχετικά άρθρα των τελευταίων 3 ημερών.

Κάθε μία από αυτές τις ενέργειες έχει άμεση επίδραση στον τρόπο με τον οποίο ανανεώνονται οι λέξεις κλειδιά στο προφίλ του χρήστη. Κάποιες ενέργειες θεωρούνται πιο σημαντικές από άλλες και έτσι δεν αυξάνονται ομοιόμορφα οι συχνότητες των λέξεων κλειδιών του προφίλ του χρήστη. Όπως έχουμε ήδη δει η διαμόρφωση του προφίλ συνίσταται στην καταγραφή λέξεων κλειδιών με κάποιο βάρος. Η αρχική τιμή αυτού του βάρους συλλέγεται από τις λέξεις κλειδιών των κατηγοριών ενώ στην πορεία από τον πίνακα που περιέχει τις λέξεις κλειδιά του συγκεκριμένου άρθρου μαζί με τα βάρη τους.

Σύμφωνα με τα παραπάνω και βάση των ενεργειών του χρήστη, οι λέξεις κλειδιά στο προφίλ του χρήστη διαμορφώνονται βάσει του παρακάτω πίνακα.

Πίνακας 5: Τρόπος ανανέωσης των βαρών των λέξεων κλειδιών στο προφίλ του χρήστη

Ενέργεια	Επίδραση	Πολλαπλασιαστής (επί του relative frequency)
Μη επιλογή Άρθρου	**Αρνητική**	**-0,25**
Επιλογή άρθρου	**Θετική**	**+1**
Διάβασμα κυρίου σώματος άρθρου	**Θετική**	**+0,25**
Σύνδεσμος στο δικτυακό τόπο που φιλοξενεί το άρθρο	**Θετική**	**+0,25**
Εκτύπωση Άρθρου	**Θετική**	**+0,15**
Αποστολή σε φίλο	**Θετική**	**+0,15**
Ταυτόσημα Άρθρα	**Θετική**	**+0,20**
Παρόμοια Άρθρα	**Θετική**	**+0,17**
Σχετικά Άρθρα	**Θετική**	**+0,15**

Σύμφωνα με τον παραπάνω πίνακα, αν ένας χρήστης επιλέξει ένα άρθρα, διαβάσει το κύριο σώμα του (εκτός από την περίληψη), το τυπώσει, το στείλει σε ένα φίλο του, μεταβεί στη σελίδα με τα ταυτόσημα άρθρα και ακολουθήσει το σύνδεσμο προς το δικτυακό τόπο που φιλοξενεί το άρθρο τότε συνολικά για κάθε λέξη κλειδί που είναι στο άρθρο θα έχουμε μια ανανέωση/προσθήκη στο προφίλ του χρήστη της τάξης του 2x(relative frequency).

9

ΤΟ ΣΥΣΤΗΜΑ ΣΕ ΠΛΗΡΗ ΛΕΙΤΟΥΡΓΙΑ

Στο κεφάλαιο αυτό περιγράφονται τα πειράματα που έγιναν στο σύστημα καθώς αυτό είναι σε πλήρη λειτουργία.

9. ΤΟ ΣΥΣΤΗΜΑ ΣΕ ΠΛΗΡΗ ΛΕΙΤΟΥΡΓΙΑ

Η ανάπτυξη του συστήματος που έγινε στα πλαίσια της παρούσας εργασίας έγινε τμηματικά με κάθε module αυτού να αναπτύσσεται ξεχωριστά από τα υπόλοιπα. Την ανάπτυξη του καθενός τμήματος ακολουθούσε και μια διαδικασία αξιολόγησης του ώστε: α) να εντοπισθεί η αποτελεσματικότητά του ως ξεχωριστή οντότητα και β)να προσδιοριστούν οι απαραίτητες παράμετροι που πρέπει να χρησιμοποιηθούν σε κάθε βήμα ώστε ο μηχανισμός, ως σύνολο, να παράγει το βέλτιστο αποτέλεσμα. Ακολουθεί μια αναλυτική παρουσίαση των πειραματικών διαδικασιών και αξιολογήσεων που έλαβαν μέρος και που αφορούν στα βασικά υποσυστήματα του μηχανισμού: τη διαδικασία του keyword extraction, τους μηχανισμούς κατηγοριοποίησης και περίληψης μαζί με τις μεταξύ τους αλληλεπιδράσεις, και τέλος, το υποσύστημα παρουσίασης πληροφορίας στο χρήστη συσκευής μικρού μεγέθους.

9.1. Μηχανισμός εξαγωγής λέξεων κλειδιών

Σε αρχικό στάδιο υλοποίησης του μηχανισμού, εξετάσθηκε η αποτελεσματικότητα της διαδικασίας εξαγωγής keywords από διάφορες μορφές κειμένου. Με αυτό τον τρόπο, προσπαθήσαμε να αξιολογήσουμε τη διαδικασία αλλά και να θέσουμε κάποιες αρχικές παραμέτρους οι οποίες θα χρειαστούν για την λειτουργία του μηχανισμού ως σύνολο.

Δεδομένου ότι ο μηχανισμός εξαγωγής keywords είναι ένα ανεξάρτητο υποσύστημα, ο τύπος των κειμένων εισόδου μπορεί να διαφέρει κατά πολύ. Έτσι χρησιμοποιήθηκαν e-mails, άρθρα νέων αλλά και ερευνητικές εργασίες papers ως είσοδος. Για κάθε μία από αυτού του είδους την είσοδο, διεξάγαμε πειραματική διαδικασία ώστε να εντοπιστεί ποιο είναι το ελάχιστο δυνατό μήκος από keywords του αρχικού κειμένου που πρέπει να κρατηθούν, ώστε το αποτέλεσμα που προκύπτει να μη χάνει σημαντικά το νόημα του κειμένου. Για την διαδικασία αυτή, αξιολογήθηκαν δύο παράγοντες:

- ποιο είναι το ελάχιστο μήκος λέξεων που πρέπει να κρατηθεί
- τι ποσοστό των τελικών keywords πρέπει να κρατηθεί.

Για να «μετρηθεί» η διαφορά του νοήματος μεταξύ δύο κειμένων (δηλ. εκείνου στο οποίο έχουμε ελάχιστο μήκος λέξεων 4 και εκείνου που έχουμε ελάχιστο μήκος λέξεων 6), χρησιμοποιήθηκε μια απλή έκδοση του SVM αλγορίθμου.

Αν υποθέσουμε ότι έχουμε έναν πίνακα με όλα τα keywords και τις συχνότητές τους για το κείμενο Α, και έναν πίνακα του κειμένου Β, τότε μπορούμε να υπολογίσουμε τη συσχέτιση μεταξύ των δύο κειμένων ως:

$x = a*b$

$y = |a|*|b|$

$z = x/y$

$r = \sin(z)$

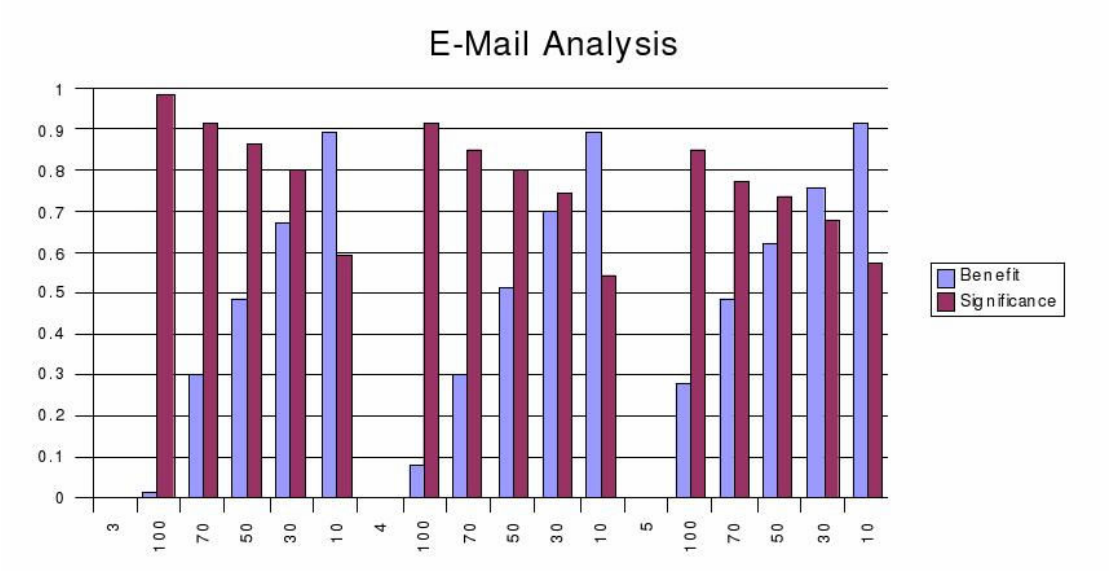
όπου x είναι το εσωτερικό γινόμενο των πινάκων a και b και y το γινόμενο των νορμών (2) του Α και του Β. Όπως μπορούμε να δούμε από τις προηγούμενες εξισώσεις, το r κινείται μεταξύ των τιμών μηδέν και ένα. Όταν το r είναι μηδέν, τότε οι πίνακες a και b είναι εντελώς ασυσχέτιστοι μεταξύ τους, ενώ όταν το r είναι ένα, οι πίνακες είναι εντελώς όμοιοι. Αυτό σημαίνει ότι όταν το r είναι κοντά στο ένα, τότε έχουμε υψηλή συσχέτιση μεταξύ των κειμένων που αναπαρίστανται μέσω των πινάκων a και b.

Με σκοπό να περιοριστεί ακόμη περισσότερο ο αριθμός των keywords του κειμένου, κρατήσαμε μόνο ένα ποσοστό αυτών και επανυπολογίσαμε από τη σχέση

r=sin(z) την συσχέτιση μεταξύ των keywords του αρχικού κειμένου και του ποσοστού των keywords που κρατήθηκε.

9.1.1. Πειραματισμός με τα κείμενα των e-mails

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα που προέκυψαν από την πειραματική διαδικασία με κείμενα ηλεκτρονικού ταχυδρομείου. Κατά τη διάρκεια της πειραματικής διαδικασίας χρησιμοποιήθηκε ελάχιστο μήκος λέξεων τριών, τεσσάρων και πέντε γραμμάτων. Τα αποτελέσματα συνοψίζονται στην γραφική απεικόνιση του παρακάτω σχήματος.



Εικόνα 23: Ανάλυση κειμένων ηλεκτρονικού ταχυδρομείου

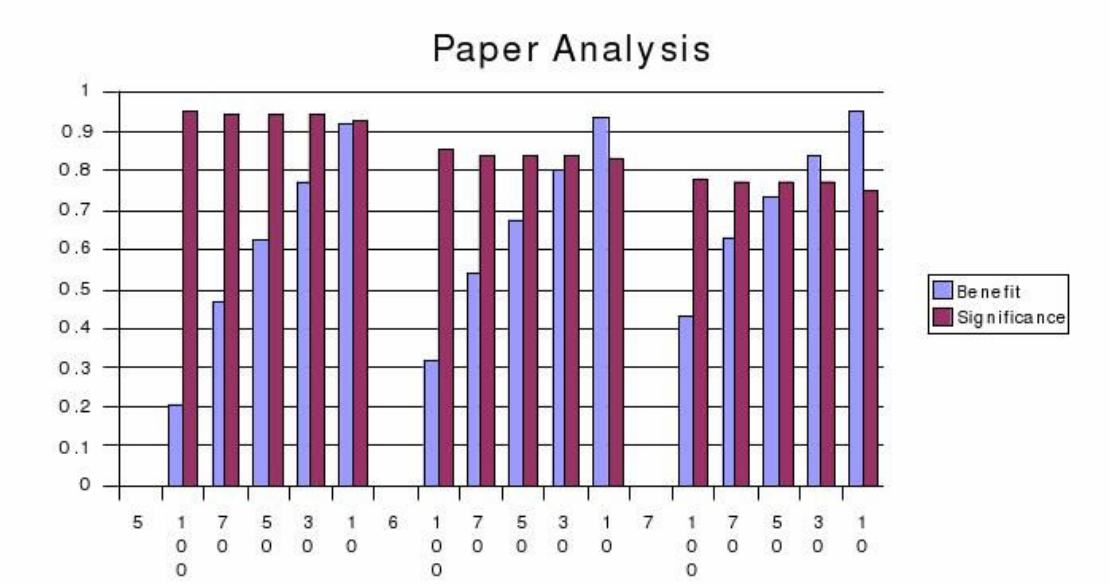
Όπως φαίνεται και στο σχήμα, έχουμε περιορίσει το ελάχιστο μήκος των λέξεων σε 3, 4, 5 και περισσότερους χαρακτήρες και κρατήσει ένα ποσοστό των keywords που απομένουν. Μειώνοντας το ελάχιστο μήκος λέξεων σε 3 γράμματα και κρατώντας το 70% των εξαγόμενων keywords, έχουμε ένα όφελος περίπου 30% των keywords του αρχικού κειμένου και η ομοιότητα των δύο κειμένων είναι πάνω από 90%.

Αυτό που μας ενδιαφέρει είναι η συσχέτιση μεταξύ του αρχικού κειμένου και των εξαγομένων keywords. Έτσι αποφασίσαμε να κρατήσουμε το επίπεδο της συσχέτισης στο 85% αφού είναι προφανές ότι τα keywords που απομένουν είναι αντιπροσωπευτικά του αρχικού κειμένου. Ο περιορισμός αυτός σημαίνει ότι το ελάχιστο μήκος λέξεων και το ποσοστό των keywords που προκύπτουν από το προηγούμενο διάγραμμα, μπορεί να είναι: 3/100%, 3/70%, 3/50%, 4/100%, 4/70% και 5/100% αντίστοιχα. Το όφελος από τα ζευγάρια αυτά είναι 1%, 29%, 48%, 8%, 30% και 28% αντίστοιχα. Ο λόγος όφελος / ομοιότητα είναι 0.01, 0.33, 0.56, 0.09, 0.35 και 0.33 για καθένα από τα ζεύγη που αναφέρθηκαν. Αυτό σημαίνει ότι το καλύτερο ζεύγος μοιάζει να είναι το 3/50% για την ανάλυση κειμένων ηλεκτρονικού ταχυδρομείου, μειώνουμε δηλαδή το ελάχιστο μήκος λέξεων σε 3 γράμματα και κρατάμε τις μισές από τις κωδικολέξεις που προκύπτουν από την ανάλυση. Πρέπει να αναφερθεί επίσης ότι τα keywords βρίσκονται σε φθίνουσα σειρά διάταξης σε σχέση με τη συχνότητα εμφάνισης, πριν κρατηθεί το κατάλληλο ποσοστό.

9.1.2. Πειραματισμός με εξόρυξη λέξεων κλειδιών από papers

Σε αυτή την ενότητα παρουσιάζουμε τα αποτελέσματα του μηχανισμού προεπεξεργασίας όταν επεξεργάζεται papers. Στην ανάλυση χρησιμοποιήθηκε ελάχιστο μήκος λέξεων 5, 6, 7 και περισσοτέρων γραμμάτων. Στο παρακάτω

σχήμα παρουσιάζονται τα αποτελέσματα που προέκυψαν μέσω της πειραματικής διαδικασίας.



Εικόνα 24: Ανάλυση κειμένων δημοσιεύσεων

Όπως μπορούμε να δούμε από τη γραφική παράσταση του σχήματος, κρατήθηκε ελάχιστο μήκος λέξεων 5, 6, 7 και περισσότεροι χαρακτήρες και στη συνέχεια κρατήθηκε ένα ποσοστό των keywords για καθένα από τον περιορισμό μήκους λέξεων. Όπως μπορούμε να δούμε, τα αποτελέσματα δεν επηρεάζονται (σημαντικά) από τον παράγοντα ποσοστού κράτησης των λέξεων. Αυτό μπορεί να εξηγηθεί ως εξής: τα κείμενα που επεξεργάζεται ο μηχανισμός εξαγωγής keywords σε αυτή την περίπτωση, περιέχουν περισσότερες από 900 μοναδικές λέξεις οι οποίες εμφανίζονται πολλές φορές μέσα στο κείμενο και αυτό γιατί τα papers έχουν ένα συγκεκριμένο θεματικό πεδίο, με αποτέλεσμα, η επαναληπτικότητα των όρων είναι αναπόφευκτη. Το όριο της συσχέτισης ώστε να θεωρηθεί ότι το κείμενο δεν έχει χάσει το νόημά του, επιλέχθηκε να είναι το 80%. Αυτό σημαίνει ότι ο περιορισμός μήκους λέξεων για 7 ή περισσότερους χαρακτήρες μοιάζει να μην επιτυγχάνει το στόχο. Αντίθετα, με ελάχιστο μήκος λέξεων 5 ή 6 χαρακτήρων, το κείμενο που προκύπτει ξεπερνά σε συσχέτιση με το αρχικό κείμενο το όριο του 80% για την ομοιότητα.

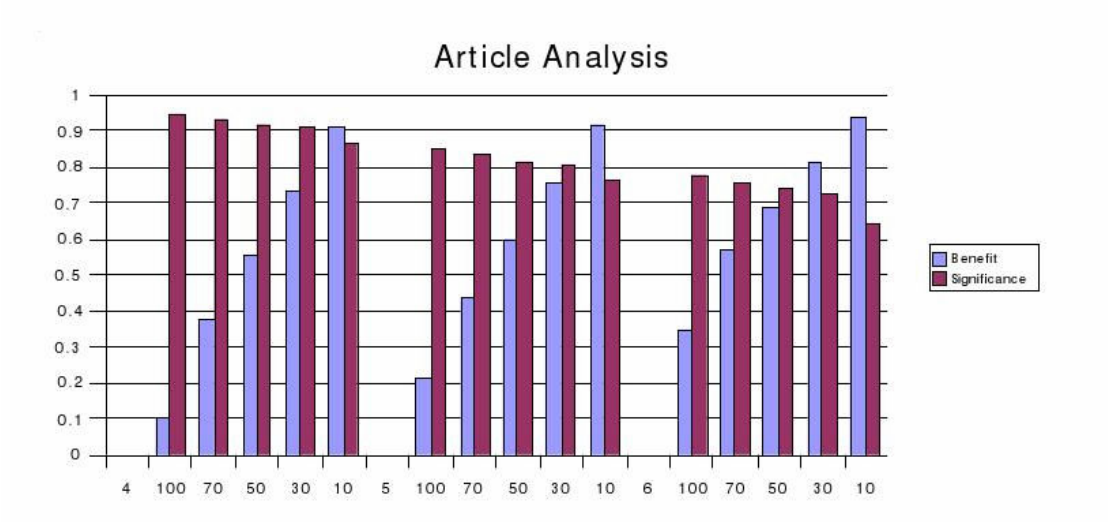
Το ζεύγος που αξιολογήθηκε ως βέλτιστο για να κρατηθεί, είναι το 6/10%, δηλαδή 6 χαρακτήρες ως ελάχιστο μήκος λέξεων και 10% των εξαγόμενων keywords, το οποίο μας οδηγεί σε 83% ομοιότητα και πάνω από 90% όφελος.

9.1.3. Πειραματισμός με εξόρυξη λέξεων κλειδιών από άρθρα

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα που προέκυψαν από την ανάλυση άρθρων ειδήσεων του διαδικτύου. Σε αυτή την περίπτωση κρατάμε ελάχιστο μήκος 4, 5, 6 και περισσοτέρων χαρακτήρων, και κρατάμε ένα ποσοστό των εξαγόμενων keywords για να βρούμε το καλύτερο ζεύγος ελάχιστου μήκους λέξης / ποσοστού των keywords το οποίο έχει καλά αποτελέσματα για την ομοιότητα και το όφελος που προκύπτει.

Το όριο για την ομοιότητα που τέθηκε είναι το 85%, κάτι που προέκυψε ύστερα από πειραματική διαδικασία με χρήση πολλών άρθρων και λειτουργία όλου του μηχανισμού (όχι μόνο του υποσυστήματος εξαγωγής κωδικολέξεων αλλά και των υποσυστημάτων περίληψης / κατηγοριοποίησης κειμένων). Ανεβάζοντας αυτό το ποσοστό στο 90%, οδηγούμαστε σε πάρα πολλά keywords κάτι που υπερφορτώνει τη βάση δεδομένων αλλά και τους μηχανισμούς εξαγωγής πληροφορίας που ακολουθούν.

Τα ζεύγη που μπορούν να περάσουν το όριο του 85%, μπορούν να βρεθούν μόνο στις περιπτώσεις που κρατούνται 4 και 5 χαρακτήρες ως ελάχιστο μήκος λέξεων. Πιο συγκεκριμένα, όλα τα ζεύγη που προκύπτουν από την χρήση 4 χαρακτήρων και η πρώτη επιλογή από τη χρήση 5 χαρακτήρων ικανοποιούν το όριο που αναφέρθηκε. Η πρώτη επιλογή από τη χρήση 5 χαρακτήρων, έχει πολύ μικρό όφελος (21%). Αντίθετα το ζεύγος 4/10% μας δίνει ομοιότητα πάνω από 85% και όφελος που ξεπερνάει το 90%. Αυτό σημαίνει ότι κόβουμε το 90% των μοναδικών keywords και αποθηκεύουμε μόνο ένα 10% αυτών που μας δίνουν πάνω από 85% ομοιότητα του τελικού κειμένου σε σχέση με το αρχικό. Τα παραπάνω συνοψίζονται και στο διάγραμμα του παρακάτω σχήματος.



Εικόνα 25: Ανάλυση κειμένων άρθρων

9.1.4. Γενικά Αποτελέσματα πρώτων πειραμάτων

Ύστερα από τον πειραματισμό με διάφορα είδη κειμένων, μπορούμε να αντιληφθούμε ότι τα διάφορα είδη κειμένων χρειάζονται διαφορετική αντιμετώπιση από τον μηχανισμό προεπεξεργασίας. Η απλή δομή και περιεκτικότητα των μηνυμάτων ηλεκτρονικού ταχυδρομείου είναι πολύ διαφορετική από την πολύπλοκη δομή των papers. Κάπου ενδιάμεσα βρίσκονται τα άρθρα ειδήσεων από το διαδίκτυο που μας απασχολούν και στην συγκεκριμένη εργασία.

Όπως μπορούμε να δούμε από τα αποτελέσματα που προέκυψαν, στα e-mails πρέπει να κρατηθούν όλα τα keywords με μικρό μάλιστα ελάχιστο μήκος λέξεων. Αντίθετα, στις δημοσιεύσεις, όπου οι λέξεις που χρησιμοποιούνται είναι συνήθως επίσημες και μεγάλες σε μήκος, μπορούμε να ωφεληθούμε από αυτό και να θέσουμε υψηλότερα το ελάχιστο μήκος λέξεων και να κρατήσουμε ένα σχετικά μικρό ποσοστό των keywords που προκύπτουν για να αναπαραστήσουμε το κείμενο.

Αναμέναμε ότι κερδίζοντας σε σημαντικότητα τις τελικής λίστας keywords θα οδηγούμασταν σε μείωση του οφέλους. Αντίθετα, από τα αποτελέσματα προέκυψε ότι μπορούμε να κρατήσουμε ένα υψηλό ποσοστό και για δύο αυτές παραμέτρους. Αυτό σημαίνει ότι καταφέραμε, για τα διάφορα είδη κειμένων, να καταλήξουμε σε ένα τελικό μέγεθος λίστας keywords το οποίο ήταν 80% περίπου μικρότερο από την αρχική λίστα των keywords και συσχετίζονταν με αυτή σε ποσοστό πάνω από 80%. Με άλλα λόγια, για ένα κείμενο 5000 λέξεων, κρατώντας μόνο 20% αυτών (100 λέξεις) έχουμε μια καλή αναπαράσταση του αρχικού κειμένου η οποία μπορεί να αποθηκευθεί στη βάση δεδομένων για να αξιοποιηθεί από τους μηχανισμούς ανάκτησης πληροφορίας που ακολουθούν (περίληψη, κατηγοριοποίηση). Επομένως, δεν είναι αναγκαία η δεικτοδότηση ολόκληρου του αρχικού κειμένου και άρα, με τη χρήση ενός μικρού μόνο μέρους του, μειώνουμε α) τις απαιτήσεις

για αποθήκευση δεδομένων και β) την πολυπλοκότητα και τους χρόνους εκτέλεσης των μηχανισμών που ακολουθούν.

9.2. Μηχανισμοί Κατηγοριοποίησης και Περίληψης

Κάθε μια από τις εξισώσεις (1),(6) και (8) που είδαμε στο κεφάλαιο 8 για την βαθμολόγηση των προτάσεων ελέγχθηκε σε κάποια προκατηγοριοποιημένα (από ανθρώπους) κείμενα. Τα αποτελέσματα του μηχανισμού δείχνουν να είναι επαρκή σε σύγκριση με ήδη υπάρχοντα συστήματα. Ο βασικός μας στόχος είναι να παρουσιάζουμε μια προσωποποιημένη περίληψη άρθρων στον τελικό χρήστη και επομένως οι περιλήψεις που προκύπτουν βάσει των σχέσεων (1) και (6) δεν θα πρέπει να παράγουν περιλήψεις που διαφέρουν πολύ από ήδη υπάρχοντες αλγορίθμους. Η διαδικασία προσωποποίησης στην περίληψη δεν μπορεί να αξιολογηθεί σε σχέση με μια πρωτότυπη, ανθρώπινα παραγόμενη περίληψη αφού κάθε τέτοια εμπεριέχει τον υποκειμενικό ανθρώπινο παράγοντα. Ο μόνος πραγματικός εκτιμητής του συστήματος είναι ο τελικός χρήστης ο οποίος διαβάζει τις περιλήψεις.

Για την αξιολόγηση του αλγόριθμου περίληψης, εκτελέσθηκε πειραματική διαδικασία για την σύγκρισή του με τον MEAD αλγόριθμο περίληψης ο οποίος χρησιμοποιείται από την εφαρμογή του Microsoft Word. Οι προσωποποιημένες περιλήψεις που προέκυψαν από το σύστημα αξιολογήθηκαν από πέντε διαφορετικούς χρήστες οι οποίοι επιθυμούσαν να λάβουν μέρος στη δοκιμή.

9.2.1. Αξιολόγηση Μηχανισμού Εξαγωγής Αυτόματης Περίληψης

Για να εξασφαλίσουμε ότι η διαδικασία πριν την εφαρμογή του παράγοντα προσωποποίησης παράγει επαρκή αποτελέσματα για τις περιλήψεις, αξιολογήσαμε τον μηχανισμό σε σχέση με τα αποτελέσματα από τον περιλήπτη του Microsoft Word. Τα αποτελέσματα συγκρίνονται με εξαγωγές του MEAD περιλήπτη σε 30 άρθρα συγκεντρωμένα από βασικά portals των Η.Π.Α και της Βρετανίας. Οι μετρικές που χρησιμοποιήθηκαν για τον υπολογισμό των αποτελεσμάτων είναι η ακρίβεια και η ανάκληση.

Πίνακας 6: Σύγκριση του αλγορίθμου περίληψης του συστήματος με τον περιλήπτη του MS Word

	MS Word		Proposed Mechanism	
	Precision	Recall	Precision	Recall
Article 1	0,33	0,12	0,66	0,75
Article 2	0,12	0,25	0,75	0,66
Article 3	0,25	0,12	0,5	0,66
Article 4	0,25	0,12	0,75	0,5
Article 5	0,33	0,5	0,66	1
Article 6	0,33	0,25	0,66	0,75
Article 7	0,25	0,33	0,75	0,66

Από τα αποτελέσματα συνεπάγεται ότι ο μηχανισμός περίληψης που υλοποιήθηκε παράγει επαρκή αποτελέσματα συγκρινόμενος με δοκιμές που έγιναν με τον MEAD περιλήπτη, και σαφώς καλύτερα αποτελέσματα από τον περιλήπτη του MS Word. Προσθέτοντας τον παράγοντα κατηγοριοποίησης στη διαδικασία περίληψης, καταφέρνουμε να λάβουμε λίγο καλύτερα αποτελέσματα. Παρατηρούμε ότι η συνολική αύξηση είναι περίπου 10% σε σχέση με τα προηγούμενα αποτελέσματα όσον αφορά τις μετρικές της ακρίβειας και ανάκλησης. Η διαφορά οφείλεται στην διαδικασία κατηγοριοποίησης και, πιο συγκεκριμένα, στην προσθήκη της παραμέτρου στην εξίσωση εξαγωγής περίληψης. Η παράμετρος αυτή, επιτρέπει την υψηλότερη βαθμολόγηση των προτάσεων που περιέχουν keywords αντιπροσωπευτικά της κατηγορίας στην οποία

ανήκει το άρθρο. Εάν ένα άρθρο δεν περιέχει πολλά keywords από την κατηγορία στην οποία ανήκει, δεν συμβαίνουν αλλαγές. Σε αυτή την περίπτωση, είναι αξιοσημείωτο να σημειωθεί ότι ύστερα από λίγο χρόνο (και ενώ νέα keywords προστίθενται στο σύστημα), όταν κάποιος προσπαθεί να έχει πρόσβαση στην περίληψη του συγκεκριμένου άρθρου, αυτή ανανεώνεται και οι μετρικές της ακρίβειας και ανάκλησης μετρώνται υψηλότερα σε σχέση με την πρώτη φορά της εξαγωγής περίληψης. Στον επόμενο πίνακα οι μετρικές της ακρίβειας και ανάκλησης παρουσιάζονται για ένα συγκεκριμένο άρθρο και πως μεταβάλλονται όταν νέα άρθρα κατηγοριοποιούνται και πιο αντιπροσωπευτικά keywords για την κατηγορία προστίθενται στο σύστημα. Τα άρθρα «καταφτάνουν» στο σύστημα κάθε μία ώρα αφού τα σημαντικά news portal ανανεώνουν το περιεχόμενό τους πολύ συχνά.

Πίνακας 7: Αλλαγές στην ακρίβεια και την ανάκληση για την περίληψη ενός άρθρου ύστερα από την προσθήκη πιο αντιπροσωπευτικών για την κατηγορία στην οποία το άρθρο ανήκει.

Time (after arrival)	Articles added to category (sum)	Proposed Mechanism	
		Precision	Recall
10 min	0	0,5	0,66
8 hours	8	0,5	0,66
24 hours	31	0,66	0,5
36 hours	43	0,66	0,66
48 hours	59	0,66	0,66
62 hours	88	0,75	0,75
78 hours	103	0,75	0,8

Από τα προηγούμενα στατιστικά στοιχεία, φαίνεται ότι ο μηχανισμός δεν είναι στατικός. Αντίθετα το σύστημα μπορεί να προσαρμόζεται δυναμικά και να ανανεώνει τις περιλήψεις που εξάγονται. Παράλληλα, είναι αναμενόμενο το γεγονός ότι μετά την δημοσίευση ενός άρθρου κάποιου σημαντικού νέου, πολλά ακόμη άρθρα σχετικά με αυτό θα ακολουθήσουν. Αυτό σημαίνει ότι στα επόμενα 103 άρθρα μιας κατηγορίας που συλλέγονται από τον μηχανισμό στις επόμενες 78 ώρες, τουλάχιστον ένα θα είναι παρόμοιο με το πρώτο άρθρο είτε ως επανέκδοσή του είτε ως συμπλήρωμά του.

9.2.2. Αξιολόγηση του μηχανισμού εξαγωγής προσωποποιημένης περίληψης

Η αξιολόγηση μιας δυναμικά εξαγόμενης προσωποποιημένης περίληψης κειμένου δεν είναι μια διαδικασία που μπορεί να γίνει με χρήση μέτρων σύγκρισης. Το μέτρο που χρησιμοποιείται για να αξιολογηθούν οι εξαγόμενες περιλήψεις είναι η συσχέτιση μεταξύ της περίληψης και του άρθρου που παρατηρείται από τους χρήστες του μηχανισμού. Η διαδικασία που ακολουθήθηκε για να αξιολογηθούν τα αποτελέσματα της πειραματικής διαδικασίας ήταν: (α) δώσε στους χρήστες το πλήρες κείμενο του άρθρου, (β) δώσε στους χρήστες τις περιλήψεις που προέκυψαν τόσο από την εξίσωση (6), όσο και από την εξίσωση (8), και (γ) άφησε τους χρήστες να επιλέξουν ποια περίληψη θεωρούν ως περισσότερο αντιπροσωπευτική για το άρθρο που διάβασαν. Η αντίστροφη διαδικασία εξετάστηκε επίσης, δόθηκαν δηλαδή πρώτα οι περιλήψεις στους χρήστες, στη συνέχεια το κείμενο και τέλος οι χρήστες αποφάνθηκαν για το ποια περίληψη θεωρούν ως περισσότερο αντιπροσωπευτική για το πλήρες άρθρο που διάβασαν. Και στις δύο περιπτώσεις που αναφέρθηκαν οι απαντήσεις ήταν οι ίδιες.

Οι χρήστες που έλαβαν μέρος στην πειραματική διαδικασία μπορούν να χωριστούν σε τρεις ομάδες: (α) νέοι χρήστες του συστήματος, (β) παλιοί χρήστες

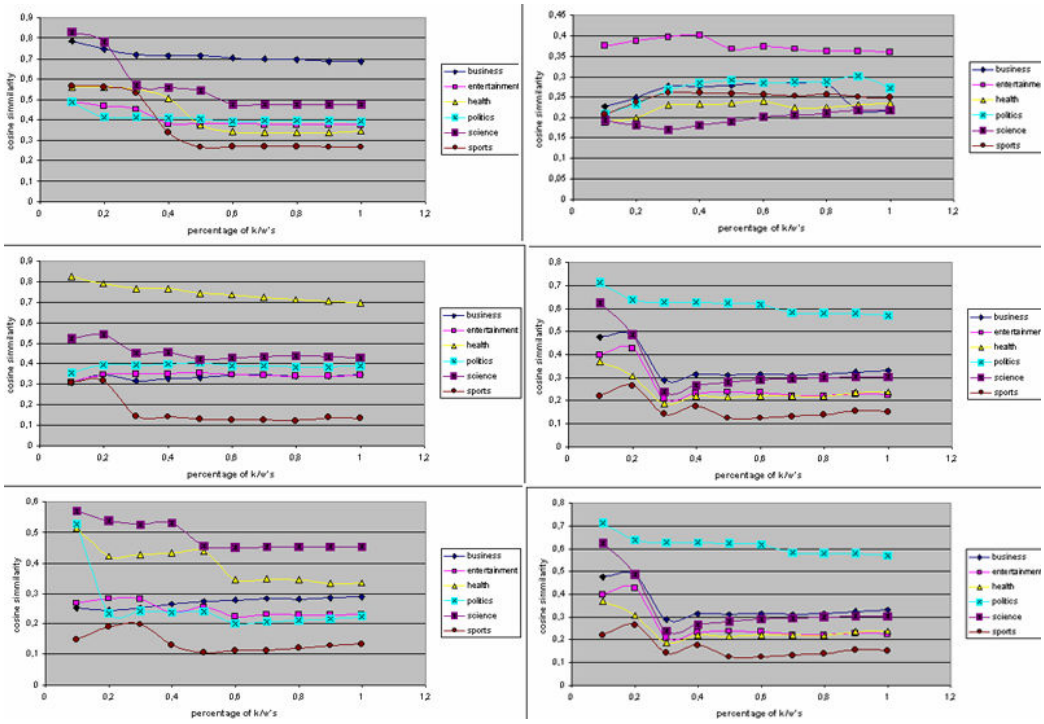
του συστήματος αλλά με μικρή δραστηριότητα (το οποίο σημαίνει λίγα δεδομένα για προσωποποίηση), και (γ) προχωρημένοι χρήστες του συστήματος με υψηλή καθημερινή δραστηριότητα (το οποίο σημαίνει πολλά δεδομένα για προσωποποίηση). Σύμφωνα με αυτές τις κατηγορίες, τρεις διαφορετικές καταστάσεις παρατηρήθηκαν. Οι νέοι χρήστες του συστήματος εξέφρασαν την άποψη ότι οι περιλήψεις που τους δόθηκαν ήταν όμοιες, κάτι που είναι μια λογική παρατήρηση εφόσον το σύστημα δεν έχει αρκετή πληροφορία για την διαδικασία προσωποποίησης και επομένως, η βαθμολόγηση των προτάσεων για την περίληψη δεν επηρεάζεται από τον παράγοντα k_4 (που χρησιμοποιείται για την προσωποποίηση της περίληψης). Οι χρήστες της δεύτερης ομάδας επέλεξαν, με ποσοστό μεγαλύτερο του 80% των άρθρων, την περίληψη που εξήχθη από την εξίσωση (6) (χωρίς τον παράγοντα προσωποποίησης). Αυτό ήταν επίσης αναμενόμενο αφού το προφίλ των χρηστών αυτών (με μικρή συμμετοχή) δεν ήταν πλήρες και περιείχε πολλά keywords που στην πραγματικότητα ήταν χαμηλής σημασίας τόσο για το άρθρο όσο και για την κατηγορία. Τα πλέον σημαντικότερα αποτελέσματα πηγάζουν από την τρίτη ομάδα χρηστών, τα μέλη της οποίας θεωρούνται από τους πιο «έμπειρους» στη χρήση του συστήματος με σχεδόν σταθεροποιημένα προφίλ ύστερα από χρήση του συστήματος για μακρύ χρονικό διάστημα. Η σταθερότητα και η πληρότητα του προφίλ των χρηστών αυτών δίνει τη δυνατότητα προσωποποίησης στο μηχανισμό εξαγωγής περίληψης. Τα μέλη αυτής της ομάδας επέλεξαν σε ποσοστό μεγαλύτερο του 90% των άρθρων, την προσωποποιημένη περίληψη ως πιο αντιπροσωπευτική του άρθρου και μόνο 3% των περιλήψεων αξιολογήθηκαν ως «όμοιες». Είναι σημαντικό να τονιστεί ότι τα περισσότερα από τα υπολειπόμενα άρθρα (7%), αξιολογήθηκαν από τον μηχανισμό κατηγοριοποίησης του συστήματος ως «ανήκοντα σε κάποια κατηγορία αλλά με ασθενή συσχέτιση». Αυτό σημαίνει ότι αυτά ήταν άρθρα τα οποία προστέθηκαν στη συγκεκριμένη κατηγορία με την «υποσημείωση» ότι το σύστημα δεν μπόρεσε με απόλυτη βεβαιότητα να τα κατατάξει σε κάποια κατηγορία, αλλά η κατηγορία στην οποία τελικά εισήχθηκαν είναι η πιο «κοντινή» για αυτά τα άρθρα.

9.2.3. Αλληλεπίδραση μεταξύ της διαδικασίας περίληψης και κατηγοριοποίησης

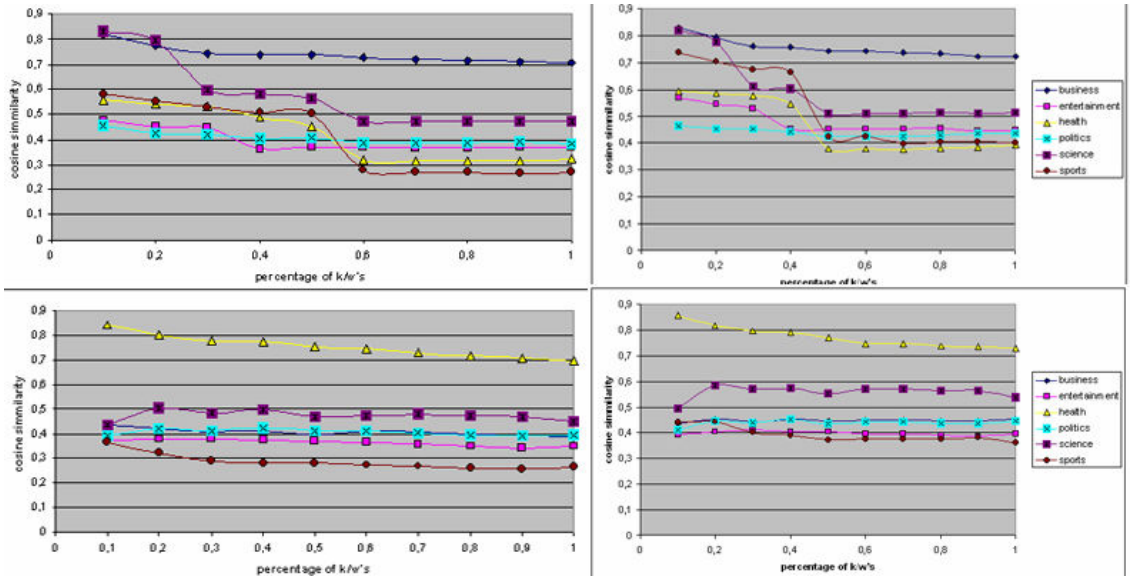
Με σκοπό να εκτιμηθεί η αλληλεπίδραση μεταξύ των μηχανισμών περίληψης και κατηγοριοποίησης, διεξάγαμε πειραματική διαδικασία. Για να έχουμε για αρχική βάση γνώσης (ακόμα και μια μικρή), συγκεντρώθηκαν άρθρα νέων από ορισμένα σημαντικά news portals. Ορίσθηκαν 6 διαφορετικές κατηγορίες νέων: business, entertainment, health, politics, science, και sports. Τα κείμενα που κρατήθηκαν, οργανώθηκαν σε αυτές τις κατηγορίες (περίπου 180 σε κάθε μια). Στη συνέχεια, χρησιμοποιώντας τους μηχανισμούς εξαγωγής κειμένου και κατηγοριοποίησης, κρατήθηκε το 50% των keywords για κάθε κείμενο και κάθε keyword συσχετίστηκε με κάθε κατηγορία χρησιμοποιώντας την απόλυτη συχνότητα εμφάνισης ως μέτρο ομοιότητας. Πιο συγκεκριμένα, διεξήχθηκαν τριών ειδών πειραματικές διαδικασίες.

Αρχικά, χρειαζόταν να καθοριστεί το ποσοστό από keywords του κειμένου το οποίο πρέπει να κρατηθεί ουτοσώστε ο μηχανισμός κατηγοριοποίησης να έχει την μεγαλύτερη αποτελεσματικότητα. Προς αυτή την κατεύθυνση, μεταβάλαμε το ποσοστό των keywords που κρατούνται από 0,1 (δηλ. 10% των keywords) σε 1 (δηλ. όλα τα keywords) με βήμα 0,1, κάνοντας χρήση ενός αντιπροσωπευτικού κειμένου για κάθε μια από τις προαναφερθέντες κατηγορίες, και το κατηγοριοποιήσαμε. Το κείμενο που επιλέχθηκε για είσοδο στον μηχανισμό κατηγοριοποίησης δεν ήταν μέρος των κειμένων που χρησιμοποιήθηκαν για την κατασκευή της βάσης γνώσης (δεν ήταν μέρος του training set). Για κάθε ποσοστό από keywords μετρήθηκε η ομοιότητα συνημιτόνου μεταξύ του κειμένου και της κάθε κατηγορίας που υπάρχει στη βάση γνώσης. Εκτελέστηκαν πειράματα χρησιμοποιώντας ελάχιστο μήκος keywords 5 και 6 γράμματα, τόσο για την βάση

γνώσης, όσο και για το κείμενο που εισήχθηκε στον μηχανισμό κατηγοριοποίησης. Ακολουθούν ορισμένα διαγράμματα που αποτυπώνουν τα αποτελέσματα.



Εικόνα 26: Ομοιότητα συνημιτόνου των κειμένων σε σχέση με τις κατηγορίες. Το training set κατασκευάζεται με χρήση του 50% των keywords (διαδικασία προεπεξεργασίας)



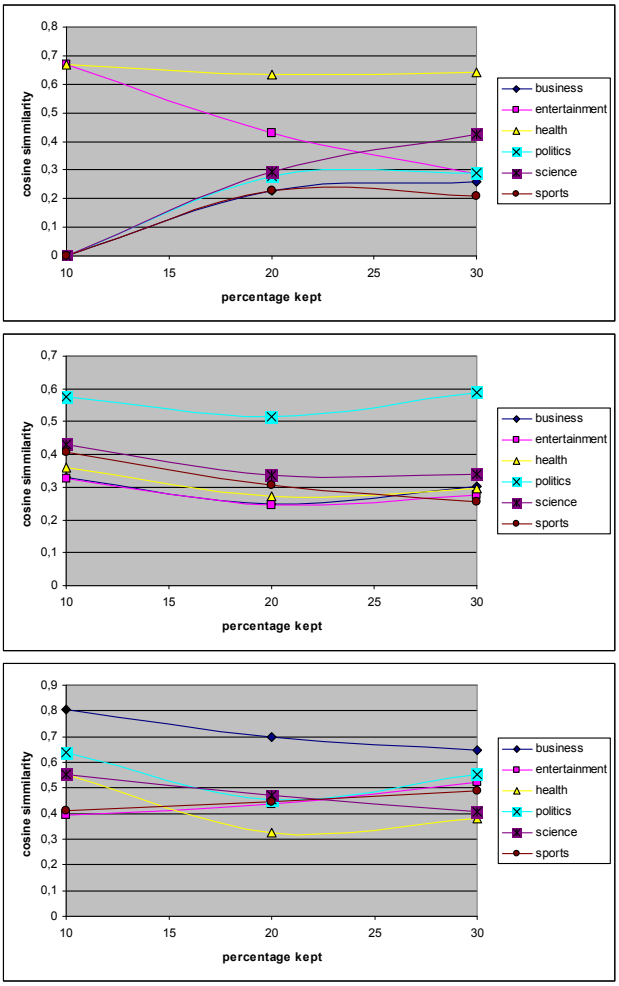
Εικόνα 27: Η πρώτη στήλη δείχνει την ομοιότητα συνημιτόνου μετρημένη χρησιμοποιώντας το 50% των keywords από το training set. Η δεύτερη στήλη δείχνει την ίδια ομοιότητα συνημιτόνου μετρημένη χρησιμοποιώντας το 100% των keywords του training set.

Από την πρώτη εικόνα (αποτελέσματα διαδικασίας κατηγοριοποίησης), προκύπτει ότι ένα ποσοστό 30% των keywords του κειμένου πρέπει να κρατηθούν από την διαδικασία κατηγοριοποίησης ώστε αυτή να είναι βέλτιστη. Αν και ένα μικρότερο ποσοστό μπορεί να είναι επαρκές ώστε να αποφασιστεί η κατηγορία του κειμένου, κρατάμε ένα ποσοστό 30% διότι, πρώτον μας δίνει σχεδόν πάντα σωστή απόφαση για την κατηγορία του κειμένου και δεύτερον, μας δίνει έναν ισχυρό διαχωρισμό (διαφορά ποσοστού) μεταξύ της σωστής κατηγορίας και των

υπολοίπων. Κατά την γνώμη μας, αυτή η διαφορά στην ομοιότητα είναι ο πιο σημαντικός παράγοντας για έναν μηχανισμό κατηγοριοποίησης, αφού μπορεί να μας δώσει σωστές απαντήσεις ακόμη και για μικρή βάση γνώσης. Για παράδειγμα είναι δυνατό, όταν η βάση γνώσης έχει πολλές κατηγορίες μερικές από τις οποίες παρόμοιες, η ομοιότητα μεταξύ ενός κειμένου και παραπάνω από μια κατηγορίες να είναι μεγάλη. Σε αυτή την περίπτωση, η διαφορά στην ομοιότητα μπορεί να είναι ένα καλύτερο μέτρο για την κατηγοριοποίησης, παρά ένα όριο απόλυτης ομοιότητας.

Όπως είναι φανερό από τη δεύτερη εικόνα, ένα κείμενο μπορεί να επιτύχει καλύτερο σκορ χρησιμοποιώντας ένα ελάχιστο μήκος 5 γραμμάτων για τα keywords και κρατώντας 50% των keyowrds που προκύπτουν. Με αυτό τον τρόπο, η βάση γνώσης είναι πιο φιλτραρισμένη, ενώ δεν μένουν έξω από τη διαδικασία keywords σημαντικά για κάποια/ες κατηγορία/ες.

Στο επόμενο βήμα της πειραματικής διαδικασίας, θέλουμε να εξεταστεί η επιρροή που έχει η διαδικασία περίληψης στο στάδιο της κατηγοριοποίησης. Για να το πετύχουμε αυτό, αρχικά περάστηκαν από το μηχανισμό περίληψης κάποια ανθρωπίνως προκατηγοριοποιημένα κείμενα τα οποία στη συνέχεια προωθήθηκαν στην διαδικασία κατηγοριοποίησης. Τελικά συγκρίναμε την έξοδο του μηχανισμού κατηγοριοποίησης (η οποία με αυτό τον τρόπο μας δίνει την ομοιότητα της περίληψης του κειμένου με τη καταγεγραμμένη κατηγορία που αυτό ανήκει), με την προκαθορισμένη κατηγορία του κειμένου.



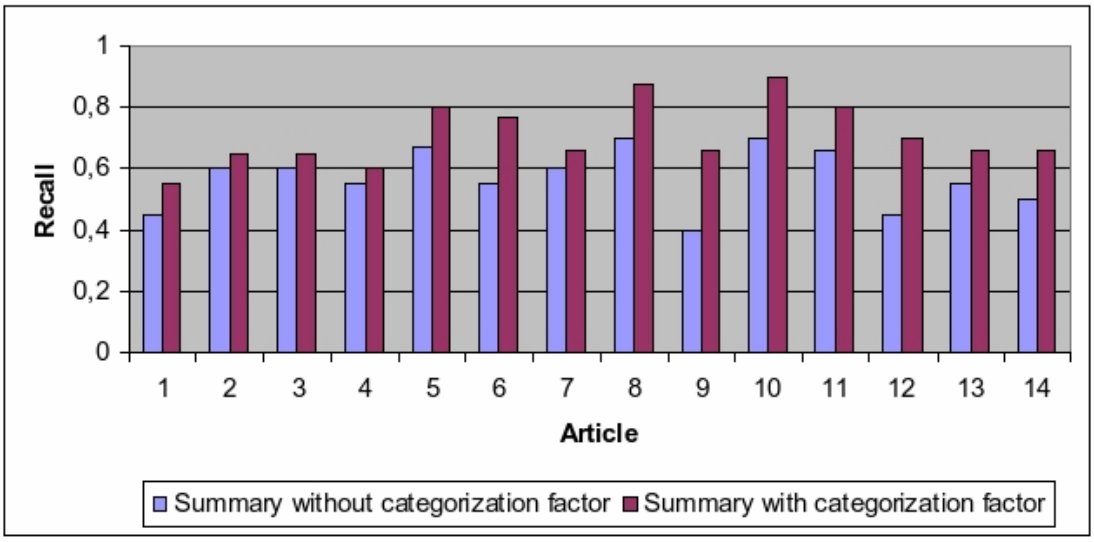
Εικόνα 28: Ομοιότητα συνημιτόνου που μετρήθηκε για την κατηγοριοποίηση περιλήψεων χρησιμοποιώντας διάφορα ποσοστά για την δημιουργία των περιλήψεων

Χρησιμοποιήθηκαν διάφορα μεγέθη περιλήψεων με σκοπό να εντοπιστεί η επίδραση που έχουν στην κατηγοριοποίηση της περίληψης. Ακολουθούν ορισμένα διαγράμματα της πειραματικής διαδικασίας χρησιμοποιώντας κείμενα που ανήκουν σε διαφορετικές κατηγορίες, τα οποία αποκαλύπτουν το ιδανικό ποσοστό των προτάσεων οι οποίες μπορούν να διαμορφώσουν μια «καλή» περίληψη.

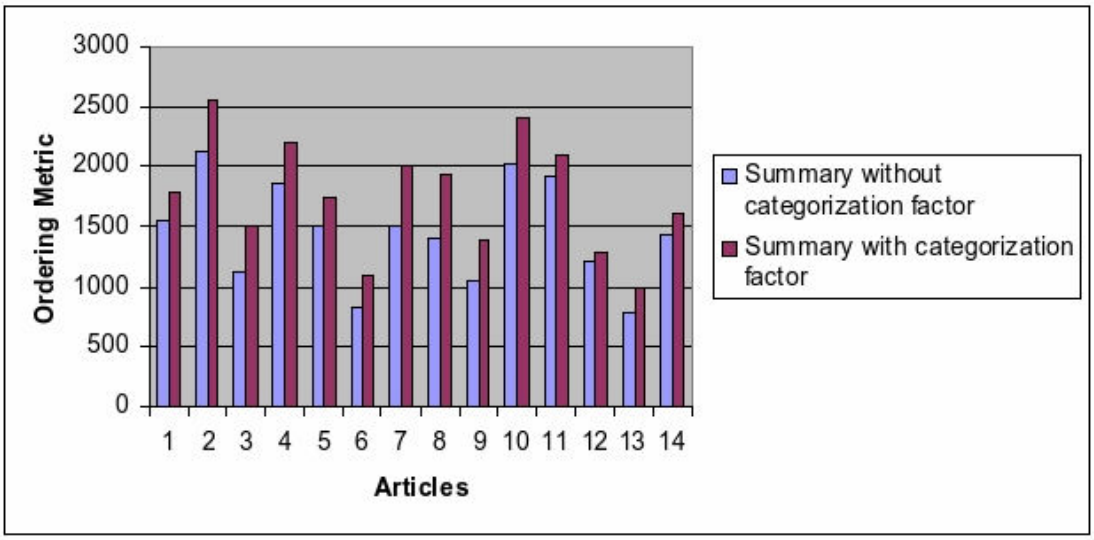
Από αυτού του είδους την πειραματική διαδικασία καταλήξαμε στο συμπέρασμα ότι κρατώντας ένα εύλογο μέγεθος από τις αρχικές προτάσεις, περίπου 20%, για την παραγωγή της περίληψης του κειμένου, μπορούμε να κατηγοριοποιήσουμε την περίληψη σωστά στην κατηγορία του κειμένου. Με αυτό τον τρόπο γλιτώνουμε ένα τεράστιο ποσοστό της δουλειάς που πρέπει να γίνει στην πλευρά της κατηγοριοποίησης, αφού η περίληψη είναι μόνο ένα μικρό μέρος του κειμένου. Αυτό το αποτέλεσμα είναι μεγάλης σημασίας για ένα γρήγορα ανταποκρινόμενο, πραγματικού χρόνου σύστημα κατηγοριοποίησης.

Ένα επιπλέον πεδίο στο οποίο έγινε πειραματισμός αφορούσε τη διερεύνηση της επίπτωσης που έχει η κατηγοριοποίησης στην διαδικασία της περίληψης. Για να αποκαλυφθεί η πιθανή συσχέτιση, κατασκευάσαμε τον μηχανισμό περίληψης ενσωματώνοντας σε αυτόν την δυνατότητα κατηγοριοποίησης. Αυτό σημαίνει πως, όταν γνωρίζουμε εκ' των προτέρων την κατηγορία του κειμένου, μπορούμε να λάβουμε υπ' όψιν αυτή την πληροφορία κατά τη διαδικασία της περίληψης ρυθμίζοντας το βάρος της κάθε πρότασης ανάλογα. Για παράδειγμα, εάν μια πρόταση περιέχει πολλά keywords άσχετα με την κατηγορία του κειμένου (εκ' των προτέρων γνώση), το σκορ της θα είναι πολύ χαμηλό, ή ακόμη και αρνητικό σε σχέση με την περίπτωση που δεν γνωρίζουμε την κατηγορία του κειμένου.

Χρησιμοποιώντας κείμενα από συλλογές κειμένων (corpus texts), αρχικά παρηγάγαμε την περίληψη του κειμένου χωρίς την χρήση του παράγοντα κατηγοριοποίησης (δηλ. =1) και μετά χρησιμοποιήσαμε αυτή την επιπλέον πληροφορία για να παράγουμε μια ακόμη περίληψη. Συγκρίναμε τις δύο περιλήψεις με την «βέλτιστη» περίληψη που είχαμε από το corpus και που παρήχθη από ανθρώπους. Τα αποτελέσματα είναι αρκετά ενθαρρυντικά αφού βρέθηκε ότι το στοιχείο της κατηγοριοποίησης βελτίωνε τα αποτελέσματα της περίληψης κατά περίπου 10% ή ακόμη παραπάνω σε ορισμένες περιπτώσεις, κάτι που σημαίνει ότι οι προτάσεις τις οποίες κράτησε ο μηχανισμός περίληψης μετά τη χρήση της πληροφορίας κατηγοριοποίησης είναι πιο κοντά στις «βέλτιστες».



Εικόνα 29: Σύγκριση της ανάκλησης των περιλήψεων οι οποίες εξήχθηκαν με και χωρίς την χρήση του παράγοντα κατηγοριοποίησης.

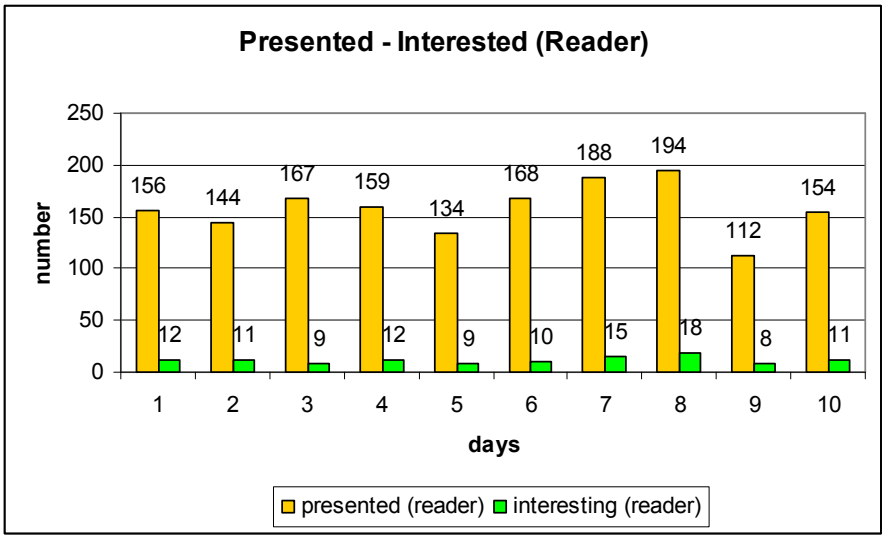


Εικόνα 30: Σύγκριση της μετρικής σειράς από περιλήψεις που εξήχθηκαν με και χωρίς τον παράγοντα κατηγοριοποίησης.

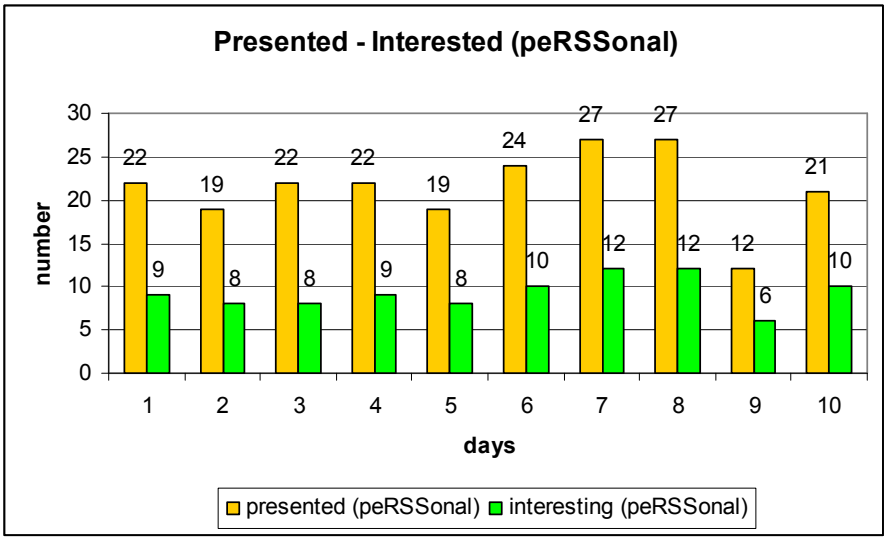
Για να συγκρίνουμε τα αποτελέσματα από τις δύο περιπτώσεις (με χρήση της πληροφορίας κατηγοριοποίησης και χωρίς), χρησιμοποιήθηκε η μετρική ανάκλησης, δηλαδή, πόσες από τις προτάσεις της ανθρώπινα εξαγόμενης («βέλτιστης») περίληψης ανακλήθηκαν από κάθε διαδικασία, και η μετρική σειράς των προτάσεων. Η τελευταία, χρησιμοποιήθηκε για να σημειώσει την σημασία που έχει η σειρά των προτάσεων σε μια περίληψη. Για παράδειγμα, είναι πιθανό και οι δύο τεχνικές περίληψης να επιτύχουν την ίδια ανάκληση προτάσεων αλλά η σειρά των προτάσεων να είναι καλύτερη σε μια από αυτές. Για την ακρίβεια, παρατηρήθηκε ότι η τεχνική περίληψης που κάνει χρήση της πληροφορίας κατηγοριοποίησης επιτυγχάνει όχι μόνο καλύτερη ανάκληση, αλλά και καλύτερη σειρά στις προτάσεις που επιστρέφουν.

9.3. Σύστημα παρουσίασης πληροφορίας (γενικά στοιχεία)

Στην τρέχουσα ενότητα γίνεται μια παρουσίαση και αξιολόγηση του συστήματος peRSSonal. Για να αξιολογηθεί το υποσύστημα παρουσίασης και το κατά πόσο η πληροφορία που φτάνει στο χρήστη είναι ικανοποιητική, εκτελέστηκε πειραματική διαδικασία. Κατά τη διάρκεια αυτής, δημιουργήθηκαν 10 προφίλ χρηστών με συγκεκριμένες προτιμήσεις σε κατηγορίες νέων, με τα άρθρα από τα οποία τροφοδοτείται ο μηχανισμός (από 10 RSS feeds), να τροφοδοτούνται στους 10 χρήστες. Παράλληλα τροφοδοτείται σε αυτούς και το προσωποποιημένο περιεχόμενο του συστήματος (περίληψη). Αυτό που εξετάστηκε είναι, το κατά πόσον οι χρήστες μπορούν να μείνουν ευχαριστημένοι α) από την επιλογή άρθρων που έγινε γι' αυτούς, και β) από το προσωποποιημένο περιεχόμενο που έχει να κάνει με τα συγκεκριμένα άρθρα που έλαβαν. Επίσης εκτιμάται και η μείωση στο φόρτο των χρηστών στη μία και στην άλλη περίπτωση, σε σχέση με την πληρότητα σε ενημέρωση που μπορούν να έχουν. Σε αυτό το σημείο θα πρέπει να τονιστεί ότι, ναι μεν θέλουμε το σύστημα να κάνει ένα φιλτράρισμα της υπέρογκης πληροφορίας για λογαριασμό των χρηστών, από την άλλη όμως, δεν θέλουμε να χάνονται άρθρα που θεωρούνται σημαντικά από τους χρήστες. Τα αποτελέσματα που προέκυψαν παρουσιάζονται στις γραφικές παραστάσεις του επόμενου σχήματος και αφορούν ημερήσιες τιμές.



Εικόνα 31: Τα άρθρα όπως παρουσιάζονται στους χρήστες απ' ευθείας από news portals

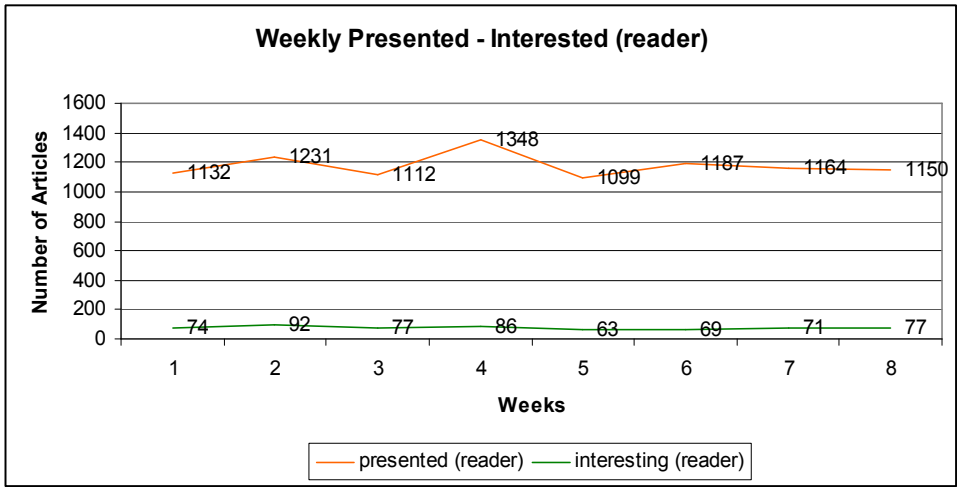


Εικόνα 32: Τα άρθρα όπως παρουσιάζονται στους χρήστες από το μηχανισμό

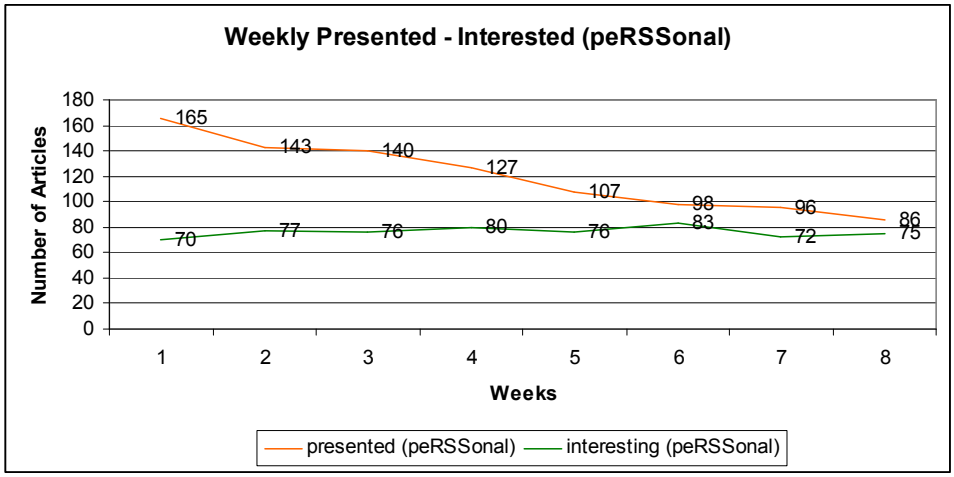
Όπως μπορούμε να δούμε, το σύστημα παρουσιάζει περίπου 85% λιγότερα άρθρα στους χρήστες ημερησίως αλλά το ποσοστό των άρθρων που μοιάζουν ενδιαφέροντα για τους χρήστες είναι πάνω από 40% ενώ στην περίπτωση άμεσης λήψης των άρθρων από news portals το ποσοστό των ενδιαφέροντων άρθρων είναι μόλις 7% των άρθρων που τους παρουσιάζονται. Αυτό σημαίνει ότι ο μηχανισμός μπορεί να επιτύχει καλύτερο «καθάρισμα» των άρθρων παρέχοντας νέα στους χρήστες που πραγματικά τους ενδιαφέρουν γλιτώνοντάς τους έτσι από τη χρονοβόρα διαδικασία του «ξεκαθαρίσματος» των άρθρων.

Ένα ακόμα πείραμα εστιάστηκε στη δυνατότητα του μηχανισμού να προσαρμόζεται σε κάθε χρήστη και από τα παρακάτω σχήματα είναι εμφανές πως προκύπτουν θεαματικά αποτελέσματα από τη χρήση του μηχανισμού, καθώς είναι σε θέση να εντοπίσει το προφίλ κάθε χρήστη. Τα RSS feeds περιέχουν το ακριβές URL της HTML σελίδας που έχει το άρθρο. Τα δικά μας feeds περιέχουν ένα URL προς το δικό μας δικτυακό τόπο, όπου ο χρήστης ανακατευθύνεται άμεσα προς το δικτυακό τόπο όπου βρίσκεται το άρθρο αφού πρώτα έχουμε συλλέξει πληροφορία για την επιλογή του χρήστη και το άρθρο που επέλεξε. Με αυτό τον τρόπο ακόμα και από τα rss feeds έχουμε τη δυνατότητα να συλλέξουμε πληροφορίες για το

προφίλ του χρήστη, για τα άρθρα που επέλεξε να διαβάσει και αυτά που έδειξε να μην τον ενδιαφέρουν. Ελέγχοντας τις κινήσεις αυτές του χρήστη είμαστε σε θέση να δημιουργήσουμε τις λίστες με τις λέξεις κλειδιά για το προφίλ του χρήστη. Οι παρακάτω εικόνες δείχνουν τη δυνατότητα εντοπισμού του προφίλ του χρήστη.



Εικόνα 33: Εβδομαδιαία παρουσίαση άρθρων από το RSS



Εικόνα 34: Εβδομαδιαία προσαρμογή του μηχανισμού στο προφίλ του χρήστη

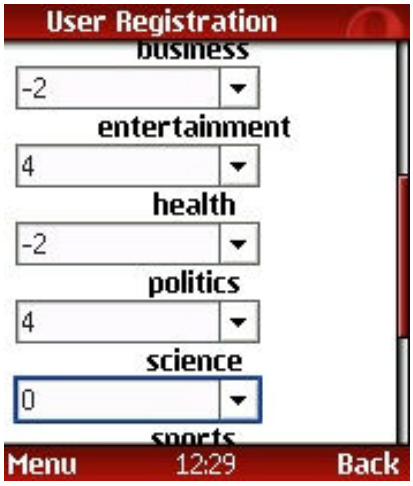
Όπως φαίνεται ξεκάθαρα από τα προηγούμενα διαγράμματα, ο μηχανισμός είναι σε θέση να προσαρμόζεται σε κάθε χρήστη ακόμα κι αν αυτός χρησιμοποιεί μόνο τα RSS Feeds και όχι το δικτυακό τόπο που έχουμε κατασκευάσει. Αυτό που θέλουμε να επιτύχουμε είναι τη σύγκλιση της γραμμής που περιλαμβάνει τα άρθρα που παρουσιάζονται στο χρήστη με τη γραμμή που παρουσιάζει τα άρθρα που ενδιαφέρουν πραγματικά το χρήστη. Από τα δύο παραπάνω διαγράμματα αποδεικνύονται δύο βασικές δυνατότητες του μηχανισμού. Πρώτα, ο μηχανισμός μπορεί να αποτελέσει ένα φίλτρο προκειμένου ένας χρήστης να λαμβάνει μόνο τα άρθρα για τα οποία ενδιαφέρεται και δεύτερον ο μηχανισμός είναι σε θέση ακόμα και με τη χρήση του RSS να συγκλινει προς το προφίλ του εκάστοτε χρήστη.

9.4. Σύστημα παρουσίασης πληροφορίας σε συσκευές μικρού μεγέθους

Σε αυτή την ενότητα γίνεται μια προσπάθεια αξιολόγησης του υποσυστήματος παρουσίασης της προσωποποιημένης περίληψης άρθρων στον τελικό χρήστη συσκευής μικρού μεγέθους. Προς την κατεύθυνση αυτή, παρουσιάζονται κάποια screenshots από τη συσκευή του τελικού χρήστη που χρησιμοποιεί το σύστημα.

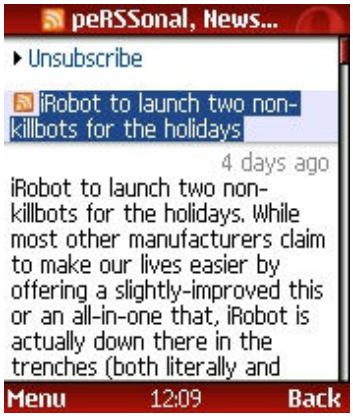
Όταν ένας νέος χρήστης καταφθάνει, οδηγείται στη φόρμα εγγραφής που φαίνεται στο παρακάτω σχήμα όπου εισάγει το όνομά του (ως απαραίτητο στοιχείο ταυτοποίησης) και τις δυνατότητες της συσκευής του (ανάλυση οθόνης). Το δεύτερο ανιχνεύεται αυτόματα από το υποσύστημα παρουσίασης, μπορεί όμως να ορισθεί και από τον χρήστη, και χρησιμοποιείται για τον καθορισμό α) του μήκους των περιλήψεων που στέλνονται στον χρήστη και β) το πλήθος των άρθρων που ταιριάζουν με το μέγεθος της συσκευής του. Ο χρήστης επίσης παρέχει τις προτιμήσεις του για τις κατηγορίες του συστήματος, σε κλίμακα από -5 έως +5.

Εικόνα 35: Εγγραφή του χρήστη (συσκευή μικρού μεγέθους)



Εικόνα 36: Επιλογές χρήστη για κάθε κατηγορία (συσκευή μικρού μεγέθους)

Όταν ένας μη-εγγεγραμμένος χρήστης ζητάει ένα RSS feed, μια προκαθορισμένη RSS απάντηση, η οποία περιέχει ορισμένες προκαθορισμένες περιλήψεις, στέλνεται πίσω στον χρήστη. Αντίθετα, αν ο χρήστης είναι ήδη εγγεγραμμένος, του στέλνεται η προσωποποιημένη απάντηση σύμφωνα με το προφίλ του όπως φαίνεται από τα παρακάτω σχήματα.



Εικόνα 37: Μια προκαθορισμένη απάντηση του συστήματος για μη-εγγρεγραμμένο χρήστη

Εικόνα 38: Προσωποποιημένη απάντηση σε εγγεγραμμένο χρήστη

Ο σημαντικός παράγοντας που πρέπει να ληφθεί υπ' όψιν, είναι ότι διαφορετικοί χρήστες το συστήματος λαμβάνουν διαφορετικές RSS απαντήσεις που ποικίλουν σε μήκος, σειρά κατάταξης, πλήθος και κατηγορία των νέων και των περιλήψεών τους. Είναι πολύ πιθανό δύο διαφορετικοί χρήστες να λαμβάνουν τα ίδια άρθρα αλλά διαφορετικές περιλήψεις αυτών (βάσει των προτιμήσεων των χρηστών)·αυτή είναι και η περίπτωση του παρακάτω σχήματος.

Εικόνα 39: πόκριση για τον χρήστη Α σχετικά με ένα άρθρο

Εικόνα 40: Απόκριση για το χρήστη Β για το ίδιο άρθρο

9.5. Ο δικτυακός τόπος personal

Ως τελικό στοιχείο σε αυτό το κεφάλαιο αφήσαμε το τελικό αποτέλεσμα του μηχανισμού συνολικά και δεν είναι άλλο από το δικτυακό τόπο που κατασκευάσαμε. Στην παρακάτω εικόνα εμφανίζεται η αρχική σελίδα του δικτυακού τόπου.

Εικόνα 41: Η αρχική σελίδα του δικτυακού τόπου

Η πρώτη σελίδα του δικτυακού τόπου περιέχει τα πιο πρόσφατα άρθρα που έχουν προστεθεί σε όλες τις κατηγορίες του συστήματος. Εμφανίζεται ο τίτλος του άρθρου όπως αυτός συλλέχθηκε από τα RSS Feeds καθώς και ένα μικρό κείμενο που προέρχεται από την περίληψη του κειμένου.

Στο αριστερό μέρος της οθόνης εμφανίζονται οι βασικές επιλογές του χρήστη:

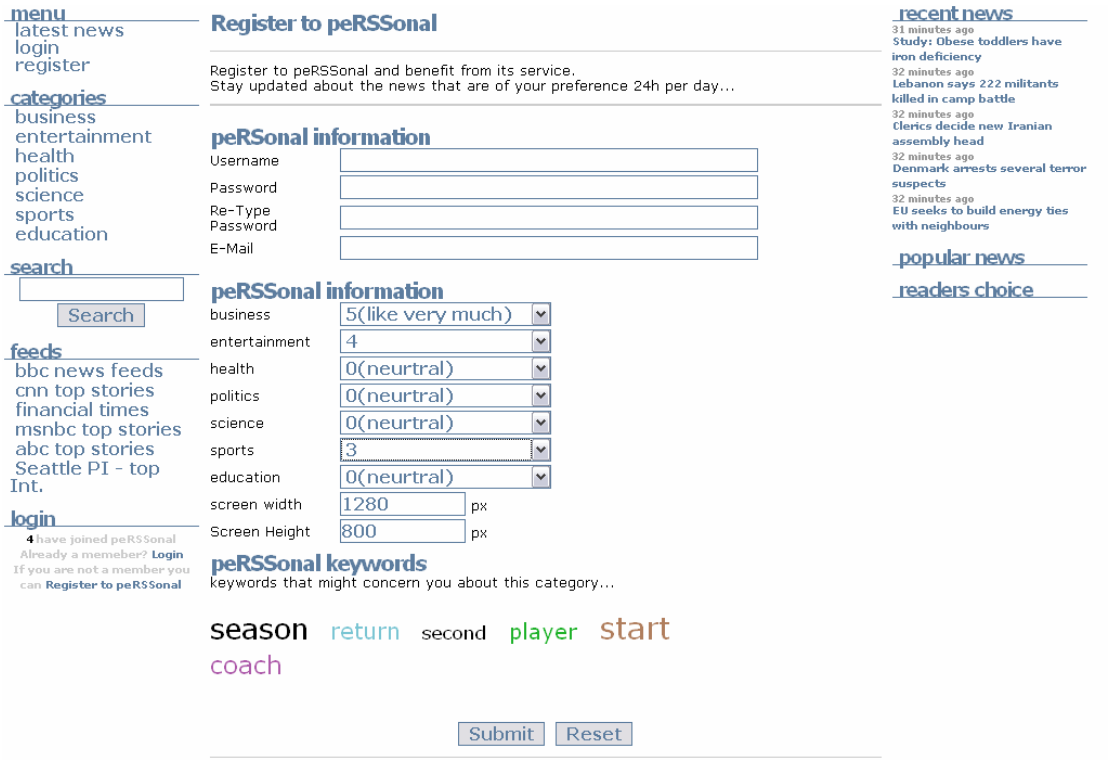
- Latest news (πρόσφατα άρθρα)

- Login (είσοδος)
- Register (εγγραφή)

Καθώς και οι βασικές κατηγορίες του συστήματος όπως και οι δικτυακοί τόποι απ' όπου συλλέγονται τα άρθρα του συστήματος:

- Business
- Entertainment
- Health
- Politics
- Science
- Sports
- Education

Ο εγγεγραμμένος χρήστης βλέπει ακριβώς τα ίδια στοιχεία με έναν απλό χρήστη με τη διαφορά πως τα άρθρα που παρουσιάζονται στον εγγεγραμμένο χρήστη είναι προσωποποιημένα στις προσωπικές του ρυθμίσεις που κάνει κατά την εγγραφή:



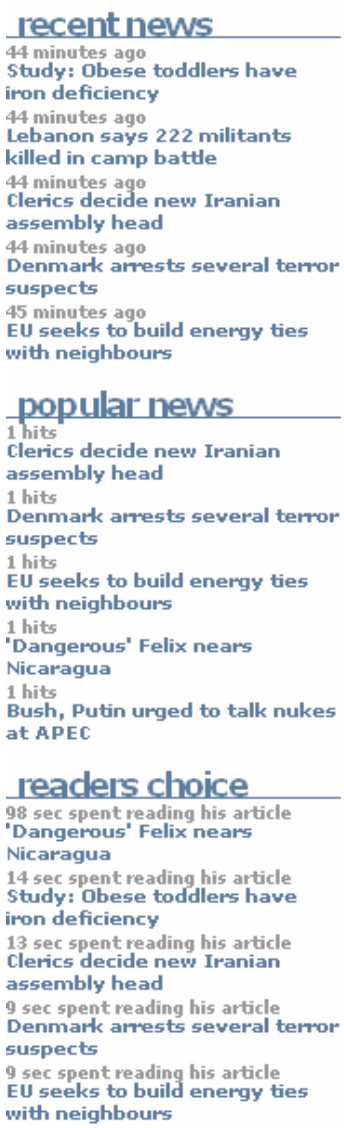
Εικόνα 42: Εγγραφή του χρήστη στο σύστημα

Κατά την εγγραφή του, ο χρήστης καλείται να συμπληρώσει στοιχεία που αφορούν το πόσο τον ενδιαφέρουν οι διαφορετικές κατηγορίες του συστήματος. Η εγγραφή στο σύστημα είναι ελεύθερη και αμέσως μετά ο χρήστης είναι σε θέση με ένα απλό login να δει την προσωποποιημένη του σελίδα.

Εικόνα 43: Σελίδα μετά από Login. Τα άρθρα που παρουσιάζονται είναι προσωποποιημένα στις ανάγκες του χρήστη

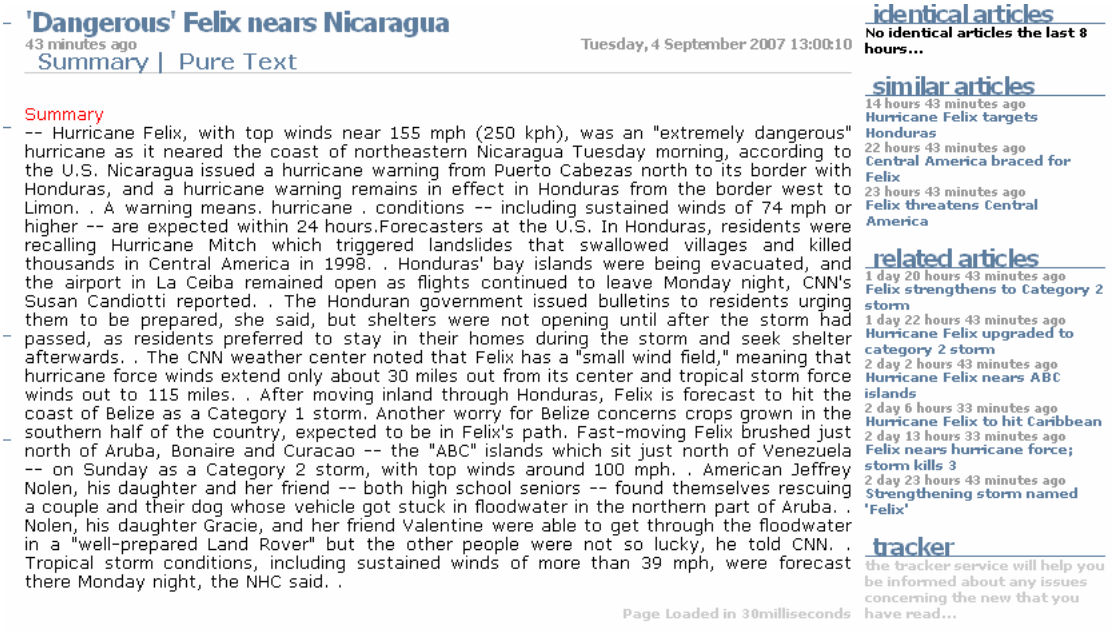
Όπως φαίνετα και από την εικόνα ο χρήστης παραμένει στο ίδιο περιβάλλον όπως ήταν και πριν κάνει login. Η φιλοσοφία είναι ο χρήστης να μην αλλάζει περιβάλλον προκειμένου να είναι σε θέση να αντιλαμβάνεται καλύτερα το συνολικό σύστημα.

Ένα πρωτοποριακό στοιχείο του συστήματος είναι το γεγονός πως στο δεξί μέρος της κεντρικής σελίδας εμφανίζονται εκτός από τα πιο πρόσφατα νέα και εκτός από αυτά που έχουν τα περισσότερα hits, αυτά στα οποία οι χρήστες έχουν ξοδέψει την περισσότερη ώρα για να τα διαβάσουν.



Εικόνα 44: Το δεξί μενού του δικτυακού τόπου. Αξίζει προσοχής το readers choice που περιέχει στοιχεία για το χρόνο τον οποίο ξόδεψαν οι χρήστες σε κάθε άρθρο.

Σημαντικά είναι επίσης τα στοιχεία που εμφανίζονται στο χρήστη όταν επιλέγει να αναγνώσει ένα άρθρο. Στην κεντρική οθόνη εμφανίζεται η προσωποποιημένη περίληψη ενώ δεξιά εμφανίζονται τρεις διαφορετικές ενότητες.



Εικόνα 45: Τρόπος απεικόνισης άρθρου στο χρήστη.

Όπως μπορούμε να δούμε ο χρήστης μπορεί να επιλέξει να αναγνώσει την περίληψη που έχουμε εξάγει γι αυτόν ή το συνολικό κείμενο του άρθρου. Αξιολογα είναι και τα στοιχεία που υπάρχουν στο δεξί μενού. Πρόκειται για τα στοιχεία identical articles, similar articles και related articles. Μέσα από πειράματα που κάναμε εντοπίσαμε πως για κάθε άρθρο υπάρχουν τρεις κατηγορίες άλλων άρθρων που σχετίζονται με αυτό. Πρόκειται για τα ταυτόσημα άρθρα, για τα παραπλήσια άρθρα και για τα σχετικά άρθρα. Η συσχέτισή τους αφορά κυρίως το περιεχόμενό τους αλλά και το χρόνο στον οποίο εμφανίζονται. Έτσι τα ταυτόσημα άρθρα είναι αυτά τα οποία έχουν ομοιότητα πάνω από 80% με το άρθρο (cosine similarity) και έχουν δημοσιευθεί μέχρι και 8 ώρες πριν και μετά από το συγκεκριμένο άρθρο, τα παραπλήσια άρθρα είναι αυτά που έχουν ομοιότητα πάνω από 70% με το άρθρο και έχουν δημοσιευθεί από 8 έως 24 ώρες πριν και μετά από το άρθρο, ενώ τα σχετικά άρθρα είναι αυτά που έχουν ομοιότητα πάνω από 65% με το άρθρο και έχουν δημοσιευθεί από 24 έως 48 ώρες πριν και μετά από το άρθρο. Φυσικά, θα πρέπει να έχουμε υπόψη μας πως υπάρχουν και άρθρα που έχουν δημοσιευθεί πριν και μετά από αυτό το διάστημα των 48 ωρών. Τα άρθρα αυτά ελέγχονται με απλή προσέγγιση στις λέξεις κλειδιά του άρθρου και όχι με την cosine similarity μετρική.

Θα πρέπει να τονίσουμε πως ο έλεγχος με τη μετρική cosine similarity σε τόσα πολλά άρθρα χρειάζεται περίπου 5 δευτερόλεπτα να δώσει απάντηση για κάθε κατηγορία σχετιζόμενων άρθρων. Αν, λοιπόν, επιλέγαμε να εμφανίζουμε απ' ευθείας τα αποτελέσματα θα παρατηρούσαμε μία καθυστέρηση μεγέθους 15 δευτερολέπτων στη φόρτωση της σελίδας. Για το λόγο αυτό, η εμφάνιση των σχετιζόμενων άρθρων έχει υλοποιηθεί αποκλειστικά με χρήση AJAX.

Τέλος, για τους χρήστες του συστήματος προσφέρεται μία επιπλέον υπηρεσία που ονομάζεται tracker. Αν ο χρήστης επιλέξει να κάνει track ένα συγκεκριμένο άρθρο τότε ενημερώνεται μέσω e-mail για όλα τα σχετιζόμενα άρθρα με αυτό για διάστημα 72 ωρών. Εν συνεχεία ενημερώνεται και πάλι με e-mail πως έχει περάσει ο χρόνος tracking και ο χρήστης μπορεί εφόσον επιθυμεί να συνεχίσει τον έλεγχο για άλλες 72 ή περισσότερες ώρες.

10

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στο κεφάλαιο αυτό περιγράφονται τα συμπεράσματα από τη χρήση του μηχανισμού

10. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εργασία που εκπονήθηκε στα πλαίσια της μεταπτυχιακής εργασίας «*Προσωποποιημένη Προβολή Περιεχομένου του Διαδικτύου με τεχνικές Προεπεξεργασίας, Αυτόματης Κατηγοριοποίησης και Αυτόματης Εξαγωγής Περίληψης*» περιλαμβάνει την κατασκευή ενός μηχανισμού που θα είναι σε θέση ξεκινώντας από συλλογή άρθρων από το διαδίκτυο να τα παρουσιάσει πίσω στους χρήστες με τρόπο ο οποίος θα είναι προσαρμοσμένος στις ανάγκες του.

Η εργασία περιλαμβάνει την κατασκευή πολλών υποσυστημάτων καθένα από τα οποία είναι κομμάτι διαφορετικού ερευνητικού πεδίου. Μέσα από την εργασία προσεγγίσαμε διαφορετικά πεδία της επιστήμης των υπολογιστών προσθέτωντας το δικό μας λιθαράκι στην περεταίρων εξέλιξή τους.

Το διαδίκτο έχει λάβει χαοτικές διαστάσεις και η πληροφορία που διακινείται σε αυτό είναι υπέρογκη. Στην εποχή μας και με τα μέσα που διαθέτει ακόμα και ο απλός χρήστης, η προσθήκη περιεχομένου στο Διαδίκτυο από τον καθένα είναι μία διαδικασία το ίδιο εύκολη και απλή με την περιαγωγή στο χώρο του παγκόσμιου ιστού (fora, blogs, web 2.0). Το πρόβλημα που δημιουργεί αυτή η ανεξέλεγκτη κατάσταση είναι ότι ακόμα και οι πιο έμπειροι χρήστες καταναλώνουν πολύ χρόνο στην προσπάθεια εύρεσης πληροφορίας και συγκεκριμένα πηγών ενημέρωσης για τα θέματα που τους ενδιαφέρουν.

Εστιάζοντας στο συγκεκριμένο πρόβλημα δημιουργήσαμε το σύστημα που περιγράφηκε από τη συγκεκριμένη εργασία προσπαθώντας να περιορίσουμε με κάποιον τρόπο το πρόβλημα που δημιουργείται από την αναζήτηση άρθρων σε μεγάλα ειδησεογραφικά πρακτορείο. Κατασκευάσαμε ένα μηχανισμό που μπορεί να συλλέγει άρθρα από το διαδίκτυο, να απομονώνει το χρήσιμο κείμενο από αυτά, να πραγματοποιεί μεθόδους εξαγωγής των λέξεων κλειδιών από τις συγκεκριμένες σελίδες καθώς και να τις κατηγοριοποιεί και να εξάγει αυτόματα περίληψη. Τέλος, το πιο σημαντικό στοιχεία και αυτό που είναι ουσιωδώς ορατό στους χρήστες είναι ο δικτυακός τόπος που εμφανίζει τα άρθρα στους τελικούς χρήστες.

10.1. Συμπεράσματα

Ο μηχανισμός συλλογής σελίδων με άρθρα από το διαδίκτυο βασίζεται στην απλή ιδέα πως τα links για αυτές τις σελίδες μπορούν εύκολα να εντοπιστούν από τα RSS feeds και γενικότερα τα κανάλια επικοινωνίας που προσφέρουν πλέον όλα τα μεγάλα ειδησεογραφικά πρακτορεία. Ο μηχανισμός συλλογής των σελίδων δουλεύει πλήρως παρουσιάζοντας πολύ καλά αποτελέσματα

Ο μηχανισμός απομόνωσης χρήσιμου κειμένου βασίζεται στην ιδέα του web clipping. Μέσα από μία αλγοριθμική διαδικασία γίνεται προσπάθεια εντοπισμού των στοιχείων μίας ιστοσελίδας που περιέχουν αρκετή ποσότητα κειμένου συγκριτικά με το κείμενο της σελίδας αλλά και γενικά συγκρικά με άλλες περιοχές μικρότερης πυκνότητας σε κείμενο. Ο μηχανισμός δουλεύει πλήρως δίνοντας καλά αποτελέσματα ενώ υπάρχει δυνατότητα βελτίωσης προκειμένου να συλλέγονται μόνο τα στοιχεία κειμένου και όχι επιπρόσθετα στοιχεία της σελίδας.

Ο μηχανισμός εξαγωγής των λέξεων κλειδιών βασίζεται σε γνωστούς αλγορίθμους γλωσσολογικής ανάλυσης κειμένου ενώ τα αποτελέσματά του είναι πολύ καλά. Παράλληλα, στην εξαγωγή των λέξεων κλειδιών εφαρόζονται επιπλέον αλγόριθμοι προκειμένου να διατηρούνται όσο λιγότερες λέξεις κλεδιά χωρίς να χάνεται το νόημα του κειμένου.

Ο μηχανισμός κατηγοριοποίησης βασίζεται σε ευρεστικές μεθόδους και πιο συγκεκριμένα στη σύγκριση κάθε κειμένου με πρότυπες κατηγορίες που υπάρχουν για το σύστημα (training set). Η κατηγοριοποίηση εξαρτάται άμεσα από τα κείμενα των πρότυπων κατηγοριών και έτσι η βελτίωσή της έγκειται στην κατασκευή καλύτερου και πληρέστερου training set.

Ο μηχανισμός αυτόματης εξαγωγής περίληψης βασίζεται επίσης σε ευρεστικές μεθόδους και η λειτουργία του ενοπίζεται στη βαθμοδότηση των προτάσεων προκειμένου να κρατηθούν αυτές με τη μεγαλύτερη βαθμολογία. Στη συγκεκριμένη εργασία προτείνουμε ένα νέο αλγόριθμο αυτόματης εξαγωγής περίληψης και πιο συγκεκριμένα έναν αλγόριθμο βαθμοδότησης των προτάσεων ο οποίος παρουσιάζει καλά αποτελέσματα, και σίγουρα καλύτερα από αρκετούς αλγορίθμους και μηχανισμούς που έχουν προταθεί στο παρελθόν.

Οι μηχανισμοί κατηγοριοποίησης και περίληψης μπορούν να λειτουργήσουν και παράλληλα στο σύστημα προκειμένου να έχουμε πληρέστερα και ακριβέστερα αποτέλεσματα για καθέναν από τους δύο αυτούς μηχανισμούς. Μάλιστα παρατηρήσαμε πως σε πολλές περιπτώσεις που δεν είναι εφικτή η κατηγοριοποίηση ενός κειμένου, είναι εφικτή η κατηγοριοποίηση της περίληψης που εξάγουμε γι αυτό.

Τέλος, ο μηχανισμός προσωποποίησης στο χρήστη είναι ένας πρωτοποριακός δικτυακός τόπος ο οποίος περιέχει πολλά στοιχεία που θα μπορούσαν να εφαρμοστούν σε σύγχρονα news portals προκειμένου να παρέχουν περισσότερες υπηρεσίες στους χρήστες τους.

11

ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Στο κεφάλαιο αυτό ιδέες για μελλοντική εργασία βελτίωσης του μηχανισμού

11. ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Μέσα από την εκπόνηση της συγκεκριμένης εργασίας ίσως το πιο σημαντικό συμπέρασμα που εντοπίστηκε είναι το γεγονός πως η επιστήμη δεν έχει όρια. Έτσι, μέσα από την υλοποίηση των μηχανισμών του συστήματος, εντοπίστηκαν στοιχεία που μπορούν να αλλαξουν και να βελτιωθεί συνολικά το σύστημα, ιδέες που μπορούν να εφαρμοστούν για ένα πιο ολοκληρωμένο μηχανισμό.

Δεδομένου ότι ο μηχανισμός συλλογής ιστοσελίδων άρθρων λειτουργεί άψογα προχωρούμε στον επόμενο μηχανισμό για να εντοπίσουμε στοιχεία που μπορούν να βελτιωθούν. Ο επόμενος μηχανισμός είναι ο μηχανισμός εξαγωγής χρήσιμου κειμένου από HTML σελίδες. Ο μηχανισμός αυτός υλοποιείται αποκλειστικά σε JAVA και μάλιστα γίνεται χρήση βιβλιοθηκών που δεν είναι ενδεδειγμένες για ανάλυση HTML (χρησιμοποιούνται κυρίως για XML). Η βελτίωση που θα γίνει στο συγκεκριμένο μηχανισμό έχει να κάνει με τον τρόπο υλοποίησης και όχι με τους αλγορίθμους που χρησιμοποιούνται. Τα όποια σφάλματα του μηχανισμού δεν προκύπτουν από τους αλγορίθμους αλλά από αστοχία της JAVA να αναγνωρίσει συγκεκριμένα HTML tags. Έτσι στο μέλλον ο μηχανισμός αυτός θα μετατραπεί σε C++ και θα χρησιμοποιεί βιβλιοθήκες ανάλυσης HTML σελίδων καθώς και αναγνώριση όλων των HTML tags προκειμένου να μην εισάγονται «ξένα» στοιχεία στο σώμα του άρθρου από λάθος της γλώσσας προγραμματισμού.

Εν συνεχεία ακολουθεί ο μηχανισμός εξαγωγής των λέξεων κλειδιών. Από προγραμματιστικής άποψης θα μπορούσαν να βελτιωθούν τα εξής στοιχεία:

- Χρήση πληρέστερης λίστας από stopwords η οποία δε θα απορρίπτει σημαντικές λέξεις και δε θα κρατά μη χρήσιμες. Θα μπορούσε επίσης να χρησιμοποιηθεί δυναμική λίστα η οποία θα προσαρμόζεται ανάλογα με τη θεματολογία του κειμένου.
- Χρήση ενός καλύτερου stemmer Που θα βασίζεται σε κανόνες και θα έχει ελάχιστα λάθη συγκριτικά με τον porter stemmer που χρησιμοποιείται και βασίζεται κυρίως σε «έτοιμες» καταλήξεις και όχι σε κανόνες.
- Χρήση λεξικού για εντοπισμό ορθογραφικών λαθών πριν λάθος λέξεις κλειδιά θεωρηθούν από το σύστημα σαν πραγματικά keywords.
- Εντοπισμός μόνο των ουσιαστικών του κειμένου που περιέχουν όλο το νόημα του κειμένου. Παράλληλα θα πρέπει να γίνει ανανέωση της βάσης γνώσης ώστε να έχει μόνο ουσιαστικά
- Αναγνώριση της γλώσσας του κειμένου και επέκταση συνολικά του μηχανισμού σε ένα ενοποιημένου πολυγλωσσικό περιβάλλον τεχνικών ανάκτησης και επεξεργασίας πληροφορίας.

Όσον αφορά τη μετατροπή του μηχανισμού σε πολυγλωσσικό σύστημα επεξεργασίας κειμένων θα πρέπει να περιληφθούν λίστες με stopwords σε όλες τις γλώσσες, κανόνες stemming για κάθε γλώσσα καθώς και λεξικά για κάθε γλώσσα. Οι πληροφορίες θα αποθηκεύονται κεντρικά στη βάση δεδομένων και θα είναι διαθέσιμες όταν ζητηθούν κατά την εκτέλεση του μηχανισμού και αφού αναγνωριστεί η γλώσσα στην οποία είναι γραμμένο το κείμενο. Σε αρχική φάση εστιάζουμε την προσοχή μας στην επέκταση του μηχανσιμού στην ελληνική γλώσσα.

Όσον αφορά τα υποσυστήματα κατηγοριοποίησης και αυτόματης εξαγωγής περίληψης, στοχεύουμε σε ανάπτυξη πολλαπλών διαφορετικών αλγορίθμων προκειμένου να υπάρχει η δυνατότητα επιλογής του αλγορίθμου κατηγοριοποίησης και εξαγωγής περίληψης. Παράλληλα, θα αναπτυχθεί γραφικό περιβάλλον ελέγχου

της λειτουργίας αυτών των μηχανισμών που θα επικοινωνεί απ’ ευθείας με εντολές πυρήνα του συστήματος στο οποίο λειτουργούν τα συγκεκριμένα συστήμα.

Τέλος ο δικτυακός τόπος θα βελτιωθεί προκειμένου να υποστηρίζει ένα πολύγωσσο σύστημα υποστήριξης. Για το δικτυακο τόπο υπάρχουν πολλά στοιχεία που θα μπορούσαν να βελτιωθούν και αφορούν:

- Τον τρόπο παρουσίασης των δεδομένων (γραφιστικά)
- Τον τρόπο επιλογής των σχετιζόμενων άρθρων
- Τη διαφάνεια των διαδικασιών προς τον τελικό χρήστη
- Περισσότερους τρόπους αξιολόγησης του προφίλ του χρήστη

Ο δικτυακός τόπος είναι ένα πολύ σημαντικό εργαλείο και η σωστότερη ανάπτυξή του μπορεί να οδηγήσει ακόμα και σε εμπορικό πρόγραμμα χρήσης από εξειδικευμένους χρήστες που χρειάζονται διαρκή και συνεχή ενημέρωση για νέα και ειδήσεις.

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΒΙΒΛΙΑ / ΔΗΜΟΣΙΕΥΣΕΙΣ

[1] Mooers, C. N. 1952. Information Retrieval Viewed as Temporal Signaling. In Proceedings of the International Conference of Mathematicians, Cambridge, Massachusetts. American Mathematical Society, σελίδες 572-573.

[2] Doyle L. B. 1961. Semantic Road Maps for Literature Searchers. In Journal of the Association for Computing Machinery, 8, σελίδες 553-578.

[3] Salton, G. 1968. Automatic Information Organization and Retrieval. New York: McGraw-Hill.

[4] Shneiderman, B., Byrd, D. and Croft, B. 1998. Sorting out Searching: a User-Interface Framework for Text Searches. In Communications of the ACM, 41(4), σελίδες 95-98.

[5] Salton, G. and Buckley, C. 1988. Improving Retrieval Performance by Relevance Feedback. In Journal of the American Society for Information Science, 41, σελίδες 288-297.

[6] Cleverdon, C. W. 1972. The Cranfield Tests on Index Language Devices. In Aslib Proceedings, 19, σελίδες 173-192.

[7] Belkin, N. J., and Croft, W. B. 1992. Information Filtering and Information Retrieval: Two Sides of the Same Coin? In Communications of the ACM, 35(12), σελίδες 29-38.

[8] Quinlan, J. R. 1986. Induction of Decision Trees. In Machine Learning I.

[9] Chickering, D., Heckerman, D. AND MECK C. 1997. A Bayesian Approach for Learning Bayesian Networks with Local Structure. In Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence.

[10] Belur, V. D. 1991. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. McGraw-Hill Computer Science Series. IEEE Computer Society Press.

[11] Vapnik, V. 1995. The Nature of Statistical Learning Theory, Springer-Verlag.

[12] Cortes, C. and Vapnik, V. 1995. Support-Vector Networks. In Machine Learning, 20, σελίδες 273-297.

[13] Belur, V. D. 1991. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. McGraw-Hill Computer Science Series. IEEE Computer Society Press.

[14] Bouras, C., and Konidaris, A. 2001. Web Components: A Concept for Improving Personalization and Reducing User Perceived Latency on the World Wide Web. The 2nd International Conference on Internet Computing (IC2001),Las Vegas, Nevada, USA, Vol. 2, σελίδες 238-244.

[15] Bouras, C., Kapoulas, V., and Misedakis, I. 2004. Web Page Fragmentation for Personalized Portal Construction. IEEE International Conference on Information Technology: Coding and Computing - ITCC 2004 (Web/IR Track), The Orleans, Las Vegas, Nevada, USA, σελίδες 332 – 336.

[16] Πανεπιστημιακό Σύγγραμμα «Ανάκτηση Πληροφορίας». Δρ. Χρήστος Μακρής, Ε. Θεοδωρίδης, Ι. Παναγής, Α. Περδικούρη, Ε. Χριστοπούλου.

[17] Πανεπιστημιακές Παραδόσεις «Προηγμένα Πληροφοριακά Συστήματα». Α. Τσακαλίδης, Β. Βασιλειάδης, Ε. Σακκόπουλος.

[18] Nuno Miguel de Sousa Maria, "Theme-Based Retrieval of Web News", PhD Thesis.

[19] H. Luhn. The automatic creation of literature abstracts. Originally published in IBM Journal of R&D.

[20] S. Myaeng and D. Jang. Development and evaluation of a statistically based document summarization system.

[21] H. Edmundson. New methods in automatic extracting. Originally published in Journal of the ACM.

[22] J. Pollock and A. Zamora. Automatic abstracting research at chemical abstracts service. Originally published in Journal of Chemical Information and Computer Science.

[23] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. Originally published in Proceedings of SIGIR.

[24] G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. Originally published in Information Processing and Management.

[25] C. Aone, M. Okurowski, J. Gorlinsky, and B. Larsen. A trainable summarizer with knowledge acquired from robust NLP techniques.

[26] A. Berker and V. Mittal. OCELOT: a system for summarizing web pages. In Proceedings of SIGIR, pages 144–151, 2000.

[27] M. Witbrock and V. Mittal. Ulta-summarization : A statistical approach to generating highly condensed non-extractive summaries. In Proceedings of SIGIR, pages 315–316, 1999. Poster description.

[28] M. Maybury. Generating summaries from event data. Originally published in Information Processing and Management.

[29] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producting summaries. In Proceedings of SIGIR, pages 335–336, 1998

[30] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. Originally published in Information Retrieval.

[31] I. Mani and M. Maybury, editors. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, 1999.

[32] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction, pages 40–48.

[33] E. Hovy and C. Lin. Automated text summarization in SUMMARIST.

[34] K. McKeown and D. Radev. Generating summaries of multiple news articles.

[35] D. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. pages 21–30.

[36] F. Fukumoto and Y. Suzuki. Extracting key paragraph based on topic and event detection. pages 31–39.

[37] Pinkerton, B. (1994). Finding what people want: Experiences with the WebCrawler[dead link – history]. In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

[38] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7):107–117.

[39] Heydon, A. and Najork, M. (1999). Mercator: A scalable, extensible Web crawler. World Wide Web, 2(4):219–229.

[40] Edwards, J., McCurley, K. S., and Tomlin, J. A. (2001). "An adaptive model for optimizing performance of an incremental web crawler". In Proceedings of the Tenth Conference on World Wide Web: 106-113.

[41] Marc Najork and Janet L. Wiener. Breadth-first crawling yields high-quality pages[dead link – history]. In Proceedings of the Tenth Conference on World Wide Web, pages 114–118, Hong Kong, May 2001. Elsevier Science.

[42] Zeinalipour-Yazti, D. and Dikaiakos, M. D. (2002). Design and implementation of a distributed crawler and filtering processor. In Proceedings of the Fifth Next Generation Information Technologies and Systems (NGITS), volume 2382 of Lecture Notes in Computer Science, pages 58–74, Caesarea, Israel. Springer.

[43] Boldi, P., Codenotti, B., Santini, M., and Vigna, S. (2004a). UbiCrawler: a scalable fully distributed Web crawler. Software, Practice and Experience, 34(8):711–726.

[44] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. Communications of the ACM, 43(8):142--151, 2000.

[45] T. Joachims, D. Freitag, and T. Mitchell. 1997. Webwatcher: A tour guide for the World Wide Web. In Proc. IJCAI-97. http://citeseer.ist.psu.edu/joachims96webwatcher.html

[46] Dick Hardt. How SXIP Works (whitepaper). https://sxip.org/docs/specs/how-sxip-works.pdf 2004.

[47] Proposal for an Open Profiling Standard. W3C Note – 02 June 1997. http://www.w3.org/TR/NOTE-OPS-FrameWork

[48] PIDL - Personalized Information Description Language. W3C Note - 09 Feb 1999. http://www.w3.org/TR/1999/NOTE-PIDL-19990209

[49] Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0. W3C Recommendation 15 January 2004. http://www.w3.org/TR/2004/REC-CCPP-structvocab-20040115/

[50] The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. W3C Recommendation 16 April 2002. http://www.w3.org/TR/P3P/

[51] Project Liberty. Introduction To The Liberty Alliance Identity Architecture (whitepaper). 2003. Available from https://www.projectliberty.org/resources/whitepapers/LAP%20Identity%20Architecture%20Whitepaper%20Final.pdf

[52] Gary Ellison, Jeff Hodges, Susan Landau (2002) Security and Privacy Concerns of Internet Single Sign-On: Risks and Issues as They Pertain to Liberty AllianceVersion 1.0 Specifications. Technical Report.

[53] Dick Hardt. How SXIP Works (whitepaper). https://sxip.org/docs/specs/how-sxip-works.pdf 2004.

[54] Jude Shavlik et al : An Instructable, Adaptive Interface for Discovering and Monitoring Information on the World Wide Web. Proceedings of the 1999 International Conference on Intelligent User Interfaces, pp. 157 - 160, Redondo Beach, CA.

[55] Dwi H. Widyantoro, Thomas R.Ioerger and John Yen, Learning User Interest Dynamics with a Three-Descriptor Representation. Journal of the American Society for Information Science, 52(3):212-225.

[56] Philip Chan: Constructing Web User Profiles: A Non-invasive Learning Approach. KDD-99 Workshop on Web Usage Analysis and User Profiling, pp. 7-12, 1999.

[57] Michael Pazzani et al : Syskill & Webert : Identifying interesting Web sites. M. Pazzani, J. Muramatsu, and D. Billsus, "Syskill & Webert: Identifying

interesting Web sites," in Proceedings of the 13th National Conference on Artificial Intelligence (AAA196), 1996, pp. 54--61.

[58] Jeremy Goecks, Jude Shavlik: Automatically Labeling Web Pages Based on Normal User Actions. In Proceedings of the IJCAI Workshop on Machine Learning for Information Filtering, Stockholm, Sweden, July 1999.

[59] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In ANLP/NAACL Workshop on Summarization, Seattle, WA, April 2000.

[60] Dragomir Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Celebi, Hong Qi, Daniu Liu, and Elliott Drabek. Evaluation challenges in large-scale multidocument summarization: the MEAD project. Johns Hopkins University CLSP Workshop Final Report, 2001.

[61] Eduard Hovy , Chin-Yew Lin, Automated text summarization and the SUMMARIST system, Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998, October 13-15, 1998, Baltimore, Maryland.

[62] H. Arimura, A.Wataki, R. Fujino, and S. Arikawa. A fast algorithm for discovering optimal string patterns in large text databases. Proc. the 8th International Workshop on Algorithmic Learning Theory, 1501:247–261.

[63] M. Montes-y G´omez, A. Gelbukh, and A. L´opez-L´opez. Mining the News: Trends, Associations, and Deviations. Computaci´on y Sistemas, 5(1):14–24, 2001.

[64] K. Hoang and P. Do. Discovering Motiv Based Association Rules in a Set of DNA sequences. RSCTC, pages 386–390, 2000.

[65] Colleen E. Crangle. Text summarization in data mining. In Soft-Ware 2002: Proceedings of the First International Conference on Computing in an Imperfect World, pages 332–347, London, UK, 2002. Springer-Verlag.

[66] N. Collier, C. Nobata, and J. Tsujii. Extracting the names of genes and gene products with a hidden Markov model. Proceedings of the 18th conference on Computational linguistics-Volume 1, pages 201–207, 2000.

[67] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. Tuning Support Vector Machines for Biomedical Named Entity Recognition. Proc. of the Workshop on Natural Language Processing in the Biomedical Domain (at ACL'2002), pages 1–8, 2002.

[68] C. Nobata, N. Collier, and J. Tsujii. Automatic term identification and classification in biology texts. Proc. of the 5th NLPRS, pages 369–374, 1999.

[69] K.S. Jones. Exhaustivity and specificity. Journal of Documentation, 28(1):11–21, 1972.

[70] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning, 97, 1997.

[71] D. Mladenic and M. Grobelnik. Word sequences as features in text-learning. Proceedings of ERK-98, the Seventh Electrotechnical and Computer Science Conference, pages 145–148, 1998.

[72] J.C. French, A.L. Powell, J. Callan, C.L. Viles, T. Emmitt, K.J. Prey, and Y. Mou. Comparing the performance of database selection algorithms. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 238–245, 1999.

[73] M. Lennon. Pierce. D., Tarry, B.. & Willett, P.(198 1). An evaluation of the stemming algorithms.

[74] J.B. Lovins. Development of a Stemming Algorithm. 1968.

[75] M. Porter. The Porter Stemming Algorithm. Accessible at http://www. tartarus. org/martin/PorterStemmer.

[76] C.D. Paice. Another stemmer. ACM SIGIR Forum, 24(3):56–61, 1990.

[77] R. Krovetz. Viewing morphology as an inference process. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 191–202, 1993.

[78] W.B. Frakes and R. Baeza-Yates. Information retrieval: data structures and algorithms. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1992.

[79] R. Krovetz. Viewing morphology as an inference process. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 191–202, 1993.

[80] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. Proceedings of the seventh international conference on Information and knowledge management, pages 148–155, 1998.

[81] Y. Yang and C.G. Chute. An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems (TOIS), 12(3):252–277, 1994.

[82] B. Masand. Lino, G., &Waltz, D.(1992). Classifying news stories using memory based reasoning. Proceedings of 506 15th ACM SIGIR international conference on research and development in information retrieval, pages 59–65.

[83] Y. Yang. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 13–22, 1994.

[84] K. Tzeras and S. Hartmann. Automatic indexing based on Bayesian inference networks. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 22–35, 1993.

[85] D.D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. Third Annual Symposium on Document Analysis and Information Retrieval, pages 81–93, 1994.

[86] C. Apt´e, F. Damerau, and S.M. Weiss. Towards language independent automated learning of text categorization models. Springer-Verlag New York, Inc. New York, NY, USA, 1994.

[87] W.W. Cohen. Text categorization and relational learning. Proceedings of ICML-95, 12th International Conference on Machine Learning, pages 124–132, 1995.

[88] H.T. Ng, W.B. Goh, and K.L. Low. Feature selection, perception learning, and a usability case study for text categorization. Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, pages 67–73, 1997.

[89] T. Joachims. Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. Springer-Verlag London, UK, 1998.

[90] PC Reghu Raj and S. Raman. Content identification and semantic indexing of text documents. Proc. Of the Indo European Conference on Multilingual Communication Technologies (IEMCT-02), pages 203–217, 2002.

[91] M.F. Caropreso, S. Matwin, and F. Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. Text Databases and Document Management: Theory and Practice, pages 78–102, 2001.

[92] J. Furnkranz, T. Mitchell, and E. Riloff. A case study in using linguistic phrases for text categorization on the WWW. Learning for Text Categorization: Proceedings of the 1998 AAAI/ICML Workshop, pages 98–05, 1998.

[93] E. Riloff and J. Shepherd. A corpus-based approach for building semantic lexicons. Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, pages 117–124, 1997.

[94] C. Jacquemin. Spotting and Discovering Terms Through Natural Language Processing. MIT Press, 2001.

[95] H. Berger and D. Merkl. A Comparison of Text-Categorization Methods applied to N-Gram Frequency Statistics. Proc. of the 17th Australian Joint Conf. on Artificial Intelligence, 2004.

[96] B. Sankaran. Tamil Search Engine.

[97] K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artificial Intelligence, 139(1):91–107, 2002.

[98] R. Barzilay and L. Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. Proceedings of HLT-NAACL 2004, pages 113–120, 2004.

[99] G. Salton, J. Allan, C. Buckley, and A. Singhal. Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts. Science, 264(5164):1421, 1994.

[100] M. Saravanan, Pc Reghu Raj, and S. Raman. Summarization and Categorization of text data in high-level data cleaning for information retrieval. Applied Artificial Intelligence, 17(5):461–474, 2003.

[101] M. Saravanan and S. Raman. The term distribution model for summarization of multiple documents. Proceedings of the Indo European Conference on Multilingual Communication Technologies (IEMCT 2002), pages 182–192, 2002.

[102] R. Evans, R. Gaizauskas, L. Cahill, J.Walker, J. Richardson, and A. Dixon. POETIC: a system for gathering and disseminating traffic information. Journal of Natural Language Engineering, 1(4), 1995.

[103] D. Marcu. The rhetorical parsing of natural language texts. Proceedings of the 35th annual meeting on Association for Computational Linguistics, pages 96–103, 1997.

[104] R.C. Schank. Reading and Understanding: Teaching from the Perspective of Artificial Intelligence. Lawrence Erlbaum Associates, 1982.

[105] WA Woods and JG Schmolze. The KL-ONE family. Semantic Networks in Artificial Intelligence, Pp133-178, 1992.

[106] D. Fum, G. Guida, and C. Tasso. Forward and backward reasoning in automatic abstracting. Proceedings of the 9th conference on Computational linguistics-Volume 1, pages 83–88, 1982.

[107] PS Jacobs and L.F. Rau. SCISOR: extracting information from on-line news. Communications of the ACM, 33(11):88–97, 1990.

[108] U. Hahn and U. Reimer. Semantic Parsing and Summarizing of Technical Texts in the TOPIC System. Informations linguistik, pages 153–193, 1986.

[109] L.A. Mather and J. Note. Discovering Encyclopedic Structure and Topics in Text. Sixth ACM SIGKDD.

[110] I. Mani and G. Wilson. Robust temporal processing of news. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pages 69–76, 2000.

[111] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, R. Hoberman, and D. Caputo. Topic-based novelty detection: 1999 summer workshop at CLSP, final report, 1999.

[112] R. Swan and J. Allan. Automatic generation of overview timelines. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 49–56, 2000.

[113] PeRSSonal's core functionality evaluation: Enhancing text labeling through personalized summaries. Data and Knowledge Engineering Journal, Elsevier Science, 2007, C. Bouras, V. Poulopoulos, V. Tsogkas, 2007, (to appear)

[114] Efficient Summarization Based On Categorized Keywords. The 2007 International Conference on Data Mining (DMIN07), Las Vegas, Nevada, USA, C. Bouras, V. Poulopoulos, V. Tsogkas, 25 - 28 June 2007

[115] Personalizing text summarization based on sentence weighting. IADIS European First International Conference Data Mining (ECDM 2007), Lisbon, Portugal, C. Bouras, V. Poulopoulos, V. Tsogkas, 3 - 8 July 2007

[116] The importance of the difference in text types to keyword extraction: Evaluating a mechanism. 7th International Conference on Internet Computing 2006 (ICOMP 2006), Las Vegas, Nevada, USA, C. Bouras, C. Dimitriou, V. Poulopoulos, V. Tsogkas, 26 - 29 June 2006, pp. 43 - 49

[117] Scalability of text classification. 2nd International Conference on Web Information Systems and Technologies (WEBIST 2006), Setubal, Portugal, I. Antonellis, C. Bouras, V. Poulopoulos, A. Zouzias, 19 - 22 April 2006, pp. 408 - 413

[118] Personalized News Categorization through Scalable Text Classification. The Eight Asia Pacific Web Conference (APWeb – 06), Harbin, China, I. Antonellis, C. Bouras, V. Poulopoulos, 16 - 18 January 2006, pp. 391 - 401

ΔΙΚΤΥΑΚΟΙ ΤΟΠΟΙ

[119] Μηχανή αναζήτησης Google. http://www.google.com

[120] Μηχανή αναζήτησης Altavista. http://www.altavista.com

[121] Portal Yahoo! http://www.yahoo.com

[122] Διεθνές ειδησεογραφικό πρακτορείο CNN. http://www.cnn.com

[123] Διεθνές ειδησεογραφικό πρακτορείο BBC. http://www.bbc.co.uk

[124] Διεθνές ειδησεογραφικό πρακτορείο Reuters. http://www.reuters.com

[125] Διεθνές ειδησεογραφικό πρακτορείο FoxNews http://www.foxnews.com/

[126] MySQL, Βάση Δεδομένων ανοιχτού κώδικα. http://www.mysql.com

[127] PostgreSQL, Βάση Δεδομένων ανοιχτού κώδικα. http://www.postrgresql.org

[128] Ελεύθερη εγκυκλοπαίδεια WikiPedia. Θέμα C++ (C Plus Plus). http://en.wikipedia.org/wiki/C_Plus_Plus

[129] Το χρονικό της Java. http://ils.unc.edu/blaze/java/javahist.html.

[130] Ελεύθερη εγκυκλοπαίδεια WikiPedia. Θέμα Perl. http://en.wikipedia.org/wiki/Perl

[131] Επίσημος Δικτυακός τόπος της PHP. http://www.php.net/

[132] Ελεύθερη εγκυκλοπαίδεια WikiPedia. Θέμα JSP. http://en.wikipedia.org/wiki/JavaServer_Pages

[133] GNU Wget – GNU Project – Free SoftWare Foundation http://www.gnu.org/software/wget/

[134] Heritrix - Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project. http://crawler.archive.org/

[135] ht://Dig – Internet search engine software. http://www.htdig.org/

[136] HTTrack Website Copier - Offline Browser. http://www.httrack.com/

[137] Larbin web crawler. http://larbin.sourceforge.net/index-eng.html

[138] Methabot web crawler. http://bithack.se/methabot/

[139] Nutch open source web search engine. http://lucene.apache.org/nutch/

[140] WebSPHINX: A Personal, Customizable Web Crawler. http://www.cs.cmu.edu/~rcm/websphinx/

[141] Web Information Retrieval Environment (WIRE). http://www.cwr.cl/projects/WIRE/

[142] http://www.passport.net

[143] http://www.amazon.com

[144] http://www.gartner.com/

[145] Copernic Summarizer. http://www.copernic.com/en/products/summarizer/index.html

[146] MS Office Autosummarize. http://office.microsoft.com/en-us/word/HP052334521033.aspx