Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Α Τ Ρ Ω Ν
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ & ΠΛΗΡΟΦΟΡΙΚΗΣ

Δ Ι Δ Α Κ Τ Ο Ρ Ι Κ Η Δ Ι Α Τ Ρ Ι Β Η

Μηχανισμοι και Τεχνικες Διαχειρισης Επεξεργασιας, Αναλυσης, Κατηγοριοποιησης, Εξαγωγης Περιληψης και Προσωποποιησης Συχνα Ανανεωσιμων Δεδομενων του Παγκοσμιου Ιστου για Παρουσιαση σε Σταθερες και Κινητες Συσκευες

Βασιλειος Ν. Πουλοπουλος
Μηχανικος Ηλεκτρονικων Υπολογιστων και Πληροφορικης, MSc
Α.Μ.: 442

Πατρα 2010

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Μηχανισμοι και Τεχνικες Διαχειρισης Επεξεργασιας, Αναλυσης, Κατηγοριοποιησης, Εξαγωγης Περιληψης και Προσωποποιησης Συχνα Ανανεωσιμων Δεδομενων του Παγκοσμιου Ιστου για Παρουσιαση σε Σταθερες και Κινητες Συσκευες

Επιβλεπων Καθηγητης:

Χρήστος Μπούρας, Καθηγητής

Τριμελης Επιτροπη:

Ιωάννης Γαροφαλάκης, Αναπληρωτής Καθηγητής

Χρήστος Μπούρας, Καθηγητής

Δημήτριος Χριστοδουλάκης, Καθηγητής

Επταμελης Επιτροπη:

Ευστράτιος Γαλλόπουλος, Καθηγητής

Ιωάννης Γαροφαλάκης, Αναπληρωτής Καθηγητής

Χρήστος Μακρής, Επίκουρος Καθηγητής

Βασίλης Μεγαλοοικονόμου, Αναπληρωτής Καθηγητής

Χρήστος Μπούρας, Καθηγητής

Αθανάσιος Τσακαλίδης, Καθηγητής

Δημήτριος Χριστοδουλάκης, Καθηγητής

στη Νινέτα μου
στο Λεωνίδα και την Έρη
στο Νίκο και την Ελένη
τους στυλοβάτες της ζωης μου

ΠΕΡΙΕΧΟΜΕΝΑ

ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

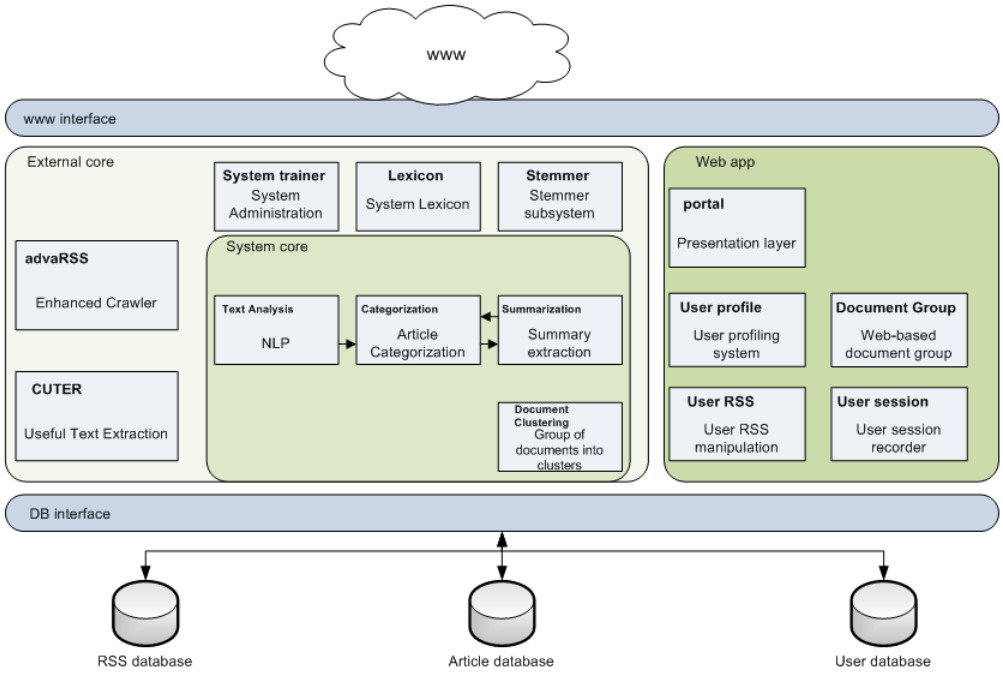
EXECUTIVE SUMMARY

We live an era of technology advances and huge technological steps where the Internet becomes a basic place of demonstration of the technology trends. Nevertheless, the way of operation and construction of the WWW is extremely uneven and this results in dead-ends when the users are trying to locate information. Besides the existence of billions of domains leads to difficulties in difficulties in recording all this information. The research that we are doing, is focused on websites that are sources of information and specifically news portals and informational blogs. A simple search on the Internet led to more than 40 large scale press agencies in America. This means that when trying to search for information and more specifically a news article in all its existences somebody has to visit all the websites. This problem, or at least this tedious task is of major concern of the research community. Many solutions were proposed in order to overcome the aforementioned issues with usage of RSS feeds or personalized microsites, or even analytical search applications. In any occasion there are many disadvantages that lead the user to a dead-end again. The RSS feeds do not filter information and they feed the user's RSS readers with large amounts of information that most of it is not of the user's concern. For example, a simple addition of 2 rss feeds from large Greek portals led to receipt of more that 1000 news articles within a day! On the other side, the usage of microsites that many websites support is a solution if and only if the user visits every single website and of course have and maintain an account to each one of them. The search engines are an alternative but lately, due to the expansion of the WWW, the results to simple queries are often million or the first results retrieved are outdated. Finally, the websites of the major news agencies are not directly constructed to offer extensive searching facilities and thus they usually offer search results through support of a large well-known search engine (eg. Google). According to the aforementioned the research that we are conducting is furthermore focused on the study of techniques and mechanisms that try to give a solution to the everyday issue of being informed about news and having a spherical opinion about an issue. The idea is simple and lies on the problem of the Internet: instead of letting the user do all the search of the news and information that meet their needs we collect all the information

and present them directly to the user, presenting only the information that meet their profile. This sounds pretty simple and logical, but the implementation we have to think of a number of prerequisites. The constraints are: the users of the Internet speak different languages and they want to see the news in their mother language and the users want access to the information from everywhere. This implies that we need a mechanism that would collect news articles from many – if not all – news agencies worldwide so that everybody can be informed. The news articles that we collect should be furthermore analyzed before presented to the users. In parallel we need to apply text pre-processing techniques, categorization and automatic summarization so that the news articles can be presented back to the user in a personalized manner. Finally, the mechanism is able to construct and maintain a user profile and present only articles that meet the profile of the user and not all the articles collected by the system. As it is obvious this is not a simple procedure. Substantially it a multilevel modular mechanism that implements and uses advanced algorithm on every level in order to achieve the required result. We are referring to eight different mechanisms that lead to the desired result. The systems are:

1. Retrieve news and articles from the Internet –advaRSS system
2. HTML page analysis and useful text extraction – CUTER system.
3. Preprocess and Natural Language Processing in order to extract keywords.
4. Categorization subsystem in order to construct ontologies that assigns texts to categories
5. Article Grouping mechanism (web application level)
6. Automatic Text Summarization
7. Web based User Personalization Mechanism
8. Application based User Personalization Mechanism

The subsystems and system architecture is presented in figure 1: The procedure of fetching articles and news from the WWW is a procedure that includes algorithms that fetch data of the large database that is called internet. In this research we have included algorithms for instant retrieval of articles and the mechanism has furthermore mechanism for fetching HTML pages that include news articles. As a next step and provided that we own HTML pages with articles we have procedures for efficient useful text extraction. The HTML pages include the body of the article and information that are disrelated to the article like advertisements. Our mechanism introduces algorithms and systems for extraction of the original body of the text out of the aforementioned pages and omitting any irrelevant information. As a furthermore procedure of the same mechanism we try and extract multimedia related to the article. The aforementioned mechanism are communicating directly with the Internet. The next procedures



Σχήμα 1: peRSSonal Architecture

are that core procedures that are the lexical analysis, the keyword extraction, the categorization and summarization mechanisms. The lexical analysis is based on basic pre-processing techniques that include among others: locating the article's language, spelling check, deletion of stopwords and common words, tagging of important words and finally extracting the stem of the words in order to locate the keywords. For each of the keyword we store information about its grammar, its stem, the frequency in the text, the sentence in which it appears and generally any other semantic data. The categorization procedure follows and its scope is to assign a probability with which each article belongs to one of the system's seven categories. For this procedure we study multicategorization algorithms that are based on keywords extracted from the text and techniques of interaction of the categorization mechanism with summarized text or personalized input. In parallel we study algorithms that are possibly valuable for constructing an ontology tree for the system categories in order to achieve subcategory construction, basically based on the users' profiles. Another mechanism that is studied is the summarization procedure. For the summarization procedure we study algorithms and we construct a mechanism that is able to create a summary based on sentences of the original text. The summarization is based on lexical analysis, categorization and personalization procedures while the system is able to create summaries per user. Each user is considered to have a different profile and thus the users may want to be faced with different summaries. Finally, as an outcome of all the procedures we

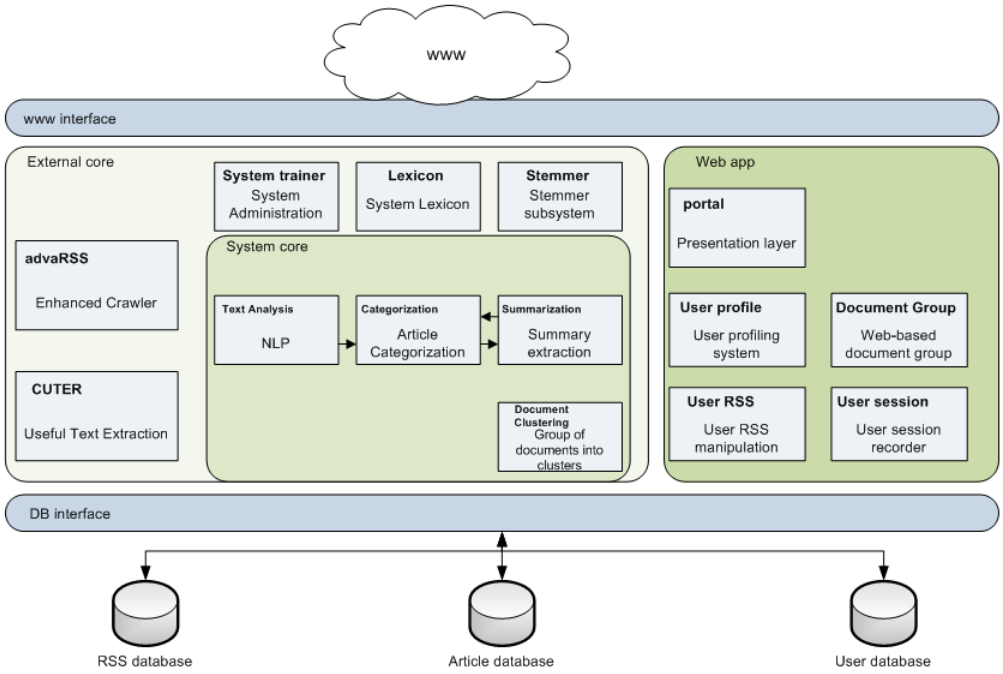
study and implement mechanism for presenting information. The presentation mechanisms are accessible through the internet and through desktop applications. Finally, we are experimenting with plig-ins and add-ons that can be possibly be used into browsers in order to overcome the procedure of visiting the website. The website applies fully personalized procedures, constructs and maintains a fully dynamic, always changing profile that represent the user's characteristics. All the aforementioned procedures and mechanisms are analyzed thoroughly and their algorithms are optimized in order to achieve maximum quality in results. The techniques that were studied led to a multilingual, multilevel, interacting, self trained and maintained system that is able to reconstruct web information to personalized human readable easily accessible data. Finally, we have conducted extensive experimental evaluation to each of the systems in order to assure smooth execution and efficient results.

ΕΠΙΤΕΛΙΚΗ ΣΥΝΟΨΗ

Ζούμε μία εποχή τεχνολογικών εξελίξεων και τεχνολογικών αλμάτων με το Διαδίκτυο να γίνεται ένας από τους βασικότερους εκφραστές των νέων τεχνολογικών τάσεων. Ωστόσο, ο τρόπος λειτουργίας του και δόμησής του παρουσιάζει εξαιρετικά μεγάλη ανομοιογένεια με αποτέλεσμα οι χρήστες να βρίσκονται συχνά μπροστά από αδιέξοδο στην προσπάθεια αναζήτησης πληροφορίας. Άλλωστε η ύπαρξη εκατομμυρίων domains οδηγεί σε δυσκολίες κατά την αναζήτηση πληροφορίας. Η έρευνα που πραγματοποιείται επικεντρώνεται στους δικτυακούς τόπους που αποτελούν πηγές ενημέρωσης και πιο συγκεκριμένα στα ειδησεογραφικά πρακτορεία ειδήσεων, αλλά και στα blogs. Μία απλή αναζήτηση αποκάλυψε περισσότερους από 40 δικτυακούς τόπους από μεγάλα ειδησεογραφικά πρακτορεία στην Αμερική. Αυτό σημαίνει πως στην προσπάθεια αναζήτησης μίας είδησης και δη, όλων των πτυχών της, κάποιος θα πρέπει να επισκεφθεί αν όχι όλους, τους περισσότερους από αυτούς τους δικτυακούς τόπους για να εντοπίσει στοιχεία για το θέμα που τον ενδιαφέρει. Σε αυτό το «πρόβλημα» ή έστω σε αυτή την επίπονη διαδικασία, έχει γίνει προσπάθεια να δοθούν λύσεις μέσα από τη χρήση των καναλιών επικοινωνίας RSS και μέσα από προσωποποιημένους δικτυακούς τόπους που διαθέτουν τα μεγάλα ειδησεογραφικά πρακτορεία ή ακόμα και από τους μηχανισμούς αναζήτησης που αυτοί διαθέτουν. Σε κάθε περίπτωση όμως, υπάρχουν σημαντικά μειονεκτήματα που συχνά οδηγούν και πάλι το χρήστη σε αδιέξοδο. Τα κανάλια επικοινωνίας δε φιλτράρουν πληροφορίες, τροφοδοτώντας τους RSS readers των χρηστών με πληθώρα πληροφοριών που δεν αφορούν τους χρήστες ή ακόμα είναι ενοχλητικές για αυτούς. Για παράδειγμα η προσθήκη δύο (2) μόνον καναλιών από Ελληνικά μεγάλα ειδησεογραφικά portals μας οδήγησε στη λήψη περισσότερων από 1000 ειδήσεων καθημερινά. Από την άλλη, η χρήση των microsites που έχουν οι δικτυακοί τόποι επιβάλει στους χρήστες την επίσκεψη σε όλους τους δικτυακούς τόπους που τους ενδιαφέρουν. Όσον αφορά στη χρήση των μηχανών αναζήτησης, ακόμα και οι πιο μεγάλες από αυτές συχνά επιστρέφουν εκατομμύρια αποτελέσματα στα ερωτήματα των χρηστών ή πληροφορίες που δεν είναι επικαιροποιημένες. Τέλος, επειδή οι δικτυακοί τόποι των ειδησεογραφικών πρακτορείων δεν έχουν κατασκευαστεί για να προσφέρουν εκτενείς υπηρεσίες αναζήτησης ειδήσεων, είναι

συχνό το φαινόμενο είτε να μην προσφέρουν καθόλου υπηρεσία αναζήτησης, είτε η υπηρεσία που προσφέρουν να μη μπορεί να απαντήσει με δομημένα αποτελέσματα και αντί να βοηθά τους χρήστες να εντοπίσουν την πληροφορία που αναζητούν, να τους αποπροσανατολίζει. Βάσει όλων των παραπάνω, η έρευνα που πραγματοποιείται εστιάζει στη μελέτη μηχανισμών και τεχνικών που θα μπορούν να δώσουν λύση στο πρόβλημα της αναζήτησης ειδήσεων και στο πρόβλημα της καθημερινής σφαιρικής ενημέρωσης για την ειδησεογραφία που επιθυμούν πολλοί από τους χρήστες. Η ιδέα είναι απλή και βασίζεται στο πρόβλημα το οποίο υφίσταται στο χώρο του διαδικτύου: αντί να τοποθετούμε το χρήστη στη διαδικασία ανεύρεσης των ειδήσεων που τον απασχολούν, συλλέγουμε τις ειδήσεις και τις εμφανίζουμε με τον τρόπο που επιθυμεί, παρουσιάζοντας μόνο τις ειδήσεις που ταιριάζουν στο προφίλ του. Η ιδέα ακούγεται απλή και λογική, ωστόσο για να πραγματοποιηθεί αυτό λαμβάνουμε υπόψη μας συγκεκριμένους παράγοντες. Οι παράγοντες αυτοί είναι οι εξής: οι χρήστες του διαδικτύου μιλούν διαφορετικές γλώσσες και προφανώς ενδιαφέρονται να βλέπουν τις ειδήσεις που τους αφορούν στη γλώσσα τους. Έτσι, υπάρχει κάποιος μηχανισμός που θα συλλέγει όλες τις ειδήσεις από πολλά – αν όχι όλα – τα ειδησεογραφικά πρακτορεία παγκοσμίως για να είναι εφικτή η ενημέρωση όλων των χρηστών του συστήματος παγκοσμίως. Σε αυτές τις ειδήσεις που έχουν συγκεντρωθεί, υλοποιούνται τεχνικές για τον εντοπισμό των ταυτόσημων ειδήσεων προκειμένου για να μην εμφανίζεται πολλές φορές η ίδια είδηση. Παράλληλα, πραγματοποιούνται διεργασίες προ-επεξεργασίας κειμένου, κατηγοριοποίησης και αυτόματης εξαγωγής περίληψης προκειμένου να μην αναπαράγονται απλώς οι ειδήσεις όπως αυτές συλλέγονται (η βασική ιδέα του web clipping) και τέλος ο μηχανισμός είναι σε θέση να αναγνωρίζει το προφίλ του χρήστη και να του εμφανίζει αποκλειστικά και μόνο τα άρθρα που τον ενδιαφέρουν και όχι όλα τα άρθρα που συλλέγονται από το σύστημα. Παρατηρούμε, λοιπόν, πως δεν πρόκειται για μία απλή διαδικασία. Ουσιαστικά πρόκειται για ένα μηχανισμό ο οποίος λειτουργεί σε πολλαπλά επίπεδα, υλοποιεί και εφαρμόζει αναλυτικούς αλγορίθμους σε κάθε επίπεδο, προκειμένου για να επιτευχθεί το επιθυμητό αποτέλεσμα. Ουσιαστικά μιλούμε για 8 διαφορετικούς μηχανισμούς οι οποίοι οδηγούν στο επιθυμητό αποτέλεσμα. Πρόκειται για τους:

1. Ανάκτηση Άρθρων και Ειδήσεων από το Διαδίκτυο – σύστημα advaRSS
2. Ανάλυση HTML σελίδων για εξαγωγή του σώματος του άρθρου – σύστημα CUTER
3. Προεπεξεργασία και Ανάλυση κειμένου (στη φυσική του γλώσσα) για εξαγωγή λέξεων κλειδιών. Υποστηρίζονται Αγγλικά και τα Ελληνικά βρίσκονται στο στάδιο ενεργοποίησης
4. Εφαρμογή αλγορίθμων κατηγοριοποίησης για τη δημιουργία οντολογιών που θα κατατάσσει τα εισερχόμενα στο σύστημα κείμενα
5. Μηχανισμός εύρεσης ταυτόσημων άρθρων (web application level)
6. Εφαρμογή αλγορίθμων αυτόματης εξαγωγής περίληψης



Σχήμα 2: Αρχιτεκτονική του peRSSonal

7. Μηχανισμός Προσωποποίησης στο χρήστη μέσα από το Διαδίκτυο
8. Μηχανισμός Προσωποποίησης στο χρήστη μέσα από stand alone εφαρμογές

Η αρχιτεκτονική και η διάταξη όλων των μηχανισμών εμφανίζεται στο σχήμα 2:

Όσον αφορά στην ανάκτηση άρθρων και ειδήσεων από το διαδίκτυο, είναι μία διαδικασία που περιλαμβάνει αλγορίθμους που αφορούν λήψη δεδομένων από την τεράστια βάση δεδομένων που αποτελεί το διαδίκτυο. Στο μηχανισμό που μελετάμε έχουν ενσωματωθεί στοιχεία προκειμένου για να επιτυγχάνεται άμεση ανάκτηση άρθρων από κανάλια επικοινωνίας του διαδικτύου. Ο συγκεκριμένος μηχανισμός έχει σαν κύριο σκοπό την εφαρμογή τεχνικών για συλλογή των HTML σελίδων που περιέχουν τα νέα άρθρα. Στην επόμενη διαδικασία του μηχανισμού μελετούμε αλγορίθμους που πραγματοποιούν αποδοτική εξαγωγή κειμένου και πολυμέσων από τις HTML σελίδες που έχει συλλέξει ο παραπάνω μηχανισμός. Συγκεκριμένα, γνωρίζουμε πως οι HTML σελίδες που περιέχουν ένα άρθρο, περιέχουν συν τοις άλλοις και πληροφορίες που δεν αφορούν το άρθρο όπως διαφημίσεις, μενού πλοήγησης κλπ. Στο μηχανισμό εξαγωγής χρήσιμου κειμένου μελετούμε τεχνικές και αλγορίθμους που επιτυγχάνουν την απομόνωση του κειμένου και των εικόνων του άρθρου από τις HTML σελίδες, χωρίς να περιλαμβάνονται στο κυρίως σώμα πληροφορίες της HTML σελίδας που δεν είναι σχετικές με το άρθρο. Παράλληλα, σε επόμενο επίπεδο του μηχανισμού πραγματοποιείται περαιτέρω ανάλυση των άρθρων του συστήματος,

με σκοπό να εξαλειφθούν εντελώς άρθρα που περιέχουν ανούσια πληροφορία. Οι μηχανισμοί και οι τεχνικές που αναφέρθηκαν λειτουργούν σε επίπεδο επικοινωνίας με το διαδίκτυο. Ακολουθούν οι διαδικασίες πυρήνα. Η αρχική διαδικασία πυρήνα περιλαμβάνει τη λεξικολογική ανάλυση του κειμένου ώστε να εξαχθούν οι λέξεις κλειδιά. Η λεξικολογική ανάλυση βασίζεται σε βασικές διαδικασίες προ-επεξεργασίας κειμένου που περιλαμβάνουν: εντοπισμό της γλώσσας του άρθρου, ορθογραφικό έλεγχο του άρθρου, διαγραφή όλων των λέξεων που είναι κοινότυπες, σηματοδότηση των ουσιαστικών που σύμφωνα με μελέτες περιέχουν όλη τη χρήσιμη πληροφορία του κειμένου και εύρεση της ρίζας των λέξεων για να οδηγηθούμε τελικά στην εξαγωγή δεδομένων που αφορούν τις λέξεις κλειδιά του κειμένου. Η διαδικασία κατηγοριοποίησης είναι μία διαδικασία που αναθέτει έναν γενικό αφαιρετικό τίτλο σε ένα άρθρο προκειμένου να το εντάξει σε μία προκαθορισμένη οντολογία. Για τη διαδικασία κατηγοριοποίησης μελετούνται αλγόριθμοι πολλαπλής κατηγοριοποίησης κειμένου που βασίζονται τόσο στις λέξεις κλειδιά που έχουν εξαχθεί από το κείμενο, σε τεχνικές αλληλεπίδρασης της διαδικασίας κατηγοριοποίησης με τις διαδικασίες αυτόματης εξαγωγής περίληψης όσο και με τη διαδικασία προσωποποίησης του χρήστη. Παράλληλα εξετάζονται αλγόριθμοι που μπορούν να δομήσουν αυτόματα το δέντρο της οντολογίας που αφορά στις κατηγορίες του μηχανισμού, ούτως ώστε να είναι εφικτή η αυτόματη δημιουργία υποκατηγοριών, πέραν των βασικών κατηγοριών του συστήματος και οι οποίες βασίζονται κυρίως στα προφίλ των χρηστών. Ένας ακόμα μηχανισμός ο οποίος μελετάται στο πλαίσιο της διδακτορικής εργασίας είναι ο μηχανισμός εξαγωγής περίληψης. Για την εξαγωγή περίληψης κειμένου πραγματοποιείται μελέτη υπαρχόντων αλγορίθμων και κατασκευή ενός μηχανισμού που είναι σε θέση να επιτύχει περίληψη που βασίζεται στους παράγοντες κατηγοριοποίησης κειμένου, στη λεξικολογική ανάλυση του κειμένου και ίσως το πιο σημαντικό, είναι σε θέση να δημιουργεί προσωποποιημένες περιλήψεις για τους χρήστες του συστήματος δεδομένου ότι κάθε χρήστης είναι μία ξεχωριστή προσωπικότητα και συνεπώς έχει άλλα ενδιαφέροντα για το περιεχόμενο κάθε κειμένου. Τέλος, σαν αποτέλεσμα όλων των παραπάνω επεξεργασιών που έχουν γίνει πάνω στα άρθρα που έχουν συγκεντρωθεί από το διαδίκτυο, μελετώνται και αναπτύσσονται μηχανισμοί προβολής των άρθρων στους χρήστες. Οι μηχανισμοί προβολής που είναι προσπελάσιμοι μέσα από το διαδίκτυο (web portal), αλλά και μέσω εφαρμογών που μπορούν να λειτουργήσουν σαν αυτόνομες εφαρμογές στα πιο κοινά λειτουργικά συστήματα (personalized RSS readers). Τέλος, ερευνώνται και εφαρμογές που μπορούν να χρησιμοποιηθούν σαν plug-ins ή add-ons σε προγράμματα και δικτυακούς τόπους για την προσθήκη προσωποποιημένων στοιχείων σε αυτούς. Η προσωποποίηση στο χρήστη βασίζεται σε ανάλυση όλων των ενεργειών που πραγματοποιεί ο χρήστης καθώς χρησιμοποιεί τις προαναφερθείσες εφαρμογές, αλλά και στη διαμόρφωση ενός προφίλ που θα αντιπροσωπεύει το χαρακτήρα του. Σε όλες τις παραπάνω διαδικασίες και εφαρμογές, πραγματοποιείται εκτενής ανάλυση και μελέτη αλγορίθμων προκειμένου για να επιτευχθεί βέλτιστη λειτουργία κάθε μηχανισμού αλλά και βέλτιστη λειτουργία του συνόλου των μηχανισμών. Οι τεχνικές που μελετήθηκαν κατέληξαν σε ένα μηχανισμό που θα είναι ένα πολύγλωσσο, πολυεπίπεδο, διαδραστικό σύστημα και θα έχει

σαν σκοπό τη βελτιστοποίηση της λειτουργίας του διαδικτύου και πιο συγκεκριμένα την αναδόμηση της πληροφορίας, ώστε να παρέχει στους χρήστες πληροφορίες και δεδομένα με συνοχή και βασισμένα σε μία ενιαία οντολογία. Τέλος, πραγματοποιούνται πειραματικές διαδικασίες για την εξαγωγή αποτελεσμάτων για κάθε υποσύστημα αλλά και στο μηχανισμό σαν σύνολο, ενώ ο μηχανισμός δοκιμάζεται εκτενώς και σε χρήστες που ενδιαφέρονται να τον χρησιμοποιήσουν.

Δημοσιεύσεις σε Διεθνή Περιοδικα

[1] C. Bouras, V. Poulopoulos, V. Tsogkas. “Adaptation of RSS feeds based on the user profile and on the end device”, Journal of Network and Computer Applications, 2010, (to appear)

Περίληψη: Την τελευταία δεκαετία, η ανάπτυξη της τεχνολογίας σε συνδυασμό με την ευκολία πρόσβασης στην πληροφορία έχουν αλλάξει αυτό που ονομάζουμε Internet. Το Internet είναι ένα μέσο εύρεσης χρήσιμης πληροφορίας και συχνά άρθρων. Παράλληλα, ολοένα και περισσότεροι άνθρωποι θέλουν να χρησιμοποιήσουν τις κινητές συσκευές τους προκειμένου να μπορούν να ενημερώνονται από το διαδίκτυο. Η παραπάνω κατάσταση δημιουργεί ένα σημαντικό πρόβλημα, το οποίο είναι η εύρεση χρήσιμης πληροφορίας από τους χρήστες του διαδικτύου σε καθημερινή βάση και πιο συγκεκριμένα στο χώρο που είναι διαθέσιμος σε μία συσκευή μικρής οθόνης. Στη εργασία μας προτείνουμε ένα μηχανισμό ο οποίο χρησιμοποιώντας RSS feeds είναι σε θέση να προσωποποιήσει την πληροφορία βάσει των αναγκών των χρηστών και των δυνατοτήτων που έχουν οι συσκευές τους, προκειμένου για να τους παρουσιάσει μόνο ένα κομμάτι της πληροφορίας, το οποίο είναι δυνητικά αυτό που τους ενδιαφέρει. Αυτός ο μηχανισμός μοιάζει να είναι η απόλυτη λύση για κάθε μηχανισμό ανάκτησης κειμένου. Στο πλαίσιο δημιουργίας αυτού του συστήματος, κατασκευάσαμε το peRSSonal, έναν μηχανισμό που μπορεί να δημιουργήσει προσωποποιημένα προ-κατηγοριοποιημένα, δυναμικά κατασκευασμένα RSS feeds ώστε να μπορούν να παρουσιαστούν σε συσκευές μικρού μεγέθους. Το σύστημα είναι βασισμένο σε αλγορίθμους που περιλαμβάνουν τον ίδιο το χρήστη σε όλες τις παραπάνω διαδικασίες.

[2] I. Antonellis, C. Bouras, V. Poulopoulos. "Scalable Text Classification as a tool for Personalization", Computer Systems Science & Engineering, CRL Publishing Ltd., 2009 Vol. 6, pp. 51 – 60

Περίληψη: Αναλύουμε θέματα κλιμάκωσης αναφορικά με το πρόβλημα κατηγοριοποίησης κειμένου, όπου, χρησιμοποιώντας προ-κατηγοριοποιημένα κείμενα, προσπαθούμε να χτίσουμε τρόπους κατηγοριοποίησης ώστε να είναι εφικτή η κατηγοριοποίηση ενός κειμένου σε πολλαπλές «ταμπέλες». Μία νέα τάξη προβλημάτων κατηγοριοποίησης που ονομάζονται κλιμακωτά παρουσιάζεται με εφαρμογές της στην εξόρυξη γνώσης από το διαδίκτυο. Η μέθοδος αυτή χρησιμοποιεί προκατηγοριοποιημένη πληροφορία προκειμένου για να αυξήσει την αποτελεσματικότητα μελλοντικών διαδικασιών κατηγοριοποίησης και να εντοπίσει αλλαγές στην αναπαράσταση διαφορετικών θεμάτων. Επιπλέον, επιτρέπεται ο ορισμός διαφορετικών κλάσεων και έτσι μπορούμε να επιτύχουμε κατηγοριοποίηση ανά χρήστη. Αυτή η μέθοδος αποτελεί καινοτόμα μεθοδολογία για τη δημιουργία προσωποποιημένων εφαρμογών. Αυτό οφείλεται στο γεγονός ότι ο χρήστης γίνεται κομμάτι της συνολικής διαδικασίας κατηγοριοποίησης. Ερευνούμε λύσεις για την κλιμάκωση κατηγοριοποίησης κειμένου και παρουσιάζουμε έναν αλγόριθμο που ενσωματώνει μία νέα τεχνική ανάλυσης κειμένου που αποδομεί έγγραφα σε

αναπαράσταση πίνακα των προτάσεών τους, ανάλογα με τους χρήστες που τα προσπελαύνουν.

[3] C. Bouras, V. Poulopoulos, V. Tsogkas. PeRSSonal's core functionality evaluation: Enhancing text labeling through personalized summaries, Data and Knowledge Engineering Journal, Elsevier Science, 2008, Vol. 64, Issue 1, pp. 330 - 345

Περίληψη: 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. 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 για να ενισχυθεί η κατηγοριοποίηση με τη χρήση προσωποποιημένων περιλήψεων.

Δημοσιεύσεις σε Διεθνή Συνέδρια

[1] C. Bouras, V. Poulopoulos. "Dynamic User Context Web Personalization in Meta - Portals". The Fifteenth IEEE Symposium on Computers and Communications (ISCC'10), Riccione, Italy, June 22 - 25 2010, pp. 925 - 930

ΕΛΑΒΕ ΒΡΑΒΕΙΟ ΚΑΛΥΤΕΡΟΥ PAPER

Περίληψη: Παρουσιάζουμε το μηχανισμό δυναμικής προσωποποίησης χρηστών σε επίπεδο διαδικτύου καθώς και ένα μηχανισμό για ομαδοποίηση κειμένων για το μηχανισμό peRSSonal, ένα ολοκληρωμένο σύστημα συλλογής άρθρων από δικτυακές ενημερωτικές πύλες και blogs σε όλο τον κόσμο. Ένα meta-portal, όπως είναι το peRSSonal, είναι ένας πληροφοριακός κόμβος που συγκεντρώνει πληροφορία από πολλές διαφορετικές πηγές και την παρουσιάζει πίσω στους χρήστες προσαρμοσμένες στις ανάγκες τους. Ο μηχανισμός μας βασίζεται στη δημιουργία και διατήρηση δυναμικού προφίλ χρήστη σύμφωνα με τις επιλογές που κάνει ο χρήστης κατά τη διάρκεια της εγγραφής και περιαγωγής στο δικτυακό τόπο. Θεωρώντας ότι έχουμε όλη την προαπαιτούμενη πληροφορία παρουσιάζουμε έναν πρωτοποριακό μηχανισμό που μπορεί να δημιουργήσει και να διατηρήσει προφίλ χρήστη σε περιβάλλον διαδικτύου. Παράλληλα παρουσιάζουμε το μηχανισμό ομαδοποίησης άρθρων πραγματικού χρόνου που βασίζεται στο προφίλ των χρηστών.

[2] G. Adam, C. Bouras, V. Poulopoulos. “Efficient extraction of news articles based on RSS crawling”. International Conference on Machine and Web Intelligence (ICMWI2010), USTHB University, Algiers, October 3- 5 2010 (to appear).

(INVITED PAPER)

Περίληψη: Η εξάπλωση του παγκόσμιου ιστού έχει οδηγήσει σε μία κατάσταση όπου πληθώρα χρηστών του διαδικτύου πρέπει να αντιμετωπίσουν το πρόβλημα εύρεσης χρήσιμης πληροφορίας. Είναι γνωστό ότι καθημερινά δημιουργούνται εκατοντάδες χιλιάδες νέες σελίδες με άρθρα ή δημοσιεύσεις σε blog. Το πρόβλημα που υπάρχει με τον ολοένα αυξανόμενο, σε μέγεθος δεδομένων, παγκόσμιο ιστό είναι πως ακόμα και οι έμπειροι χρήστες αδυνατούν να βρουν χρήσιμες πληροφορίες ή ακόμα καλύτερα να εντοπίσουν πληροφορίες που ταιριάζουν στο προφίλ τους. Πολλοί μηχανισμοί έχουν προταθεί και αρκετοί είναι διαθέσιμοι στους διαδικτυακούς χρήστες. Σε αυτή την εργασία παρουσιάζουμε ένα μηχανισμό που ανακτά σελίδες από το διαδίκτυο που περιέχουν άρθρα ή ειδήσεις με αποδοτικό τρόπο χρησιμοποιώντας αλγορίθμους προσαρμογής στο ρυθμό ανανέωσης μίας σελίδας.

[3] G. Adam, K. Asimakis, C. Bouras, V. Poulopoulos. "An Efficient Mechanism for Stemming and Tagging: the Case of Greek Language". Advanced Knowledge - based Systems, Invited Session of the 14th International Conference on Knowledge - based and Intelligent Information & Engineering Systems, Cardiff Wales, UK, September 2010 (to appear)

Περιληψη:Σε μία εποχη όπου, η αναζήτηση στο Διαδίκτυο είναι μία επίπονη διαδικασία, είναι προφανές ότι προτίστως οι μηχανές αναζητησης αλλα και άλλοι μηχανισμοί data mining πρέπει να εφοδιαστούν με χαρακτηριστικά, όπως NLP, προκειμένου να μπορούν να αναλύσουν και να αναγνωρίζουν τις ανάγκες των χρηστών. Παρουσιάζουμε ένα μηχανισμό για stemming και tagging για την Ελληνική γλώσσα. Ο μηχανισμός μας είναι κατασκευασμένος με τέτοιον τρόπο ώστε να μπορεί εύκολα να ενσωματωθεί σε υπάρχοντα συστήματα. Αναλύουμε τις διαδικασίες του μηχανισμού και μελετούμε την ακρίβεια του συστήματος συγκριτικά με ήδη υπάρχοντα συστήματα.

[4] G. Adam, C. Bouras, V. Poulopoulos. Image extraction from online text streams. The 2010 IEEE International Symposium on Mining and Web (MAW 2010), Perth, Australia, April 20 – 23 2010, pp. 609 - 613.
Περίληψη: Σε αυτή την εργασία παρουσιάζουμε ένα σύστημα ανάλυσης HTML σελίδων που περιέχουν άρθρα με σκοπό την εξαγωγή όσων εικόνων σχετίζονται με το σώμα του άρθρου. Ο προτεινόμενος μηχανισμός δεν εξαρτάται από τον τύπο και τη δομή της ιστοσελίδας και μπορεί να εφαρμοστεί σε κάθε σελίδα. Σαν σχετικές με το σώμα του άρθρου εικόνες, ορίζουμε αυτές οι οποίες αφορούν το ίδιο το άρθρο που περιέχεται σε μία σελίδα. Προκειμένου για να εξάγουμε την εικόνα ή τις εικόνες, αποδομούμε τη σελίδα στο DOM μοντέλο της και με κατάλληλους αλγορίθμους βαθμολόγησης κόμβων, αξιολογούμε ποιες εικόνες είναι πιθανά σχετικές με το άρθρο. Ο μηχανισμός αυτός εφαρμόζεται και αξιολογείται πάνω στο σύστημα peRSSonal.

[5] C. Bouras, V. Poulopoulos, G. Tsichritzis. “Trash Article Detection using Categorization Techniques”. IADIS International Conference Applied Computing, Rome, Italy, November 19 - 21 2009, pp. 51 - 58
Περίληψη: Παρουσιάζουμε τεχνικές με τις οποίες μπορούμε να εντοπίζουμε άρθρα που περιέχουν «ανούσια» πληροφορία, χρησιμοποιώντας αλγρίθμους κατηγοριοποίησης κειμένου. Η πληροφορία που υπάρχει στον παγκόσμιο ιστό είναι τεράστια με αποτέλεσμα να αποπροσανατολίζονται οι χρήστες στην προσπάθεια αναζήτησης χρήσιμης πληροφορίας. Προκειμένου για να αποφύγουμε την πληθώρα ανούσιων δεδομένων, εξετάζουμε μία σειρά από προτεινόμενες μεθοδολογίες. Το βασικό πρόβλημα είναι πως η πληροφορία εντοπίζεται και αποθηκεύεται από crawlers και συνήθως δεν υπάρχει μεγάλη ανάλυση για την εγκυρότητα της και αναφερόμαστε κυρίως σε κείμενα που περιέχουν κείμενο άσχετο με το περιεχόμενο της σελίδας (π.χ. διαφημίσεις). Αυτό μπορεί να οδηγήσει τους χρήστες να χάσουν την εμπιστοσύνη τους στα συστήματα παροχής δεδομένων. Σε αυτή την εργασία αναλύουμε τις εξειδικευμένες πληροφορίες που περιέχονται στα επονομαζόμενα trashy articles (ανούσια άρθρα) και πληροφορίες για το πώς θα μπορούσαμε να χρησιμοποιήσουμε υπάρχουσες τεχνικές για να εντοπίσουμε τα άρθρα αυτά. Αρχικά αναλύουμε τον προτεινόμενο αλγόριθμο και στη συνέχεια τον εφαρμόζουμε σε ένα ήδη υπάρχον σύστημα παρουσιάζοντας τον τρόπο ενίσχυσης.

[6] G. Adam, C. Bouras, V. Poulopoulos. Monitoring RSS feeds. International Conference on Knowledge Management and Knowledge Technologies (i-know09), Gratz, Austria, 2 - 4 September 2009

Περίληψη: Ένα από τα βασικότερα προβλήματα του Παγκόσμιου Ιστού σήμερα είναι η εύρεση χρήσιμης πληροφορίας. Για τη λύση αυτού του προβλήματος έχουν προταθεί πολλοί μηχανισμοί που βασίζονται σε crawlers που διατρέχουν το διαδίκτυο και κατεβάζουν σελίδες. Σε αυτή την εργασία παρουσιάσουμε το advaRSS, έναν crawler που έχει σαν σκοπό να αποτελέσει ένα σύστημα συλλογής δεδομένων από το διαδίκτυο σε πραγματικό χρόνο. Σε αντίθεση με τους κλασσικούς crawlers, ο advaRSS βασίζεται στην εξαγωγή άρθρων από το διαδίκτυο και εκμεταλλεύεται τα κανάλια επικοινωνίας που κυριαρχούν σε αυτό που ονομάζουμε web2.0. Τα άρθρα παράγονται τυχαία και σε καθημερινή βάση, με αποτέλεσμα ο μηχανισμός να πρέπει να συμπεριφέρεται διαφορετικά στον τρόπο ανανέωσης δεδομένων από κάθε διαφορετικό κανάλι επικοινωνίας.

[7] G. Adam, C. Bouras, V. Poulopoulos. Utilizing RSS feeds for crawling the Web. The Fourth International Conference on Internet and Web Applications and Services - ICIW 2009, 24 - 28 May 2009, pp. 211 - 216

Περίληψη: Παρουσιάζουμε το μηχανισμό advaRSS, έναν crawler που έχει δημιουργηθεί για να υποστηρίξει το σύστημα peRSSonal. Εστιάζουμε στη συλλογή άρθρων από ειδησεογραφικά portals και blogs. Η πρόκληση είναι μεγάλη, αν αναλογιστεί κανείς πως τα άρθρα δημοσιεύονται στο διαδίκτυο με τυχαίο τρόπο και έτσι, για να υπάρχει μία επικαιροποιημένη συλλογή, ο μηχανισμός θα πρέπει να λαμβάνει άρθρα πολύ συχνά. Για να μπορέσουμε να το επιτύχουμε αυτό, χρησιμοποιούμε αλγορίθμους καθώς και κανάλια επικοινωνίας (RSS feeds). Οι αλγόριθμοι επικαιροποίησης της συλλογής βασίζονται στο γεγονός ότι ο crawler είναι σε θέση να προσαρμοστεί στο ρυθμό ανανέωσης του κάθε feed.

[8] G. Adam, C. Bouras, V. Poulopoulos. CUTER: An Efficient Useful Text Extraction Mechanism. The 2009 IEEE International Symposium on Mining and Web (WAM09), Bradford, UK, , 26 - 29 May 2009, pp. 703 - 708

Περίληψη: Παρουσιάζουμε τον CUTER, ένα μηχανισμό επεξεργασίας HTML σελίδων, προκειμένου για να εξάγουμε χρήσιμο κείμενο από αυτές. Ο μηχανισμός εστιάζεις σε σελίδες που περιέχουν άρθρα και άρα μιλούμε κυρίως για ενημερωτικά portals και blogs. Προκειμένου για να μπορέσουμε να εξάγουμε το σώμα ενός άρθρου, αποδομούμε την HTML σελίδα σύμφωνα με το DOM μοντέλο και εφαρμόζουμε μία σειρά αλγορίθμων ώστε να εντοπίσουμε τα στοιχεία του μοντέλου που πιθανά περιέχουν χρήσιμο κείμενο. Ο CUTER είναι ένα υποσύστημα του

μηχανισμού peRSSonal, ένα εργαλείο που συλλέγει άρθρα από το διαδίκτυο και τα παρουσιάζει επεξεργασμένα και προσωποποιημένα στους χρήστες. Ο ρόλος του CUTER είναι να εμπλουτίζει το peRSSonal με το σώμα των άρθρων που εξάγονται, προκειμένου για να ακολουθήσουν ασφαλώς τα βήματα της επεξεργασίας κειμένου.

[9] C. Bouras, V. Poulopoulos, V. Tsogkas. Evaluating PeRSSonal: A Medium for Personalized Dynamically Created News Feeds. IADIS International Conference WWW/Internet Freiburg, Germany, 13 - 15 October 2008

Περίληψη: Σε αυτή την εργασία παρουσιάζουμε την αξιολόγηση του peRSSonal ενός συστήματος που παράγει προσωποποιημένα δυναμικά κατασκευασμένα RSS feeds και επικεντρώνεται σε συσκευές μικρού μεγέθους. Το σύστημα αυτό εντοπίζει άρθρα που δημοσιεύονται σε ενημερωτικά portals, τα επεξεργάζεται και τα παρουσιάζει στους χρήστες σύμφωνα με το προφίλ που αυτοί διαμορφώνουν. Σε μία εποχή που τα κανάλια επικοινωνίας (π.χ. RSS feeds) γίνονται κομμάτι της καθημερινότητας το peRSSonal μοιάζει να είναι η λύση που μπορεί να παρέχει εύκολη πρόσβαση στην πληροφορία για όλους. Το σύστημα βασίζεται σε αλγορίθμους που εμπλέκουν το χρήστη στις διαδικασίες και με αυτό τον τρόπο η πληροφορία που παράγεται είναι μοναδική και προσαρμοσμένη στον κάθε χρήστη

[10] C. Bouras, V. Poulopoulos, V. Tsogkas. Creating dynamic personalized RSS summaries. 8th Industrial Conference on Data Mining – ICDM 2008, , Leipzig, Germany, 16 - 18 July 2008, pp. 1 - 15

Περίληψη: Η αυτοματοποιημένη δημιουργία περίληψης υψηλής ποιότητας είναι δύσκολο, τόσο να δημιουργηθεί, όσο και να αξιολογηθεί, κυρίως λόγω του ότι αφενός τα κείμενα παρουσιάζουν μεγάλες διαφοροποιήσεις μεταξύ τους και αφετέρου κάθε άνθρωπος επιθυμεί κάτι διαφορετικό από κάθε κείμενο. Σε αυτή την εργασία, προτείνουμε ένα μοντέλο το οποίο χρησιμοποιεί τα RSS feeds, προκειμένου για να προσωποποιήσει στις ανάγκες ενός χρήστη, αλλά και στην τελική συσκευή χρήστη, ώστε να παρουσιάσει ένα κομμάτι μόνο άρθρων και ειδήσεων. Οι παραγόμενες περιλήψεις χρησιμοποιούν αλγορίθμους που πραγματοποιούν διαβάθμιση προτάσεων και επιλογή αυτών για την περίληψη του κειμένου. Ο μηχανισμός έχει αξιολογηθεί βάσει της κλασσικής μετρικής ανάκλησης/ακρίβειας και του συνδυασμού αυτών.

[11] C. Bouras, V. Poulopoulos, V. Tsogkas "Efficient Summarization Based On Categorized Keywords", C. Bouras, V. Poulopoulos, V. Tsogkas, The 2007 International Conference on Data Mining (DMIN'07), Las Vegas, Nevada, USA, June 25 - 28, 2007, pp. 285 - 291

Περίληψη: 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 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 εργαλεία που θα τον βοηθήσουν στην ευκολότερη εύρεση πληροφορίας.

Παράλληλες Δημοσιεύσεις σε Περιοδικα

[1] A. Jurgelionis, P. Fechteler, P. Eisert, F. Bellotti, H. David, J. Laulajainen, R. Carmichael, V. Poulopoulos, A. Laikari, P. Perala. Platform for Distributed 3D Gaming, A. Gloria, C. Bouras, International Journal of Computer Games Technology, Hindawi Publishing Corporation, 2009

Περίληψη: Τα παιχνίδια υπολογιστών συνήθως κατασκευάζονται για να εκτελούνται σε πλατφόρμες Windows με χρήση DirectX και παράλληλα χρειάζονται ισχυρές CPU και GPU. Για να πετύχουμε εκτεταμένη χρήση παιχνιδιών σε διάφορα περιβάλλοντα όπως σπίτια, ξενοδοχεία, ή internet cafés, είναι απαραίτητο να μπορούμε να εκτελούμε τα παιχνίδια σε συσκευές μικρού μεγέθους, προσπαθώντας να ξεφύγουμε από τους περιορισμούς που υπάρχουν στα ίδια τα παιχνίδια για πόρους συστήματος. Σε αυτή την εργασία παρουσιάζουμε μία νέα καθολική πλατφόρμα που βασίζεται σε κατανεμημένο 3D gaming σε διάφορα περιβάλλοντα δικτύου (ενσύρματα ή ασύρματα). Παρουσιάζουμε μία καινοτόμο αρχιτεκτονική και πρωτόκολλα που χρησιμοποιούνται για να μεταφερθούν οι εντολές των γραφικών παιχνιδιού πάνω από ένα δίκτυο στις τελικές συσκευές. Παράλληλη εκτέλεση των παιχνιδιών σε έναν κεντρικό εξυπηρετητή και ένας καινούριος τρόπος streaming 3D γραφικών προς πολλαπλές συσκευές επιτρέπουν την πρόσβαση σε παιχνίδια από συσκευές πολύ χαμηλού κόστους όπως STB και SSD (small-screen devices).

[2] C. Bouras, E. Giannaka, T. Karounos, A. Priftis, V. Poulopoulos, T. Tsiatsos. A Unified Framework for Political Parties to Support e-Democracy Practices: the case of a Greek Party. International Journal of Electronic Democracy (IJED), Interscience Publishers, Vol. 1, 2008, pp. 98 – 117

Περίληψη: Η ηλεκτρονική διακυβέρνηση και η ηλεκτρονική δημοκρατία αποτελούν ένα κεντρικό θέμα στην κοινωνία της πληροφορίας σε κάθε επίπεδο: τοπικό, περιφερειακό εθνικό, ευρωπαϊκό αλλά και παγκόσμιο. Σε αυτή την κατεύθυνση, πληθώρα προσπαθειών έχουν πραγματοποιηθεί και πολλά συστήματα έχουν δημιουργηθεί. Σε αυτή την εργασία προτείνουμε μία μεθοδολογία για το σχεδιασμό και την υλοποίηση web-based υπηρεσιών για να υποστηριχθούν ηλεκτρονικές υπηρεσίες διακυβέρνησης. Επιπλέον, παρουσιάζουμε την προσπάθεια που πραγματοποιήθηκε σε ένα κόμμα της Ελλάδας προς αυτή την κατεύθυνση, προς το σχεδιασμό δηλαδή ενός ενοποιημένου περιβάλλοντος για να προσφέρονται και να υποστηρίζονται υπηρεσίες ηλεκτρονικής δημοκρατίας.

[3] I. Antonellis, C. Bouras, V. Kapoulas, V. Poulopoulos "Enhancing a Web Based Community: the case of SIG – GLUE ", International Journal of Web Based Communities (IJWBC), Inderscience Publishers, Vol. 2, No 1, 2006, pp. 112 - 130

Περίληψη: 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.

Παράλληλες Δημοσιεύσεις - Κεφάλαια σε Βιβλία

[1] C. Bouras, V. Poulopoulos, V. Tsogkas. "Squeak Etoys: Interactive and Collaborative Learning Environment", Chapter 37, Handbook of Research on Social Interaction Technologies and Collaboration Software: Concepts and Trends, IGI Global, 2009

Περίληψη: Παρουσιάζουμε το περιβάλλον eToys, ένα μέσο επικοινωνίας και συνεργασίας για το OLPC (One Laptop Per Child). Το OLPC είναι μία ιδέα που δημιουργήθηκε για να φέρει ένα αποτελεσματικό και χρήσιμο τεχνολογικό εργαλείο το οποίο μπορεί εύκολα να υιοθετηθεί από μαθητές και δασκάλους για την ολοκλήρωση των εκπαιδευτικών τους δραστηριοτήτων. Εφόσον το OLPC έρχεται με το περιβάλλον Squeak eToys προ-εγκατεστημένο, εστιάζουμε στον συν-δυασμό του Squeak με κάθε έκφανση του συνεργατικού περιβάλλοντος που προσφέρεται μαζί με την δυνατότητα διαμοιρασμού εργασιών του OLPC (activity sharing). Το περιβάλλον Squeak είναι εξαιρετικά ευέλικτο, διαθέσιμο για κάθε πλατφόρμα και παρέχει ένα απλό framework για την εύκολη ανάπτυξη και debugging νέου εκπαιδευτικού λογισμικού. Προσφέρει ένα εργαλείο για projects που έχουν να κάνουν με πολυμεσικές εφαρμογές, εκπαιδευτικές πλατφόρμες, ακόμα και εφαρμογές web. Το framework που παρέχεται είναι ανεξάρτητο του υλικού και του λειτουργικού συστήματος. Παρουσιάζουμε λοιπόν, βασικές εφαρμογές που μπορούν να αξιοποιηθούν μέσα από το περιβάλλον του Squeak και που έχουν ιδιαίτερο ενδιαφέρον για την εκπαιδευτική διαδικασία.

Παράλληλες Δημοσιεύσεις - Άρθρα σε Εγκυκλοπαίδειες

[1] I. Antonellis, C. Bouras, V. Poulopoulos "Content Transformation Techniques", Encyclopedia of Mobile Computing & Commerce, Vol. 1, IDEA Group Publishing, 2007, pp 119 - 123

Παράλληλες Δημοσιεύσεις - Διεθνή Συνέδρια

[1] C. Bouras, V. Poulopoulos, P. Silintziris. Date - based dynamic caching mechanism. IADIS European Conference on Data Mining 2009, Algarve, Portugal, June 18 - 20 2009.

Περίληψη: Τα ενημερωτικά portals που βασίζονται σε κανάλια επικοινωνίας όπως το RSS πληθαίνουν ολοένα και περισσότερο. Οι μηχανές αναζήτησης οι οποίες αποτελούν τους υποστηρικτικούς μηχανισμούς πολλών portals, αλλά και η λειτουργία τους σαν αυτόνομοι μηχανισμοί, βασίζονται στην καθημερινή ανάγνωση εκατομμυρίων ερωτημάτων και στην ανάλυση τους. Και ενώ τα ερωτήματα παράγονται από πολλούς και διαφορετικούς χρήστες, οι μελέτες αποδεικνύουν πως ερωτήματα που μοιάζουν μεταξύ τους παράγονται συνήθως από ανθρώπους με κοινές ανάγκες. Παράλληλα, μία ακόμα αλήθεια είναι πως οι χρήστες τείνουν να επιλέγουν τα ίδια ερωτήματα, ξανά και ξανά, κάθε φορά που χρησιμοποιούν μία μηχανή αναζήτησης. Συνδυάζοντας τα παραπάνω παρουσιάζουμε έναν αλγόριθμο caching τον οποίο εφαρμόζουμε σε ένα σύστημα προσωποποιημένων RSS feeds. Χρησιμοποιώντας ελάχιστους πόρους συστήματος, είμαστε σε θέση να πραγματοποιήσουμε cache σε ερωτήματα που πραγματοποιούν οι χρήστες.

[2] C. Bouras, V. Poulopoulos, P. Silintziris. Personalized News Search in WWW: Adapting on user's behavior. The Fourth International Conference on Internet and Web Applications and Services - ICIW 2009, Venice, Italy, , 24 - 28 May 2009, pp. 125 - 130

Περίληψη: Η προσωποποιημένη αναζήτηση στο Διαδίκτυο γίνεται ολοένα και σημαντικότερη ειδικά στον τομέα της ανάκτησης πληροφορίας και φυσικά στις μηχανές αναζήτησης, καθότι είναι δυνατόν να ενισχύσει την ποιότητα των αποτελεσμάτων αλλά και την εμπειρία των χρηστών. Σε αυτή την εργασία παρουσιάζουμε και αξιολογούμε ένα σύστημα το οποίο εφαρμόζει προσωποποιημένη αναζήτηση στο σύστημα peRSSonal, έναν δικτυακό μηχανισμό ανάκτησης και παρουσίασης άρθρων στο διαδίκτυο. Η τεχνική που παρουσιάζουμε χρησιμοποιεί πληροφορίες που λαμβάνει από το προφίλ χρηστών, όπως και πληροφορίες που ανακτά καθώς ένας χρήστης πραγματοποιεί αναζητήσεις. Χρησιμοποιώντας αυτή την προσέγγιση μπορούμε να παρουσιάσουμε στο χρήστη ποιοτικά καλύτερα αποτελέσματα σε πολύ πιο γρήγορο χρόνο απ' ότι μία κανονική αναζήτηση.

[3] C. Bouras, V. Poulopoulos, V. Tsogkas. Networking Aspects for the Security of Game Input. 5th IEEE International Workshop on Networking Issues in Multimedia Entertaiment - NIME09, Las Vegas, USA, 13 January 2009

Περίληψη: Ακολουθώντας τις προκλήσεις της εποχής μας, το ISP project Games at Large εισάγει μία καινοτόμα πλατφόρμα για την εκτέλεση διαδραστικών, πλούσιων σε πολυμεσικό

περιεχόμενο εφαρμογών πάνω από ασύρματα δίκτυα WLAN. Η αποστολή του Games at Large project είναι να παρέχει μία νέα αρχιτεκτονική για διαδραστικά πολυμέσα η οποία θα δώσει τη δυνατότητα σε υπάρχουσες συσκευές όπως τα Set Top Boxes (STB), συσκευές μικρού μεγέθους, κ. α., οι οποίες δεν έχουν αρκετή επεξεργαστική ισχύ και απόδοση γραφικών, να παρέχουν μία πλούσια εμπειρία στους χρήστες τους. Σε αυτή τη δημοσίευση παρουσιάζουμε το τμήμα μετάδοσης των εντολών εισόδου που παρέχει δυνατότητες κρυπτογράφησης για να εξασφαλίσει την ασφάλεια των μεταδιδόμενων ευαίσθητων δεδομένων. Συνοπτικά, το λογισμικό στον client όταν συλλαμβάνει είσοδο από μία συσκευή πληκτρολογίου, κρυπτογραφεί τις εντολές που μεταφέρονται από το ασύρματο δίκτυο χρησιμοποιώντας σχήμα κρυπτογράφησης δημοσίου κλειδιού. Ο Server είναι υπεύθυνος για την αποκωδικοποίηση και την εκτέλεση των εντολών στο αντίστοιχο παράθυρο της εφαρμογής.

[4] C. Bouras, V. Poulopoulos, I. Sengounis, V. Tsogkas. Networking Aspects for Gaming Systems. Third International Conference on Internet and Web Applications (ICIW 2008), Athens, Greece, 8 - 13 June 2008, pp. 650 - 655

ΕΛΑΒΕ ΒΡΑΒΕΙΟ ΚΑΛΥΤΕΡΟΥ PAPER

Περίληψη: Καθώς η εξέλιξη της τεχνολογίας υπολογιστών εισάγει νέες βελτιώσεις στα δίκτυα, ανάμεσα σε όλα τα υπόλοιπα, τα διαδικτυακά παιχνίδια κερδίζουν όλο και περισσότερο έδαφος. Ακολουθώντας αυτές τις επιταγές, το ISP Project, Games at Large, εισάγει μία καινοτόμα πλατφόρμα για την εκτέλεση διαδραστικών, πλούσιων σε πολυμέσα εφαρμογών για την εκτέλεσή τους πάνω από ασύρματα δίκτυα. Η αποστολή του Games at Large project είναι να παρέχει μία νέα αρχιτεκτονική για διαδραστικά πολυμέσα η οποία θα δώσει τη δυνατότητα σε υπάρχουσες συσκευές όπως τα Set Top Boxes (STB), συσκευές μικρού μεγέθους, κ. α., οι οποίες δεν έχουν αρκετή επεξεργαστική ισχύ και απόδοση γραφικών, να παρέχουν μία πλούσια εμπειρία στους χρήστες τους. Σε αυτή τη δημοσίευση παρουσιάζουμε το υποσύστημα των controllers που υλοποιείται στο πλαίσιο του project Games at Large. Πιο συγκεκριμένα, παρουσιάζεται η γενική αρχιτεκτονική του συνολικού συστήματος και εστιάζουμε στα τμήματα «σύλληψης και εκτέλεσης» των εντολών σε διάφορους clients, καθώς και στην απομακρυσμένη εκτέλεση των εντολών στη μεριά του server. Το λογισμικό του χρήστη συλλαμβάνει την είσοδο από διάφορες συσκευές εισόδου, τις στέλνει πάνω από ένα ασύρματο δίκτυο (WLAN) και ο server είναι υπεύθυνος για την λήψη και εκτέλεση των εντολών στην εκάστοτε εφαρμογή.

[5] C. Bouras, V. Poulopoulos, I. Sengounis, V. Tsogkas "Input here – Execute there through networks: the case of gaming", The 15th Workshop on Local and Metropolitan Area Networks (LANMAN 2007), Princeton, NJ, USA, June 10 – 13, 2007

Περίληψη: 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.

[6] I. Antonellis, C. Bouras, V. Kapoulas, V. Poulopoulos. "Design and Implementation of a Game Based Learning Related Community", IADIS International Conference – Web Based Communities 2005, Algarve, Portugal, February 23 – 25, 2005, pp. 215 - 222

Περίληψη: 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.

[7] C. Bouras, G. Kounenis, I. Misedakis, V. Poulopoulos. "A Web Clipping Service's Information Extraction Mechanism", 3rd International Conference on Universal Access in Human – Computer Interaction, Las Vegas, Nevada, USA, July 22 – 27, 2005

Περίληψη: 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.

[8] I. Antonellis, C. Bouras, V. Poulopoulos. "Game Based learning for Mobile Users", The 6th International GAME – ON Conference on the theme: Computer Games: AI and Mobile Users, Louisville, Kentucky, USA, July 27 – 30, 2005

Περίληψη: 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.

[9] C. Bouras, V. Poulopoulos, A. Thanou. "Creating a Polite Adaptive and Selective Incremental Crawler", IADIS International Conference WWW/INTERNET 2005, Lisbon, Portugal, October 19 – 22, 2005

Περίληψη: 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.

[10] C. Bouras, E. Giannaka, T. Karounos, A. Priftis, V. Poulopoulos, T. Tsiatsos. "A Unified Framework for Political Support e – Democracy Practices", IADIS International Conference WWW/INTERNET 2005, Lisbon, Portugal, October 19 – 22, 2005

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.

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

Περίληψη: 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.

[12] C. Bouras, C. Dimitriou, V. Poulopoulos, V. Tsogkas. "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, June 26 – 29, 2006, pp. 43 - 49

Περίληψη: Η πληροφορία υπάρχει και κάθε άποψη της ζωής μας. Η επέκταση του παγκοσμίου ιστού έχει βοηθήσει προς αυτή την κατεύθυνση. Ο παγκόσμιος ιστός μας ``ταϊζει" με τεράστια ποσότητα πληροφορίας και η εκτεταμένη χρήση των υπολογιστών και άλλων συσκευών μας

έχει οδηγήσει σε μια κατάσταση όπου παρότι έχουμε πολύ διαθέσιμη πληροφορία στα χέρια μας, τις περισσότερες φορές μας είναι άχρηστη. Οι άνθρωποι δυσκολεύονται να εντοπίσουν πληροφορία που χρειάζονται και στην ουσία κατέχουν. Πόσες φορές δεν έχουμε προσπαθήσει να εντοπίσουμε ένα συγκεκριμένο άρθρο ή ένα συγκεκριμένο μήνυμα που λάβαμε ή κάποιο SMS που γνωρίζουμε ότι κατέχουμε. Γι' αυτούς τους λόγους πολλές τεχνικές ανάκτησης πληροφορίας έχουν προταθεί και πολλοί μηχανισμοί εξαγωγής πληροφορίας έχουν δημιουργηθεί. Σε αυτή τη δημοσίευση παρέχουμε την πειραματική αξιολόγηση ενός μηχανισμού εξαγωγής κωδικολέξεων και παρουσιάζουμε την διαφορετική αντιμετώπιση που έχουμε για τα διάφορα είδη κειμένων (άρθρα νέων, δημοσιεύσεις και e-mails). Αυτός ο μηχανισμός εξαγωγής κωδικολέξεων είναι μέρος ενός πλήρους συστήματος που περιλαμβάνει υποσυστήματα ανάκτησης πληροφορίας, εξαγωγή πληροφορίας, κατηγοριοποίησης και δημοσίευση πληροφορίας σε προσωποποιημένο portal.

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

Περίληψη: 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.

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

Περίληψη: 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 κατηγοριο-

ποίησης 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.

Αναφορές

I. Antonellis, C. Bouras, V. Kapoulas, V. Poulopoulos. "Design and Implementation of a Game Based Learning Related Community", IADIS International Conference – Web Based Communities 2005, Algarve, Portugal, February 23 – 25, 2005, pp. 215 - 222

1. An Invstigation of Learning Style Differences and Attidudes toward Digital Game – based Learning among Mobile Users, C. Chao, 4th IEEE International Workshop on Wireless, Mobile and Ubiquitous Technology in Education, 2006, pp. 29 – 31

Antonellis, I., Bouras, C., Poulopoulos, V. "Game based learning for mobile users, (2005) Paper presented at the 6th International GAME - ON Conference on the theme: Computer Games: AI and Mobile Users 27-30 July, Louisville, Kentucky, USA

2. 2D barcode and augmented reality supported english learning system, Liu, T.-Y., Tan, T.-H., Chu, Y.-L., 2007, Proceedings - 6th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2007; 1st IEEE/ACIS International Workshop on e-Activity, IWEA 2007, art. no. 4276349, pp. 5-10

3. Mobile game based learning – Taxonomy and student experience, Cisic, D., Tijan, E., Kurek, A., 2007, Proceedings of the International Conference on Information Technology Interfaces, ITI, art. no. 4283787, pp. 299-305

4. Computer Science students can help to solve problems of multiplayer mobile games, CI Sedano, E Kuts, E Sutinen, In Proc. Seventh Baltic Sea Conference on Computing Education Research (Koli Calling 2007), Koli National Park, Finland. CRPIT, 88. Lister, R. and Simon, Eds. ACS. 213-216.

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

5. PolyNews: Delivering Multiple Aspects of News to Mitigate Media Bias, S. Park, S. Kang, S. Chung, S. Choe, J. Song, KAIST Technical Report, CS – TR – 2006 – 263, November 2006

6. ASNA: An Intelligent Agent for retrieving and Classifying News on the Basis of Emotion – Affinity, S. Masum, T. Islam, M. Ishizuka, International Conference on Computational Intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA – IAWTIC'06), 2006

7. Implementation of Affect Sensitive News Agent (ASNA) for Affectively Classifying of News Summary, S. Masum, T. Islam, M. Ishizuka, H. Prendinger, International Conference on Computer and Information Technology – ICCITT 2006

8. A Study of Local and Global Thresholding Techniques in Text Categorization, N. Wanas, D.

Said, N. Hegazy, N. Darwish, Fifth Australasian Data Mining Conference (AusDm2006)
9. Emotion Sensitive News Agent: An Approach Towards User Centric Emotion Sensing from the News, NM Wanas, DA Said, NH Hegazy, NM Darwish, 2007, Proceedings of the fifth Australasian conference on Data mining and analystics - Volume 61 pp. 91 – 101
10. Mining frequent generalized patterns for web personalization in the presence of taxonomies, Giannikopoulos, P., Varlamis, I., Eirinaki, M., International Journal of Data Warehousing and Mining Volume 6, Issue 1, January 2010, pp 58-76

"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
11. A Virtual Teacher Community to Facilitate Professional Development, D Ratcheva, E Stefanova, I Nikolova, Proceedings of the Second International Conference" Informatics in Secondary Schools: Evolution and Perspectives", Vilnius, Lithuania 2006

"Platform for Distributed 3D Gaming", A. Jurgelionis, P. Fechteler, P. Eisert, F. Bellotti, H. David, J. Laulajainen, R. Carmichael, V. Poulopoulos, A. Laikari, P. Perala. A. Gloria, C. Bouras, International Journal of Computer Games Technology, Hindawi Publishing Corporation, 2009, Article ID 231863
12. A server-assisted approach for mobile-phone games, Arsov, I., Preda, M., Preteux, F., Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Volume 5960 LNCS, 2010,pp 170-187

ΠΡΟΛΟΓΟΣ

Όταν το 2000 ξεκινούσα τα πρώτα μου βήματα στο χώρο του Πανεπιστημίου δε θα μπορούσα να σκεφτώ πως σήμερα, 10 χρόνια μετά, θα έχω στα χέρια μου ένα κείμενο που θα κρύβει μέσα του 10 χρόνια δουλειάς που έγινε με πολλή αγάπη και πολύ μεράκι, όπως κάθε τι που καθορίζει τη ζωή του καθενός. 10 χρόνια μετά, μπορώ να νοιώσω ευτυχισμένος που κλείνω ένα κεφάλαιο της ζωής μου και με αυτό να μπορέσω να ανοίξω πολλά, αμέτρητα άλλα. Το σημερινό αποτέλεσμα σίγουρα μπορεί να μην ήταν μία ξεκάθαρη σκέψη στο μυαλό ενός 18χρονου ήταν όμως μία μεγάλη πίστη και μία ατελείωτη προσπάθεια, που είχε σαν αποτέλεσμα, εκτός από την απόλυτη προσωπική ευχαρίστηση, τη συμβολή σε πληθώρα ερευνητικών θεμάτων. Η εργασία αυτή, προφανώς, δεν είναι ατομική. Υπάρχουν άνθρωποι που συντέλεσαν στο να πραγματοποιηθεί όλο αυτό το εγχείρημα είτε με την προσωπική τους εργασία, είτε με τη βοήθειά τους. Κι όσο κι αν ο πρόλογος έχει γραφτεί για να εκφραστεί η προσωπική ικανοποίηση άλλο τόσο υπάρχει για να εκφράσει την απέραντη ευγνωμοσύνη στους ανθρώπους που έπραξαν τα μέγιστα για το τελικό αποτέλεσμα. Γι' αυτό, αρχικά, θέλω να ευχαριστήσω θερμά τους Γιώργο Αδάμ, Κωνσταντίνο Ασημάκη και Γιώργο Τσιχριτζή. Οι μεγαλύτερες και πιο θερμές μου ευχαριστίες στον καθηγητή Χρήστο Μπούρα ο οποίος, σα δάσκαλος, αγκάλιασε κάθε προσπάθεια που έγινε όλα αυτά τα χρόνια, είτε θετική είτε αρνητική, είτε ερευνητική είτε προσωπική, πάντα με έναν ιδιαίτερο τρόπο και πάντοτε κοιτώντας μπροστά.

Η εργασία είναι αφιερωμένη στους γονείς μου, Νίκο και Ελένη, που μου δείχνουν όχι μόνο να κοιτάζω μπροστά αλλά να κοιτάζω και ψηλά. Στα αδέρφια μου Λεωνίδα και Έρη που έχουν έναν ευχάριστο τρόπο να σου προσφέρουν απλόχερα την φροντίδα τους. Στους συναδέλφους Δημήτρη, Αντώνη, Ανδρέα, Βασίλη, Γιάννη, Γιώργο, Άκη και όλους όσους ξεχνάω.

Τέλος, στη γυναίκα μου που μου δείχνει τι σημαίνει ζωή και τι σημαίνει μαζί, στη Νινέτα.

Πάτρα, Καλοκαίρι 2010

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Όταν αναζητάς κατι, ψάξε πρώτα να το βρεις εκεί που θα το έκρυβες εσυ

(Ανώνυμος)

Το εισαγωγικό κεφάλαιο παρουσιάζει γενικά στοιχεία για την εργασία που πραγματοποιήθηκε, δίνει γενικά στοιχεία για τις ερευνητικές περιοχές που αναλυσαμε και παραθέτει τη δομή της εργασιας.

1.1 Εισαγωγικά Στοιχεία

Ζούμε μία εποχή τεχνολογικών εξελίξεων και τεχνολογικών αλμάτων με το Διαδίκτυο να γίνεται ένας από τους βασικότερους εκφραστές των νέων τεχνολογικών τάσεων. Ωστόσο, ο τρόπος λειτουργίας του και δόμησής του παρουσιάζει εξαιρετικά μεγάλη ανομοιογένεια με αποτέλεσμα οι χρήστες να βρίσκονται συχνά μπροστά σε αδιέξοδο στην προσπάθεια αναζήτησης πληροφορίας. Άλλωστε η ύπαρξη εκατομμυρίων domains οδηγεί σε δυσκολίες κατά την αναζήτηση πληροφορίας. Η έρευνα που πραγματοποιείται επικεντρώνεται στους δικτυακούς τόπους που αποτελούν πηγές ενημέρωσης και πιο συγκεκριμένα στα ειδησεογραφικά πρακτορεία ειδήσεων, αλλά και στα blog. Μία απλή αναζήτηση αποκάλυψε περισσότερους από 40 δικτυακούς τόπους από μεγάλα ειδησεογραφικά πρακτορεία στην Αμερική. Αυτό σημαίνει πως στην προσπάθεια αναζήτησης μίας είδησης και δη, όλων των πτυχών της, κάποιος θα πρέπει να επισκεφθεί αν όχι όλους, τους περισσότερους από αυτούς τους δικτυακούς τόπους για να εντοπίσει στοιχεία για το θέμα που τον ενδιαφέρει. Σε αυτό το «πρόβλημα» ή έστω σε αυτή την επίπονη διαδικασία, έχει γίνει προσπάθεια να δοθούν λύσεις μέσα από τη χρήση των καναλιών επικοινωνίας RSS και μέσα από προσωποποιημένους δικτυακούς τόπους που διαθέτουν τα μεγάλα ειδησεογραφικά πρακτορεία ή ακόμα και από τους μηχανισμούς αναζήτησης που αυτοί διαθέτουν. Σε κάθε περίπτωση όμως, υπάρχουν σημαντικά μειονεκτήματα που συχνά οδηγούν και πάλι το χρήστη σε αδιέξοδο. Τα κανάλια επικοινωνίας δε φιλτράρουν πληροφορίες, τροφοδοτώντας τους RSS Reader των χρηστών με πληθώρα πληροφοριών που δεν αφορούν τους χρήστες ή ακόμα είναι ενοχλητικές για αυτούς. Για παράδειγμα η προσθήκη δύο (2) μόνον καναλιών από Ελληνικά μεγάλα ειδησεογραφικά portal μας οδήγησε στη λήψη περισσότερων από 1000 ειδήσεων καθημερινά. Από την άλλη, η χρήση των microsites που έχουν οι δικτυακοί τόποι επιβάλει στους χρήστες την επίσκεψη σε όλους τους δικτυακούς τόπους που τους ενδιαφέρουν. Όσον αφορά στη χρήση των μηχανών αναζήτησης, ακόμα και οι πιο μεγάλες από αυτές συχνά επιστρέφουν εκατομμύρια αποτελέσματα στα ερωτήματα των χρηστών ή πληροφορίες που δεν είναι επικαιροποιημένες. Τέλος, επειδή οι δικτυακοί τόποι των ειδησεογραφικών πρακτορείων δεν έχουν κατασκευαστεί για να προσφέρουν εκτενείς υπηρεσίες αναζήτησης ειδήσεων, είναι συχνό το φαινόμενο είτε να μην προσφέρουν καθόλου υπηρεσία αναζήτησης, είτε η υπηρεσία που προσφέρουν να μη μπορεί να απαντήσει με δομημένα αποτελέσματα και αντί να βοηθά τους χρήστες να εντοπίσουν την πληροφορία που αναζητούν, να τους αποπροσανατολίζει.

Βάσει όλων των παραπάνω, η έρευνα που πραγματοποιείται εστιάζει στη μελέτη μηχανισμών και τεχνικών που θα μπορούν να δώσουν λύση στο πρόβλημα της αναζήτησης ειδήσεων και στο πρόβλημα της καθημερινής σφαιρικής ενημέρωσης για την ειδησεογραφία που επιθυμούν πολλοί από τους χρήστες. Η ιδέα είναι απλή και βασίζεται στο πρόβλημα το οποίο υφίσταται στο χώρο του διαδικτύου: αντί να τοποθετούμε το χρήστη στη διαδικασία ανεύρεσης των ειδήσεων που τον απασχολούν, συλλέγουμε τις ειδήσεις και τις εμφανίζουμε με τον τρόπο που επιθυμεί, παρουσιάζοντας μόνο τις ειδήσεις που ταιριάζουν στο προφίλ του. Η ιδέα ακούγεται απλή και

λογική, ωστόσο για να πραγματοποιηθεί αυτό λαμβάνουμε υπόψη μας συγκεκριμένους παράγοντες. Οι παράγοντες αυτοί είναι οι εξής: οι χρήστες του διαδικτύου μιλούν διαφορετικές γλώσσες και προφανώς ενδιαφέρονται να βλέπουν τις ειδήσεις που τους αφορούν στη γλώσσα τους. Έτσι, υπάρχει κάποιος μηχανισμός που θα συλλέγει όλες τις ειδήσεις από πολλά – αν όχι όλα – τα ειδησεογραφικά πρακτορεία παγκοσμίως για να είναι εφικτή η ενημέρωση όλων των χρηστών του συστήματος παγκοσμίως. Σε αυτές τις ειδήσεις που έχουν συγκεντρωθεί, υλοποιούνται τεχνικές για τον εντοπισμό των ταυτόσημων ειδήσεων προκειμένου για να μην εμφανίζεται πολλές φορές η ίδια είδηση. Παράλληλα, πραγματοποιούνται διεργασίες προ-επεξεργασίας κειμένου, κατηγοριοποίησης και αυτόματης εξαγωγής περίληψης προκειμένου να μην αναπαράγονται απλώς οι ειδήσεις όπως αυτές συλλέγονται (η βασική ιδέα του web clipping) και τέλος ο μηχανισμός είναι σε θέση να αναγνωρίζει το προφίλ του χρήστη και να του εμφανίζει αποκλειστικά και μόνο τα άρθρα που τον ενδιαφέρουν και όχι όλα τα άρθρα που συλλέγονται από το σύστημα.

1.2 Περιγραφή της υπάρχουσας κατάστασης

Όπως ήδη περιγράψαμε η κατάσταση που επικρατεί αυτή τη στιγμή στον κόσμο του διαδικτύου μοιάζει χαοτική. Αυτό συμβαίνει κυρίως γιατί ξαφνικά το διαδίκτυο έγινε χώρος έκφρασης και χώρος ελεύθερης διάδοσης ιδεών χωρίς να υπάρχει κανένα απολύτως όριο στον τρόπο με τον οποίο γίνεται αυτό. Έτσι, σαν αποτέλεσμα έχουμε μία τελείως άναρχη δομή η οποία από τη μία επιτρέπει στον κάθε ένα να γίνει κομμάτι του Διαδικτύου αλλά από την άλλη δημιουργεί τεράστιες δυσκολίες όταν κανείς προσπαθεί να αναζητήσει πληροφορία. Φυσικά, όπως κάθε άναρχα δομημένος οργανισμός υπάρχουν οάσεις μέσα από τις οποίες μπορεί κανείς να φτάσει σε κάθε άκρη αυτού το χαοτικού συστήματος.
Στην ουσία το Διαδίκτυο σήμερα αποτελεί μία από τις πιο σημαντικές πηγές ενημέρωσης. Στην ενημέρωση θα στηριχθούμε σε όλη την εργασία μας και μάλιστα θα επικεντρωθούμε στις ειδήσεις που δημοσιεύονται στο διαδίκτυο. Μέχρι πριν λίγα χρόνια, οι πηγές ειδήσεων του διαδικτύου ήταν είτε μεγάλα ειδησεογραφικά πρακτορεία, διεθνή ή εθνικά και δικτυακοί τόποι εφημερίδων με online παρουσία στο διαδίκτυο, διαδικτυακών πυλών, που συνήθως υποστηρίζονταν από εκδοτικούς οίκους αλλά και κάποιες μεμονωμένες παρουσίες δικτυακών τόπων περιορισμένης συνήθως εμβέλειας που περιείχαν άρθρα περιορισμένης θεματολογίας. Η κατάσταση σήμερα είναι εντελώς διαφορετική συγκριτικά με αυτό που μόλις περιγράψαμε. Μέσα σε λίγα χρόνια οι αλλαγές στο Διαδίκτυο είναι τεράστιες. Σχεδόν κάθε είδους έντυπο ή γενικότερα μέσο ενημέρωσης έχει διαδικτυακή παρουσία. Πέραν τούτου, η ανανέωση του υλικού των Δικτυακών τόπων των μέσων ενημέρωσης γίνεται συνεχώς, συχνά σε όλη τη διάρκεια της ημέρας. Επιπλέον, εκτός από τα μέσα ενημέρωσης, λόγο στην ενημέρωση έχει και ο κάθε πολίτης καθότι αυτό επιτάσσει η νέα μόδα των προσωπικών – ενημερωτικών blog. Όλα τα παραπάνω οδηγούν στην κατάσταση στην οποία έχουμε ήδη αναφερθεί. Αν μάλιστα το δούμε από τη μεριά του τελικού χρήστη θα

διαπιστώσουμε πως όλη αυτή η πληροφορία είναι υπερβολική για κάποιον ο οποίος προσπαθεί να ενημερωθεί. Συχνά, λοιπόν, παρατηρείται το φαινόμενο εφόσον κάποιος προσπαθεί να ενημερωθεί σφαιρικά για κάποιο θέμα ή γενικότερα για την ημερήσια ειδησεογραφία να πρέπει να καταναλώσει πολύ χρόνο για να μπορέσει να βρει διαφορετικές πηγές πληροφορίας ώστε να καλύψει τη θεματολογία που τον ενδιαφέρει. Από την άλλη, οι λύσεις που προτείνονται μοιάζουν περιορισμένης εμβέλειας για τις ανάγκες που έχει σήμερα ένας χρήστης του Διαδικτύου. Οι πιο κοινές λύσεις περιλαμβάνουν RSS feeds ή προσωποποιημένα Microsite τα οποία λύνουν εν μέρει το πρόβλημα. Η χρήση των RSS feed επιτρέπει στους χρήστες να έχουν άμεση πρόσβαση στην πληροφορία που επιθυμούν χωρίς να χρειάζεται να επισκέπτονται συνέχεια τους δικτυακούς τόπους. Ωστόσο, αυτό σημαίνει πως θα πρέπει ο χρήστης να έχει βρει πρώτα όλα τα RSS feeds που υπάρχουν. Από την άλλη τα microsites ακριβώς επειδή δεν έχουν διαφημιστεί πολύ και επειδή οι πρώτες εμφανίσεις τους ήταν απογοητευτικές δε χρησιμοποιούνται ευρέως και μάλλον είναι περισσότερο συχνή η χρήση τους από τους ίδιους τους δημοσιογράφους. Έτσι λοιπόν καταλήγουμε στην πραγματική κατάσταση που επικρατεί στις μέρες μας στο χώρο του διαδικτύου. Κάθε κομμάτι του διαδικτύου αποτελεί πηγή πληροφόρησης και με τη μεγάλη εξάπλωση των blogs κάθε κομμάτι του Internet δύναται να αποτελέσει πηγή ενημέρωσης. Στεκόμαστε στις πηγές ενημέρωσης γιατί με αυτό θα ασχοληθούμε στην εργασία μας. Οι πηγές ενημέρωσης λοιπόν είναι αμέτρητες και ο κάθε χρήστης θα πρέπει να ακολουθήσεις τους βασικούς τρόπους που προσφέρονται για να μπορέσει να λάβει πλήρη ενημέρωση από το διαδίκτυο. Αυτοί είναι είτε να επισκεφθεί μία σειρά από σελίδες της προτίμησής του και να λάβει πληροφορίες για θέματα που τον ενδιαφέρουν, είτε να έχει συγκεντρώσει από τις σελίδες που τον ενδιαφέρουν τα RSS feeds που τον ενδιαφέρουν και να τα παρακολουθεί. Ωστόσο, ούτε η μία λύση, ούτε η άλλη λύση όπως είδαμε έχουν πάντοτε επιθυμητά αποτελέσματα. Για την ακρίβεια πολλές φορές μπορεί να φτάσουν ένα χρήστη στα όριά του.

Ένα ακόμα σημαντικό θέμα, το οποίο έχει προκύψει πρόσφατα σχετίζεται με την πρόσβαση στην πληροφορία από συσκευές μικρού μεγέθους. Έχει παρατηρηθεί πως ολοένα και περισσότεροι χρήστες μπαίνουν, ή δοκιμάζουν τουλάχιστον να το κάνουν, στους δικτυακούς τόπους από συσκευές μικρού μεγέθους με αυξημένες δυνατότητες (κυρίως δικτύωσης). Αυτοί οι χρήστες αναμένουν ότι τόσο η συσκευή τους όσο και οι δικτυακοί τόποι τους οποίους επισκέπτονται θα τους προσφέρουν ίδια εμπειρία πλοήγησης στον ιστό με αυτή ενός συμβατικού υπολογιστή. Στην πραγματικότητα όμως τα πράγματα δεν είναι έτσι και στην πλειονότητα των περιπτώσεων, αν ένας χρήστης καταφέρει να ανοίξει μία σελίδα του διαδικτύου χωρίς αυτή να εμφανίσει προβλήματα τότε έρχεται αντιμέτωπος με ένα σοβαρό πρόβλημα. Η πληροφορία είναι τόσο εκτενής που είναι πολύ δύσκολο για κάποιον να τη διαβάσει. Φυσικά, αν κανείς αναζητά κάτι πολύ συγκεκριμένο με τη βοήθεια του τηλεφώνου του τότε σίγουρα δε θα ενοχλήσει το μέγεθος της πληροφορίας. Η αναγνωση ειδήσεων όμως δεν είναι ένα θέμα στο οποίο ο χρήστης θέλει να αναλωθεί. Για την ακρίβεια θέλει να πάρει όσο περισσότερες πληροφορίες γίνεται όσο το δυνατόν πιο γρήγορα και αν επιστρέψει στην εργασία του.

1.3 Περιγραφή της εργασίας

Βασισμένοι στα προβλήματα που περιγράψαμε παραπάνω, πραγματοποιήσαμε μία εργασία η οποία περιλαμβάνει μία σειρά από μηχανισμούς προκειμένου να επιλύσει όσο το δυνατόν καλύτερα όλα τα θέματα που υπάρχουν. Στη συγκεκριμένη εργασία επικεντρωνόμαστε στην πληροφορία η οποία αποτελεί άρθρα και ειδήσεις που δημοσιεύονται στο Διαδίκτυο. Όπως είναι γνωστό, οι ειδήσεις και τα άρθρα που δημοσιεύονται στο διαδίκτυο μπορεί να προκύψουν τόσο σε δικτυακούς τόπους εφημερίδων, μέσων ενημέρωσης που δραστηριοποιούνται μέσα από το διαδίκτυο αλλά και πλέον στα πολύ γνωστά μας blogs.
Αυτό που θέλουμε να επιτύχουμε είναι να φτιάξουμε ένα μηχανισμό που θα μπορεί να παρέχει στους χρήστες όλη την πληροφορία που υπάρχει σε όποιο ειδησεογραφικό σύστημα συναντάμε με έναν τρόπο μεθοδικό και προσωποποιημένο. Για να το επιτύχουμε αυτό δημιουργούμε μία σειρά από συστήματα τα οποία μπορούν να λειτουργήσουν αρμονικά και συνεργαζόμενα. Κάθε σύστημα έχει μία συγκεκριμένη εργασία και όλα μαζί δημιουργούν το τελικό αποτέλεσμα που ονομάζουμε peRSSonal [65].
Τα υποσυστήματα που χρησιμοποιούμε πραγματοποιούν εργασίες οι οποίες ξεκινούν από την ανάκτηση όλων των άρθρων και των ειδήσεων και καταλήγουν στην παρουσίαση αυτών στο χρήστη μέσω ενός meta-portal. Η ανάκτηση των άρθρων πραγματοποιείται σε δύο στάδια και βασίζεται σε έναν RSS crawler και σε έναν απλό HTML downloader. Ακολουθεί ένα σύστημα εξαγωγής χρήσιμου κειμένου (χρήσιμο κείμενο) από HTML σελίδες το οποίο μας βοηθά να απομονώσουμε το σώμα ενός άρθρου από το υπόλοιπο κείμενο της HTML και από όλα τα στοιχεία που δεν μας είναι χρήσιμα. Ακολουθούν συστήματα πυρήνα τα οποία πραγματοποιούν λεξικολογική ανάλυση στα κείμενα προκειμένου να εξαχθεί πληροφορία που θα βοηθήσει στην κατηγοριοποίηση και εξαγωγή περίληψης που ακολουθεί. Τέλος και αφού έχουν γίνει όλα τα παραπάνω βήματα καταλήγουμε στην παρουσίαση της πληροφορίας στο χρήστη η οποία πραγματοποιείται μέσα από το peRSSonal portal. Το τελευταίο σύστημα πραγματοποιεί μία σειρά από διαδικασίες οι οποίες περιλαμβάνουν δημιουργία, διατήρηση και ενημέρωση του προφίλ ενός χρήστη, προσωποποίηση των δεδομένων στο χρήστη, καταγραφή πληροφοριών σχετικά με τη συμπεριφορά του χρήστη αλλά και επιπλέον στοιχεία τα οποία περιέχουν microapplications για το χρήστη προκειμένου να υπάρχει ολοκληρωμένη εμπειρία πλοήγηση στις σελίδες του συστήματος.

1.4 Δομή της εργασίας

Η εργασία μας έχει την ακόλουθη δομή: πέραν της εισαγωγής που μόλις κάναμε στο σύστημα που θα αναπτύξουμε προσδιορίζουμε πολύ αναλυτικά το πρόβλημα το οποίο εντοπίσαμε και με το οποίο έρχόμαστε σε επαφή κατά την εργασία μας. Στη συνέχεια πραγματοποιούμε ανασκόπηση της ερευνητικής περιοχής αναφορικά με τα αντικείμενα τα οποία θα μελετήσουμε.

Το επόμενο κεφάλαιο κάνει μία πρώτη εισαγωγή στο σύστημα peRSSonal το οποίο θα δούμε αναλυτικά στα κεφάλαια που ακολουθούν. Το κεφάλαιο που ακολουθεί περιγράφει αναλυτικά την αρχιτεκτονική του μηχανισμού καθώς και αρκετά τεχνικά στοιχεία που σχετίζονται με την υλοποίηση του μηχανισμού. Ακολουθεί ανάλυση των αλγορίθμων και εκτενής πειραματική διαδικασία που προκύπτει τόσο από πειράματα που έχουν δημοσιευθεί όσο και εκτενή πειράματα που έχουν γίνει και είναι αδημοσίευτα. Η εργασία κλείνει με συμπεράσματα και προτεινόμενη μελλοντική εργασία.

ΚΕΦΑΛΑΙΟ 2

ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

We can't solve problems by using the same kind of thinking we used when we created them.

(Albert Einstein)

Σε αυτή την ενότητα κάνουμε εκτενή παρουσίαση του προβλήματος το οποίο παρατηρήσαμε και μας έδωσε έναυσμα για να πραγματοποιήσουμε την εργασία μας. Πρόκειται για διαδικασίες της καθημερινής μας ζωής που παρουσιάζουν δυσκολίες. Στεκόμαστε στο συνδυασμό αυτών που δημιουργεί προβλήματα στην καθημερινή χρήση του διαδικτύου και καθιστούμε σαφές τι ακριβώς προσπαθούμε να επιλύσουμε.

Η μεγάλη ανάπτυξη των νέων τεχνολογιών, κυρίως την τελευταία δεκαετία που διανύουμε στον 21ο πλέον αιώνα, έχει εισάγει την τεχνολογία και τα παράγωγά της σχεδόν σε κάθε τομέα ακόμα και της καθημερινής μας ζωής. Στο χορό της ανάπτυξης τεχνολογιών αυτό που ονομάζεται Διαδίκτυο έχει παίξει πρωταρχικό ρόλο καθότι η λειτουργία του, η ανάπτυξή του αλλά ακόμα και η ίδια η ύπαρξή του έχει ενεργήσει με τέτοιο τρόπο ώστε να βοηθήσει τα μέγιστα στην εξάπλωση των τεχνολογιών σε όλο τον κόσμο. Αυτό έχει οδηγήσει τους λάτρεις των υπολογιστών αλλά και των gadgets, που έχουν πλέον μία σαφή κατεύθυνση προς το Διαδίκτυο, να βλέπουν και να αντιμετωπίζουν το Διαδίκτυο σαν ένα αγαθό. Το αγαθό που ονομάζεται Διαδίκτυο αυξάνεται με ραγδαίους ρυθμούς με την πάροδο του χρόνου. Η αύξηση αυτή συνοδεύεται και με τεράστια αλλαγή στον τρόπο λειτουργίας του Διαδικτύου αλλά και στις προσφερόμενες από αυτό υπηρεσίες. Ξεκίνησε σαν ένα μέσο προβολής, διαφήμισης, ενημέρωσης που δημιουργείτο από λίγους και απευθυνόταν σε ακόμα λιγότερους και σήμερα πρόκειται για ένα από τα βασικότερα μέσα παραγωγής πληροφορίας. Το αγαθό, λοιπόν, που ονομάζεται διαδίκτυο δημιουργεί από μόνο του, με τον τρόπο που είναι δομημένο, μία κοινωνία που αποτελείται από εκατομμύρια “σπίτια”. Τα “σπίτια” αυτά είναι οι δικτυακοί τόποι, σημεία συνάντησης, ενημέρωσης, σημεία πηγής αλλά και κατανάλωσης πληροφορίας.

Ο χαρακτηρισμός ενός δικτυακού τόπου είναι μια διαδικασία που λαμβάνει μέρος καθημερινά και με κάθε τρόπο στις υπηρεσίες του διαδικτύου και είναι σαφές πως οι χώροι αυτοί μπορούν να κατηγοριοποιηθούν βάση πληθώρας οντολογιών ανάλογα με το ύφος τους, τις υπηρεσίες που προσφέρουν αλλά και το περιεχόμενό τους. Υπάρχουν, λοιπόν, οι χώροι κοινωνικής δικτύωσης, που είναι της μόδας το τελευταίο διάστημα, οι χώροι προσφοράς υπηρεσιών, όπως ενημέρωση ή αναζήτηση, οι χώροι προβολής και διαφήμισης, οι χώροι προσωπικών διαδικτυακών ημερολογίων και τόσοι άλλοι αμέτρητοι πια που γεννώνται κάθε λεπτό που περνάει. 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. Όλα τα παραπάνω οδηγούν σε ένα γενικευμένο συμπέρασμα που ωστόσο μπορεί να θεωρηθεί ασφαλές. Ο Παγκόσμιος Ιστός, όπως υπαγορεύει και το ίδιο του το όνομα γίνεται μία καθολική κοινωνία, χωρίς σαφείς κανόνες ή υποχρεώσεις αλλά με αμέτρητες ελευθερίες, όπου ο καθένας μπορεί να συμμετάσχει, μπορεί να διαμορφώσει καταστάσεις, μπορεί να εκφράσει την άποψή του, μπορεί να χρησιμοποιήσει πολλές ελεύθερες υπηρεσίες και γενικά μπορεί να βρει ένα χώρο που πραγματικά φαίνεται να κινείται απόλυτα στον όρο δημοκρατία. Ωστόσο, όπως κάθε κοινωνική δομή έτσι και το Διαδίκτυο έχει πολλά και μεγάλα προβλήματα. Πηγή αυτών των προβλημά-

των 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.

Στην εργασία μας δε θα αναλωθούμε στην καταγραφή των προβλημάτων του Διαδικτύου αλλά θα επικεντρώσουμε την προσοχή μας σε ένα μόνο κομμάτι των προβλημάτων. Τα προβλήματα που θα αναφέρουμε και θα αναλύσουμε αφορούν την την αέναη, καθημερινή και καταιγιστική δημιουργία δεδομένων και πληροφοριών. Για να είμαστε πιο σαφείς, εστιάζουμε την προσοχή μας στην πληροφορία που πηγάζει από τις ενημερωτικές δικτυακές πύλες. Αναφερόμαστε στα γνωστά πλέον news portals και blogs. Οι δικτυακοί αυτοί τόποι ξεκίνησαν έχοντας σίγουρα άλλες βλέψεις αλλά ακολούθησαν, θέλοντας και μή, τις ανάγκες που γέννησε η ίδια η ροή του Διαδικτύου. Αυτό συμβαίνει περισσότερο με τα blogs παρά με τα news portals σε σημείο που σήμερα παρατηρείται το φαινόμενο τα blogs να αποτελούν βασικότερη πηγή ενημέρωσης ακόμα και από μεγάλες και γνωστές πύλες ενημέρωσης. Μερικές από τις πιο γνωστές πύλες ενημέρωσης ειναι οι: CNN [12] , το BBC [10], το Reuters [32], το FoxNews [17], το Bloomberg [11], το Associated Press [9], καθώς και οι υπηρεσίες που προσφέρονται από τους πολυπληθείς και από τους πλέον αναγνωρίσιμους δικτυακούς τόπους Google [19] και Yahoo [33].

Οι ενημερωτικές πύλες αλλά και τα blogs έχουν σαν σκοπό να ενημερώσουν τους χρήστες τους για το τι συμβαίνει καθημερινά σε ολόκληρο τον πλανήτη. Και ενώ αρχικά η εμβέλειά τους μπορεί να ήταν τοπική ή εθνική, σήμερα, λόγω του τρόπου εξάπλωσης του Διαδικτύου γίνεται παγκόσμια. Ο τρόπος δόμησης της πληροφορίας είναι συγκεκριμένος και σε ελάχιστες περιπτώσεις διαφέρει. Τα νέα/άρθρα παρουσιάζονται με δομημένο τρόπο σε συγκεκριμένες σελίδες (καθένα σε μία σελίδα), ωστόσο το πλήθος είναι τέτοιο που είναι ανθρωπίνως αδύνατο να μπορέσει κανείς να παρακολουθήσει όλες τις ειδήσεις που δημοσιεύονται. Αυτό οφείλεται, όπως αναφέρθηκε ήδη, στο γεγονός ότι υπάρχουν πολλές πηγές πληροφορίας αλλά και στο γεγονός ότι κάθε πηγή πληροφορίας δημοσιεύει καθημερινά μεγάλο αριθμό άρθρων. Βέβαια, ο κάθε ενημερωτικός δικτυακός τόπος είναι χωρισμένος σε κατηγορίες και μάλιστα οι κατηγορίες αυτές είναι κοινές για όλους σχεδόν τους δικτυακούς τόπους. Έτσι θα μπορούσε κανείς να παρακολουθεί μεμονωμένες κατηγορίες αλλά και πάλι απαιτείται συνεχής παρακολούθηση κάθε δικτυακού τόπου προκειμένου να υπάρχει πλήρης και σφαιρική ενημέρωση. Αυτό διότι, πολλές ειδήσεις, παρουσιάζονται σε κάθε δικτυακό τόπο με διαφορετική οπτική και χάνεται, συχνά, το κριτήριο της αντικειμενικότητας της είδησης. Απόρροια των παραπάνω είναι ότι: οι χρήστες του διαδικτύου έρχονται αντιμέτωποι με ένα χείμαρρο πληροφορίας. Ακόμα κι αν προσπαθήσουν να γίνουν πολύ συγκεκριμένοι αναφορικά με την είδηση που αναζητούν είναι πολύ πιθανό να έχουν απέναντί τους δεκάδες παραπλήσιες ειδήσεις από πολλές διαφορετικές πηγές. Έτσι, μπορεί να παρατηρηθεί πολύ συχνά πλέον το φαινόμενο οι χρήστες να καταναλώνουν περισσότερη ώρα στην αναζήτηση μίας είδησης που τους ενδιαφέρει παρά στην ανάγνωση της ίδιας της είδησης. Είναι, λοιπόν, υποχρεωμένοι να πραγματοποιήσουν ένα εκτενέστατο φιλτράρισμα πληροφορίας

προτού φτάσουν στο σημείο που επιθυμούν. Ενδεχόμενα σε όλη αυτή την προσπάθεια εκτός του ότι θα καταναλώσουν χρόνο για την αναζήτηση μίας είδησης, θα βρεθούν και στο σημείο να διαβάσουν ειδήσεις για να διαπιστώσουν εκ των υστέρων πως δεν τους ενδιέφεραν. Μία άλλη οπτική στο παραπάνω παράδειγμα είναι και οι χρήστες που αναζητούν πληροφορία προκειμένου να ενημερωθούν σφαιρικά γύρω από ένα θέμα. Αυτοί είναι υποχρεωμένοι να διατρέξουν όλες τις πιθανές πηγές πληροφορίας προκειμένου να συγκεντρώσουν στοιχεία που επιθυμούν. Πόσο μάλλον αυτοί που εργάζονται σαν δημοσιογράφοι και επιθυμούν να παρακολουθήσουν ένα θέμα από την ώρα που προκύπτει και όσο εξελίσσεται. Ας βάλουμε σε όλα τα παραπάνω και έναν επιπλέον άξονα για να ολοκληρώσουμε την περιγραφή του προβλήματος. Δεδομένης της σημερινής κατάστασης όλα τα παραπάνω μπορεί να κληθεί να τα πραγματοποιήσει κανείς από συσκευές μικρού μεγέθους, των οποίων οι δυνατότητες είναι ήδη τέτοιες που να επιτρέπουν πλήρη εργασία από αυτές. Ερχόμαστε, λοιπόν, αντιμέτωποι με τα εξής προβλήματα. Επιθυμούμε πλήρη, καθολική και σφαιρική ενημέρωση χωρίς κόπο, σε μικρό χρονικό διάστημα και έχοντας κατά νου ότι η πηγή πρόσβασης στην πληροφορία μπορεί να είναι οποιαδήποτε. Λύσεις για όλα τα παραπάνω και έχουν προταθεί και κάποιες από αυτές έχουν εφαρμοστεί με επιτυχία. Φυσικά, όπως όλες οι λύσεις που προτείνονται και εφαρμόζονται, έτσι και αυτές που υπάρχουν έχουν τα τρωτά τους σημεία.

Η παρουσία των RSS [31], που σε ελεύθερη μετάφραση θα μπορούσαμε να ονομάσουμε «Περίληψη του Δικτυακού Τόπου», έρχεται να δώσει μία πρώτη λύση στο δυσβάστακτο 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 [37]) ενώ στην πορεία ακόμα και αυτό δεν είναι αναγκαίο καθότι φυλλομετρητές του Διαδικτύου έχουν τη δυνατότητα ανάλυσης του XML [15], [7] 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 που δημιουργείται στο Διαδίκτυο φαίνεται να έχει ως εξής: οι χρήστες, μην έχοντας κάποια εναλλακτική, καταφεύγουν στην εκτεταμένη χρήση RSS feeds ή χρήση δικτυακών τόπων που τουλάχιστον συγκεντρώνουν υπηρεσίες προσωποποιημένης πρόσβασης σε υπηρεσίες. Η εναλλακτική είναι ίσως το πιο συνηθισμένο φαινόμενο που παρατηρείται στο Διαδίκτυο: περιαγωγή σε όλες τις σελίδες ειδησεογραφικού περιεχομένου και αναζήτηση σε όλα τα άρθρα για τον εντοπισμό αυτών που περιέχουν την πληροφορία που αναζητά ο εκάστοτε χρήστης. Η παραπάνω κατάσταση δεν είναι και τόσο απλή όσο φαίνεται. Τα RSS feeds από τη μία βοηθούν τους χρήστες να έχουν πρόσβαση σε υπηρεσίες χωρίς να χρειάζεται να επισκέπτονται τους δικτυακούς τόπους. Από την άλλη, είναι ζωτικής σημασία για τους δικτυακούς τόπους να έχουν μεγάλη επισκεψιμότητα προκειμένου να μπορούν να επιβιώσουν (πουλώντας ακριβές διαφημίσεις). Η διαμάχη είναι μεγάλη κι ας μην είναι εμφανής. Ο ανθρωπιστικός χαρακτήρας των νέων τεχνολογιών συχνά έρχεται σε δεύτερη μοίρα όταν εμφανίζονται στοιχεία που έχουν να κάνουν με οικονομικά συμφέροντα. Ωστόσο, οσο 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 του 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 τριάντα μερικώς 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.

Αν δοκιμάσουμε να κρατήσουμε κάποιες λέξεις-κλειδιά από την παραπάνω ανάλυση μπορούμε να καταλήξουμε σε μία σειρά στοιχείων που καταγράφουν την υπάρχουσα κατάσταση και μας δίνουν στοιχεία για το πρόβλημα που καλούμαστε να αναλύσουμε.

- Καθημερινά δημοσιεύονται αμέτρητες ποσότητες άρθρων κάθε είδους
- Οι χρήστες του διαδικτύου αδυνατούν να έρθουν σε επαφή με όλη την πληροφορία που τους ενδιαφέρει
- Λύσεις που έχουν δοθεί είναι

– Προσωποποιημένες Σελίδες

είναι συχνά δύσχρηστες και 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 (καθώς φαίνεται σκόπιμα)

Όλα τα παραπάνω συνηγορούν στο ότι είναι δύσκολο αν όχι ακατόρθωτο για κάποιο χρήστη να μπορέσει να παρακολουθήσει αποκλειστικά και μόνο ειδήσεις που τον ενδιαφέρον και ως εκ τούτου αναμένει πως θα έρθει αντιμέτωπος με πληροφορία η οποία του είναι ενδεχομένως άχρηστη. Η κατεύθυνση που φανταζόμαστε και υποστηρίζουμε θεωρεί τους χρήστες πυρήνα του συστήματος από τους οποίους πηγάζει τόσο ο τρόπος κατηγοριοποίησης των ειδήσεων όσο και ο τρόπος προβολής, ομαδοποίησης ακόμα και ο τρόπος με τον οποίο εμφανίζονται αποτελέσματα αναζήτησης. Έχοντας το χρήστη στο κέντρο του συστήματος μελετούμε όλες τις διαδικασίες προκειμένου να επιτευχθεί η χρηστοκεντρική προσέγγιση.

Οι διαδικασίες που είναι απαραίτητες για να φτάσουμε στο επιθυμητό αποτέλεσμα είναι εν μέρει δεδομένες. Θα πρέπει αρχικά να συλλέγεται αυτόματα όλη η πληροφορία σε ένα κεντρικό σημείο (ή έστω να εντοπίζεται), εν συνεχεία να γίνεται ανάλυση, διασύνδεση, κατηγοριοποίηση και εξαγωγή περίληψης και τέλος θα πρέπει η πληροφορία να παρουσιάζεται στους χρήστες με τρόπο σαφή και δομημένο και πάντοτε ανάλογα με τις επιθυμίες και ανάγκες του κάθε χρήστη. Συνεπώς, το αρχικό πρόβλημα (στόχος) που αναγνωρίζουμε για την παρούσα κατάσταση που επικρατεί στον κόσμο του Διαδικτύου ανάγεται σε ένα πρόβλημα που διασπάται στις πολλές διαφορετικές διαδικασίες που πρέπει να εκτελεστούν προκειμένου να προκύψει το τελικό, επιθυμητό αποτέλεσμα. Οι διαδικασίες του συστήματος εκτελούνται στην ουσία σειριακά αλλά και ανεξάρτητα η μία από την άλλη. Έτσι, σε αυτό το σημείο είναι αρκετό να μπορέσουμε να εντοπίσουμε τα προβλήματα και τις δυσκολίες κάθε μιας από τις διαδικασίες. Περιληπτικά, οι διαδικασίες του συστήματος είναι οι εξής:

- εντοπισμός νέων/άρθρων ειδήσεων
- λήψη άρθρων ειδήσεων
- εξαγωγή χρήσιμου κειμένου από σελίδες του διαδικτύου

- λεξικολογική ανάλυση κειμένου για εξαγωγή stemmed keywords
- αυτόματη κατηγοριοποίηση κειμένου
- αυτόματη εξαγωγή περίληψης κειμένου (μεταβλητού μεγέθους)
- προσωποποίηση στις ανάγκες του χρήστη
- δημιουργία και διατήρηση δυναμικού προφίλ χρήστη σε Δικτυακό περιβάλλον

Συνεπώς θα πρέπει να δούμε εν τάχει την επιμέρους ανάλυση του προβλήματος αναφορικά με τις διαδικασίες που πρέπει να εκτελεστούν για την ολοκλήρωση του συστήματος. Οι βασικές διαδικασίες που μας ενδιαφέρουν και μπορούν να αναλυθούν αναφορικά με τον προσδιορισμό του προβλήματος έχουν να κάνουν με την αναζήτηση και γρήγορη εύρεση πληροφορίας, με την παρουσίαση πληροφορίας αλλά και με την προβολή σε κάθε πιθανή συσκευή.

2.1 Ανάκτηση Πληροφορίας

Ένα από τα σημαντικότερα κομμάτια του μηχανισμού αλλά και του προβλήματος που έχουμε να αντιμετωπίσουμε είναι ο τρόπος συλλογής της πληροφορίας που υπάρχει στο διαδίκτυο και αφορά ειδήσεις και άρθρα. Πρόκειται γενικότερα για τη βάση όλων των μηχανισμών που εμφανίζουν κάποιο είδος αλληλεπίδρασης με πληροφορία και συγκεκριμένα μηχανισμοί αναζήτησης πληροφορίας (search engines [123], [131], [128]), μηχανισμοί αποθήκευσης πληροφορίας αλλά και κάθε είδους υπηρεσία που υπάρχει στον κόσμο των υπολογιστών. Όπως ήδη αναφέρθηκε στο μηχανισμό που εξετάζουμε η πληροφορία που έχουμε να συλλέξουμε είναι αποκλειστικά νέα και άρθρα. Ωστόσο πρόκειται για ένα "περίεργο" είδος πληροφορίας με ιδιαίτερα χαρακτηριστικά. Δεδομένου ότι ελέγχουμε ένα "είδος πληροφορίας" αυτό μας διευκολύνει να προσαρμόσουμε το μηχανισμό μας στα ιδιαίτερα χαρακτηριστικά της πληροφορίας αυτής. Έτσι, γνωρίζουμε για το συγκεκριμένο είδος ότι:

- η πληροφορία μπορεί να προκύψει ανά πάσα στιγμή χωρίς καμία περιοδικότητα (τυχαίες αφίξεις, Gaussian κατανομή [28], [56], [89])
- η πληροφορία συνήθως αποτελείται από κείμενο και αρκετά συχνά συνοδεύεται από μία ή περισσότερες εικόνες. Πλέον είναι αρκετά συνηθισμένο και το ηχητικό ντοκουμέντο ή και βίντεο, ωστόσο στην πλειοψηφία τους τα άρθρα αποτελούνται απο κείμενο και εικόνα.
- η πληροφορία μπορεί να δημοσιευθεί σε πολλά κανάλια επικοινωνίας ταυτόχρονα είτε αυτούσια είτε προβεβλημένη από διαφορετική οπτική γωνία (διατηρώντας τα βασικά χαρακτηριστικά της είδησης)

Επιπλέον των παραπάνω που αποτελούν κομμάτι της πληροφορίας, έχουμε να προσθέσουμε και το γεγονός ότι η πληροφορία αυτή έχει συγκεκριμένες πηγές στο διαδίκτυο. Οι βασικές πηγές είναι ειδησεογραφικά πρακτορεία, δικτυακοί τόποι εφημερίδων ή γενικότερα Μέσων Μαζικής Ενημέρωσης και πλέον βασική πηγή αποτελούν και τα blog όπου μάλιστα γεννήτορες των ειδήσεων είναι οι ίδιοι οι αναγνώστες. Ταυτόχρονα και λόγω της ανάπτυξης του διαδικτύου η πληροφορία είναι προσβάσιμη μέσω καναλιών επικοινωνίας που παρουσιάζουν κοινή δομή και οργάνωση. Συνεπώς, τα προβλήματα που έχουμε να αντιμετωπίσουμε είναι αφενός ο εντοπισμός όλων των πηγών πληροφορίας και των καναλιών επικοινωνίας που αυτοί έχουν αλλά και ένας τρόπος ώστε να λαμβάνουμε τα νέα που προκύπτουν από τα διαφορετικά κανάλια επικοινωνίας. Η πρώτη διαδικασία θεωρείται τετριμμένη, συνεχιζόμενη και ολοένα αυξανόμενη. Έτσι, η λίστα των καναλιών επικοινωνίας δημιουργείται από τα κανάλια που προσφέρονται από τους δικτυακούς τόπους και αυξομειώνεται ανάλογα με τις ανάγκες τόσο του συστήματος όσο και των χρηστών. Το βασικό κομμάτι που εμπεριέχει αλγοριθμικές διαδικασίες έχει να κάνει με τις διαδικασίες εντοπισμού των νέων και ειδήσεων που προκύπτουν στα κανάλια επικοινωνίας. Επιπλέον, ο τρόπος λειτουργίας των καναλιών επικοινωνίας είναι τέτοιος που δε μας επιτρέπει να παρουσιάσουμε την είδηση αλλά κομμάτι αυτής και συνεπώς αν θέλουμε να παρουσιάσουμε ολόκληρη την είδηση τότε θα πρέπει να προχωρήσουμε και σε ένα επιπλέον βήμα στη διαδικασία συλλογής ειδήσεων. Αυτό έχει να κάνει με την εξαγωγή του τίτλου και του κυρίως σώματος μίας είδησης. Αυτό διότι οι ειδήσεις έτσι όπως παρουσιάζονται στους δικτυακούς τόπους από τους οποίους πηγάζουν παρουσιάζονται παράλληλα με πληροφορίες που μπορεί να είναι "άχρηστες" για τους χρήστες που διαβάζουν το άρθρο. Η "άχρηστη" πληροφορία έχει να κάνει με διαφημίσεις, μενού πλοήγησης, εικόνες, βίντεο, συνδέσμους κ.α. Έτσι, λοιπόν, θα πρέπει να προχωρήσουμε σε φιλτράρισμα πληροφορίας αμέσως μετά την εξαγωγή των άρθρων και των ειδήσεων.

2.1.1 Φιλτράρισμα Πληροφορίας

Το φιλτράρισμα πληροφορίας είναι μία διαδικασία στην οποία επιδίδεται ο κάθε άνθρωπος σε κάθε προσωπική του εργασία. Όταν μάλιστα προσπαθεί να βρει κάποιο άρθρο που τον ενδιαφέρει μέσα από το διαδίκτυο τότε το φιλτράρισμα πληροφορίας γίνεται ολοένα και πιο έντονο. Στο σύστημα που αναπτύσσουμε καλούμαστε μέσω του μηχανισμού μας να προχωρήσουμε σε αυτοματοποιημένο φιλτράρισμα πληροφορίας. Αυτό γίνεται σε πολλαπλά επίπεδα αλλά στη διαδικασία στην οποία αναφερόμαστε το φιλτράρισμα είναι συγκεκριμένο και αφορά: εξαγωγή μόνο του χρήσιμου κειμένου από τη σελίδα διαδικτύου που περιέχει ένα άρθρο. Ως χρήσιμο κείμενο μίας σελίδας που περιέχει ένα άρθρο θεωρούμε αποκλειστικά και μόνον τον τίτλο και το σώμα του άρθρου και ενδεχομένως ότι άλλα στοιχεία αποτελούν το άρθρο και μπορεί να είναι: σχετικές εικόνες, βίντεο, ηχητικά ντοκουμέντα. Στο εξής όλη η παραπάνω πληροφορία

που αποτελεί ένα άρθρο θα αναφέρεται σαν χρήσιμο κείμενο [39].

2.2 Ανάλυση Πληροφορίας και Επεξεργασία Πληροφορίας

Ένα πολύ μεγάλο κομμάτι του μηχανισμού περιλαμβάνει όλα εκείνα τα στοιχεία που μεσολαβούν από το σημείο στο οποίο έχουμε λάβει την πληροφορία και έως το σημείο στο οποίο την παρουσιάζουμε πίσω στο χρήστη. Πολλά συστήματα δεν πραγματοποιούν κανενός είδους ανάλυση και δε διαθέτουν κάποιο βήμα σε αυτό το σημείο. Στο δικό μας μηχανισμό, τα κομμάτια που περιλαμβάνονται σε αυτό το σημείο κάνουν ίσως την πιο σημαντική εργασία αν σκεφτεί κανείς τις διαδικασίες και τις υπηρεσίες που προσφέρονται μέσα από το σύστημά μας. Σε αυτό το σημείο υπάρχουν οι μηχανισμοί λεξικολογικής ανάλυση του κείμενου από τις οποίες προκύπτει ο πυρήνας του συστήματος που δεν είναι άλλο από τις λέξεις κλειδιά. Οι λέξεις κλειδιά είναι ζωτικής σημασίας για το σύστημα που αναπτύσσουμε καθότι πάνω σε αυτές βασίζονται οι κατηγορίες του συστήματος, ανάλυση των κειμένων, οι κάθε είδους συσχετίσεις αλλά και το προφίλ των χρηστών

2.3 Παρουσίαση Πληροφορίας

Ένα σημαντικό κομμάτι του προβλήματος με το οποίο ερχόμαστε αντιμέτωποι είναι η παρουσίαση πληροφορίας στο χρήστη. Ως παρουσίαση πληροφορίας δεν εννοούμε τη γραφιστική προσέγγιση που ακολουθείται για να παρουσιαστεί η πληροφορία στο χρήστη αλλά το γεγονός ότι αφενός στους δικτυακούς τόπους παραγωγής και παρουσίασης πληροφορίας συναντάται το φαινόμενο της παράλληλης προβολής και πληροφορίας που δεν αφορά το χρήστη ενώ από την άλλη έχει παρατηρηθεί το φαινόμενο περιορισμένης χρήσης των δυνατοτήτων των καναλιών επικοινωνίας με αποτέλεσμα να μην παρουσιάζεται ολόκληρο το άρθρο σε ένα χρήστη που διαβάζει τα feeds. Ως εκ τούτου είναι σημαντικό να ελέγχουμε το είδος της πληροφορίας με την οποία έρχεται σε επαφή ο χρήστης. Στο σύστημα το οποίο αναλύουμε έχουμε μία τεράστια πολυτέλεια. Στο σημείο προβολής της πληροφορίας διαθέτουμε τεράστιο όγκο δεδομένων με αποτέλεσμα να είμαστε σε θέση να επιλέξουμε την πληροφορία που θα δώσουμε στο χρήστη. Φυσικά, οι ίδιες οι πηγές της πληροφορίας διαθέτουν τα ίδια και παραπάνω δεδομένα αλλά όπως γνωρίζουμε η πρακτική τους είναι διαφορετική σε ότι σχετίζεται με την παροχή υπηρεσιών μέσω του διαδικτύου.

2.4 Συμμετοχή του χρήστη στις διαδικασίες

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.
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.

2.5 Πρόσβαση από κάθε μέσο

Ένα πολύ σοβαρό κομμάτι το οποίο μας απασχολεί στην παρούσα εργασία είναι πως στις μέρες μας έχουν αλλάξει εντελώς και οι συνθήκες πρόσβασης στην πληροφορία. Έτσι, οι χρήστες επιχειρούν να δουν την πληροφορία από κάθε μέσο κάτι το οποίο σημαίνει δύο βασικά πράγματα. Πρώτον, περιορισμένος χώρος για την παρουσίασης της πληροφορίας και από την άλλη περιορισμένοι πόροι για την παρουσίαση της πληροφορίας. Σε αυτό το θέμα έχουμε να δώσουμε δύο απαντήσεις: (α) εκτενής αλλά και σωστή χρήση RSS feeds προσαρμοσμένα στις συσκευές μικρού μεγέθους, (β) δυναμική εμφάνιση μεγέθους κειμένου ανάλογα με τη συσκευή πρόσβασης. Αυτή η απλή προσέγγιση μπορεί να δώσει λύση στα σημαντικά θέματα που απασχολούν τους χρήστες εναλλακτικών συσκευών όπως οι συσκευές μικρού μεγέθους αλλά και τα set-top box. Στην εργασία μας έχουμε λάβει υπόψη μας τα παραπάνω στοιχεία και η παρουσίαση της πληροφορίας βασίζεται σε μεγάλο βαθμό στον τρόπο πρόσβασης σε αυτή.

2.6 Ενοποίηση Τεχνολογικών Στοιχείων

Όλα όσα περιγράφηκαν παραπάνω δίνουν σαφώς την ταυτότητα του προβλήματος το οποίο καλούμαστε να δούμε, να αναλύσουμε, να αντιμετωπίσουμε και να προτείνουμε κάποιες, ίσως ριζοσπαστικές λύσεις, προκειμένου να το κάνουμε λιγότερο πρόβλημα. Ωστόσο θα πρέπει κανείς να αναλογιστεί πως όλα τα παραπάνω καλύπτουν ένα τεράστιο φάσμα της έρευνα και ίσως πολλές ερευνητικές περιοχές. Παράλληλα με αυτό εισάγονται και στοιχεία που έχουν να κάνουν με συνένωση όλων των παραπάνω στοιχείων για ένα τελικό, ποιοτικά αποδεκτό αποτέλεσμα. Η έρευνα που κάνουμε σε όλο το διάστημα κατά το οποίο εκπονείται η εργασία μας οδηγεί στο συμπέρασμα πως οδηγούμαστε στην εποχή των καθολικών λύσεων και οι καθολικές λύσεις δυστυχώς ή ευτυχώς απαιτούν καθολικά συστήματα. Και λέμε δυστυχώς ή ευτυχώς γιατί από τη μία αποζητάμε συστήματα που να προσφέρουν ολοκληρωμένες λύσεις από την άλλη η τεχνολογία γίνεται τόσο αχανής που είναι σχεδόν αδύνατο να υπάρχει ένα καθολικό σύστημα κοινής αποδοχής που να είναι απόλυτο και αρκούντως ποιοτικό έστω και στην πλειοψηφία των στοιχείων του. Από την άλλη, πρέπει να τεθεί ένας στόχος και αυτός ο στόχος για μας είναι ένας και μοναδικός: ο ανθρώπινος παράγοντας.

Η οπτική μας και ο τρόπος αντιμετώπισης των τεχνολογικών προκλήσεων είναι να προσπαθήσουμε να στρέψουμε όλες τις διαδικασίες σε ανθρωποκεντρικές διαδικασίες προκειμένου να μπορέσουμε να βάλουμε τον ανθρώπινο, υποκειμενικό παράγοντα μέσα σε κάθε διαδικασία του συστήματος. Με αυτό τον τρόπο καταφέρνουμε να προσαρμόσουμε κάθε κομμάτι του συστήματος ακριβώς σε αυτό που ενδιαφέρει το χρήστη ή τις ομάδες χρηστών με κοινά ενδιαφέροντα. Από τις πολύ απλές διαδικασίες ανάκτησης πληροφορίας όπου η επιλογή του feed URL για τον crawler εξαρτάται από τη δημοφιλία του URL μέχρι τα στοιχεία προβολής πληροφορίας, όπως για παράδειγμα αυτό των συναφών άρθρων τα οποία εμφανίζονται όχι μόνο σε αντιστοιχία με το άρθρο αλλά και με το προφίλ χρήστη, έχουμε τοποθετήσει το χρήστη σαφώς στις διαδικασίες συστήματος και αυτό γίνεται μία πραγματική πρόκληση. Συνεπώς, ο φορέας μας είναι ο άνθρωπος, από εκεί ξεκινάμε και εκεί καταλήγουμε στην πορεία της εργασίας μας.

ΚΕΦΑΛΑΙΟ 3

ΑΝΑΣΚΟΠΗΣΗ ΕΡΕΥΝΗΤΙΚΗΣ ΠΕΡΙΟΧΗΣ

Don't be pushed by your problems.
Be led by your dreams.

(Ανώνυμος)

Η παρούσα κατάσταση περιγράφει που βρισκόμαστε ***αλγοριθμικά*** αναφορικά με τα συστήματα με τα οποία θα ασχοληθούμε. Είναι σαν state-of-the-art. (SoA). Σε αυτό το κεφάλαιο θα ενσωματώσουμε και τους μηχανισμούς που έχουν αναπτυχθεί για κάθε θέμα πέραν από την αλγοριθμική ανάλυση.

3.1 Το WWW σε νούμερα

Το διαδίκτυο σήμερα αποτελεί ενδεχομένως τη μεγαλύτερη πηγή πληροφοριών. Τεράστιος όγκος δεδομένων δημιουργείται, αλλάζει, διαγράφεται, αναζητείται, ανταλλάσσεται και επεξεργάζεται μέσω του Παγκόσμιου Ιστού. Επειδή ο όγκος των δεδομένων του ιστού έχει πάρει χαοτικές διαστάσεις ενώ παράλληλα είναι αδύνατον να δημιουργηθεί και να εφαρμοστεί ένας ενιαίος τρόπος οργάνωσης αυτής της πληροφορίας, η διαχείριση του περιεχομένου καθίσταται αδύνατη. 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 [55]. 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 Πληροφορίας 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. Πιο πρόσφατες μελέτες δείχνουν πως τα τελευταία 8 χρόνια έχουν διπλασιαστεί τα δεδομένα που υπάρχουν στο διαδίκτυο ενώ αναμένεται σε 8 χρόνια να διπλασιάζονται κάθε 8

ώρες. 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 [126].
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 πληροφοριακά συστήματα 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 [181].

- Ασάφεια / Διφορούμενες έννοιες: ανάκτηση μη σχετικών λόγω ασάφειας της ερώτησης ή λόγω ύπαρξης διφορούμενων εννοιών [159].

- 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 [140].

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 [144] και [148]). 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 Ανάκτηση Δεδομένων και Ανάκτηση Πληροφορίας από το Διαδίκτυο

Εξόρυξη πληροφορίας από το Διαδίκτυο ονομάζεται κάθε διαδικασία που έχει σαν αποτέλεσμα ανάκτηση πληροφορίας (Information Retrieval [139], [102], [192]) 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 [151] 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 [83] [171]. 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 [179].

Στην παρούσα φάση το IR αντιμετωπίζει μία σειρά από θέματα. Αρχικά, εφαρμόστηκε σε Βάσεις Δεδομένων 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. [173]. 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 (επί τοις εκατό) των σχετικών κειμένων εν συγκρίσει με αυτά που ανακτήθηκαν ενώ η ανάκληση μας δίνει το ποσοστό (επί τοις εκατό) 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 [77].
Τα τρία κλασσικά μοντέλα στην glsap είναι το Boolean [122], το Vector Space[170] και το Πιθανοτικό [169]. 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 [188]) και τα δίκτυα πεποίθησης (belief networks [99]). 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 [155]).

3.2.1 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 που ανήκει στο Q και μια αναπαράσταση κειμένου dj που ανήκει στο 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 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 δεκάδες 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, script) 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 link 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 τις δύο ή τρεις πρώτες σελιδες 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 υποσυστήματα 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 (κείμενο 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.

3.2.3 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 Management).

3.2.4 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.

3.2.5 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 και Matheus [93]:

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 εξαγόμενη 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:

1. Την ανάπτυξη και κατανόηση της περιοχής της εφαρμογής, της σχετικά προγενέστερης γνώσης του προς εξέταση τομέα και τους στόχους του τελικού χρήστη.

2. 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.

3. Τη δημιουργία του στόχου-συνόλου δεδομένων. Επιλογή του συνόλου δεδομένων (δηλαδή μεταβλητές, δείγματα δεδομένων) στο οποίο η διαδικασία εξόρυξης πρόκειται να εκτελεστεί.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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 τυχόν 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.

3.2.6 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:

1. 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.

2. 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.

3.2.7 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.

3.2.8 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 [47], 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. [196]. 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 [116]. Μελετούν υποακολουθείες που εμφανίζονται πολύ συχνά στο σύνολο των ακολουθιών DNA, για την ανακάλυψη εκείνων των κανόνων συσχέτισης, που βασίζονται στην επανάληψη.

3.2.9 Ανάκτηση γνώσης από βάσεις δεδομένων

Η ανάκτηση γνώσης από βάσεις δεδομένων (Knowledge Discovery in Databases) 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 Knowledge Discovery in Databases 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.

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

3.2.10 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 Crawler [82], [51]) 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 συγκεκριμένους crawlers που αξίζουν προσοχής για τα ιδιαίτερα χαρακτηριστικά που παρουσιάζουν.

WebCrawler

Πρόκειται για έναν από τους πρώτους crawlers που κατασκευάστηκαν από τον Pinkerton το 1994 [124]. 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.

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 [69]. 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.

Mercator

Ο Mercator [108], [154] 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.

WebFountain

Πρόκειται για έναν κατανεμημένο τμηματικό crawler παραπλήσιο του mercator, με τη διαφορά ότι είναι γραμμένος σε C++ [86]. 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.

WebRACE

Πρόκειται για έναν crawler ο οποίος είναι γραμμένος σε Java και αποτελεί ένα κομμάτι ενός γενικότερου συστήματος που ονομάζεται eRACE [200]. 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.

Ubicrawler

Ο Ubicrawler είναι ένας κατανεμημένος crawler γραμμένος σε Java και δε διαθέτει κεντρικοποιημένη διαδικασία [57]. 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.

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 [18]
- Heritrix [21]
- ht://Dig [22]
- HTTrack [24]
- Larbin [25]
- Methabot [27]

- Nutch [29]
- WebSPHINX (Miller and Bharat, 1998) [36]
- WIRE - Web Information Retrieval Environment (Baeza-Yates and Castillo, 2002) [35]

3.3 Εξαγωγή Χρήσιμης Πληροφορίας από σελίδες του Παγκοσμίου Ιστού

Υπάρχει ένας μεγάλος όγκος εργασιών ως προς την αναγνώριση περιεχομένου και την ανίχνευση πληροφορίας, που προσπαθεί να επιλύσει παρόμοια προβλήματα χρησιμοποιώντας διάφορες τεχνικές. Ο Finn et al. [90] παρουσιάζει ορισμένες μεθόδους για την εξαγωγή περιεχομένου, όπου οι πηγές είναι ενιαία άρθρα και το περιεχόμενό τους είναι ένα απλό ενιαίο σώμα. Η αποτελεσματικότητα αυτής της προσέγγισης είναι αποδεδειγμένη για κείμενα ενιαίου σώματος. Ωστόσο, η δομή HTML [23] καταστρέφεται εντελώς. Επιπρόσθετα, η προσέγγιση αυτή είναι αναποτελεσματική για κείμενα πολλαπλών σωμάτων, όπου το περιεχόμενο κατατέμνεται σε πολλαπλά μικρότερα κομμάτια, φαινόμενο που απαντάται πολύ συχνά στα ιστολόγια (“blog).
Ο McKeown et al.[147] παρουσιάζει μια διαφορετική μέθοδο, όπου το μεγαλύτερο σώμα κειμένου σε μια ιστοσελίδα ανιχνεύεται και κατηγοριοποιείται ως περιεχόμενο. Όπως και η προηγούμενη μέθοδος, αυτή η προσέγγιση μοιάζει αποτελεσματική για απλές σελίδες και σελίδες με άρθρα ενιαίου σώματος. Ο αλγόριθμος παράγει ανακριβή αποτελέσματα όταν χειρίζεται κείμενα πολλαπλών σωμάτων, ειδικά με τυχαία διαφήμιση και τοποθέτηση εικόνων. Μια παρεμφερής διαδικασία με ανάλογα αποτελέσματα προτείνεται από τον Wacholder et al [193].
Μια ενδιαφέρουσα τεχνική προτείνεται από τον Rahman et al. [166], όπου χρησιμοποιείται δομική ανάλυση, ανάλυση περικειμένου και περίληψη, ούτως ώστε να γίνει εξαγωγή του χρήσιμου κειμένου από τη σελίδα HTML. Αυτή η προσέγγιση έχει περισσότερους περιορισμούς από την ευελιξία της υλοποίησης, καθώς δεν περιλαμβάνει επαρκείς αλγόριθμους για περίληψη και είναι αναποτελεσματική κατά τη συγχώνευση κομματιών κειμένων πολλαπλών σωμάτων.
Ο Gupta et al. [103] δουλεύουν με το Document Object Model tree [14] και διεξάγουν εξαγωγή περιεχομένου, προκειμένου για να αναγνωρίσουν και να διατηρήσουν τα αρχικά δεδομένα. Αυτή η μέθοδος έχει αποτελέσματα στην εξαγωγή περιεχομένου και στη διαχείριση χρήσιμων δεδομένων που έχουν εξαχθεί, όταν αναλύει σελίδες ενιαίου κειμένου. Παρόλα αυτά, δεν επικεντρώνει στην εξαγωγή άρθρων, κατά συνέπεια το αποτέλεσμα είναι πιθανόν να περιέχει άχρηστο περιεχόμενο, κατά την ανάλυση σελίδων πολλαπλών σωμάτων.
Τα συστήματα που σχεδιάζονται για web clipping μπορεί να θεωρηθούν παρόμοια με τους μηχανισμού εξαγωγής χρήσιμου κειμένου. Οι υπηρεσίες web clipping στοχεύουν σε χρήστες που

έχουν συγκεκριμένες ανάγκες πληροφοριών, χωρίς όμως να έχουν το χρόνο ή τη γνώση να ψάξουν στον ιστό. Τρεις από τις πιο αντιπροσωπευτικές απόπειρες για υπηρεσίες web clipping είναι η WebCQ [132], η THOR[72] και η IEPAD[74]. Το WebCQ είναι ένα πρωτότυπο σύστημα ανίχνευσης και ειδοποίησης που βασίζεται σε διακομιστή για την παρακολούθηση αλλαγών σε αυθαίρετες ιστοσελίδες. Το THOR είναι ένα εξελικτικό και αποδοτικό σύστημα εξόρυξης για την ανακάλυψη και την εξαγωγή QAPagelets από το Deep Web [54],[135]. Το QAPagelet αναφέρεται συνήθως στην περιοχή περιεχομένου σε μια δυναμική σελίδα που περιέχει το αποτέλεσμα της αναζήτησης που διεξήχθη. Το IEPAD είναι ένα σύστημα εξαγωγής πληροφορίας, σχεδιασμένο να αναγνωρίζει πατέντες σε κείμενα Web.

3.4 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.

3.5 Προ-επεξεργασία κειμένου

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.
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 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.

3.5.1 Αφαίρεση σημείων στίξης

Τα σημεία στίξης (punctuation) ενός κειμένου δεν προσδίδουν σημασιολογική πληροφορία σε αυτό και άρα δεν δεικτοδοτούνται. Είναι επομένως αναγκαίο, ένα σύστημα ανάκτη-

σης 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

3.5.2 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.

3.5.3 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.

3.6 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 HMM [79], η Naive Bayes [157] και τα SVM [115] 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, 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 [113]. Σχετικές τεχνικές, οι οποίες περιλαμβάνουν άλλες στατιστικές που πηγάζουν από το σύνολο των κειμένων, έχουν επίσης προταθεί τα πρόσφατα χρόνια· π. χ. κέρδος πληροφορίας [199], odds ratio [149], CORI [94], κλπ. Οι τεχνικές αυτές προσφέρουν μια βελτιωμένη προσέγγιση.

3.6.1 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 (NLP). 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 [127] και περιλαμβάνουν αφαίρεση της κατάληξης, τμηματοποίηση λέξης και λεξιλογική μορφοποίηση. Δύο από τους διασημότερους αλγορίθμους, ο Lovins [133] και ο Porter [164], βασίζονται στην αφαίρεση της κατάληξης. Ο αλγόριθμος Lovins βρίσκει το μακρύτερο ταίριασμα από μια μεγάλη λίστα καταλήξεων, ενώ ο Porter χρησιμοποιεί έναν επαναληπτικό αλγόριθμο με μικρότερο αριθμό καταλήξεων και μερικούς κανόνες. Ένας ακόμη αλγόριθμος, ο Paice/Husk [160], χρησιμοποιεί αποκλειστικά ένα σύνολο κανόνων ενώ ακολουθεί επαναληπτική προσέγγιση.
Στο [118] 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 [49] 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 [119] 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.

3.7 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 [80], 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 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.

3.7.1 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 [134] 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 [152]. 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 [85], [163], [121], [174]. 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 [45], [52].
Πολλές τεχνικές για αυτόματη εξαγωγή περίληψης βασίζονται σε NLP και σε πληροφορίες 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 [100], [109]. 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 [70], [137].
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 [136]. Μάλιστα, δεδομένου ότι οι ειδήσεις και τα άρθρα αναφέρονται συχνά σε γεγονότα πολλές προ-

σπάθειες 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[146], [120], [165], [165].
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.

3.7.2 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

Σχήμα 3.1: Διαδικασία Εξαγωγής Περίληψης

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.

3.7.3 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.

3.7.4 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.

3.7.5 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.

3.7.6 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:

$$cos(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum (x_i)^2 + \sum (y_i)^2}} \tag{3.1}$$

όπου τα και βασίζονται στο μοντέλο διανυσματικού χώρου.

3.7.7 Συσχέτιση ομοιότητας

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

3.7.8 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).

3.8 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 [134], 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 [117], [50], [172]. 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 [175],[176] 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 [88] 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 [141] προσεγγίζουν το πρόβλημα με Poisson και αρνητικές διωνυμικές κατανομές ή με χρήση του k-mixture μοντέλου [177] 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 [195], 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 [96]. 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.

3.8.1 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 [111]. 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 [104]. Αυτές οι αρχές χρησιμοποιήθηκαν σε συστήματα περίληψης όπως τα: SUSY [145], SCISOR [138] και TOPIC [43] Αλλά ακόμη και αυτού του είδους τα συστήματα αδυνατούσαν να εξάγουν αποτελεσματικά αξιόλογες μεταφράσεις.

3.8.2 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 [183], 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.

3.8.3 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.

3.8.4 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 item 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.

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.

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.

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 11 και 11

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:
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.

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.

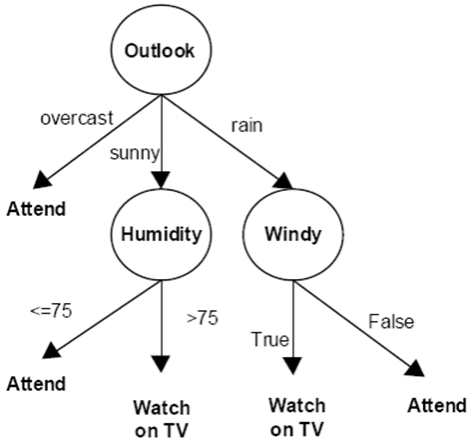
Δημιουργία

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 [109]

3.9 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



Σχήμα 3.2: Δέντρο Απόφασης

ανάλογα με την τεχνική εκμάθησης που χρησιμοποιείται, μερικά ή όλα από τα στοιχεία εισόδου εισάγονται στην πληροφορία εκπαίδευσης.

3.9.1 Αλγόριθμοι για κατηγοριοποίηση πληροφορίας

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

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 γράφημα 3.2 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.

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 [75].

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:

$$P(C = c_k|\vec{x}) = \frac{P(\vec{x}|C = C_k)P(C = C_k)}{P(\vec{x})} \tag{3.2}$$

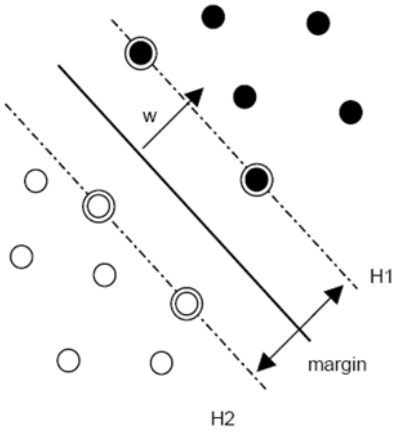
Ο πρώτος όρος του αριθμητή είναι συνήθως δύσκολο να υπολογιστεί χωρίς να απλοποιηθεί η παράσταση. Για το συγκεκριμένο μηχανισμό κατηγοριοποίησης, υποθέτουμε πως τα χαρακτηριστικά X1 … Xn είναι ανεξάρτητα υπό όρους, δεδομένης μίας μεταβλητής κατηγορίας C. Αυτή η υπόθεση απλοποιεί την παραπάνω παράσταση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

$$P(C = c_k|\vec{x}) = \prod_i P(\vec{x}|C = C_k) \tag{3.3}$$

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.

k-Nearest Neighbor (κοντινότερος γείτονας)

Ο αλγόριθμος kNN, είναι μία ακόμα στατιστική προσέγγιση στην αναγνώριση μοτίβου και κατηγοριοποίηση πληροφορίας [110]. 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.



Σχήμα 3.3: Γραμμικά Χωρισμένα Υπερεπίπεδα

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 σχήμα 3.3 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. 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.

3.9.2 Συστήματα Αυτοματης Κατηγοριοποίησης και Εξαγωγής Περιληψης

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 [84], [198], πιθανοτικές προσεγγίσεις με μεθόδους Bayes [143], [197], επαγωγική εκμάθηση κανόνων [189], [129], νευρωνικά δίκτυα [46], on-line εκμάθηση [78] και Support Vector Machines [156]. 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 11. Άλλες εναλλακτικές επιλογές, όπως συντακτικές και στατιστικές εκφράσεις έχουν επίσης χρησιμοποιηθεί [167], [71], [95]. Το βασικό θέμα της αναγνώρισης του θέματος με χρήση NLP έχει αναλυθεί διεξοδικά στο [168].

Άλλες επαναστατικές τεχνικές, όπως η χρήση κωδικών ελέγχου [184], η χρήση αιτιολογικών δικτύων έχουν προταθεί και αποτελούν ουσιαστικά μια τροποποιημένη έκδοση του Bayes αλγορίθμου του [53] που αποδίδουν καλά σε εργασίες κατηγοριοποίησης κειμένων. Καμία από τις προηγούμενες τεχνικές δεν αντιμετωπίζει τα σημασιολογικά θέματα.

3.10 Προσωποποίηση Πληροφορίας

Προσωποποίηση στον τομέα της τεχνολογίας ονομάζεται κάθε διαδικασία η οποία εντοπίζει τις διαφορές μεταξύ ατόμων. Αρχικά χρησιμοποιείται σαν έννοια στο Διαδίκτυο αλλά στην πορεία φαίνεται πως είναι ένας σημαντικός παράγοντας που παίζει τεράστιο ρόλο στην τεχνολογία συνολικά αλλά ακόμα και στην εκπαίδευση, στην ψηφιακή τηλεόραση αλλά και σε στοιχεία που έχουν να κάνουν με σχέσεις b2b και b2c. Αναφορικά με τις σελίδες του διαδικτύου αυτό που ονομάζεται προσωποποίηση μπορεί να γίνει σε μία πληθώρα από χαρακτηριστικά όπως κοινωνικά στοιχεία, δημογραφικά στοιχεία, και φυσικά το πιο σύνηθες τα προσωπικά ενδιαφέροντα. Η προσωποποίηση εφαρμόζεται κυρίως σε στοιχεία που ενδιαφέρουν άμεσα την προβολή δεδομένων προς το χρήστη. Από την άλλη υπάρχει και ο όρος customization ο οποίος αναφέρεται καλύτερα στη δομή της πληροφορίας και όχι στα δεδομένα καθεαυτά. Από την άλλη, ο όρος της προσωποποίησης μπορεί να εφαρμοστεί και σε intranets κυρίως σε εταιρικά πληροφοριακά συστήματα όπου εκεί αφορά συγκεκριμένα χαρακτηριστά του χρήστη τα οποία δεν

πηγάζουν άμεσα από το χαρακτήρα αλλά από το τμήμα και τη θέση την οποία κατέχει. Ο όρος customization σε αυτή την περίπτωση αναφέρεται στη δυνατότητα αλλαγής του layout μίας σελίδας αλλά και του τρόπου παρουσίασης των δεδομένων.
Για να γίνει σαφής ο διαχωρισμός ανάμεσα στους δύο όρους θα αναλωθούμε σε ένα απλό παράδειγμα. Προσωποποίηση είναι όταν ένας χρήστης επιλέγει να βλέπει από ένα ειδησεογραφικό portal τις ειδήσεις των οποίων το περιεχόμενο αφορά ποδόσφαιρο. Customization ονομάζεται η δυνατότητα προβολής των άρθρων σε blocks όπου φαίνεται ο τίτλος και το κείμενο ή σε λίστα που φαίνεται απλά και μόνο ο τίτλος.
Θα μπορούσαμε να πούμε ότι η προσωποποίηση μπορεί να χωριστεί σε κάποιες βασικές κατηγορίες. Έτσι έχουμε την προσωποποίηση που βασίζεται σε ένα προσωπικό προφίλ ή στο προφίλ που έχει ένα γκρουπ χρηστών, την προσωποποίηση που βασίζεται στη «συμπεριφορά» που έχει ένας χρήστης αλλά και προσωποποίηση που βασίζεται στη συνεργατικότητα. Τα μοντέλα προσωποποίησης στο διαδίκτυο περιλαμβάνουν συνήθως απλούς κανόνες που πραγματοποιούν φιλτράρισμα πληροφορίας. Αυτό που συχνά συμβαίνει σαν κομμάτι της προσωποποίησης είναι ο συνδυασμός στοιχείων της συμπεριφοράς του χρήστη, με στοιχεία που έχουν ομάδες χρηστών αλλά και φιλτράρισμα συνόλου δεδομένων πολλών χρηστών προκειμένου να εξαχθεί πληροφορία που αφορά αποκλειστικά το άτομο το οποίο πραγματοποιεί περιαγωγή.
Τα βασικά μοντέλα προσωποποίησης στο διαδίκτυο είναι τα άμεσα, έμμεσα και τα υβριδικά. Όπως ίσως είναι σαφές από τα ονόματα των μεθόδων, στην πρώτη περίπτωση ο χρήστης δηλώνει ρητά τον τρόπο προσωποποίησης, στη δεύτερη περίπτωση ο δικτυακός τόπος επιλέγει αυτόματα και διάφανα στοιχεία για το χρήστη ανάλογα με τη συμπεριφορά αυτού ενώ τα υβριδικά μοντέλα χρησιμοποιούν και τους δύο τρόπους.
Η προσωποποίηση θεωρείται σαν η πανάκεια για χρήση από εμπορικές εφαρμογές που επιθυμούν να βελτιώσουν την εμπειρία των χρηστών κατά τη λειτουργία της εφαρμογής [66]. Αξίζει να αναφερθούμε επίσης και στον τρόπο λειτουργίας των μηχανών αναζήτησης όπου το ζήτημα της προσωποποίησης τίθεται συχνά, όσο συχνά τίθεται βέβαια και η προστασία των προσωπικών δεδομένων. Χαρακτηριστικό παράδειγμα αποτελεί η Google [20] η οποία είναι η πρώτη εταιρία που χρησιμοποιεί χαρακτηριστικά προσωποποίησης σε δεδομένα μεγάλης κλίμακα. Είναι χαρακτηριστικό πως οι όροι που τίθενται στα ερωτήματα που πραγματοποιεί ένας χρήστης της Google βαθμολογούνται πριν παρουσιαστούν τα αποτελέσματα με τρόπο ο οποίος εξαρτάται από το ιστορικό του χρήστη τα google bookmarks που διαθέτει, τη συμπεριφορά του στις υπηρεσίες και την κοινότητα της google αλλά ακόμα και από το rating που έχουν ορισμένες σελίδες που έχει επιλέξει ο χρήστης. Η υπηρεσία είναι διαθέσιμη για όσους διαθέτουν λογαριασμούς στη Google [182], [130].

Είναι σαφές από την ανάλυση που κάνουμε πως η προσωποποίηση των δεδομένων στο χρήστη και η αλλαγή συμπεριφοράς ενός συστήματος ανάλογα με τις ανάγκες ενός χρήστη έχει πολλά πλεονεκτήματα. Ας δούμε μερικά από αυτά για να κατανοήσουμε τις βασικές

διαφορές που υπάρχουν από τη χρήση προσωποποιημένων υπηρεσιών.

- Λιγότερος Χρόνος: Με τη χρήση προσωποποίησης ελαττώνουμε το χρόνο που χρειάζεται κάποιος για να βρει πληροφορίες.

- Λιγότερα Χρήματα: Γνωρίζοντας τις ανάγκες που έχει κάποιος χρήστης ή και πελάτης τις περισσότερες φορές είμαστε σε θέση να ελαχιστοποιήσουμε κόστη

- Λιγότερος Φόρτος Δικτύου: γνωρίζοντας τι στοιχεία χρειάζεται να δει ένα χρήστης μειώνουμε κατά πολύ την πληροφορία που διακινούμε καθότι προβάλουμε μόνο ότι είναι απαραίτητο για το χρήστη.

- Εμπειρία Διαδικτύου: η προσωποποίηση υπόσχεται ένα πιο φιλικό διαδίκτυο προς το χρήστη, ένα διαδίκτυο που σταματά να είναι χαοτικό και προσαρμόζεται στις ανάγκες των χρηστών.

Όλα αυτά ακούγονται πολύ όμορφα και απλά ωστόσο υπάρχουν και πολλές αδυναμίες της προσωποποίησης ακόμα και όταν αυτή είναι απόλυτα πετυχημένη. Το πιο βασικό πρόβλημα που δημιουργείται με την προσωποποίηση και μάλιστα έχει δημιουργήσει ολόκληρα κινήματα εχθρών στον τρόπο λειτουργίας του διαδικτύου. Πρόκειται για την ανωνυμία του διαδικτύου που όπως πολλοί υποστηρίζουν καταστρατηγείται βάναυσα από την προσωποποίηση. Είναι ένα θέμα που μπορεί από μόνο του να αποτελέσει διδακτορική διατριβή τόσο από τεχνικής πλευράς όσο και από νομική πλευρά. Πολλοί υποστηρίζουν πως για να επιτευχθεί ποιοτική προσωποποίηση χρειάζεται συλλογή πληροφορίας η οποία θεωρείται προσωπικά δεδομένα του χρήστη. Από την άλλη, προβλήματα δημιουργούνται από την κακή λειτουργία της προσωποποίησης όπου δηλαδή έχουμε προβληματικά αποτελέσματα. Σε αυτή την περίπτωση μπορούμε ακόμα και να αποπροσανατολίσουμε το χρήστη αν όχι να του δημιουργήσουμε εκνευρισμό.
Ακόμα πιο πίσω στη λίστα με τα προβλήματα αλλά εξίσου σημαντικά είναι θέματα που αφορούν ασφάλεια των δεδομένων καθότι για να επιτευχθεί η κατηγοριοποίηση πρέπει να αποθηκεύεται πληθώρα δεδομένων που αφορούν το χρήστη, τη συμπεριφορά του κατά τη διάρκεια της περιήγησης ακόμα και δημογραφικά ή προσωπικά του στοιχεία. Από την άλλη μερικές φορές πολύ απλά είναι αδύνατον να επιτευχθεί προσωποποίηση είτε γιατί υπάρχει έλλειψη δεδομένων είτε γιατί η φύση των εφαρμογών ή δικτυακών τόπων δεν το επιτρέπει. Έτσι, θα πρέπει να τονιστεί πως ναι μεν η προσωποποίηση καλείται να λύσει και επιλύει ορισμένα προβλήματα που υπάρχουν στο χαοτικό και απρόσωπο διαδίκτυο αλλά από την άλλη δεν είναι πάντα η πανάκεια στα προβλήματα που υπάρχουν.
Σύμφωνα με τον Mobasher [150], «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 [30] [107], προσωποποίηση στη μεριά του χρήστη [112] 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) [3] 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) [4] 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) [13] 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) [6] 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 [8] 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 [30]. 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.

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) [178] 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 [162] 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 [73] 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 [5] 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 [98] 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.

3.10.1 Προφίλ Χρηστη

Σημαντικό ζήτημα για να επιτευχθεί η προσωποποίηση είναι η μοντελοποίηση του χρήστη σε γλώσσα μηχανής και ακόμα περισσότερο σε περιβάλλοντα διαδικτύου. Το προφίλ χρήστη λοιπόν νοείται σαν μία συλλογή προσωπικών πληροφοριών που σχετίζονται με ένα χρήστη. Στην ουσία πρόκειται για την ψηφιακή αναπαράσταση της ταυτότητας ενός χρήστη και των ιδιαίτερων χαρακτηριστικών που αντιπροσωπεύουν αυτό το χρήστη. Στο προφίλ ενός χρήστη αποθηκεύονται συχνά χαρακτηριστικά της προσωπικότητας του χρήστη. Φυσικά αυτό γίνεται με τρόπο με τον οποίο κατανοεί η μηχανή αλλά και με τρόπο ο οποίος να μπορεί να αναπαρασταθεί σε μηχανή. Η πληροφορία μπορεί να εξαχθεί είτε άμεσα από το χρήστη με στοιχεία που ζητάμε από το χρήστη (π.χ. ηλικιακό γκρουπ) ή με πληροφορία που μπορεί να εξαχθεί εμμέσως με τρόπο ο οποίος ορίζεται από κάθε μηχανισμό. Το προφίλ του χρήστη δεν είναι μία καινούρια έννοια στον κόσμο των υπολογιστών. Η ιδέα του προσωπικού υπολογιστή (Personal Computer) από μόνη της αναφέρεται σε έναν προσωποποιημένο υπολογιστή που ανήκει στο χρήστη και αντιπροσωπεύει πλήρως τα δικά του χαρακτηριστικά. Έτσι προφίλ χρήστη μπορούμε να εντοπίσουμε σε λειτουργικά συστήματα (operating system) εφαρμογές υπολογιστων (computer applications), ή, αυτό που μας ενδιαφέρει περισσότερο, σε δυναμικούς δικτυακούς τόπους. Μάλιστα το τελευταίο γίνεται πολύ χαρακτηριστικό όταν αναφερόμαστε σε εργαλεία κοινωνικής δικτύωσης (social networking tools).
Στην ουσία η ιδέα του οργανωμένου προφίλ χρήστη και μάλιστα με αυτή την έννοια (user profile) συναντάται στις αρχές του 2000 οπότε και είδαμε μεγάλη έξαρση στα forums (γνωστά ως τότε ως bulletin boards). Στα forum κάθε χρήστη συμμετείχε αφότου έφτιαχνε ένα προσωπικό προφίλ χρήστη. Σε πολλές περιπτώσεις τα στοιχεία που έπρεπε να δώσει ένας χρήστης έφταναν σε μεγάλο βάθος με τους χρήστες να δίνουν προσωπικές πληροφορίες ακόμα και για χόμπι του, φωτογραφίες του. Σήμερα η ιδέα του προσωπικού προφίλ θεωρείται δεδομένη και μάλιστα υπάρχει σε κάθε δικτυακό τόπο ακόμα κι αν αυτός δεν έχει δημιουργηθεί για τέτοιο σκοπό (π.χ. ακόμα και στο youtube υπάρχει user profile!).
Όπως ήδη αναφέραμε το προφίλ χρήστη είναι πλέον δεδομένο και φυσικά είναι η βάση λειτουργίας των υπηρεσιών κοινωνικής δικτύωσης (social networking services) όπως είναι το Facebook [16], η Google [20], το LinkedIn[26] και πρόσφατα το Twitter [34], όπου οι χρήστες έχουν τη δυνατότητα να δημιουργήσουν ένα πολύ αναλυτικό προφίλ του εαυτού τους που θα τους προσφέρει αναλυτικές λειτουργίες στα εργαλεία που χρησιμοποιούν. Φυσικά, είναι πολύ συχνό το φαινόμενο οι χρήστες να μη θέλουν να δώσουν προσωπικές τους πληροφορίες με αποτέλεσμα τα στοιχεία που υπάρχουν και υφίστανται σαν προφίλ χρηστών να μην είναι σε καμία περίπτωση αξιόπιστα [105], [191] και [91].

Σε αυτό το σημείο αξίζει να δούμε κάτι πολύ ενδιαφέρον από τη Google η οποία έχει εισβάλει ή εστω κάνει προσπάθειες για το σκοπό αυτό σε κάθε τομέα της τεχνολογίας. Έτσι λοιπόν το API OpenSocial έχει κατασκευαστεί από τη Google για να προσφέρει πληροφορίες για τον τρόπο αναπαράστασης, προβολής και πρόσβασης στο προφίλ ενός χρήστη. Αυτό βέβαια θέτει ένα νέο ζήτημα που αφορά το χρήστη στη γλώσσα μηχανής και δεν είναι τίποτα πε-ρισσότερο από τον τρόπο αναπαράστασης του προφίλ ενός χρήστη. Αυτό ονομάζεται μοντελοποίηση του χρήστη (user modeling) και είναι πολύ σημαντικό.
Η μοντελοποίηση χρήστη εντάσεται στην ερευνητική περιοχή της αλληλεπίδρασης ανθρώπου υπολο-γιστή. Σε αυτή την ερευνητική περιοχή η αναπαράσταση νοείται σαν ένα μοντέλο υπολογιστή που μπορεί να αποθηκεύσει, να αναπαραστάσει, να κατανοήσει και να χρησιμοποιεί στοιχεία που αφορούν ένα χρήστη για να μπορεί να διατηρεί ένα ηλεκτρονικό μοντέλο του. Τα μοντέλα αυτά έχουν τη δυνα-τότητα πρόβλεψης στοιχείων που αφορούν το χρήστη και μπορούν να χρησιμοποιηθούν τόσο για να μειωθούν τα λάθη που κάνει κάποιος χρήστης αλλά και να δημιουργηθούν εύκολα και γρήγορα μοντέλα μάθησης. Μάλιστα, η μοντελοποίηση και αναπαράσταση χρηστών χρησιμοποιείται κατά κόρων από τους σχεδιαστές εφαρμογών, διαδικτυακών και μη, προκειμένου να μπορέσουν να σχεδιάσουν και να υλοποιήσουν όσο το δυνατόν πιο «κοντά» στο χρήστη. Όπως έχει ήδη αναφερθεί η μοντελοποίηση χρηστών χρησιμοποιείται κυρίως σε αυτό που ονομάζουμε adaptive hypermedia, στην προσωποποίηση τόσο σε επίπεδο εφαρμογής όσο και σε επίπεδο διαδικτύου, στον τομέα του e-Learning και αλλού [161], [48] και [68].
Προφανώς η μοντελοποίηση του χρήστη σχετίζεται άμεσα με το προφίλ ενός χρήστη και το αντίστροφο. Για τη μοντελοποίηση του χρήστη όπως ήδη αναφέρθηκε υπάρχουν συγκεκριμένοι τρόποι ή όπως ονομάζονται representation formats and standards. Κάποια πολύ χαρακτηριστικά όπως αναφέρεται και στο [185]

- IMS-LIP (IMS - Learner Information Packaging, used in e-Learning)
- HR-XML (Used in human resource management)
- JXDM (Justice with the Global Justice Extensible Markup)
- Europass (the Europass online CV)

Όλα τα παραπάνω που έχουν αναφερθεί τόσο για την προσωποποίηση, για το προφίλ χρήστη αλλά και για τη μοντελοποίηση αυτού οδηγούν σε αυτό το οποίο χρειαζόμαστε για τους χρήστες των συστημάτων μας: μία μοναδική ταυτότητα χρήστη και μιας και αναφερόμαστε στο διαδίκτυο μιλούμε για μία Online ταυτότητα του χρήστη ή ταυτότητα Διαδικτύου. Αν και πολλοί επιλέγουν να χρησιμοποιήσουν τα πραγματικά τους στοιχεία στην Δικτυακή του Ταυτότητα, εντούτοις πολλοί επιλέγουν την ανωνυμία που κρύβεται πίσω από ψευδώνυμα που αποκαλύπτουν μέρη των πραγματικών τους στοιχείων. Σε κάποιες περιπτώσεις είναι εφικτό οι χρήστες να αναπαραστήσουν τον εαυτό τους με ένα avatar το οποίο είτε είναι μία δισδιάστατη εικόνα ή

πλέον ακόμα και 3D γραφικά σε Δικτυακούς Εικονικούς Κόσμους.
Το κομμάτι της προσωπικής ταυτότητας στο Διαδίκτυο είναι ένα κομμάτι έρευνας τόσο από της ψυχολογία όσο και από την κοινωνιολογία, ωστόσο εμείς προφανώς θα εστιάσουμε στον τομέα της τεχνολογίας και στον τρόπο δημιουργίας και διατήρησης μίας ταυτότητας χρήστη. Ο παγκόσμιος ιστός αποτελεί πλέον μία μεγάλη κοινωνία κάτι που εκφράζεται μέσα από τα εργαλεία κοινωνικής δικτύωσης που έχουν πληθύνει το τελευταίο διάστημα. Μάλιστα πολλοί βρίσκουν τη δυνατότητα τόσο να εκφράσουν όσο και να προβάλλουν την προσωπικότητά τους μέσα από αυτά τα εργαλεία [142]. Τρανταχτό παράδειγμα οι εκατομμύρια άνθρωποι οι οποίοι εκφράζουν με όσο το δυνατόν αναλυτικότερο τρόπο τα προσωπικά τους στοιχεία μέσα από προφίλ σε δικτυακούς τόπους όπως το Facebook και το LinkedIn ή ακόμα και σε υπηρεσίες Διαδικτυακών ραντεβού [180].
Η αποκάλυψη των προσωπικών στοιχείων ενός ατόμου μπορεί να δημιουργήσει πολλά προβλήματα που σχετίζονται με την ιδιωτικότητα ή με ανεπιθύμητη προβολή των στοιχείων [153]. Ωστόσο, και επειδή πολλές φορές οι διαδικασίες των δικτυακών τόπων δεν καλύπτουν τις ανάγκες των χρηστών οι χρήστες χρησιμοποιούν πολλές τακτικές προκειμένου να ρυθμίσουν τα επίπεδα πρόσβασης στις προσωπικές τους πληροφορίες [187].
Όπως αναφέραμε και παραπάνω, τίθεται ένα πολύ μεγάλο ζήτημα που σχετίζεται με την αλήθεια των στοιχείων που διαμοιράζονται οι χρήστες στα προσωπικά τους προφίλ στο διαδίκτυο. Οι ταυτότητες λοιπόν των χρηστών, κυρίως σε υπηρεσίες διαδικτυακών ραντεβού έχει αποδειχθεί ότι είναι ανακριβείς και δεν αντικατοπτρίζουν την πραγματικότητα [106] και [87]. Στην περίπτωση των σελίδων κοινωνικής δικτύωσης, όπως το γνωστό σε όλους Facebook, υπάρχουν εταιρίες που προτείνουν την «πώληση φίλων» σαν ένα μέσο να αυξηθεί η δημοτικότητα ενός χρήστη με έμμεσο τρόπο δίνοντας με αυτό τον τρόπο τη δυνατό-τητα προβολής μέσω ψευδών στοιχείων (ψευδή στοιχεία για τους φίλους) [125].
Φυσικά όλα τα παραπάνω έχουν κάποιο σκοπό για τον οποίο πραγματοποιούνται και φυσικά κανείς δεμπορεί να αρνηθεί ότι υπάρχουν σημαντικά κέρδη από τη δημιουργία και χρήση προσωποποιημένων κοινοτήτων με τις πραγματικές ταυτότητες χρηστών. Το μέσο που ονομάζεται υπολιστής για έναν ιδια-ίτερο λόγο απελευθερώνει τους ανθρώπους. Η απουσία επαφής και άμεσων αντιδράσεων για το «είναι» του καθενός επιτρέπει στον καθένα να εκφράσει ελεύθερα και χωρίς αναστολές σκέψεις, ιδέες, ακόμα και την ίδια την προσωπικότητά τους. Η ελευθερία αυτή δίνει νέες δυνατότητες στο κοινωνικό σύνολο και φυσικά δίνει τη δυνατότητα σε ανθρώπους να εκφράσουν προσωπικά τους στοιχεία που ενδεχόμενα να αποτελούν ταμπού σε κάποιες κοινωνίες. Ωστόσο, σε όλα αυτά θα πρέπει να υπάρχει κάποιο όριο. Δεδομένου ότι η νομοθεσία για το διαδίκτυο είναι ακόμα ανεπαρκής για την πλειονότητα των παγκόσμιων κοινοτήτων, είτε οι ίδιοι οι χρήστες ή οι διαδικτυακές κοινότητες θα πρέπει να θέσουν τα όρια λειτουργίας. Οι Online Ταυτότητες οι οποίες είναι προσβάσιμες από όλους δε θα μπορούσαν να είναι πλήρως απελευθερωμένες και να μην ακολουθούν κάποιες συμβάσεις του πραγματικού κόσμου. Η ιδέα του να ξεφύγεις από την κοινωνική δικαιοσύνη και ιεραρχία δεν είναι, όσο και να φαίνεται πε-

ρίεργο, χαρακτηριστικό του Διαδικτύου.
Ας δούμε όμως τομείς του Διαδικτύου οι οποίοι βασίζουν σημαντικά κομμάτια τους στις ταυτότητες χρηστών. Πρώτα και κύρια, ένα σημαντικό κομμάτι του Διαδικτύου που στηρίχθηκε στην ταυτότητα των χρηστών ήταν τα forum και πριν από αυτά τα bulletin boards τα οποία στηρίζουν κομμάτι της ποι-ότητάς τους στις ταυτότητες των χρηστών που συμμετέχουν σε αυτά. Μάλιστα, πολλά από αυτά έχουν επίπεδα χρηστών που έχουν άμεση εξάρτηση από τα προσωπικά στοιχεία που επιλέγει κάποιος χρήστης να προσθέσεις στο προφίλ του. Στη συνέχεια ως εξέλιξη των forum ήρθαν στη ζωή του Διαδικτύου τα προσωπικά blogs τα οποία εκτός από μόδα, αποτελούν και βασικούς χώρους έκφρασης των χρηστων του διαδικτύου. Όπως και στα forum έτσι και στα blog υπάρχει το χαρακτηριστικό του προφίλ χρήστη ο οποίος συμμετέχει σα συγγραφές σε ένα blog. Στην ουσία τα blog επιτρέπουν σε μεμονωμένα άτομα να εκφράσουν τις προσωπικές τους απόψεις για ένα θέμα. Παρά το γεγονός ότι οι bloggers προτιμούν ψευδώνυμα από το να δώσουν την πραγματική τους ταυτότητα πολλοί από τους χρήστες δε διστάζουν να δώσουν τα πραγματικά τους στοιχεία. Χρησιμοποιώντας ένα ψευδώνυμο επιτρέπει σε ένα άτομο να «κρύψει» την ταυτότητά του αλλά να έχει ακόμα τη δημοτικότητα στην κοινότητα του διαδικτύου καθότι κάθε ταυτότητα τον χαρακτηρίζει [81].
Περνώντας σε κάτι πιο σύγχρονο, οι σελίδες κοινωνικής δικτύωσης επιτρέπουν στους χρήστες να έχουν μία Διαδικτυακή ταυτότητα ενώ σε αυτές τις περιπτώσεις το σημαντικό είναι πως τα κανάλια αυτά γίνονται κομμάτι της προσωπικής ζωής. Οι ταυτότητες αυτές έχει παρατηρηθεί ότι φτιάχνονται αντα-νακλώντας έναν εξιδανικευμένο εαυτό του χρήστη. Όπως και στην καθημερινότητα σημαντικά στοιχεία για τους χρήστες των εργαλείων κοινωνικής δικτύωσης αποτελούν τα κομμάτια των δικτυακών τόπων που επιτρέπουν περιορισμό προβολής της ταυτότητας και προστασία των προσωπικών δεδομένων που πλέον οι χρήστες θεωρούν σαν ένα πολύ σημαντικό κομμάτι [67] και [101].
Η ταυτότητα χρηστών σε άλλα πληροφοριακά συστήματα παίζει πολύ μεγάλο ρόλο καθότι οι αλληλεπιδράσεις είναι σημαντικές για την εξέλιξη των προγραμμάτων. Βασικός εκπρόσωπος και χαρακτηριστικό παραδειγμα αποτελούν οι εικονικές τάξεις και γενικότερα όλο το κομμάτι που αφορά e-learning. Ενώ μερικές φορές η ανωνυμία επιτρέπει στους μαθητές να νοιώσουν πιο άνετα και να εκφραστούν πιο ελεύθερα (πχ. Οι μαθητές κάνουν πιο συχνά ερωτήσεις αν δεν γνωρίζει κανείς την ταυτότητά τους), εντούτοις είναι σαφές πως ο καθηγητής πρέπει και επιβάλλεται να έχει ολοκληρωμένη εικόνα της προσωπικότητας του μαθητή για να μπορέσει να προβεί σε αξιολόγηση. Πέραν της αξιολόγησης ο καθηγητής θα πρέπει να είναι σε θέση να αναγνωρίσει και άλλα στοιχεία που αφορούν τους μαθητές και ενδεχομένες σχετίζονται με προσωπικές ή κοινωνικές καταστάσεις. Πολλές φορές, η βοήθεια του καθηγητή συνολικά και η μεμονωμένη προσοχή προς μαθητές οδηγεί σε ποιοτικά αποτελέσματα. Χωρίς αυτή την αλληλεπίδραση είναι επόμενο να πέσει η απόδοση των μαθητών και οι μαθητές να μη δείχνουν καθόλου ενδιαφέρον για το μάθημά τους [1].
Ένα πολύ σημαντικό στοιχείο της εργασία είναι το προφίλ χρήστη σε δυναμικό περιβάλλον. Εί-

ναι 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.

3.11 Συστήματα Αποδελτίωσης του Παγκόσμιου Ιστού

Η διακίνηση της Διαδικτυακής πληροφορίας όπως ήδη έχει αναφερθεί στην εργασία μας είναι τεράστια και καθημερινά ανταλάσσεται μεταξύ των μηχανών που είναι συνδεδεμένες στο Διαδίκτυο τεράστιος όγκος πληροφορίας. Ακριβώς επειδή οι πηγές πληροφορησης και ανταλλαγής δεδομένων έχουν πολλαπλασιαστεί, και επειδή ο καθένας μας μπορεί να γίνει πηγή πληροφόρησης έχουν προκύψει νέα δεδομένα και ανάγκες. Η σκέψη μας να πραγματοποιήσουμε κάθετη αποδελτίωση πληροφορίας δεν είναι εντελώς πρωτοποριακή αν και πρωτόγνωρη για τα ελληνικά δεδομένα. Ωστόσο, υπάρχουν στο Διαδίκτυο πολλά συστήματα που λειτουργούν σα συστήματα αποδελτίωσης και indexing πληροφορίας και σαν σκοπό έχουν τη συγκέντρωση πληροφορίας σε ένα κοινό σημείο από όπου οι χρήστες θα μπορούν να λάβουν ενημέρωση για όλες ή έστω για πληθώρα δημοσιεύσεων του Διαδικτύου. Ως εκ τούτου μπορεί κανείς αναζητώντας στο αχανές διαδίκτυο να εντοπίσει εκτός των άλλων και πολλά πειραματικά και μη συστήματα που πραγματοποιούν αποδελτίωση του ημερήσιου τύπου.

3.11.1 Newsme

Πρόκειται για ένα σύστημα το οποίο εφαρμόζει τεχνικές προσωποποίησης στο χρήστη με διαδικασίες που είναι εντελώς διάφανες στο χρήστη [194] . Αυτό που γίνεται στην ουσία μοιάζει με τις διαδικασίες του συστήματός μας και βασίζεται στην καταγραφή των κινήσεων του χρήστη χωρίς να του ζητείται ρητά σε κανένα σημείο είσοδος για το σύστημα προσωποποίησης. Προφανώς είναι μία τακτική η οποία αφήνει απόλυτα ευχαριστημένο το χρήστη ενέχοντας πάντοτε τον κίνδυνο κάποια τυχαία, ή λάθος κίνηση από το χρήστη να μεταφραστεί λάθος από το σύστημα. Στο χρήστη προβάλλονται δύο ειδών πληροφορίες, νέα άρθρα και προτεινόμενα άρθρα (βάσει του προφίλ). Ο χρήστης από τη μεριά του μπορεί να προσθέσει ένα άρθρο σε αυτά που αναγιγνώσκει προκειμένου να πληροφορηθεί το σύστημα για τις επιλογές του και αντίστοιχα μπορεί να απορρίψει ένα άρθρο, δίνοντας προφανώς εμμέσως κάποια πληροφορία στο σύστημα. Όπως είναι φυσικό η προσωποποίηση βασίζεται στις πιο πρόσφατες ενέργειες που κάνει ο χρήστης και όχι στο σύνολο των ενεργειών από την έναρξη χρήσης του συστήματος κάτι φυσικά θεμιτό καθότι το προφίλ του χρήστη πρέπει να αλλάζει δυναμικά.

3.11.2 GoogleNews

Όπως και σε πολλά άλλα θέματα που αφορούν τον τομέα της πληροφορικής έτσι και σε αυτό, αν όχι την πρωτοπορία ή πληρότητα, τουλάχιστον τη δημοφιλία έχει αυτή τη στιγμή το σύστημα ηλεκτρονικής αποδελτίωσης της Google [19]. Πρόκειται για ένα σύστημα που θα ονομάζαμε πλήρες αναφορικά με τις βασικές ανάγκες που έχει ένας χρήστης. Όπως με κάθε

εργαλείο της Google έτσι και με αυτό το layout που προσφέρεται στο χρήστη είναι απλό προκειμένου το μάτι να επικεντρώνεται απόλυτα στην πληροφορία. Η υπηρεσία φαίνεται να κάνει αποδελτίωση από ειδησεογραφικά πρακτορεία ενώ παράλληλα είναι εμφανές ότι γίνεται προσπάθεια αποφυγής συγκέντρωσης πληροφορίας από blog. Όπως κάθε υπηρεσία που διατίθεται από τη google έτσι και αυτή η υπηρεσία εμφανίζει ιδιαίτερα χαρακτηριστικά αναζήτησης που είναι και το μεγάλο ατού της εταιρίας. Παράλληλα, μπορεί κανείς να λαμβάνει alerts για νέα που δημοσιεύονται ενώ προσφέρονται και προσωποποιημένα RSS feeds.

3.11.3 Newsjunkies

Το newsjunkie είναι ένα σύστημα το οποίο βασίζεται στον εντοπισμό άρθρων που παρουσιάζουν «ενδιαφέρον» [97]. Πρόκειται για μία τεχνική η οποία έχει αποκτήσει πολλούς υποστηρικτές τον τελευταίο καιρό διότι παρουσιάζει στοιχεία του σημασιολογικού ιστού άμεσα εφαρμοσμένα στην πράξη. Ο όρος που χρησιμοποιείται προς αυτή την κατεύθυνση ονομάζεται Information Novelty. Στην ουσία εντοπίζονται Updates για κάποιο προϋπάρχον θέμα με την προσπάθεια να αποφεύγονται τυχόν επαναλήψεις πληροφορίας. Για τον εντοπισμό κειμένων παραπλήσιων με κάποιο προϋπάρχον θέμα χρησιμοποιούνται συγκρίσεις που βασίζονται στην εξόρυξη named-entities (ονόματα ανθρώπων, τοποθεσιών, οργανισμών, επιχειρήσεων, κλπ.). Δυστυχώς το πολλά υποσχόμενο σύστημα δεν είναι διαθέσιμο για δοκιμές.

3.11.4 PersoNews

Το σύστημα PersoNews [114] βασίζεται σε μία προκαθορισμένη οντολογία από την οποία ο χρήστης επιλέγει ένα πεδίο το οποίο τον ενδιαφέρει. Από την επιλογή του χρήστη εξαρτώνται και τα στοιχεία που του προβάλλονται καθότι όπως φαίνεται κάθε RSS feed έχει αντιστοιχηθεί σε στοιχεία της οντολογίας. Παρά τις προδιαγραφές που υπάρχουν για το σύστημα η online έκδοσή του http://news.csd.auth.gr δε φαίνεται να είναι τίποτα περισσότερο από ένας RSS Reader.

ΚΕΦΑΛΑΙΟ 4

ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΜΗΧΑΝΙΣΜΟΥ

Αρχιτεκτονική είναι η αποτύπωση των οραμάτων

(Ανώνυμος)

Στο κεφάλαιο αυτό παρουσιάζουμε την αρχιτεκτονική την οποία σχεδιάσαμε για το σύστημά μας και στην οποία βασιστήκαμε για να προχωρήσουμε σε ανάπτυξη των συστημάτων του peRSSonal

4.1 Μηχανισμός peRSSonal

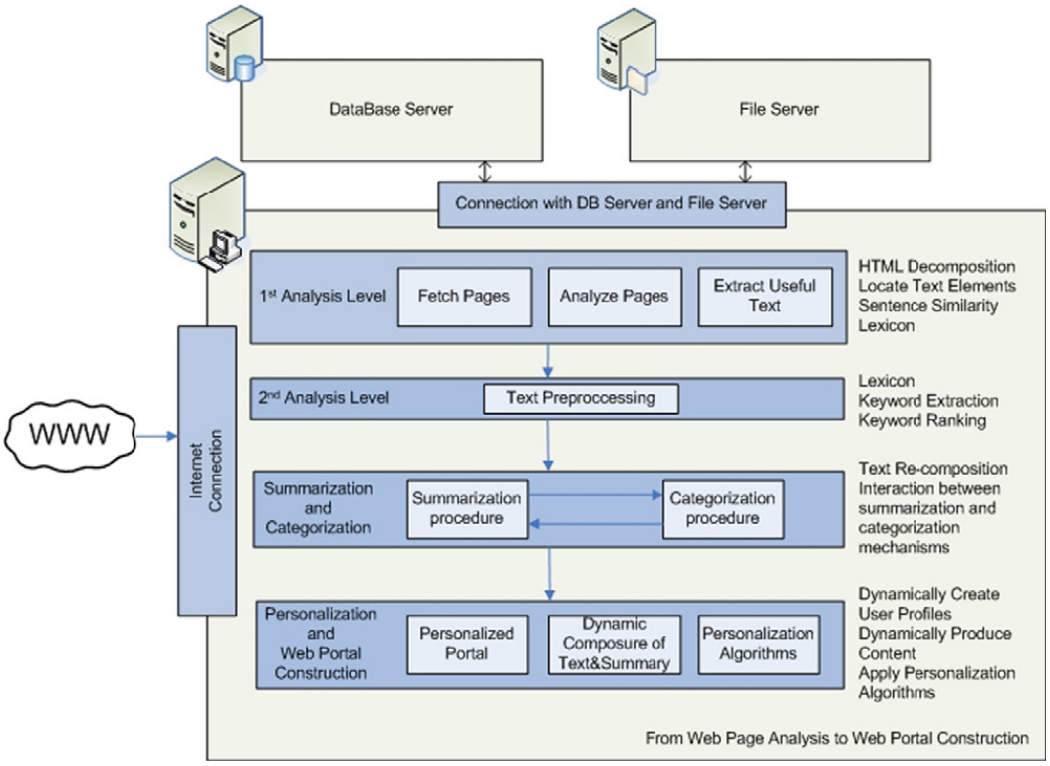
Η κεντρική ιδέα που βρίσκεται πίσω από την κατασκευή του μηχανισμού στηρίζεται στο γεγονός ότι η καθημερινή ζωή στο διαδίκτυο έχει αλλάξει με τέτοιον τρόπο που κάθε γωνιά του Διαδικτύου έχει μετατραπεί σε μία πηγή πληροφορίας. Και λέγοντας πηγή πληροφορίας να αναφέρουμε για ακόμα μία φορά πως αυτό που μας ενδιαφέρει είναι άρθρα και ειδήσεις που πηγάζουν από κανάλια επικοινωνίας του διαδικτύου. Από την άλλη, η αναζήτηση πληροφορίας και εν προκειμένω η αναζήτηση άρθρων και ειδήσεων γίνεται ολοένα και πιο δύσκολη εργασία εφόσον αυτό που επιθυμούμε είναι η αναζήτηση καθολικής και πλήρους ενημέρωσης. Περαν της απλής αναζήτησης, ορισμένες φορές αυτό που είναι εξίσου σημαντικό είναι αυτό που ονομάζεται data refining, το ξεκαθάρισμα δηλαδή της πληροφορίας που παρουσιάζεται στο χρήστη προκειμένου να περιέχει μόνο πληροφορίες που ενδιαφέρουν το χρήστη. Στην ουσία το πρόβλημα του ξεκαθαρίσματος πληροφορίας εν μέρει λύνεται ή γίνεται προσπάθεια να λυθεί με τη βοήθεια των μηχανών αναζήτησης (π.χ. Google, Yahoo, κ.α.), με τη βοήθεια οντολογιών του διαδικτύου (DMOZ) αλλά και γενικότερα με προσπάθειες για προσωποποίηση δικτυακών τόπων (bbc). Το πρόβλημα γίνεται τεράστιο βέβαια αν αναλογιστεί κανείς πως κάθε χρήστης του διαδικτύου έχει ιδιαίτερες ανάγκες και περιμένει διαφορετική πληροφορία από το μέσο που ονομάζεται Διαδίκτυο. Η προσωπική μας εμπειρία μας οδήγησε σε σκέψεις που βρίσκονται πίσω από όλες τις ως άνω προβληματικές. Διαφαίνεται, λοιπόν, πως δημιουργείται και ένα επιπλέον πρόβλημα. Το διαδίκτυο αυξάνεται με ραγδαίους ρυθμούς και μάλιστα αυτό που αυξάνεται είναι η πληροφορία που περιέχεται σε αυτό και πιο συγκεκριμένα η ενημερωτική πληροφορία.

Οι χρήστες του διαδικτύου, νέοι και παλαιοί, έχουν κατανοήσει πως είναι εφικτό να μείνουν ενημερωμένοι για όσα συμβαίνουν στον κόσμο από το διαδύκτο και μάλιστα σε πραγματικό χρόνο. Δεν είναι τυχαίο να δει κανείς πως ανάμεσα στους 100 πιο δημοφιλείς δικτυακούς τόπους, πέραν των μηχανών αναζήτησης είναι οι ενημερωτικοί δικτυακοί τόποι παγκοσμίως. Παρά το γεγονός ότι οι χρήστες έχουν αυτή την απεριόριστη ελευθερία να βρίσκουν σε πραγματικό χρόνο σχεδόν οτιδήποτε επιθυμούν, όλο αυτό σημαίνει πως ο κάθε χρήστης θα πρέπει να επισκεφθεί τον κάθε δικτυακό τόπο ξεχωριστά ώστε να μπορέσει να λάβει πληροφορίες. Τα RSS feeds ή γενικότερα τα κανάλια επικοινωνίας αλλά και αυτά που ονομάζονται personalized microsites μπορούν να λύσουν μέρος του προβλήματος. Στην πρώτη περίπτωση οι χρήστες δε χρειάζεται να επισκέπτονται συνέχεια κάποιον δικτυακό τόπο αλλά αρκεί να έχουν συλλέξει όλα αυτά τα κανάλια επικοινωνίας που επιθυμούν να παρακολουθούν. Αυτό βέβαια σημαίνει πως ο κάθε χρήστης θα πρέπει να ανακαλύψει τους δικτυακούς τόπους από τους οποίους επιθυμεί να λαμβάνει πληροφόρηση και από την άλλη το φιλτράρισμα πληροφορίας περιορίζεται σε επίπεδο γενικής κατηγορίας και τίποτα περισσότερο. Αξιόλογες προσπάθειες έχουν παρουσιαστεί σε αυτή την κατεύθυνση από εργασίες που σα σκοπό έχουν να δημιουργήσουν RSS readers που επιτρέπουν στους χρήστες να θέτουν λέξεις κλειδιά προκειμένου να γίνεται φιλτράρισμα της

πληροφορίας που εμφανίζεται στο χρήστη. Από την άλλη, η λύση των microsites μπορεί να είναι μία λύση που προσαρμόζει την πληροφορία στις ανάγκες των χρηστών, ωστόσο και σε αυτή την περίπτωση ο χρήστης θα πρέπει να επισκέπτεται τους δικτυακούς τόπους από τους οποίους επιθυμεί να ενημερωθεί.

Παράλληλα με τα παραπάνω, από το 2002, η ενασχόλησή μας με το διαδίκτυο και τη ραγδαία αύξησή του είχε αρχίσει να αποκτά κάποια πιο ιδιαίτερη μορφή. Εργασίες που έχουν παρουσιασθεί στο παρελθόν από την ομάδα μας αποδεικνύουν τον ιδιαίτερο προβληματισμό σχετικά με την κατάσταση του διαδικτύου. Χαρακτηριστικά παραδείγματα αποτελούν οι εργασίες πάνω σε web caching techniques αλλά και σε web clipping. Αναφορικά με το web-caching, η ομάδα έχει παρουσιάσει εργασίας που αφορούν προεπεξεργασία rough data του διαδικτύου προκειμένου να γίνεται πιο εύκολα, γρήγορα και αξιόπιστα πλοήγηση στο διαδίκτυο, ενώ αναφορικά με τις υπηρεσίες web clipping η ομάδα έχει προτείνει μηχανισμούς εφαρμογής του web clipping σε σελίδες του διαδικτυου και συγκεκριμένα σε ενημερωτικούς δικτυακούς τόπους. Ξεκινώντας την αναζήτησή μας για το διαδίκτυο από τα παραπάνω και προσπαθώντας να εντοπίσουμε το πρόβλημα που υπάρχει σήμερα και να προτείνουμε τις δικές μας λύσεις καταλήξαμε σε ιδέες για ένα καθολικό σύστημα που θα μπορεί να ενσωματώσει πληθώρα χαρακτηριστικών και διαφορετικών ερευνητικών πεδίων προκειμένου να καταφέρουμε να ολοκληρώσουμε διαδικασίες οι οποίες και χρονοβώρες είναι αλλά και απαιτητικές ακόμα και για τους έμπειρους χρήστες του διαδικτύου. Έτσι λοιπόν, ξεκινώντας από το 2004 προσπαθήσαμε και αναπτύξαμε ένα ολοκληρωμένο σύστημα που επιτρέπει στους χρήστες να μπορούν να δουν ποιοτική πληροφορία που έχει συλλεχθεί από σελίδες του διαδικτύου. Για την ακρίβεια, δημιουργήσαμε ένα μηχανισμό που σα σκοπό έχει να λύσει όλα τα προαναφερθέντα προβλήματα μέσα από ένα και μοναδικό δικτυακό τόπο που παρουσιάζει προσωποποιημένη πληροφορία στους τελικούς χρηστες, πληροφορία που πηγάζει από το διαδίκτυο στο σύνολό του. Προκειμένου να επιτύχουμε αυτό είδαμε πως είναι απαραίτητο να γίνουν μία σειρά από διαδικασίες.

Αρχικά πρέπει να συλλέξουμε όλη την πληροφορία που θέλουμε να εμφανίζεται στους χρήστες του συστήματός μας. Σε αυτό το σημείο το σύστημα είναι πολύ περιοριστικό αλλά από την άλλη εξασφαλίζει data integrity. Η πληροφορία εισόδου ειναι αποκλειστικά και μόνο κανάλια επικοινωνίας του διαδικτύου που σε πρώτη φάση εμείς οι ίδιοι έχουμε συλλέξει από όλες τις πιθανές πηγές που ανακαλύψαμε και συνεχίζουμε να ανακαλύπτουμε. Εν συνεχεία, στα άρθρα που έχουμε συλλέξει εφαρμόζουμε τεχνικές κατηγοριοποίησης και εξαγωγής περίληψης προκειμένου να μπορούμε να έχουμε αρκετά στοιχεία για να παρουσιάσουμε στους χρήστες του συστήματος. Τέλος, η πληροφορία που έχουμε παρουσιάζεται ραφιναρισμένη στους χρήστες με τρόπο μεθοδικό και συστηματικό προκειμένου να εξασφαλίσουμε ποιοτική συλλογή κειμένων για κάθε χρήστη αλλά και να μπορούμε να διαμορφώνουμε με σαφή τρόπο το προφίλ του κάθε χρήστη.



Σχήμα 4.1: Αρχιτεκτονική του Συστήματος

4.2 Αρχιτεκτονική του peRSSonal

Η αρχιτεκτονική του συστήματος βασίζεται σε αυτόνομα υποσυστήματα τα οποία μπορούν να λειτουργήσουν ανεξάρτητα ωστόσο η ροή της πληροφορίας όπως είναι αναμενόμενο είναι σειριακή. Αυτό σημαίνει πως παρά το γεγονός πως κάθε μηχανισμός έχει δική του γενική είσοδο και έξοδο για την πληροφορία και συνεπώς μπορεί ανά πάσα στιγμή να εκτελεστεί και να παράγει αποτελέσματα, οι μηχανισμοί εντούτοις θα πρέπει να εκτελεστούν με συγκεκριμένη σειρά προκειμένου να έχουμε άμεσα το επιθυμητό αποτέλεσμα. Όπως έχει ήδη αναφερθεί για μηχανισμούς μικρής κλίμακας η αρχιτεκτονική ίσως δεν παίζει σημαντικό ρόλο, ωστόσο για μηχανισμούς πολύ μεγάλης κλίμακας όπως φαίνεται να έχει μετατραπεί ο μηχανισμός μας η αρχιτεκτονική παίζει πρωτεύοντα ρόλο. Έτσι, λοιπόν, αυτή τη στιγμή έχουμε καταλήξει για το μηχανισμό μας στην αρχιτεκτονική που παρουσιάζεται στο παρακάτω σχήμα. Σε γενικές γραμμές η αρχιτεκτονική είναι κλιμακωτή και πολυεπίπεδη.

Το σύστημά μας όπως είναι εμφανές αποτελείται από μία σειρά διαφορετικών συστημάτων τα οποία είναι χωρισμένα σε πολλά διαφορετικά επίπεδα ανάλογα με τη διαδικασία του συστήματος στην οποία εμπλέκονται και έτσι δημιουργείται ένα πολυεπίπεδο σύστημα που

μας βοηθά τόσο να το αναλύσουμε όσο και να το βελτιώνουμε κλιμακωτά χωρίς να αλλάζει η λειτουργία του μηχανισμού.

Αρχικά θα πρέπει να δούμε τα περιφερειακά και βοηθητικά στοιχεία του μηχανισμού τα οποία είναι ένας Internet Connection Manager ο οποίος είναι υπεύθυνος για τις συνδέσεις που πραγματοποιούνται προς σελίδες του διαδικτύου, ένας database Server και ένας file Server, οι οποίοι προφανώς χρησιμοποιούνται για να αποθηκεύουν πληροφορία. Αναφορικά με τον Internet Connection manager, αυτός λειτουργεί κυρίως σε επίπεδο πυρήνα του κώδικα και βοηθά στο να πραγματοποιούνται ότι συνδέσεις χρειάζεται το σύστημα προς τον «εξω κόσμο» από τα συστήματα πυρήνα που είναι το κέντρο του μηχανισμού και είναι απόλυτα διάφανα ως προς τις διαδικασίες τους ακόμα και προς εμάς τους ίδιους. Ο database server και ο file server λειτουργούν σαν τα βασικά μέσα αποθήκευσης της πληροφορίας, πέραν της μνήμης του συστήματος που αποθηκεύει προσωρινά, ωστόσο ο database server είναι το κεντρικό σημείο διασύνδεσης όλων των επιπέδων. Δεδομένου ότι κάθε επίπεδο του συστήματος μπορεί να δράσει αυτόνομα και ανεξάρτητα, αυτό συνεπάγεται πως είναι εφικτό κάθε επίπεδο να λειτουργήσει χωρίς στην ουσία να «ενδιαφέρεται» για κάθε άλλο επίπεδο, ούτε για την κατάσταση εκτέλεσης στην οποία βρίσκεται αλλά ούτε και για την πληροφορία που αυτός αναλύει. Η «διασύνδεση» μεταξύ των επιπέδων έχει επιλεγεί να γίνεται χωρίς τη χρήση ενός γενικευμένου system manager αλλά με σαφές data integrity στη βάση δεδομένων. Έτσι, λοιπόν, ενώ οι μηχανισμοί προσπελαύνουν την ίδια πληροφορία στη βάση δεδομένων, αφήνουν ο καθένας το δικό του ίχνος πάνω στην πληροφορία με αποτέλεσμα να είναι εφικτό να γνωρίζουμε σε τι στάδιο βρίσκεται η πληροφορία ανά πάσα στιγμή. Πρόκειται για μία γραμμή παραγωγής, ένα δομημένο σύστημα που βασίζεται απόλυτα στο μαρκάρισμα της πληροφορίας από τους μηχανισμούς οι οποίοι τη διατρέχουν και εφαρμόζουν αλλαγές σε αυτή. Ας δούμε όμως κάθε επίπεδο του συστήματος ξεχωριστά για να προχωρήσουμε σε ανάλυση της λειτουργίας του αλλά και σε ανάλυση της ροής πληροφορίας που υπάρχει για να γίνει πιο κατανοητός ο τρόπος με τον οποίο αναπτύσσουμε το σύστημά μας και παράγουμε το τελικό αποτέλεσμα.

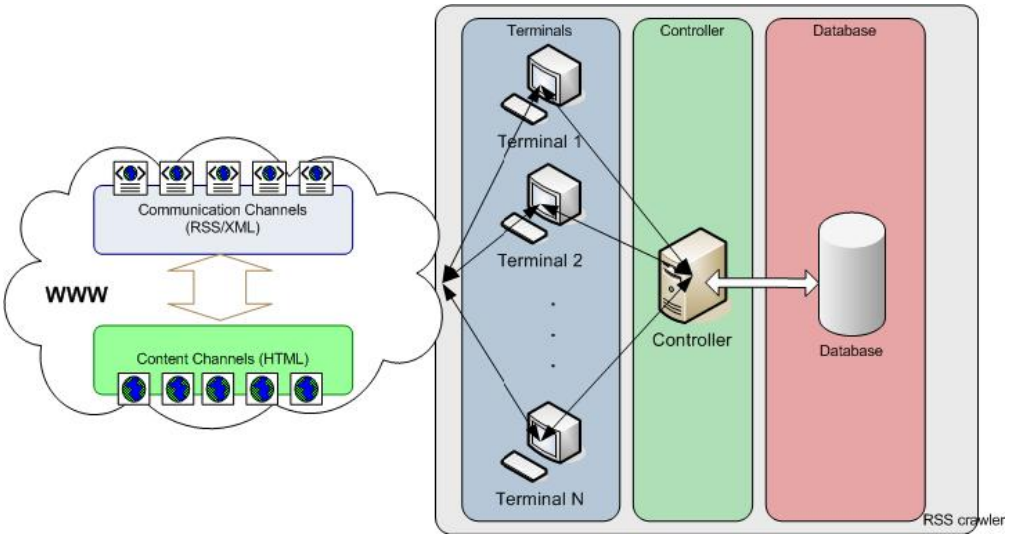
Σαν πρώτο επίπεδο του μηχανισμού μας συναντάμε το σύστημα AdvaRSS [40]. Το σύστημα αυτό είναι ένας mixed Crawler ο οποίος είναι σχεδιασμένος με τέτοιον τρόπο ώστε να λειτουργεί έχοντας σαν feed URLs κανάλια επικοινωνίας του διαδικτύου και μόνο. Στη συνέχεια και κατά τη διάρκεια ανάλυσης αυτού του συστήματος θα δούμε διεξοδικά τον τρόπο λειτουργίας. Επιγραμματικά, ο μηχανισμός αυτός ελέγχει τις λίστες από RSS feeds που διαθέτει το personal και εφόσον εντοπίσει κάποιο νέο άρθρο ή είδηση το οποίο δε βρίσκεται στη συλλογή του μηχανισμού, τότε μεταβαίνει στην HTML σελίδα και την κατεβάζει. Για το λόγο αυτό και ονομάζουμε τον advaRSS έναν mixed crawler, εφόσον δεν κατεβάζει απλά αλλά και αναλύει σελίδες προκειμένου να εντοπίσει αλλαγές. Παράλληλα με τον advaRSS και πάντα στο πρώτο επίπεδο του συστήματός μας βρίσκεται ο μηχανισμός CUTER [39] και πρόσφατα η νέα του έκδοση με όνομα mCUTER [42]. Πρόκειται για ένα σύστημα ανάλυσης HTML σελίδων και εξαγωγής του χρήσιμου κειμένου (CUTER) αλλά και των εικόνων (mCUTER) από αυτό. Στον αρχικό σχεδιασμό

του συστήματος τα advaRSS και CUTER λειτουργούσαν εντελώς ανεξάρτητα. Το AdvaRSS κατέβαζε τον HTML κώδικα και τον αποθήκευε στη βάση δεδομένων. Στην πορεία ο CUTER διάβαζε τον HTML κώδικα από τη ΒΔ και τον ανέλυε. Για λόγους ευκολίας αλλά και οικονομίας σε resources η νεότερη έκδοση του συστήματος περιέχει συνδυασμό των δύο συστημάτων τα οποία λειτουργούν και βασίζονται σε alerts. Όταν ο advaRSS εντοπίσει ένα νέο άρθρο, και αφού κατεβάσει τον HTML κώδικα, ειδοποιεί τον CUTER πως πρέπει να αναλύσει μία σελίδα. Έτσι ο CUTER αναλύει αμέσως τον κώδικα ο οποίος αποθηκεύεται απλώς προσωρινά στη μνήμη και όχι σε κάποιο κεντρικοποιημένο σημείο. Τα αποτελέσματα του CUTER αποθηκεύονται στη ΒΔ, και δεν είναι άλλο από το σώμα του άρθρου μαζί με τα URLs προς τις εικόνες του κειμένου. Το δεύτερο επίπεδο του συστήματος περιέχει αυτό που ονομάζουμε πυρήνα των διαδικασιών που πραγματοποιούνται από το personal. Αυτός ο χαρακτηρισμός δεν είναι τυχαίος καθότι σε αυτά τα επίπεδα γίνεται αφενός αυτό που ονομάζουμε «βρώμικη δουλειά» και από την άλλη τα συστήματα αυτά αποτελούν ένα μαύρο κουτί καθότι δεν έχουν καμία επικοινωνία με τον «έξω κόσμο» και κάθε επικοινωνία γίνεται μέσα από τον internet connection manager. Σε πρώτη φάση αυτό που είναι αναγκαίο να γίνει είναι η ανάλυση του κειμένου που έχει εξαχθεί από τον CUTER. Στο πλαίσιο αυτής της ανάλυσης πραγματοποιούμε προ-επεξεργασία κειμένου με όλα τα στάδια που αυτή περικλείει. Το χρήσιμο κείμενο όπως ονομάζουμε το σώμα του άρθρου που έχουμε συλλέξει περνά από μία σειρά σταδίων επεξεργασίας προκειμένου να εξάγουμε από αυτό ποιοτικά και ποσοτικά στοιχεία για να χρησιμοποιηθούν από τους μηχανισμούς που ακολουθούν (κατηγοριοποίηση, προσωποποίηση, διαμόρφωση προφίλ). Αυτό που μας ενδιαφέρει σε αυτό το σημείο είναι να μπορέσουμε να εξάγουμε ποιοτικές λέξεις κλειδιά από το κείμενο αλλά και ποσοτικά στοιχεία αυτών των λέξεων κλειδιών όπως συχνότητα εμφάνισης, σημεία εμφάνισης, κ.α. Σε αυτό το επίπεδο και για την εξαγωγή ποιοτικότερων αποτελεσμάτων χρησιμοποιείται τόσο ένα λεξικό (με αυτόματη διόρθωση λέξεων) αλλά και προκαθορισμένες λίστες λέξεων κλειδιών που θεωρούνται άνευ ουσίας για το σύστημα (stopword lists). Παράλληλα, η λειτουργία του συστήματος είναι τέτοια που για την Αγγλική γλώσσα δε θεωρεί χρήσιμα τα σημεία στίξης. Μόλις πρόσφατα προστέθηκε η λειτουργία ανάλυση ελληνικών κειμένων όπου εκτός από εξαγωγή των λέξεων κλειδιών πραγματοποιείται και αναλυτική λεξικολογική ανάλυση κειμένου με χαρακτηρισμό των λέξεων – κλειδιών βάσει της γραμματικής του υπόστασης. Σε αυτή την περίπτωση και σε αυτό το στάδιο δεν εφαρμόζονται τα στάδια αφαίρεσης των σημείων στίξης ενώ παράλληλα δεν εφαρμόζουμε κανένα stopword list για την ελληνική γλώσσα κυρίως για πειραματικούς λόγους προκειμένου να εξαχθεί πληροφορία για την ποιότητα των πρώτων αποτελεσμάτων [38].
Περνώντας σε τρίτο επίπεδο και παραμένοντας ακόμα στον πυρήνα του συστήματος συναντάμε δύο συστήματα τα οποία προχωρούν σε σύνθεση πληροφορίας καθότι μέχρι αυτό το στάδιο είχαμε σαφώς αποδόμηση της πληροφορίας που έχουμε συλλέξει [64]. Ο πρώτος μηχανισμός που συναντάμε είναι ο μηχανισμός κατηγοριοποίησης ενώ στο ίδιο επίπεδο λειτουργεί και ο μηχανισμός εξαγωγής περίληψης κειμένου, δύο μηχανισμοί οι οποίοι κάτω από κατάλληλες συνθήκες λειτουργούν παράλληλα και επικουρικά ο ένας του άλλου. Ο μηχανισμός κατηγοριοποίησης

πληροφορίας βασίζεται σε αλγορίθμους συσχέτισης πληροφορίας και η πληροφορίας που εξάγει είναι η πιθανοτική συσχέτιση ενός κειμένου με μία κατηγορία. Στην ουσία για το μηχανισμό κατηγοριοποίησης έχουμε αρχικοποίηση των κατηγοριών του συστήματος με training set documents τα οποία έχουμε προκατηγοριοποιήσει. Στη συνέχεια, για κάθε κείμενο που εισέρχεται στο σύστημα εφαρμόζουμε αλγορίθμους συσχέτισης για να βρούμε την πιθανότητα να ανήκει σε μία ή περισσότερες κατηγορίες του συστήματός μας. Μάλιστα, εφόσον κάποιο κείμενο ξεπεράσει ένα όριο συσχέτισης με μία κατηγορία τότε γίνεται και αυτό κομμάτι του training set και συνεπώς μιλάμε για συνεχώς μεταβαλλόμενες κατηγορίες συστήματος οι οποίες διαμορφώνονται από την είσοδο που λαμβάνει το σύστημα. Όπως ήδη αναφέραμε, το σύστημα διαθέτει και μηχανισμό αυτόματης εξαγωγής περίληψης κειμένου ο οποίος βασίζεται σε αλγορίθμους ανάλυσης και αξιολόγησης των προτάσεων των κειμένων. Για την ακρίβεια είναι ένας μηχανισμός που πραγματοποιεί σύνθεση ενός κειμένου από προτάσεις του αρχικού κειμένου χρησιμοποιώντας βάρη που αποδόθηκαν στις προτάσεις του κειμένου από τα ποιοτικά και ποσοτικά χαρακτηριστικά που εξήγαγε το επίπεδο 2 στο οποίο έχουμε την προεπεξεργασία του κειμένου.
Τέλος, περνώντας σε τέταρτο επίπεδο, συναντάμε μία σειρά από συστήματα τα οποία και σε αυτό το σημείο έχουν δημιουργηθεί κλιμακωτά ωστόσο λειτουργούν όλα μαζί προκειμένου να παραχθεί το τελικό αποτέλεσμα. Αυτό το επίπεδο άλλωστε αποτελεί τη διασύνδεση με τους χρήστες του συστήματος και ως εκ τούτου έχει δοθεί μεγάλη προσοχή στο σχεδιασμό. Μία σειρά από διαφορετικά συστήματα, όλο υλοποιημένα σε επίπεδο web application, είναι υπεύθυνα για τη δημιουργία του layout του personalized portal, για την εμφάνιση των άρθρων του συστήματος, για την εμφάνιση του tagging αλλά και των συναφών άρθρων, και το βασικότερο για τη δημιουργία του προφίλ των χρηστών του συστήματος [58]. Όλα τα παραπάνω λειτουργούν συνολικά και όχι μεμονωμένα καθότι το σύστημα λειτουργεί σε επίπεδο διαδικτύου και όλες οι λειτουργίες γίνεται παράλληλα. Δεδομένου ότι βρισκόμαστε σε κατάσταση που το web2.0 βρίσκεται παντού στο διαδίκτυο, η τελευταία έκδοση του συστήματος η οποία και θα παρουσιαστεί μέσα από αυτή την εργασία ακολουθεί πιστά τις επιταγές της τεχνολογίες και βασιζεται αποκλειστικά και μόνο σε ασύγχρονη επικοινωνία με τον κεντρικό server και σε χρήση αντικειμενοστραφούς προγραμματισμού για τη διατήρηση και ανάλυση στοιχείων [65]. Όπως θα δούμε και στην πορεία, το σύστημα αυτό τείνει να αποτελέσει κομμάτι του κεντρικού πυρήνα του συστήματος καθότι η προσπάθεια που γίνεται συνολικά από το σύστημα είναι ανθρωποκεντρική και προκειμένου ο χρήστης να γίνει κομμάτι των διαδικασιών του συστήματος είναι καλό να μπορούμε να εμπλακούμε σε διαδικασίες του συστήματος σε πραγματικό χρόνο.

4.3 Ροή Πληροφορίας

Στη συνέχεια θα παρουσιάσουμε μία μικρή ανάλυση για κάθε σύστημα προκειμένου να υπάρχει γενική γνώση της ροής πληροφορίας στο σύστημα προκειμένου να είναι σαφής



Σχήμα 4.2: Αρχιτεκτονική του μηχανισμού advaRSS

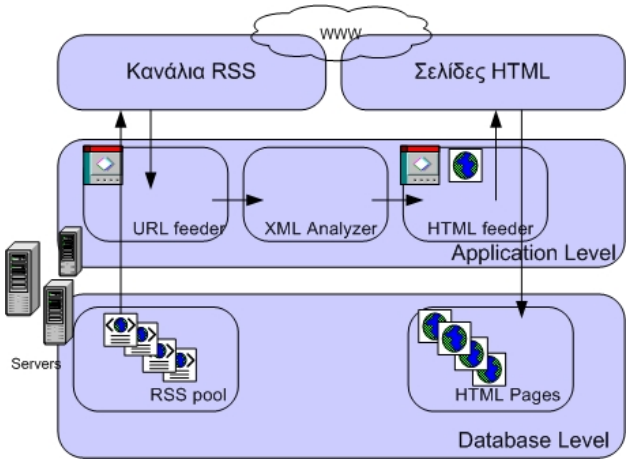
ο λόγος για τον οποίο πραγματοποιούνται οι οποίες διαδικασίες υπάρχουν για την αλγοριθμική ανάλυση.

4.3.1 Υποσύστημα advaRSS

Ο μηχανισμός advaRSS είναι κατασκευασμένος βάσει της αρχιτεκτονικής του σχήματος 4.2. Διαθέτει έναν κεντρικό controller, για τον έλεγχο της διαδικασίας που πραγματοποιείται, ο οποίος αναθέτει εργασίες στα διαφορετικά τερματικά του συστήματος.

Η πληροφορία αποθηκεύεται κεντρικά στη βάση δεδομένων ενώ τα τερματικά είναι σχεδιασμένα για να κάνουν δύο διαφορετικά είδη crawling: RSS και HTML crawling. Στην ουσία πρόκειται για έναν mixed crawler καθότι από τη μία διαθέτει αρχικό feed URL το οποίο είναι αποκλειστικά και μόνον feeds ενώ η πληροφορία που εξάγει από αυτά (normal URLs) επανέρχεται σαν είσοδος στο σύστημα το οποίο "κατεβάζει" HTML σελίδες. Το σχήμα 4.3 παρουσιάζει τη ροή πληροφορίας του μηχανισμού.

Όπως είναι εμφανές και από το παραπάνω διάγραμμα, σε πρώτη φάση ο μηχανισμός δέχεται σαν είσοδο κανάλια επικοινωνίας από το RSS pool και μετά από ανάλυση του XML κώδικα των feeds προκύπτουν link προς σελίδες HTML οι οποίες περιέχουν άρθρα που δεν έχει ανακτήσει ακόμα το σύστημα. Στη συνέχεια το σύστημα λαμβάνει σαν είσοδο τα URL από HTML σελίδες που περιέχουν άρθρα και αποθηκεύει τον HTML κώδικα στη Βάση Δεδομένων. Σε αυτό το σημείο θα πρέπει να τονίσουμε πως ο μηχανισμός έχει να αντιμετωπίσει μία σειρά από

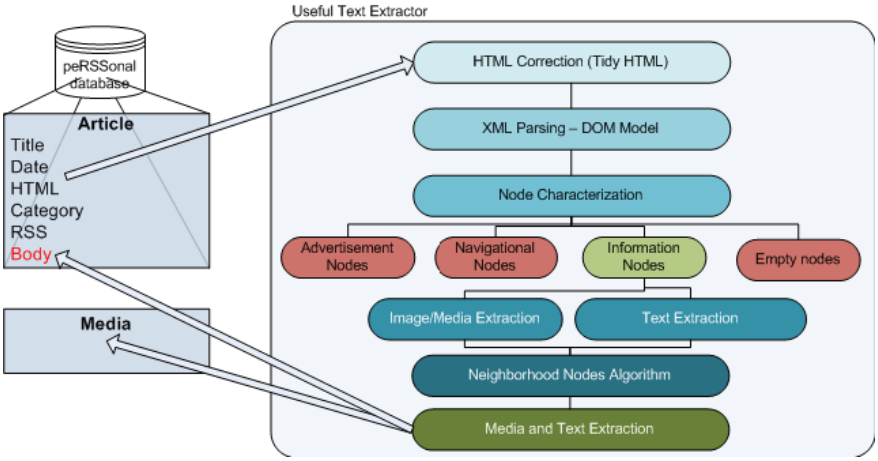


Σχήμα 4.3: Ροή Πληροφορίας του μηχανισμού advaRSS

σημαντικά ζητήματα στην προσπάθεια ανάκτησης HTML σελίδων από το διαδίκτυο τα οποία περιλαμβάνουν άρθρα και ειδήσεις. Τα βασικά προβλήματα του μηχανισμού είναι:

- τα άρθα στο Διαδίκτυο προκύπτουν κάθε στιγμή της ημέρας (θεωρητικά) και με τυχαίο τρόπο. Ο μηχανισμός όμως θα πρέπει να μένει πάντα επικαιροποιημένος με άρθρα.
- η εισοδος μπορεί να είναι πολύ μεγάλη με αποτέλεσμα να πρέπει να υπάρχει κάποια προτεραιότητα στα RSS feeds που λαμβάνει ο μηχανισμός
- ο μηχανισμός θα πρέπει να είναι επικαιροποιημένος αλλά δεν είναι "ευγενικό" προς τις σελίδες να γίνεται συνεχώς αναζήτηση αν έχουν προκύψει νέα άρθρα.
- η έρευνα μας δείχνει πως υπάρχουν συγκεκριμένα patterns σύμφωνα με τα οποία προκύπτουν άρθρα και ειδήσεις σε ένα συγκεκριμένο δικτυακο τόπο.

Από τα παραπάνω οδηγούμαστε στο γεγονός πως θα πρέπει να εφαρμόζεται μία συγκεκριμένη πολιτική αναφορικά με τον τρόπο με τον οποίο ελέγχουμε τα feed URLs (RSS feeds) για να πετύχουμε μέγιστη απόδοση στο σύστημά μας, ελάχιστη επιβάρυνση στους πόρους του δικτύου, ελάχιστη επιβάρυνση στους πόρους των server και τέλος ελάχιστο χρόνο ανάκτησης άρθρου από τη στιγμή που αυτό προέκυψε.
Ο μηχανισμός ανάκτησης HTML σελίδων εξάγει και άλλη πληροφορία που καθώς θα δούμε στην πορεία είναι πολύ χρήσιμη και για αλγοριθμικές διαδικασίες του μηχανισμού. Έτσι, εκτός από τον HTML κώδικα μίας σελίδας, ο μηχανισμός αυτός είναι σε θέση να κάνει εξαγωγή του τίτλου ενός άρθρου, της ημερομηνίας ανάρτησης του άρθρου, τη γλώσσα στην οποία είναι γραμμένο το άρθρο και τέλος κάποια μεταδεδομένα που υπάρχουν χειροκίνητα στο σύστημα όπως για παράδειγμα η κατηγορία στην οποία ανήκει το άρθρο (πολιτική, οικονομια, κ.α.).



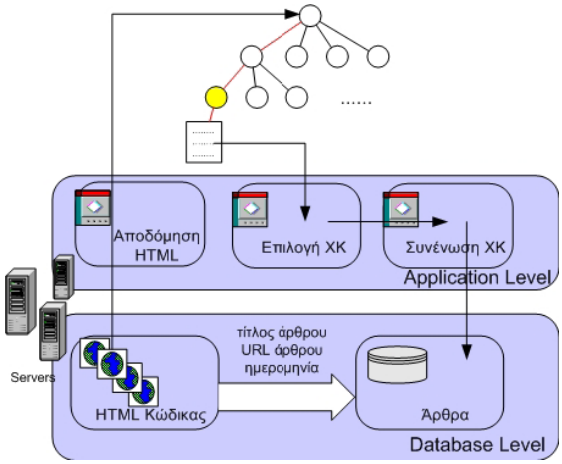
Σχήμα 4.4: Αρχιτεκτονική του μηχανισμού mCuter

4.3.2 Υποσύστημα mCUTER

Ο μηχανισμός cuter είναι υπεύθυνος για την εξαγωγή του χρήσιμου κειμένου από τις HTML σελίδες. Σαν χρήσιμο κείμενο στο δικό μας σύστημα εννοούμε αποκλειστικά και μόνο το κείμενο ενός άρθρου μαζί με ότι εικόνες το ακολουθούν. Αρχικά το σύστημα είχε δημιουργηθεί για να εξάγει κείμενο, ωστόσο στην πορεία εμπλουτίστηκε για να μπορεί να εξάγει και όλες τις εικόνες που ακολουθούν το κείμενο. Στο σχήμα 4.4 φαίνεται η αρχιτεκτονική του μηχανισμού m-CUTER.

Όπως φαίνεται και από το σχήμα, ο μηχανισμός παίρνει σαν είσοδο τον HTML κώδικα μίας σελίδες που περιέχει ένα άρθρο. Στην πορεία πραγματοποιούνται μία σειρά από διαδικασίες και συστήματα που είναι ενσωματωμένα στο μηχανισμό προκειμένου να γίνει εξαγωγή του κειμένου/σώματος του άρθρου αλλά και των εικόνων.

Σύμφωνα με τις διαδικασίες του συστήματος αυτού, σε πρώτη φάση το σύστημα καθαρίζει τον HTML κώδικα που λαμβάνει σαν είσοδο. Όπως έχει παρατηρηθεί στο διαδίκτυο μεγάλο μέρος των σελίδων περιέχουν πολλά λάθη κάτι το οποίο μπορεί να δημιουργήσει προβλήματα τόσο στους browsers που υπάρχουν όσο και σε μηχανισμούς που επιχειρούν αναλύσεις κώδικα. Μάλιστα έχουμε φτάσει σε σημείο να λέμε για sites που είναι ή δεν είναι cross-browser ή browser compatible. Για τη διόρθωση του HTML κώδικα χρησιμοποιείται το opensource εργαλείο με την ονομασία HTML Tidy. Πρόκειται για έναν HTML validator που είναι κατασκευασμένος για να διορθώνει σφάλματα που μπορεί να υπάρχουν στις σελίδες. Ο βασικός του σκοπός είναι να βοηθήσει τους κατασκευαστές δικτυακών τόπων να μπορούν να βρίσουν σφάλματα στην ανάπτυξη που κάνουν αλλά στην περίπτωσή μας είναι το εργαλείο που χρειαζόμαστε για την ανάλυση του HTML κώδικα και τη διόρθωση οποιων λαθών. Στη συνέχεια αναλύουμε τη

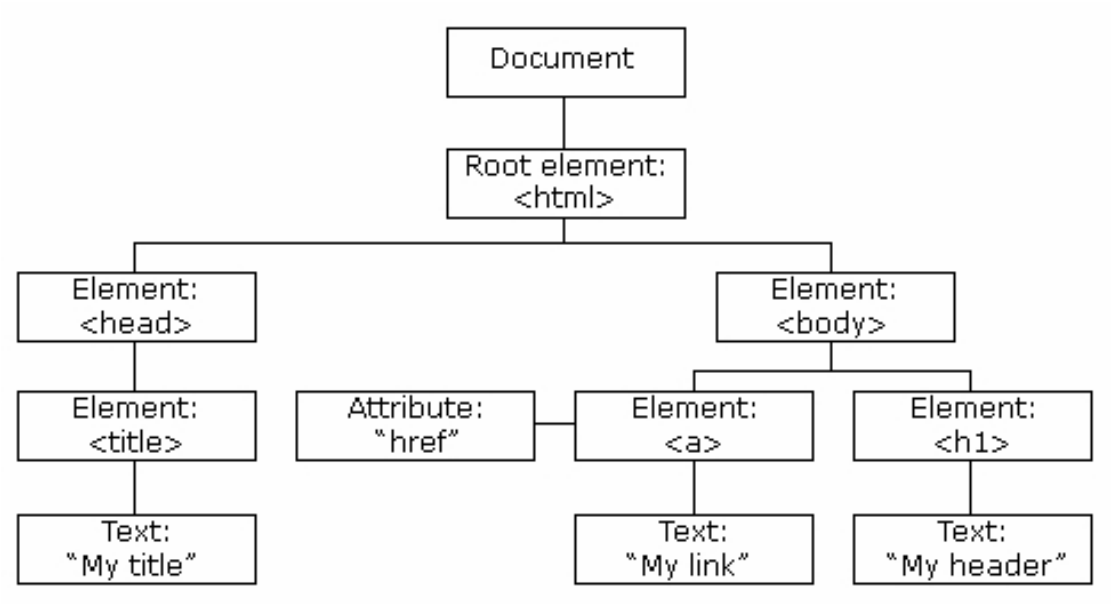


Σχήμα 4.5: Διάγραμμα ροής του μηχανισμού mCuter

σελίδα στο DOM μοντέλο της. Αυτό μας βοηθάει προκειμένου να μετατρέψουμε την HTML σελίδα σε ένα δέντρο από κόμβους και φύλλα, όπου τα φύλλα είναι συνήθως διεπαφή με το χρήστη (κείμενο, multimedia, links, κ.α.). Στη συνέχεια και ακολουθώντας έναν απλό αλγόριθμο πραγματοποιούμε χαρακτηρισμό των κόμβων του δέντρου. Σύμφωνα με το μοντέλο μας οι κόμβοι μπορεί να είναι:

- navigation
- advertisement
- information
- empty

Ως advertisement χαρακτηρίζουμε όλους αυτούς τους κόμβους που περιέχουν κάποια διαφημιστική πληροφορία. Ομολογουμένως, αρχικά δεν είχαμε σκοπό να συμπεριλάβουμε τέτοιους κόμβους, ωστόσο όπως θα δούμε και στην αλγοριθμική ανάλυση του συστήματός μας αυτός ο χαρακτηρισμός μας βοηθά στο να εντοπίσουμε το χρήσιμο κείμενο. Οι navigation κόμβοι είναι αυτοί που περιέχουν πληθώρα από intra και inter domain links, δηλαδή συνδέσμους προς το ίδιο το site ή και συνδέσμους προς άλλους δικτυακούς τόπους. Οι άδειοι κόμβοι είναι επίσης ένα κομμάτι το οποίο παρατηρείται στις μοντέρνες σελίδες και αξίζουν αναφοράς διότι οι άδειοι κόμβοι έχουν συχνά λόγο ύπαρξης για γραφιστικούς λόγους και έτσι θα πρέπει να τους επισημάνουμε και πάλι για υποστήριξη του αλγορίθμου του συστήματος. Τέλος, και το πιο σημαντικό κομμάτι των κόμβων του μηχανισμού καθότι από αυτούς εξάγεται το χρήσιμο κείμενο είναι οι κόμβοι information. Αυτοί οι κόμβοι είναι φύλλα τα οποία έχουν κείμενο ή μορφοποιημένο κείμενο ή ακόμα και multimedia. Κομμάτι αυτών των κόμβων είναι και οι κόμβοι που

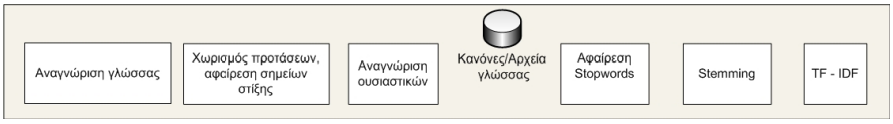


Σχήμα 4.6: DOM μοντέλο

περιέχουν το κείμενο του άρθρου. Μετά το χαρακτηρισμό των κόμβων και βάσει αλγορίθμου μπορούμε να δούμε ποιοι από τους κόμβους είναι πράγματι το σώμα του κειμένου. Αυτό εξάγεται μετά από μελέτη των χαρακτηριστικών που έχουν οι σελίδες των άρθρων αλλά και βάσει αλγορίθμου. Παράλληλα, σε αυτά συνεπικουρεί και το γεγονός ότι έχουμε γνώση τουσημείου έναρξης του άρθρου καθότι από το RSS feed έχουμε ήδη την πληροφορία για τον ακριβή τίτλο του άρθρου.

4.3.3 Υποσύστημα Προ-Επεξεργασίας

Περνώντας στις διαδικασίες πυρήνα του συστήματος μας, που όπως έχει ήδη αναφερθεί ονομάζουμε αυτές οι οποίες στην ουσία δεν έχουν καμία επικοινωνία με τον έξω κόσμο αλλά προετοιμάζουν την πληροφορία για προβολή προς τους χρήστες, βρίσκουμε αρχικά το μηχανισμό προ-επεξεργασίας κειμένου. Σύμφωνα με τις ως τώρα διαδικασίες έχουμε φτάσει έως το σημείο να εξάγουμε το χρήσιμο κείμενο από άρθρα. Ο επόμενος μηχανισμός θα πρέπει να κάνει μία σειρά αναλύσεων στο κείμενο προκειμένου να είμαστε έτοιμοι να εφαρμόσουμε τους αναλυτικούς αλγορίθμους κατηγοριοποίησης, εξαγωγής περίληψης και προσωποποιημένης προβολής πληροφορίας. Εξ αιτίας του γεγονότος ότι θα ακολουθήσουν διαδικασίες γλωσολογικής ανάλυσης του κειμένου, αυτό σημαίνει πως σε πρώτο βήμα θα πρέπει να εντοπίσουμε τη γλώσσα στην οποία είναι γραμμένο το κείμενο. Παρά το γεγονός ότι η πλειοψηφία των διαδικασιών είναι language independant, κάποιες κρίσιμες είναι language specific και συνεπώς η αναγνώριση



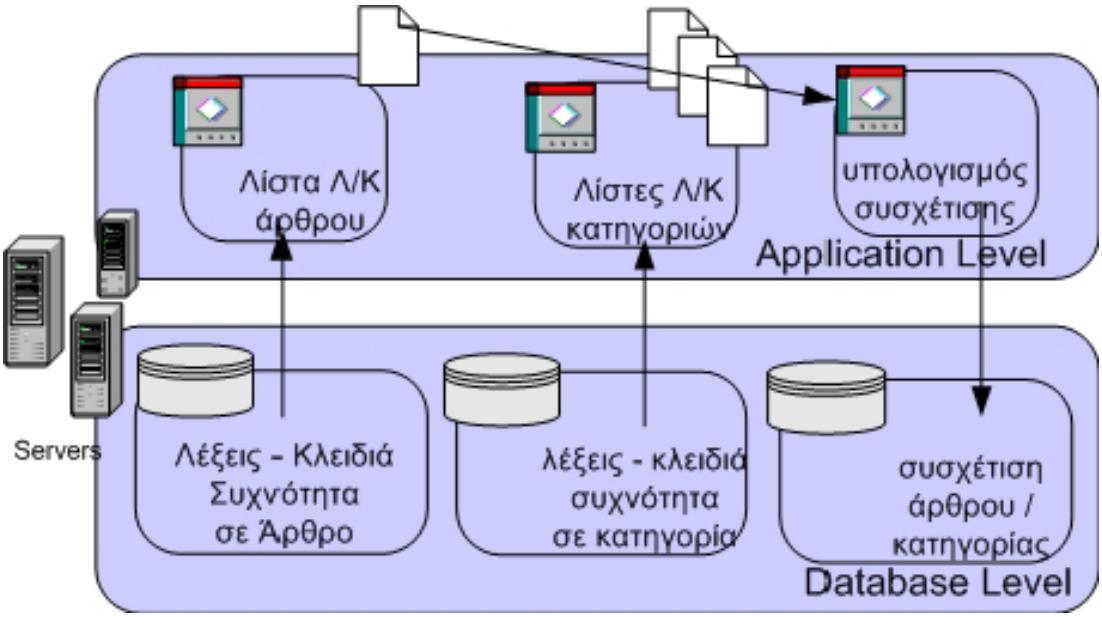
Σχήμα 4.7: Προ-Επεξεργασία και Ανάλυση Κειμένου

της γλώσσας είναι ένα σημαντικό βήμα. Ακολούθως και σύμφωνα με παραμετροποιημένα δεδομένα που δίνονται στο σύστημα ακολουθεί η διαδικασία πρώτης ανάλυσης του κειμένου. Οι παράμετροι οι οποίες μας ενδιαφέρουν είναι:

- το ελάχιστο μήκος λέξης
- τα σημεία στίξης
- λίστες από τετριμμένες λέξεις (stopwords)

Σε πρώτη φάση και ανάλογα με τις παραμέτρους, αφαιρούμε όλες τις λέξεις κάτω από ένα συγκεκριμένο μήκος, μετατρέπουμε όλο το κείμενο σε μικρά και αφαιρουμε όλα τα σημεία στίξης. Εν συνεχεία, ξεκινά η διαδικασία tagging του κειμένου. Σε μερικές περιπτώση η διαδικασία tagging μπορεί να γίνει μόνο εφόσον υπάρχουν όλα τα σημεία στίξης και ετσι οι δύο αυτές διαδικασίες μπορεί να γίνουν αντίστροφα. Στην πορεία, και ανάλογα με τη γλώσσα έχουμε τα language specific αφαίρεση όλων των stopwords και πραγματοποίηση της διαδικασίας stemming, της αφαίρεσης δηλαδή των καταλήξεων από τις λέξεις. Στο τέλος της παραπάνω διαδικασίας έχουμε καταφέρει να μετατρέψουμε το κείμενο που έχουμε σε λέξεις κλειδιά. Η ανάλυση συνεχίζεται περαιτέρω με ποσοτικά στοιχεία που είναι απαραίτητα για το μηχανισμό μας. Αυτά σχετίζονται με τη συχνότητα των λέξεων κλειδιών, με το σημείο εμφάνισής τους μέσα στο κείμενο κ.α. Συνοπτικά η έξοδος που έχουμε είναι:

- οι λέξεις κλειδιά που προέκυψαν από την ανάλυση του κειμένου
- οι θέσεις των λέξεων κλειδιών μέσα στο κείμενο, οι προτάσεις δηλαδή στις οποίες συναντάμε κάθε λέξη,
- το πλήθος καθενός keyword μέσα στο κείμενο τόσο ως απόλυτη όσο και ως σχετική συχνότητα. Αν και το ένα μπορεί να προκύψει από το άλλο, σε αυτό το σημείο μπορούμε να το υπολογίσουμε εύκολα καθότι η πληροφορία είναι στη μνήμη του συστήματος και τέλος
- ο χαρακτηρισμός/tagging της κάθε λέξης κλειδί. Η πληροφορία είναι απαραίτητη καθότι μελέτες έχουν δείξει πως τα ουσιαστικά θα πρέπει να υπολογίζονται με διαφορετικό τρόπο από άλλες λέξεις διότι προσφέρουν με πιο σημαντικό τρόπο στο νόημα ενός κειμένου.



Σχήμα 4.8: Κατηγοριοποίηση Κειμένου

4.3.4 Υποσύστημα Κατηγοριοποίησης Πληροφορίας

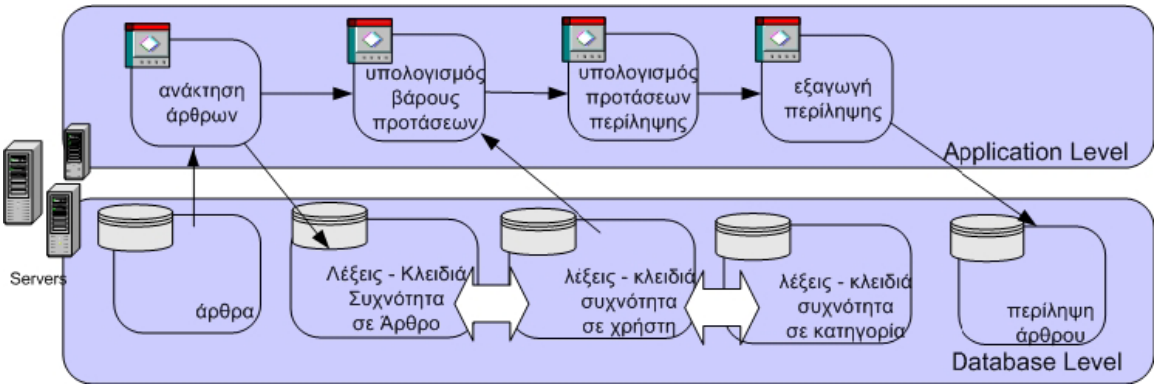
Περνώντας σε πιο περίπλοκες διαδικασίες πυρήνα, η πρώτη διαδικασία που συναντάμε είναι η διαδικασία κατηγοριοποίησης κειμένου. Αν και όπως είδαμε από τη διαδικασία ανάκτησης σελίδων η κατηγορία στην οποία ανήκει ένα κείμενο υπάρχει σαν μεταδεδομένο από την ανάκτηση πληροφορίας δε θα πρέπει να αμελούμε ότι ένα κείμενο μπορεί να ανήκει σε περισσότερες της μίας κατηγορίες. Αυτή η πληροφορία σε συνδυασμό με το γεγονός ότι το σύστημά μας στην ουσία παράγει πιθανοτικά στοιχεία για την ένταξη ενός κειμένου σε μία κατηγορία κάνει τη διαδικασία της κατηγοριοποίησης ένα πολύ σημαντικό κομμάτι του μηχανισμού. Όπως βλέπουμε και από το σχήμα, η αρχιτεκτονική του συστήματος είναι απλή όπως και η ροή πληροφορίας σε αυτή.

Για κάθε κείμενο που δεν έχει κατηγοριοποιηθεί λαμβάνουμε πληροφορία για λέξεις κλειδιά του κειμένου και τη συχνότητά τους καθώς και ότι άλλη ποσοτική και ποιοτική πληροφορία μπορεί να διαθέτουμε. Παράλληλα, δημιουργούμε (συνήθως διαθέτουμε στη μνήμη) λίστες με τις αντίστοιχες λέξεις κλειδιά κάθε κατηγορίας. Έχοντας τις λίστες αυτές είναι πολύ εύκολο και απλό με τον αλγόριθμο συσχέτισης συνημιτόνου να υπολογίσουμε την πιθανότητα με την οποία μπορεί το άρθρο να ανήκει σε κάποια κατηγορία. Αυτό που συχνά παρατηρείται είναι η κατηγορία στην οποία εντάσσεται το άρθρο να είναι ίδια με αυτή που ήδη γνωρίζουμε από τα μεταδεδομένα που έχουμε. Ωστόσο, υπάρχουν πολλές περιπτώσεις που διαφαίνεται πως το άρθρο ανήκει και σε άλλη κατηγορία, μία ή και περισσότερες. Σε αυτό το σημείο αξίζει να

πούμε πως τα στοιχεία από τη διαδικασία κατηγοριοποίησης είναι άκρως χρήσιμα για άλλες διαδικασίες του μηχανισμού οι οποίες μελετώνται και εντάσσονται σιγά σιγά συνολικά στο σύστημα. Χαρακτηριστικό παράδειγμα αποτελεί το υποσύστημα εύρεσης trash article. Σύμφωνα με μελέτη που κάναμε στα άρθρα του συστήματός μας, προκύπτουν κάποια άρθρα που φαίνεται να κατηγοριοποιούνται με κάποια μικρό συχνά ποσοστό (<10%) σε περισσότερες των τριών κατηγορίες. Έχουμε παρατηρήσει πως σχεδόν με πιθανότητα 1 αυτά τα άρθρα είναι αυτό που ονομάζουμε στο σύστημα trash articles δηλαδή άρθρα τα οποία είτε δεν είχαν κείμενο, είτε είχαν ελάχιστο κείμενο και άρα δε διαθέτουν πληροφορία, είτε είχε πραγματοποιηθεί σφάλμα στην εξαγωγή χρήσιμου κειμένου.
Ο μηχανισμός κατηγοριοποίησης λοιπόν είναι ένα σύστημα που βοηθά να εντοπίσουμε στοιχεία για τα άρθρα και ειδήσεις που έχουμε στο σύστημα και να προσθέσει μεταδεδομένα στην πληροφορία τα οποία θα βοηθήσουν στην πορεία σε πολλές διαδικασίες που έχουμε κατά την παρουσίαση πληροφορίας στο χρήστη. Τέλος, δε θα πρέπει να αμελήσουμε το γεγονός πως είναι δυνατόν να υπάρξουν άρθρα, και δεν είναι λίγα, τα οποία δεν είναι κατηγοριοποιημένα ή δεν είναι κατηγοριοποιημένα βάσει των κατηγοριών που έχουμε. Χαρακτηριστικό παράδειγμα τέτοιων άρθρων είναι αυτά τα οποία προκύπτουν από blogs ή από δικτυακούς ενημερωτικούς τόπους οι οποίοι δεν είναι μεγάλης εμβέλειας και άρα δε διαθέτουν ενιαίο τρόπο παρουσίασης και προβολής της πληροφορίας.
Όπως ήδη αναφέρθηκε ο βασικός τρόπος κατηγοριοποίησης των κειμένων ή έστω ο τρόπος ανάθεσης ενός κειμένου σε μία κατηγορία βασίζεται στη συσχέτιση συνημιτόνου. Και όπως έχει ήδη αναφερθεί ένα κείμενο δεν είναι απαραίτητο ότι θα ανατεθεί σε μία κατηγορία, απλώς θα βρεθεί ένας βαθμός συσχέτισης του κειμένου με κάθε κατηγορία.

4.3.5 Υποσύστημα Αυτόματης Εξαγωγής Περίληψης Κειμένου

Η εξαγωγή περίληψη είναι μία διαδικασία η οποία έχει σα σκοπό να δημιουργήσει ένα κομμάτι κειμένου αντιπροσωπευτικού του αρχικού κειμένου που λαμβάνουμε με τη διαδικασία εξαγωγής χρήσιμου κειμένου. Στην ουσία αυτό που μας οδήγησει να σκεφτούμε τη δημιουργία περίληψης είναι η πηγή των άρθρων μας δηλαδή τα RSS feeds. Όπως γνωρίζουμε μέσα στην πληροφορία που μπορεί να υπάρχει στα RSS feeds είναι εύκολο να τοποθετηθεί και το αμιγές κείμενο ενός άρθρου ή και ένα κομμάτι του άρθρου. Φυσικά, καταλαβαίνει κανείς πως αν οι δικτυακοί τόποι τοποθετούσαν κομμάτια κειμένου μέσα στα RSS feeds τους θα δημιουργείτο ένα τεράστιο πρόβλημα. Η επισκεψιμότητα του δικτυακού τόπου που έχει τα άρθρα θα ήταν ελάχιστη καθότι ο καθένας θα διάβαζε τα άρθρα από τον RSS reader και δε θα τον απασχολούσε να επισκεφθεί το δικτυακό τόπο. Βεβαίως, αυτός ήταν και ο αρχικός σκοπός δημιουργίας των RSS feeds, όμως, προφανώς οικονομικοί λόγοι οδήγησαν σε εναλλακτικούς τρόπους χρήσης. Το σύστημα που αναπτύσσουμε έχει σαν σκοπό να απεμπλέξει το χρήστη από κάθε διαδικασία επί-

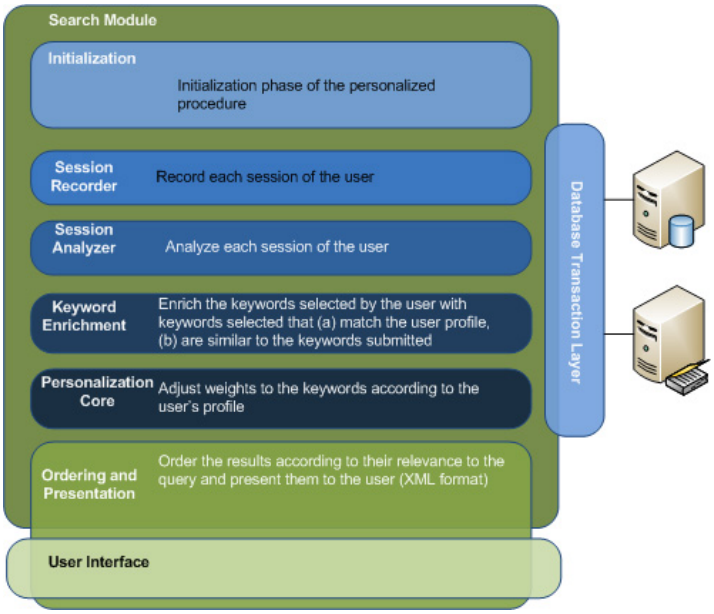


Σχήμα 4.9: Αυτόματη Εξαγωγή Περίληψης

σκεψης σε δικτυακούς τόπους και να επικεντρωθεί αποκλειστικά και μόνο στην ανάγνωση των άρθρων που τον ενδιαφέρουν. Συνεπώς, η παερίληψη που παράγουμε έχει σα σκοπό να παίξει το ρόλο του κειμένου που λείπει από τα RSS feeds, καθότι όπως θα δούμε και στην πορεία το σύστημά μας δημιουργεί και RSS feeds για τους χρήστες.
Ας επίστρέψουμε όμως στο σύστημα εξαγωγής περίληψης το οποίο και βλέπουμε στο σχήμα.

Ο τρόπος με τον οποίο παρασκευάζουμε την περίληψη του κειμένου δε βασίζεται σε γλωσσικές αναλύσεις αλλά στην επιλογή αυτών των προτάσεων του κειμένου που θα μπορούσαν να είναι αντιπροσωπευτικές του νοήματος που αυτό έχει. Για να το πετύχουμε αυτό θα πρέπει να δώσουμε ένα βάρος σε κάθε πρόταση του κειμένου. Ο αλγόριθμος που θα δούμε σε επόμενο κεφάλαιο χρησιμοποιεί ευρεστικές μεθόδους για αν αποδώσει κάποιο βάρος στις προτάσεις του κειμένου και με αυτό τον τρόπο να εξαχθεί η περίληψη. Αρχική η περίληψη του κειμένου είχε οριστεί να έχει προκαθορισμένο μέγεθος το οποίο βέβαια είχε άμεση εξάρτηση με το μήκος του αρχικού κειμένου. Ωστόσο, στην πορεία είδαμε πως είναι πολύ χρήσιμο να διαθέτουμε ολόκληρο το κείμενο ανακατασκευασμένο σύμφωνα με τα βάρη που διαθέτουν οι προτάσεις. Με αυτό τον τρόπο μπορούμε εύκολα σε επίπεδο εφαρμογής να προσαρμόσουμε το μέγεθος του κειμένου που είναι ορατό προς τους χρήστες.
Έτσι η τελική έξοδος που έχουμε είναι το κείμενο ανακατασκευασμένο από τις ίδιες του τις προτάσεις οι οποίες έχουν τοποθετηθεί σε φθίνουσα σειρά βάσει του σκορ που έχει η κάθε μία. Η βαθμολόγηση των προτάσεων όπως θα δούμε και σε ανάλυση που θα ακολουθήσει βασίζεται στις εξής παραμέτρους:

- αν λέξεις κλειδιά από τις προτάσεις υπάρχουν και στον τίτλο του άρθρου,
- αν η λέξη κλειδί είναι ουσιαστικό
- αν υπάρχει πληροφορία για την κατηγορία στην οποία ανήκει το κείμενο. (αν λέξεις κλειδιά μίας πρότασης είναι αντιπροσωπευτικές της κατηγορίας που ανήκει το κείμενο, τότε



Σχήμα 4.10: Σύστημα Παρουσίασης Πληροφορίας στο Χρήστη

θεωρητικά η πρόταση αυτή είναι πολύ σημαντική για το ίδιο το κείμενο)

- αν υπάρχει πληροφορία για τις προτιμήσεις χρηστών η συνόλου χρηστών.

Τα στοιχεία που συνθέτουν τα παραπάνω θα δούμε διεξοδικά κατά τη διαδικασία ανάλυση των αλγορίθμων του συστήματος που αναπτύξαμε.

4.3.6 Παρουσίαση Πληροφορίας στο Χρήστη

Το πιο σημαντικό κομμάτι του συστήματος μας κυρίως διότι αυτό είναι ορατό προς το χρήστη είναι το κομμάτι παρουσίασης της πληροφορίας στο χρήστη. Στην παρούσα έκδοση του συστήματος υπάρχουν όλοι οι πιθανοί τρόποι για πρόσβαση στην πληροφορία. Μπορεί κανείς να επισκεφθεί το δικτυακό τόπο του συστήματος και να δει τα άρθρα από εκεί, μπορεί να κατεβάσει την εφαρμογή που έχει δημιουργηθεί στο πλαίσιο της εργασίας του Β. Τσόγκα, ή μπορεί απλά να χρησιμοποιήσει τα προσωποποιημένα RSS feeds που προσφέρει ο δικτυακός τόπος σε μία RSS reader εφαρμογή φυσικά για να αποφύγει την επίσκεψη στο δικτυακό τόπο. Αυτή τη στιγμή βρισκόμαστε πλέον στην τρίτη έκδοση του συστήματός μας το οποίο δεδομένου ότι βρίσκεται μονίμως σε ανάπτυξη και συνεχώς προστίθενται νέα κομμάτια σε αυτό είναι σε μόνιμη κατάσταση ανάπτυξης ή όπως αλλιώς συνηθίζεται και είναι της μόδας είναι μονίμως σε κατάσταση Beta.
Η διεπαφή με το χρήστη ονομάζεται peRSSonal κάτι το οποίο προκύπτει από τη σύνθεση των

λέξεων personal και RSS και αυτό διότι το σύστημα σχεδιάστηκε για να προσφέρει προσωποποιημένα RSS feeds προς τους χρήστες. Από το όνομα της διεπαφής ονοματίστηκε και συνολικά το σύστημα όπως συνήθως συμβαίνει σε αυτές τις καταστάσεις.
Η διεπαφή ξεκίνησε από ένα απλό σύστημα και πλέον έχει αναπτυχθεί σε μεγάλο βαθμό ώστε να ακολουθεί το σημερινό web2.0 [76], [190]. Το σύστημα στη μορφή που έχει σήμερα, επιτρέπει την είσοδο μόνο σε εγγεγραμμένους χρήστες διότι πλέον ο σκοπός του είναι να παρέχει πλήρως προσωποποιημένη πληροφορία στους τελικούς χρήστες. Η εγγραφή στο σύστημα είναι εξαιρετικά απλή και γίνεται με ένα απλό username και password. Και ενώ θεωρητικά δε χρειάζεται καμία άλλη πληροφορία για την έναρξη λειτουργίας του συστήματος και ο χρήστης μπορεί να ξεκινήσει να διαβάζει νέα και άρθρα, υπάρχει μηχανισμός ο οποίος επιτρέπει την περεταίρω δημιουργία αρχικού προφίλ για τους χρήστες του συστήματος. Συγκεκριμένα, κάθε χρήστης μπορεί να δώσει πληροφορίες για τις κατηγορίες τις οποίες επιθυμεί να παρακολουθεί αλλά και λέξεις κλειδιά που τον ενδιαφέρουν ακόμα και RSS feeds που τον ενδιαφέρουν να βλέπει.
Για τη διεπαφή με το χρήστη υπάρχουν αρκετά υποσυστήματα τα οποία αναλαμβάνουν διαδικασίες τόσο παρουσίασης όσο και ανάλυσης της πληροφορίας. Τα συστήματα αυτά είναι:

- το σύστημα παρουσίασης πληροφορίας
- το σύστημα καταγραφής των session του χρήστη
- το σύστημα ανάλυσης των session και διαρκούς ενημέρωσης του προφίλ χρήστη
- συστήματα που λειτουργούν επικουρικά των παραπάνω

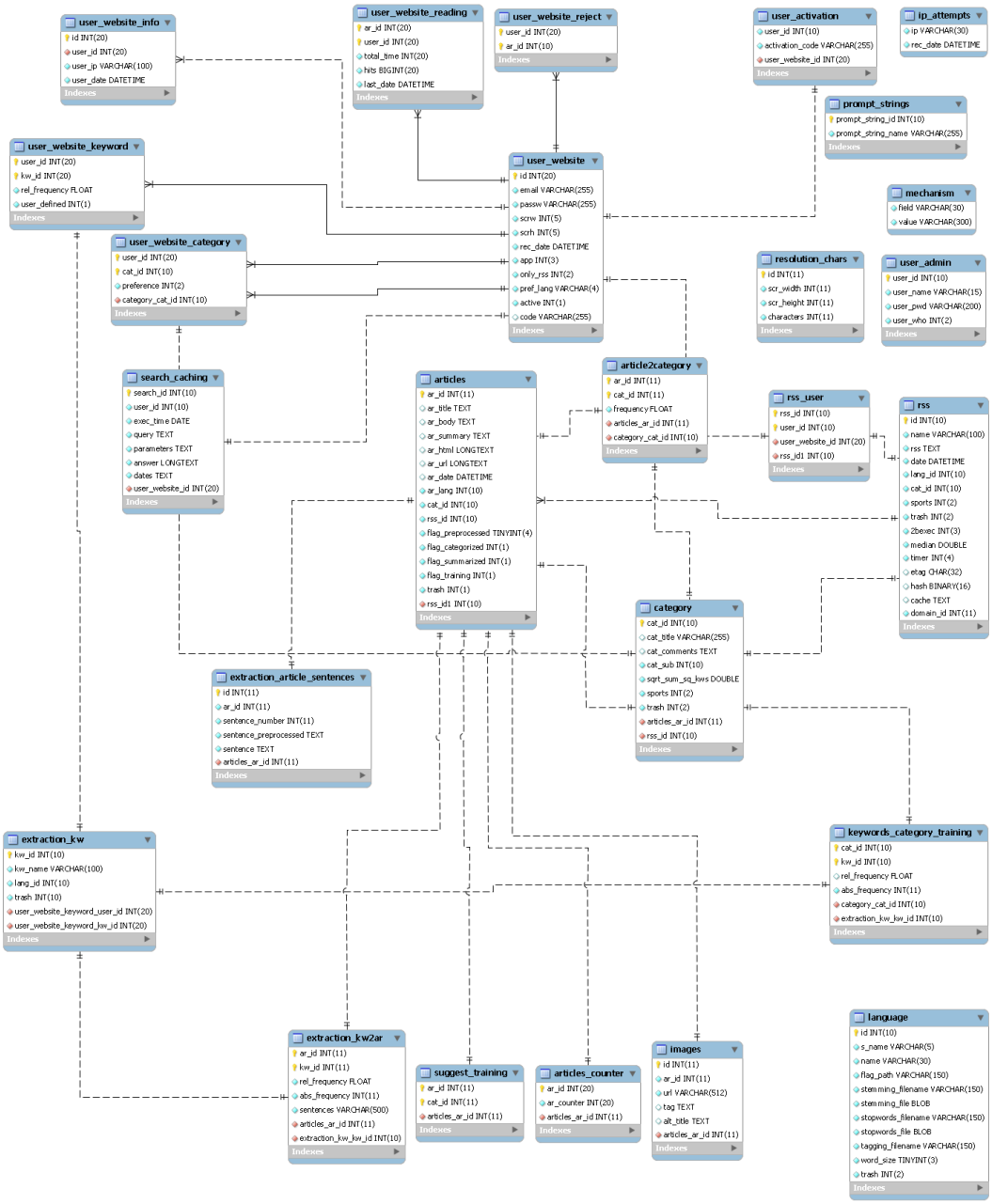
Όταν ένας χρήστης γράφεται στο σύστημα τότε ένα αρχικό προφίλ δημιουργείται προκειμένου να παρουσιαστούν στο χρήστη τα πρώτα άρθρα. Το αρχικό προφίλ (initial profile) είναι εξαιρετικά γενικευμένο και δεν περιέχει πληροφορίες που αφορούν άμεσα το χρήστη. Για την ακρίβεια, αν ο χρήστης δε δώσει καμία πληροφορία για τις επιλογές του στο σύστημα, το αρχικό προφίλ είναι σχεδόν άδειο και είναι για όλους τους χρήστες το ίδιο. Αν πάλι ο χρήστης επιλέξει να δώσει πληροφορίες στο σύστημα για τις προσωπικές του επιλογές τότε έχουμε δημιουργία ενός πιο σύνθετου προφίλ που πολλές φορές μπορεί από την αρχή να είναι αρκετά αντιπροσωπευτικό του κάθε χρήστη. Αυτό συμβαίνει διότι ο τρόπος δημιουργίας του προφίλ ενός χρήστη είναι τέτοιος ώστε να υπάρχουν 107 διαφορετικοί συνδυασμοί έναρξης του προφίλ ενός χρήστη. Θα δούμε όμως στην πορεία πως πραγματικά συμβαίνει αυτό και πως πράγματι προκύπτουν τόσοι πολλοί διαφορετικοί συνδυασμοί.
Πέραν του αρχικού προφίλ του χρήστη και αν θεωρήσουμε δεδομένου πως είναι σχετικά τετριμμένος ο τρόπος με τον οποίο παρουσιάζεται πληροφορία σε ένα χρήστη του διαδικτύου, ένα πολύ σημαντικό κομμάτι του μηχανισμού είναι η διαμόρφωση και η συνεχής ανανέωση του προφίλ ενός χρήστη. Οι μηχανισμοί οι οποίοι συμμετέχουν στη διαμόρφωση αυτού του προφίλ είναι ο μηχανισμός καταγραφής των ενεργειών του χρήστη ή όπως το ονομάζουμε session recorder και

ο μηχανισμός ενημέρωσης του προφίλ βάσει των καταγραφών που έχουν γίνει από τον session recorder.

4.4 Βάση Δεδομένων

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 σχήμα 4.11. Όπως έχει αναφερθεί και σε προηγούμενο κεφάλαιο η ΒΔ του συστήματος αποτελεί το βασικό στοιχείο επικοινωνίας του μηχανισμού. Ο μηχανισμός ο οποίος αποτελείται από κλιμακωτά αυτόνομα συστήματα έχει ανάγκη από ένα επίπεδο διασύνδεσης, ένα σημείο επαφής, ένα σημειο συντονισμού. Οι τρόποι για να επιτευχθεί αυτό είναι δύο: είτε με έναν κεντρικό μηχανισμό συντονισμού και ελέγχου ή με έναν κεντρικοποιημενο τρόπο αποθήκευσης της πληροφορίας όπου θα αποτυπώνεται σαφώς το σημείο στο οποίο βρίσκεται κάθε στιγμή η διαδικασία. Στο σύστημά μας επιλέγουμε το δεύτερο τρόπο, τη δημιουργία δηλαδή ενός κεντρικοποιημενου σημείου αποθήκευσης της πληροφορίας και αυτό δεν είναι άλλο από τη βάση δεδομένων μας. Φυσικά για να γίνει έλεγχος των διαδικασιών θα πρέπει να αποθηκεύεται σε κάθε περίπτωση πληροφορία που αφορία τις διαδικασίες του συστήματος.

Η εκτενής ανάλυση των πινάκων του συστήματος ξεφεύγει από το σκοπό της συγκεκριμένης εργασίας, ωστόσο κάποια γενικά στοιχεία για τη βάση δεδομένων είναι αναγκαία να δοθούν. Η ΒΔ του συστήματος αποτελεί και το βασικό σύστημα με το οποίο περισσότερο συντονίζονται παρά επικοινωνούν οι μηχανισμοί του συστήματος μας. Οι πίνακες του συστήματος που αφορούν τα άρθρα είναι και αυτοί οι οποίοι διατηρούν το μεγαλύτερο όγκο πληροφορίας ενώ ο πίνακας των άρθρων έχει πολλές πληροφορίες ανάλογα με το στάδιο στο οποίο βρίσκεται το σύστημα μέσα από συγκεκριμένα flags τα οποία αποθηκεύονται σε πεδία της ΒΔ για κάθε άρθρο. Εξίσου σημαντικό πίνακας είναι αυτός ο οποίος αποθηκεύει την πληροφορία που αφορά τις λέξεις κλειδιά του συστήματος και τη συσχέτιση αυτών με τις κατηγορίες. Μέσα από τις εγγραφές αυτού του πίνακα γίνεται ανάκτηση στοιχείων που αφορούν αυτό που αντιπροσωπεύει στην ουσία μία κατηγορία για το σύστημά μας. Παράλληλα, αντίστοιχοι πίνακες υπάρχουν και για τους χρήστες προκειμένου να διαμορφώνεται το προφίλ τους. Αυτό γίνεται λοιπόν μέσω από λέξεις κλειδιά οι οποίες περιέχουν θετικά και αρνητικά βάρη και αποθηκεύονται στο δυναμικό προφίλ του χρήστη. Αναφορικά με τα RSS feeds και γενικά όλες τις πληροφορίες που χρησιμοποιούνται γι αυτά, υπάρχουν οι αντίστοιχοι πίνακες. Στην ουσία όλες οι πληροφορίες που είναι απαραίτητες για να γίνεται διατήρηση ιστορικών στοιχείων για τα RSS προκειμένου να γίνεται adaptation βρίσκονται σε αυτό τον πίνακα. Ο ίδιος αποθηκεύει και πληροφορίες προκειμένου να εντοπίζονται νέα άρθρα που δεν έχουν εισαχθεί ακόμα στο σύστημα.



Σχήμα 4.11: Η Βάση Δεδομένων του Συστήματος

4.5 Τεχνολογίες Υλοποίησης

4.5.1 Τεχνολογίες Υλοποίησης Μηχανισμών Πυρήνα

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.

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.

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.

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.

Επιλογή της τεχνολογίας υλοποίησης Συστηματων Πυρήνα

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.

4.5.2 Τεχνολογίες Υλοποίησης Μηχανισμών Διεπαφής - 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.

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.

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.

4.5.3 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.

4.6 Διασύνδεση Συστημάτων

Τα συστήματα που αναπτύσουμε όπως έχει αναφερθεί πολλές φορές λειτουργούν σαν αυτόνομα υποσυστήματα εξασφαλίζοντας ανεξάρτητη λειτουργία με τη γενική είσοδο που δέχονται και με τη γενική έξοδο που παράγουν. Τεχνικά μιλώντας πριν από κάθε είσοδο του κάθε μηχανισμού και αμέσως μετά από κάθε έξοδο υπάρχουν συστήματα που κάνουν προσαρμογή της πραγματικής εισόδου που έχουμε και εξόδου που χρειαζόμαστε. Οι μηχανισμοί χρησιμοποιούν σε κάθε επίπεδο XML τόσο για την είσοδο, όσο και για την έξοδο όμως προσαρμόζεται στις ανάγκες του μηχανισμού μας σε κάθε περίπτωση.
Στην ουσία, και όπως φαίνεται και από την αρχιτεκτονική του μηχανισμού μας, η ΒΔ χρησιμοποιείται σαν ο κεντρικός μηχανισμός διασύνδεσης των συστηματών. Αυτό επιτυγχάνεται με τη ρητή δήλωση και εγγραφή στη βάση δεδομένων στοιχείων που αφορά το στάδιο στο οποίο βρίσκεται η πληροφορία. Βλέποντας λίγο τα στάδια του μηχανισμού, αρχικά έχουμε την ανάκτηση άρθρων και εξαγωγή χρήσιμου κειμένου. Από αυτό το μηχανισμό στην ουσία προκύπτουν νέα

άρθρα, με τον τίτλο τους καθώς και το κείμενο που εξάγουμε και ανήκει, θεωρητικά, στο άρθρο. Στη συνέχεια περνάμε στο μηχανισμό προεπεξεργασίας. Ο μηχανισμός αυτός χρειάζεται σαν είσοδο το κείμενο των άρθρων προκειμένου να πραγματοποιήσει όλες αυτές τις αναλύσεις για να εξαχθούν στοιχεία που αφορούν τις λέξεις κλειδιά του κειμένου αλλά και όποιες συσχετίσεις λέξεων με προτάσεις, προτάσεων με άρθρα, λέξεων με άρθρα κ.α. Τα παραπάνω γίνονται σε ένα βήμα και εξάγονται όλες οι προαναφερθείσες πληροφορίες. Στη συνέχεια έχουμε τους μηχανισμούς κατηγοριοποίησης και εξαγωγής περίληψης οι οποίοι εφαρμόζονται είτε με αυτή τη σειρά είτε συμπληρωματικά ο ένας προς τον άλλο με παράλληλο τρόπο. Τέλος, και αφού τελειώσουν αυτοί οι μηχανισμοί η πληροφορία θεωρείται έτοιμη για το μηχανισμό παρουσίασης. Όπως είναι φυσικό, παρά το γεγονός ότι κάθε μηχανισμός μπορεί να λειτουργήσει απ' ευθείας πάνω στην πληροφορία θα πρέπει τα βήματα να πραγματοποιηθούν με κάποια σειρά. Ο κάθε μηχανισμός δεν περιμένει να λάβει στοιχεία για να ξεκινήσει να λειτουργεί αλλά ούτε και κάποιο μηχανισμό να ενημερώνει για τα δεδομένα που υπάρχουν διαθέσιμα προς επεξεργασία. Αυτό που συμβαίνει είναι πως οι μηχανισμοί εκτελούνται περιοδικά. Η περίοδος του μηχανισμού που μας ενδιαφέρει να είναι συγκεκριμένη είναι αυτή του μηχανισμού ανάκτησης άρθρων και αυτό διότι θέλουμε να έχουμε όσο το δυνατόν πιο επικαιροποιημένη συλλογή άρθρων από το διαδίκτυο. Στην πορεία και στην ανάλυση των αλγορίθμων θα δούμε πολύ πιο αναλυτικά τη σπουδαιότητα αυτού του μηχανισμού και τον τρόπο λειτουργίας του.

Όταν εκτελούμε τον πρώτο μηχανισμό (κάθε έξι με δέκα λεπτά) στην ουσία γράφονται στη βάση δεδομένων στοιχεία για νέα αρθρα. Αυτά τα στοιχεία είναι: ο τίτλος, το κείμενο (rough text), η ημερομηνία του άρθρου και άλλα μεταδεδομένα. Για κάθε άρθρο υπάρχουν κάποια flags: (α) για την προεπεξεργασία, (β) για την εξαγωγή περίληψης και (γ) για την κατηγοριοποίηση. Έτσι σε αυτό το σημείο όλα τα flags είναι σβησμένα. Με κάθε εκτέλεση των μηχανισμών που έπονται της ανάκτησης πραγματοποιείται έλεγχος στα flags των άρθρων. Ο μηχανισμός προεπεξεργασίας ελέγχει και αναλύει όσα άρθρα χρειάζονται προεπεξεργασία, ο μηχανισμός εξαγωγής περίληψης αναλύει όλα τα άρθρα που έχουν υποστεί προεπεξργασία αλλά όχι ακόμα περίληψη, κ.ο.κ.

Έτσι η διασύνδεση των υποσυστημάτων του μηχανισμού επιτυγχάνεται μέσα από στοιχεία που αποθηκεύονται ρητά στην κεντρικοποιημένη βάση δεδομένων από τον κάθε μηχανισμό. Με αυτό τον τρόπο διατηρούμε την ανεξαρτησία κάθε μηχανισμού και την αυτονομία τους ενώ παράλληλα οι μηχανισμοί μπορούν και συνεργάζονται αρμονικά σημειώνοντας στοιχεία για τη λειτουργία τους στη ΒΔ. Με αυτό τον τρόπο οι μηχανισμοί είναι μαύρα κουτιά ο ένας για τον άλλο και στην ουσία αυτό που αρκεί είναι να εξασφαλίσουν ο καθένας σωστές εγγραφές στη βάση δεδομένων για τη λειτουργία τους.

Πίνακας 4.1: Πίνακες της Βάσης Δεδομένων

Πίνακας	Ανάλυση
Articles	Ο πίνακας που αποθηκεύει κάθε πληροφορία που αφορά τα άρθρα
Category	Ο πίνακας που αποθηκεύει στοιχεία για τις κατηγορίες του συστήματος
Article2category	Πίνακας που συσχετίζει άρθρα με κατηγορίες (με κάποια πιθανότητα)
Extractionkw	Ο πίνακας με τις λέξεις κλειδιά του συστήματος
Extractionkw2article	Ο πίνακας που συσχετίζει λέξεις κλειδιά με άρθρα
Keywordscategorytraining	Πίνακας που συσχετίζει λέξεις κλειδιά με κάποια κατηγορία (με κάποια συχνότητα)
Articlescounter	Πίνακας που αποθηκεύει βοηθητικές πληροφορίες για τα άρθρα (counters)
Suggesttraining	Πίνακας που αποθηκεύει πληροφορίες που σχετίζονται με άρθρα που προτείνονται να ενταχθούν στο training set
Images	Εικόνες που εξάγονται από τον mCuter και σχετίζονται με κάποιο άρθρο
Extractionarticlesentences	Πίνακας που αποθηκεύει πληροφορίες σχετικά με τις προτάσεις των άρθρων που αναλύονται
Rss	Ο πίνακας που αποθηκεύει τις λίστες με τα RSS feeds μαζί με κάθε πληροφορία που αφορά τα RSS και το μηχανισμό advaRSS
Userwebsite	Πίνακας που αποθηκεύει στοιχεία για τους χρήστες του συστήματος και γενικό configuration που έχουν για το site
Useractivation	Πίνακας που χρησιμοποιείται για το activation του χρήστη
Userwebsitereading	Πίνακας που διατηρεί στοιχεία για τα άρθρα που διαβάζει ένας χρήστης
Userwebsitereject	Πίνακας που περιέχει στοιχεία για άρθρα τα οποία ο χρήστης τοποθετεί σε blacklist
Userwebsiteinfo	Πίνακας που αποθηκεύει στοιχεία για τις συνδέσεις του χρήστη με το δικτυακό τόπο
Userwebsitecategory	Ο πίνακας που αποθηκεύει στοιχεία σχετικά με τις κατηγορίες που έχει επιλέξει να ελέγχει ένας χρήστης
Userwebsitekeywords	Ο κεντρικός πίνακας ελέγχου του προφίλ ενός χρήστη μέσα από τα keywords τα οποία παρακολουθεί.
Rssuser	Τα RSS feeds τα οποία παρακολουθεί κάθε χρήστης
Searchcaching	Πίνακας που χρησιμοποιείται από το μηχανισμό αναζήτησης του peRSSonal που πραγματοποιεί caching
Ipattempts	Πίνακας που χρησιμοποιείται βοηθητικά προκειμένου να αποφεύγονται «επιθέσεις» στο σύστημα
Mechanism	Πίνακας που αποθηκεύει κεντρικά μεταβλητές του συστήματος
Promptstrings	Πίνακας που χρησιμοποιείται για να επιτρέπεται πολυγλωσσία στο σύστημα
Language	Πίνακας που περιέχει τη λίστα με τις γλώσσες που υποστηρίζονται από το σύστημα
Useradmin	Πίνακας που περιέχει στοιχεία για την είσοδο στο διαχειριστικό σύστημα

ΚΕΦΑΛΑΙΟ 5

ΑΝΑΛΥΣΗ ΑΛΓΟΡΙΘΜΩΝ

Περί την απορρήτων μηδενί λέγε

(Ισοκράτης)

Στο κεφάλαιο αυτό γίνεται εκτενής ανάλυση των αλγορίθμων του συστήματος καθώς και παρουσίαση της εφαρμογής της αρχιτεκτονικής στην διαδικασία των συστημάτων. Για κάθε σύστημα γίνεται εκτενής αναφορά στον αλγοριθμικό τρόπο λειτουργίας η οποία αποτελεί και προοίμιο για τις πειραματικές διαδικασίες που ακολουθούν.

Ένα από τα πιο σημαντικά κομμάτια της έρευνας αποτελεί η τελική επιλογή αλγορίθμων και η πρόταση καινούριων για την επίλυση των θεμάτων με τα οποία καταπιανόμαστε στην εργασία. Οι αλγόριθμοι που χρησιμοποιούνται σε κάθε υποσύστημα του μηχανισμού ενδεχόμενα να μην είναι οι βέλτιστοι για το ερευνητικό πεδίο το οποίο αφορούν ωστόσο με τη χρήση τους παρουσιάζεται βέλτιστη απόδοση του τρόπου λειτουργίας του συστήματος ή έστω επαρκής λειτουργία του συστήματος για τις προδιαγραφές που έχουμε γι αυτό. Για να είμαστε πιο σαφής, επειδή πρόκειται για ένα σύστημα το οποίο εν πολλοίς λειτουργεί σε πραγματικό χρόνο, υπάρχουν πολλές περιπτώσεις που ο παράγοντας ταχύτητα και άρα απλότητα του αλγορίθμου έπρεπε να μπει σε μεγαλύτερη προτεραιότητα από οτιδήποτε άλλο. Έτσι, σε αρκετές περιπτώσεις ο αλγόριθμος που χρησιμοποιήσαμε είχε σαν σκοπό να μας δώσει απλώς ένα επαρκές αποτέλεσμα και όχι να μας δώσει ένα ακριβές αποτέλεσμα καταναλώνοντας πολύ χρόνο και πόρους συστήματος.
Οι αλγόριθμοι που θα παρουσιάσουμε χωρίζονται στην ουσία σε τόσα κομμάτια όσα είναι και τα υποσυστήματα του μηχανισμού καθότι κάθε υποσύστημα, ως αυτόνομη οντότητα, έχει τα δικά του χαρακτηριστικά και τους δικούς του μηχανισμούς λειτουργίας. Σε κάποιες περιπτώσεις επίσης θα εντοπίσουμε αλληλεπίδραση μεταξύ των μηχανισμών αλλά και αλγορίθμους που χρησιμοποιούν κοινά στοιχεία στους μηχανισμούς καθότι μέσα από την έρευνά μας είδαμε και αποδείξαμε πως η συνεργασία των μηχανισμών μπορεί να έχει θετικά αποτελέσματα σε ορισμένες περιπτώσεις όπως δείχνει και η πειραματική διαδικασία.
Σε αυτό το κεφάλαιο θα παρουσιάσουμε αποκλειστικά και μόνο τον τρόπο χρήσης των αλγορίθμων ενώ η πειραματική διαδικασία και ο έλεγχος της λειτουργίας των αλγορίθμων θα παρουσιαστεί σε επόμενο κεφάλαιο της εργασίας.

5.1 Υποσύστημα Ανάκτησης Πληροφορίας – advaRSS

Ο μηχανισμός advaRSS [41] και [40] όπως έχει ήδη αναφερθεί στην αρχιτεκτονική του συστήματος είναι ένας mixed crawler, και αν δε δίνουμε μεγάλη προσοχή και σημασία στον τρόπο με τον οποίο ανακτά HTML σελίδες με άρθρα, αυτό στο οποίο εστιάζουμε και πραγματοποιούμε εκτενή ανάλυση είναι ο τρόπος με τον οποίο εντοπίζει αλλαγές στα RSS feed και updates που σχετίζονται με ειδήσεις και άρθρα. Για κάθε RSS feed που έχουμε στη βάση δεδομένων διατηρούμε κάποιες μεταβλητές προκειμένου να είμαστε σε θέση να αποθηκεύουμε ποιοτικά και ποσοτικά στοιχεία σχετικά με κάθε διαφορετικό RSS feed. Εκτός, λοιπόν, όλων των βασικών χαρακτηριστικών που είναι αναγκαία για τα RSS feeds στοιχεία που διαθέτουμε επιπλέον είναι οι μεταβλητές:

1. median
2. to be executed

3. timer
4. etag
5. hash
6. cache
7. domain id

Όλα τα παραπάνω μας είναι άκρως αναγκαία για να μπορούμε να εντοπίζουμε πληροφορίες για τα RSS feeds. Σε αυτό το σημείο θα πρέπει να επανέλθουμε λίγο στον τρόπο λειτουργίας του advaRSS προκειμένου να είμαστε σε θέση να εντοπίσουμε με ποιον τρόπο πραγματοποιούμε τις διαδικασίες ανάκτησης HTML κώδικα σελίδων με άρθρα. Το σύστημά μας διαθέτει, λοιπόν, αποθηκευμένα στη ΒΔ λίστες με RSS feeds. Αυτά προέρχονται από πολλά διαφορετικά domains και αρκετά από αυτά ενδεχόμενα να ανήκουν στο ίδιο domain αλλά σε διαφορετικές κατηγορίες. Κάθε φορά που εκτελείται ο μηχανισμός (τυχαίος χρόνος από 6-10 λεπτά) λαμβάνει ένα set από RSS feeds προκειμένου να τα ελέγξει για τυχόν ενημερώσεις με άρθρα. Θα μπορούσε κανείς να πει ότι θα ήταν εξαιρετικά απλό να λάβει και να αναλύσει όλα τα RSS feeds που έχει στη βάση δεδομένων και επομένως ανά 10 λεπτά να είμαστε σε θέση να ξέρουμε τις ενημερώσεις από τα RSS feeds. Όμως, εδώ τίθενται κάποια βασικά ζητήματα.

- Ο μεγάλος όγκος RSS feeds που διαθέτουμε στη ΒΔ. Σκεφθείτε μόνο πως το cnn.com διαθέτει πάνω από 30 feeds που εντάσσει σε πολλές διαφορετικές κατηγορίες χωρίς αυτό να σημαίνει πως δεν υπάρχει επικάλυψη. Ωστόσο, αν προσθέταμε στο σύστημα πληροφορία από 10 ειδησεογραφικά θα είχαμε κιόλας 300 feeds. Αν αναλογισθούμε ότι δεν υπάρχουν 10 ειδησεογραφικά sites αλλά τουλάχιστον κάποιες δεκάδες χιλιάδες ακόμα και εκατοντάδες χιλιάδες (μία έρευνα μέσω google στην Ελλάδα μας απάντησε με αμέτρητες σελίδες ενημερωτικών sites, εφημερίδων και περιοδικών), τότε το πρόβλημα του να παρακολουθούμε εκατοντάδες χιλιάδες feeds γίνεται ξαφνικά τεράστιο.

- Αν θεωρήσουμε πως έχουμε την ισχύ να ελέγχουμε όλα αυτά τα RSS feeds ανά 5 λεπτά, τι πόρους δικτύου αλλά και πόρους των server θα καταναλώναμε; Και πάλι, να τονίσουμε ότι μιλάμε για ένα σύστημα πολύ μεγάλης κλίμακας γιατί σε κάθε περίπτωση όταν μιλάμε για μικρά νούμερα τότε το μοντέλο πλήρους ανάλυσης είναι το επικρατέστερο.

- Κατά πόσο έχει νόημα να κάνουμε έλεγχο πολύ συχνά για ενημέρωση σε ένα RSS feed το οποίο κάποιος μας έχει ενημερώσει ότι ενημερώνεται μία φορά τη βδομάδα;

- Πόσο πολύ επικαιροποιημένο set ειδήσεων θέλουμε να προσφέρουμε. Ο "πραγματικός χρόνος" που αναφέρεται σα χαρακτηριστικό του συστήματος τι απόκλιση μπορεί να έχει με τον πραγματικό χρόνο;

Καθώς διαφαίνεται από την ανάλυση που κάνουμε πάνω στο μηχανισμό ανάκτησης ειδήσεων θα πρέπει να κάνουμε έναν πολύ προσεκτικό συνδυασμό στοιχείων ώστε να πετύχουμε μία σειρά από χαρακτηριστικά χωρίς παράλληλα να αλλοιώνουμε το χαρακτήρα του μηχανισμού ενώ παράλληλα να είμαστε και "ευγενικοί" με το Διαδίκτυο και το δίκτυο. Βασικά στοιχεία για το μηχανισμό που αναπτύσσουμε έχουν να κάνουν με τα εξής στοιχεία:

- Θεωρούμε πως δεν είναι "ευγενικό" ίσως ούτε και εφικτό να ελέγχουμε ανά 5 λεπτά χιλιάδες RSS feeds για άρθρα και ειδήσεις. Ακόμα κι αν έχουμε την υπολογιστική ισχύ αποδεικνύεται ανώφελο.

- Θεωρούμε πως η απόκλιση μεταξύ χρόνο δημοσίευσης και χρόνου ανάκτησης ενός άρθρου που είναι, υποκειμενικά, ανεκτή για το σύστημα μπορεί να ποσοτικοποιηθεί και να τεθεί ίση με 15 λεπτά, ωστόσο, για RSS feeds τα οποία δεν έχουν μεγάλη δημοφιλία ο αριθμός αυτός μπορεί να αυξηθεί χωρίς να θεωρήσουμε ότι δημιουργεί πρόβλημα γενικά στο σύστημα.

- Θεωρούμε πως μπορούμε να καθορίσουμε ένα posting pattern για κάθε RSS feed το οποίο θα πρέπει να μας δίνει ποιοτικά στοιχεία για το ρυθμό ανανέωσης ενός RSS.

- Θεωρούμε πως τα RSS feeds που έχουν μεγαλύτερη δημοφιλία (τα έχουν επιλέξει πολλοί χρήστες) αξίζουν να λαμβάνουν μεγαλύτερη προτεραιότητα.

Για να μπορέσουμε να εκφράσουμε όλα τα παραπάνω, όπως είδαμε ορίσαμε μία σειρά από μεταβλητές. Άλλες μας βοηθούν ώστε να υπολογίζουμε βάσει αλγορίθμων το χρόνο που απομένει για να ελέγξουμε ένα RSS feed ενώ άλλες είναι επικουρικές και χρησιμοποιούνται για βοηθητικές εργασίες. Ας περάσουμε να δούμε, όμως, μία προς μία τις μεταβλητές του συστήματος.

- Median: εκφράζει έναν προσωρινό μέσο όρο του posting history

- to be executed: ένας απλός counter που κάνει countdown. Η τιμή του κυμαίνεται ανά πάσα στιγμή μεταξύ 0 και Median. Όταν γίνει 0 είναι η στιγμή που θα εξεταστεί το συγκεκριμένο RSS και αμέσως η τιμή του τίθεται ίση με median. Σε κάθε άλλη περίπτωση με την εκτέλεση του μηχανισμού η τιμή του μειώνεται κατά 1.

- timer: ένας counter που μετρά πόσες "εκτελέσεις" του μηχανισμού έχουν περάσει από την τελευταία φορά που εντοπίστηκε κάποια αλλαγή στο συγκεκριμένο RSS. Η έκταση της αλλαγής δεν εκφράζεται σε κάποια μεταβλητή αλλά επηρεάζει την τιμή του timer κάθε φορά που αυτός ανανεώνεται

- etag: ένας ηλεκτρονικός αριθμός που ενημερώνει για αλλαγές σε αρχεία χωρίς αυτά να κατέβουν (δεν υποστηρίζεται από όλους τους servers)

- hash: ο κώδικας hash του αρχείου που κατέβασε την τελευταία φορά ο μηχανισμός. Πρόκειται για εναλλακτικό τρόπο ελέγχου αν έχουν γίνει αλλαγές σε ένα αρχείο. Ωστόσο, προϋποθέτει το κατέβασμα του αρχείου.
- Cache: Στοιχεία που αποθηκεύονται σαν cache (κυρίως οι τίτλοι των άρθρων του RSS feed προκειμένου να εντοπίσουμε σε τυχόν αλλαγές ποια είναι τα νέα άρθρα.
- domain id: μοναδικό αναγνωριστικό του domain στο οποίο ανήκει ένα RSS feed. Είναι άκρως απαραίτητο για να αποφεύγουμε cross postings μεταξύ rss του ίδιου domain.

Όλες οι παραπάνω μεταβλητές συνθέτουν το περιβάλλον του advaRSS και προκειμένου να επιτύχουμε το επιθυμητό αποτέλεσμα πρέπει να τις συνδυάσουμε μέσω του αλγορίθμου λειτουργίας του advaRSS.

Ο βασικός αλγόριθμος του συστήματος είναι ο ακόλουθος:

```
Every X min {
  feeds_to_pars
e[] = Select RSS feeds from
database with ToE equal to zero;
  Foreach(feeds_to_parse[] as url)
    xml_code = fetch_rss(url);
    If(not modified)
      continue;
    else
      articles_in_rss[] =
extract_info(xml_code);
      Foreach(articles_in_rss[] as article)
If(title_not_found_in_last_articles(article))
         add_to_DB(article);
       End If
      End For
    End If
  End For
}
```

όπου

feeds_to_parse: RSS feeds προς ανάλυση από το μηχανισμό.
fetch_rss(): Ο XML κώδικας ενός RSS feed

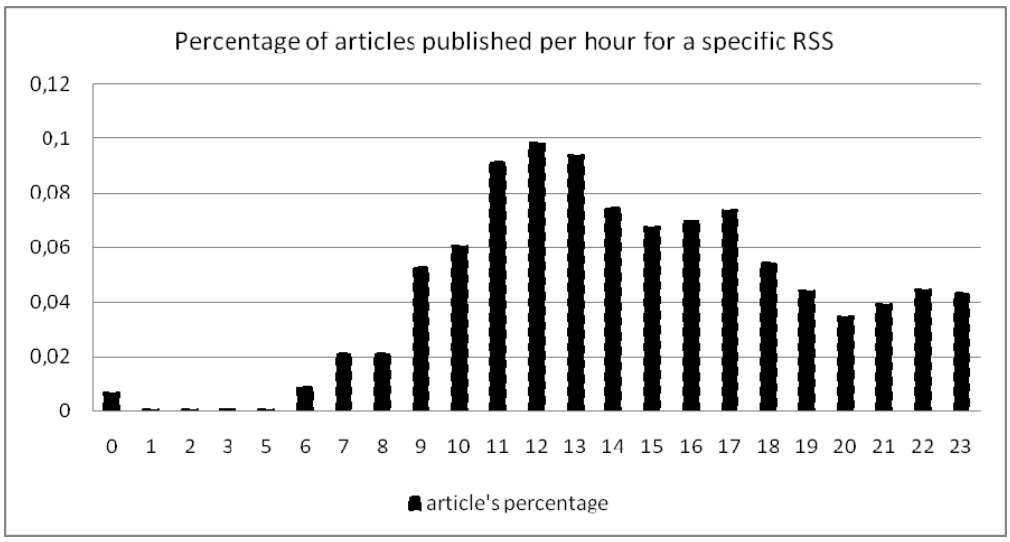
latest_articles: Τα πιο πρόσφατα άρθρα που υπάρχουν μέσα στο RSS
και δεν έχει ανακτήσει ακόμα το σύστημα.

Ο αλγόριθμος αντιπροσωπεύει τη γενική διαδικασία του advaRSS. Ωστόσο, τόσο το κομμάτι της ανάκτησης της λίστας μα τα RSS αλλά και το κομμάτι της ανάλυσης για νέα και ειδήσεις περιέχουν αναλυτικούς αλγορίθμους. Στην ουσία αυτό που θέλουμε να επιτύχουμε είναι δύο πράγματα. Από τη μία θέλουμε να ξέρουμε ποια RSS feeds θα πρέπει να ανακτήσουμε και από την άλλη, αυτά τα οποία ελέγξαμε σε πόσο χρόνο θα πρέπει να τα ελέγξουμε εκ νέου. Ο παρακάτω αλγόριθμος μας βοηθά προς τις δύο αυτές κατευθύνσεις.

```
feeds[]=Fetch_rss_having_zero_ToE();
 Foreach(feeds[] as url)
   If(not modified)
     newTimer=T+M;
     M = M + 30%T
     T = newTimer
   Else
     T=1
     M = 20%M+80%T
   End If
 ToE = M
End For
```

όπου

M (median): Ο αριθμός που μας δείχνει κάθε στιγμή την περιοδικότητα βάσει της οποίας πρέπει να ελέγχουμε ένα RSS feed. T (timer): ένας αριθμός που μας δείχνει πόση ώρα έχει περάσει από την τελευταία φορά που βρέθηκε να έχει αλλάξει το RSS και ToE : ένας μετρητής που μειώνεται σε κάθε εκτέλεση του προγράμματος. Όταν γίνει μηδέν πρέπει να αναλύσουμε το RSS.
Ο παραπάνω αλγόριθμος χρησιμοποιείται προκειμένου να εντοπίσουμε το ToE και να το υπολογίζουμε κάθε φορά ανάλογα με το ρυθμό με τον οποίο παρατηρούνται αλλαγές στο RSS feed, κάτι το οποίο μας δείχνει ο Timer. Η μεταβλητή M στην ουσία χρησιμοποιείται για να αποθηκεύει αυτό το ρυθμό αλλαγής και μεταβάλλεται διαρκώς καθότι όπως μας δείχνει η μελέτη, ένα RSS feed έχει μονίμως μεταβαλλόμενο ρυθμό αλλαγών, χωρίς αυτό να σημαίνει και ότι δε μπορεί να ακολουθήσει ένα pattern αλλαγών. Αυτό, όμως, θα το εξετάσουμε στην πορεία με τους αλγορίθμους που χρησιμοποιούμε στο σύστημα.
Ένα από τα πλέον σημαντικά κομμάτια του συστήματος είναι, λοιπόν, η ανανέωση του χρόνου βάσει του οποίου πρέπει να ελέγχουμε ένα RSS feed για αλλαγές. Ο αλγόριθμος που κατασκευάσαμε βασίζεται σε μελέτες που κάναμε για το ρυθμό αλλαγών που παρουσιάζεται στα



Σχήμα 5.1: Ποσοστό των άρθρων που δημοσιεύονται σε ένα μέσο RSS κατά τη διάρκεια μίας μέρας

RSS feeds. Αυτό που διαπιστώσαμε είναι πως κάθε domain έχει τα δικά του χαρακτηριστικά και προφανώς RSS από ίδιο domain φαίνεται πως έχει κοινά χαρακτηριστικά. Είδαμε πως υπάρχουν κάποια RSS feeds τα οποία μεταβάλλονται μία φορά ημερησίως, ώρα κατά την οποία προστίθενται όλα τα άρθρα που έχουν δημοσιευθεί. Αυτό παρατηρείται συνήθως σε περιπτώσεις ηλεκτρονικών εφημερίδων που δημοσιεύουν μία φορά τη μέρα όλα τα άρθρα τους. Από την άλλη και στην πλειοψηφία των περιπτώσεων αυτό που παρατηρείται είναι πως οι αλλαγές πραγματοποιούνται τις εργάσιμες μέρες και ώρες. Αυτή η παρατήρηση δεν είναι ούτε καινούρια, ούτε πρωτοποριακή και υφίσταται σε μελέτη των Cho και Garcia-Molina εδώ και χρόνια. Ωστόσο, αυτή η παρατήρηση μας είναι πολύ χρήσιμη για το ρυθμό με τον οποίο θέλουμε να πραγματοποιούμε αλλαγές στο ρυθμό εξέτασης ενός RSS feed. Το σχήμα 5.1 μας δείχνει το ποσοστό των άρθρων που δημοσιεύονται σε ένα μέσο RSS κατά τη διάρκεια μίας μέρας.

Όπως είναι εμφανές από το διάγραμμα, μεταξύ 09:00 και 18:00 έχουμε δημοσίευση του 70% των άρθρων ενώ στο διάστημα 00:00 έως 06:00 συνήθως δεν παρατηρείται καμία απολύτως δημοσίευση. Θα θέλαμε λοιπόν ένα σύστημα που θα μπορούσε ακόμα και να αγνοήσει το συγκεκριμένο feed κατά το διάστημα 00:00-06:00 ενώ από τις 09:00 μέχρι τις 18:00 θα θέλαμε να το κοιτάζει πολύ συχνά καθότι βλέπουμε μεγάλη εισροή πληροφορίας. Για το λόγο αυτό χρησιμοποιούμε τις μεταβλητές ToE, M και T. Ο μηχανισμός όπως έχει ήδη αναφερθεί εκτελείται ανά τυχαία χρονικά διαστήματα κάθε 6-10 λεπτά. Η βασική αρχή είναι να εκτελεστεί και να τεθεί σε κατάσταση αναμονής για ένα τυχαίο διάστημα 6-10 λεπτών μέχρι να επανεκτελστεί. Οι αλγόριθμοι που χρησιμοποιούνται είναι οι παρακάτω. Αν ToE>0

$$ToE = ToE–1 \tag{5.1}$$

στην περίπτωση που ToE=0 τότε μπορούμε να διακρίνουμε δύο διαφορετικές περιπτώσεις: (α) το RSS feed δεν έχει νέα άρθρα και (β) το RSS feed περιέχει νέα άρθρα. Στην πρώτη περίπτωση έχουμε:

$$Temp = T + M \tag{5.2}$$

$$M = M + xT \tag{5.3}$$

$$T = Temp \tag{5.4}$$

ουσιωδώς, αυξάνουμε το Timer κατά Median (ο timer δείχνει κάθε φορά το μέγεθος του χρόνου στο οποίο δεν έχουν πραγματοποιηθεί αλλαγές) και θέτουμε στο Median μία τιμή η οποία σχετίζεται με το προηγούμενο μέγεθος του Timer (όπου x ανήκει στο [0,25-0,35] βάσει πειραματικών αποτελεσμάτων).
Από την άλλη εφόσον το RSS feed βρεθεί αλλαγμένο:

$$T = 1 \tag{5.5}$$

$$M = yM + zT = yM + z(T = 1) \tag{5.6}$$

$$ToE = ceil[M] \tag{5.7}$$

με τα y και z να είναι σταθερές για τις οποίες y+z=1 όπου σύμφωνα με τα πειραματικά αποτελέσματα είδαμε πως οι τιμές που παίρνουν είναι: y ανήκει στο (0,18-0,22) και εφόσον z = 1 – y, z ανήκει στο (0,78 – 0,82). Σε γενικές γραμμές στο σύστημα χρησιμοποιούμε τη μέση τιμή για όλες τις μεταβλητές, δηλαδή:

$$x = 0,3, y = 0,2, z = 0,8 \tag{5.8}$$

Από την εξίσωση 5.1 είναι εμφανές πως σε κάθε εκτέλεση του μηχανισμού μειώνουμε το ToE κατά μία μονάδα άρα οι μονάδες του αντιπροσωπεύουν στην ουσία αριθμό εκτελέσεων του συστήματος. Επειδή το σύστημα εκτελείται τυχαία κάθε 6-10 λεπτά άρα κατά μέσο όρο ανά 8 λεπτά, μία μονάδα του ToE αντιστοιχία σε 8 λεπτά.

Για κάθε εκτέλεση του συστήματος, αν δεν παρατηρηθούν αλλαγές η μεταβλητή Timer αυξάνεται κατά M μονάδες και η μεταβλητή M αυξάνεται όπως είδαμε και πιο πάνω. Η αύξηση είναι σε γενικές γραμμές μικρή (smooth increase) όταν δε βρίσκουμε αλλαγμένο ένα RSS feed ενώ είναι μεγάλη η μείωση αν βρεθεί αλλαγμένο. Με αυτό τον τρόπο αλλαγών επιτυγχάνουμε:

α. όταν ένα RSS δεν αλλάζει τότε το ελέγχουμε λιγότερες φορές ημερησίως

β. όταν ένα RSS φαίνεται να έχει αλλάξει τότε αρχίζουμε να το ελέγχουμε αρκετά πιο συχνά για να ακολουθήσουμε τυχόν νέες αλλαγές.

Θα πρέπει να τονίσουμε πως δεν επιτρέπουμε στο Median να γίνει τόσο πολύ μεγάλο ώστε να μη μπορεί να εντοπίσει αλλαγές αλλά και παράλληλα διατηρούμε ιστορικό αλλαγών (σε αρχεία) για να δημιουργούμε patterns αλλαγών και να μπορούμε να φτιάξουμε βάσεις γνώσης για κάθε RSS feed ξεχωριστά. Στην ουσία η μέγιστη τιμή που δίνουμε στο Median εξαρτάται από το μέσο όρο αλλαγών που εντοπίζουμε για κάθε RSS. Η αρχική μέγιστη τιμή που μπορεί να πάρει ο Median είναι 25 ωστόσο μπορεί να φτάσει μέχρι και στον αριθμό 640. Πως μεταφράζεται όμως αυτό σε χρόνο. Αν το σύστημα εκτελείται ανά X λεπτά τότε με Median = 25 θα ελέγχει το RSS ανά 25xX λεπτά ή αλλιώς 58/X φορές τη μέρα στη χειρότερη περίπτωση. Δηλαδή με το X=8(λεπτά) ένα RSS feed θα ελέγχεται 7 φορές ημερησίως. Αυτό βέβαια μπορεί και να είναι πολύ μεγάλο νούμερο για ένα RSS που αλλάζει μία φορά τη μέρα ή και πολλές φορές μία φορά ανά κάποιες μέρες ή και ανά βδομάδα. Αν ο Median έφτανε το 640 αυτό μεταφράζεται σε 640x8 = 5120 λεπτά (85ώρες 3,5 μέρες), αριθμός ανεκτός για ένα RSS που ανανεώνεται σπάνια. Από την άλλη και προφανώς η ελάχιστη τιμή που μπορεί να πάρει ο Median είναι το 1, δηλαδή να γίνεται έλεγχος του RSS σε κάθε εκτέλεση του μηχανισμού, πράγμα φυσιολογικό αν βρισκόμαστε στην ώρα που ένα RSS feed ανανεώνεται με νέα και ειδήσεις.

Η παραπάνω διαδικασία που μοιάζει περισσότερο γραμμική ανάλυση κάποιων αριθμών αποτελεί το βασικό στάδιο προετοιμασίας του μηχανισμού και αποθηκεύει πληθώρα δεδομένων που μας δείχνουν στοιχεία για το ρυθμό ανανέωσης ενός RSS feed. Σε ένα δεύτερο επίπεδο και αφού συγκεντρωθεί πληροφορία κάποιων εβδομάδων για κάθε RSS feed αλλάζει ο τρόπος αλληλεπίδρασης με το μηχανισμό και συνεπώς και το set RSS feed που εξετάζει ο μηχανισμός.

Σε αυτό το δεύτερο επίπεδο ελέγχου που πραγματοποιεί ο μηχανισμός, χρησιμοποιείται το posting history που διαθέτουμε για κάθε RSS και στην ουσία αποτελεί το hourly posting rate. Σε αυτό το σημείο εισάγεται ο μηχανισμός scheduling ο οποίο χρησιμοποιεί αυτή την πληροφορία προκειμένου να διαμορφώσει το χρόνο επόμενης επίσκεψης. Σε κάθε επίσκεψη που γίνεται, χρησιμοποιούνται στοιχεία πραγματικού χρόνου αλλά και στοιχεία από το ιστορικό του RSS feed προκειμένου να γίνει πρόβλεψη του χρόνου επόμενης επίσκεψης. Ο αλγόριθμος στον οποίο καταλήξαμε έχει άμεση εξάρτηση με την ημερήσια μέση ανανέωση. Από το παραπάνω σχήμα που

δείχνει ποσοστό των άρθρων μέσα σε ένα 24ωρο επιλέγουμε ένα συγκεκριμένο 24ωρο για να δούμε το μέσο όρο των άρθρων που δημοσιεύονται όπως φαίνεται από το επόμενο σχήμα.
Όπως φαίνεται και από το σχήμα 5.1 ο μηχανισμός scheduling θα πρέπει να προγραμματίσει πιο συχνές επισκέψεις στο RSS feed τις ώρες που έχουμε μεγαλύτερο ρυθμό δημοσίευσης. Διαθέτοντας πληροφορία για το posting rate της τελευταίας ώρας χρησιμοποιούμε τον παρακάτω αλγόριθμο πρόβλεψης:

$$articles(t_{now}) = \int_{last}^{now} \frac{postingRate(t_{now} - t)}{3600} \tag{5.9}$$

Η εξίσωση 5.9 χρησιμοποιεί το μέσο όρο δημοσιεύσεων ανά δευτερόλεπτο εντοπίζοντας το κλάσμα των δημοσιεύσεων της τελευταίας ώρας δια τον αριθμό των δευτερολέπτων της ώρας. Είναι αναμενόμενο να λαμβάνουμε μεγαλύτερους αριθμούς για τα RSS feed με μεγάλα posting rates. Για λόγους που έχουμε ήδη εξηγήσει δεν είμαστε σε θέση να ελέγχουμε χιλιάδες RSS feeds ανά πάσα στιγμή. Μπορούμε ωστόσο να υπολογίζουμε και να κάνουμε χρονοπρογραμματισμό των ελέγχων. Ο αλγόριθμος προγραμματισμού λοιπόν χρησιμοποιεί την παραπάνω εξίσωση για να περιορίζει τους ελέγχους που γίνονται στο σύνολο των RSS.
Επιπλέον, επειδή το σύστημα είναι κατασκευασμένο για να παρέχει πληροφορίες προς χρήστες και επειδή βασίζεται σε μεγάλο βαθμό στους χρήστες που το χρησιμοποιούν, χρησιμοποιούμε μία ακόμα μετρική προκειμένου να κάνουμε το διαχρονισμό και χρονοπρογραμματισμό των RSS feed ακόμα πιο ελεγχόμενο και πιο κοντά στους χρήστες σύμφωνα με την εξίσωση 5.10:

$$rank(f, t) = articles_f(t) \cdot (1 + c \cdot subscribers(f)) \tag{5.10}$$

Η παράμετρος c μπορεί να τροποποιηθεί για να δείξει το βάρος το οποίο μπορεί να έχει ο αριθμός των subscribers ενός RSS και άρα και το μέγεθος της επιρροής της συγκεκριμένης τιμής στη βαθμολόγηση ενός RSS. Εφόσον θέλουμε τους subscribers να επηρεάζουν σε μικρό ποσοστό επιλέγουμε μικρές τιμές για το c ενώ εφόσον θέλουμε ο αριθμός των subscribers να ρυθμίζει σε μεγάλο βαθμό τη βαθμολόγηση των RSS feeds επιλέγουμε τιμές κοντά στο 1. Φυσικά, εφόσον το σύστημα δε λειτουργεί με subscribers για κάθε RSS (ένος τροπος λειτουργίας που επιτρέπεται από το σύστημά μας) τότε η μεταβλητή θα πρέπει να μηδενιστεί προκειμένου να μη δημιουργούνται προβλήματα consistency αν και ο μηδενικός αριθμός των subscribers δε θα δημιουργήσει πρόβλημα. Τέλος επειδή όπως φαίνεται βασιζόμαστε στο hourly posting rate της τελευταίας ώρας τιμή η οποία μπορεί να μηδενιστεί εντελώς δημιουργώντας πρόβλημα καθότι θα θεωρηθεί από το σύστημα ότι δεν πρέπει να ελέγξει πάλι αυτό το άρθρο ορίζουμε μία κατώτατη τιμή την οποία μπορεί να πάρει το posting rate προκειμένου να μη μηδενιστεί.

5.2 Εξαγωγή Χρήσιμου Κειμένου από HTML σελίδες – εξαγωγή multimedia

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 [42] 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 σε προηγούμενο κεφάλαιο 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

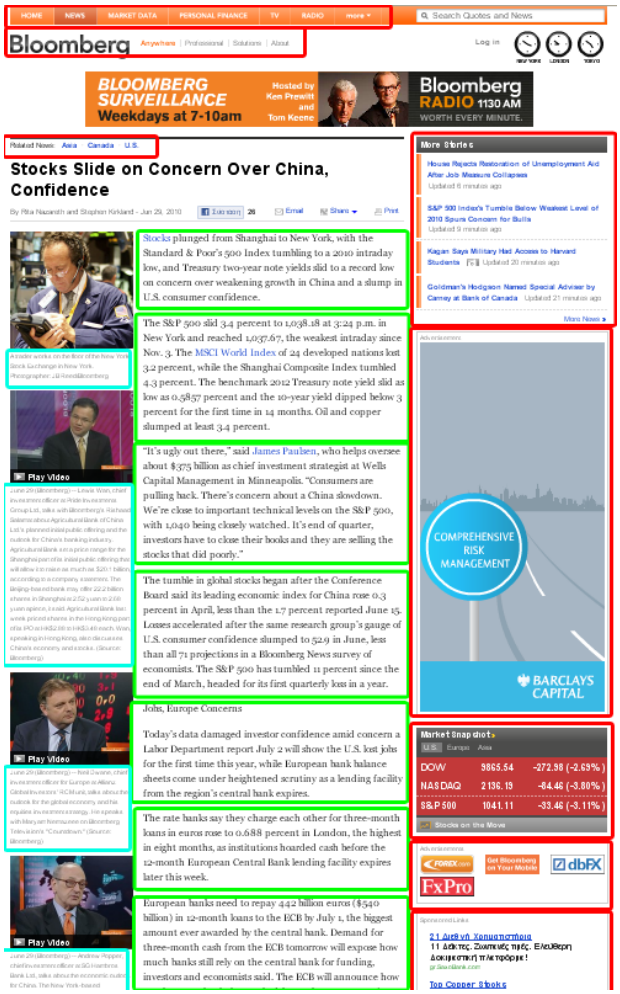
αφορά το μέγεθος κειμένου φύλλων που αφορά links. Τέλος θα πρέπει να χρησιμοποιηθούν μεταβλητές που θα εκφράζουν τη γειτονικότητα των φύλλων και συνεπώς να χρησιμοποιηθεί ένας αλγόριθμος για την αρίθμηση των κόμβων του δέντρου προκειμένου η αρίθμηση των φύλλων να είναι σειριακή. Έτσι παρά το γεγονός 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:
SH = το συνολικό μέγεθος του κειμένου σε bytes. Υπολογίζεται προσθέτωντας όλα τα SLx.
SLx = το μέγεθος κειμένου σε bytes για το φύλλο X. Υπολογίζεται μετρώντας τα bytes αλφαριθμητικών χαρακτήρων σε ένα φύλλο
SAx = το μέγεθος κειμένου του φύλλου X που περιέχεται σε tag <a> (link). Υπολογίζεται μετρώντας τα bytes αλφαριθμητικών μέσα σε tags <a> ενός φύλλου.
IX = 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:

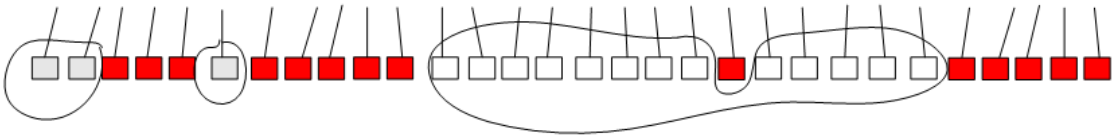
- LP = SAx / SLx. Πρόκειται για το Link Percentage το οποίο είναι μία ποσότητα που μας δείχνει πόσο από το κείμενο ενός φύλλου είναι κείμενο που βρίσκεται σε link. Αν αυτή η ποσότητα είναι μεγάλη αυτό σημαίνει πως ο συγκεκριμένος κόμβος είναι ένα navigation menu που η πλειονότητα του κειμένου του βρίσκεται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

- TP = SLx / SH. 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 IX κάθε φύλλου. DX,Y = IY – IX. 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 σχήμα 5.2 για τη λειτουργία του μηχανισμού.



Σχήμα 5.2: Χαρακτηρισμός περιοχών ιστοσελίδας από το μηχανισμό εξαγωγής χρήσιμου κειμένου



Σχήμα 5.3: Ομάδες γειτονικών φύλλων

Όπως φαίνεται και από το σχήμα 5.2 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.

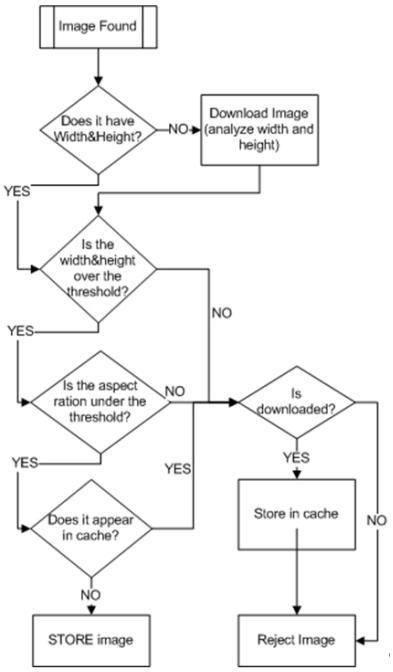
Η επιλογή βάσει γειτνίασης των φύλλων δεν είναι τόσο απλή όσο περιγράφεται παραπάνω. Ουσιαστικά περιλαμβάνει ένα σύνθετο αλγόριθμο που δημιουργεί ομάδες από γειτονικά φύλλα όπως φαίνεται στο σχήμα 5.3.

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.

5.2.1 Εξαγωγή Εικόνων

Αρχικός στόχος του μηχανισμού είναι να ανακαλύψει τα όρια του άρθρου για να προκύψει ένα μικρότερο υποσύνολο των εικόνων που πρέπει να εξεταστούν. Αυτή η διαδικασία αξιοποιεί στοιχεία που προέκυψαν από την εξαγωγή του χρήσιμου κειμένου, που στην περίπτωση του συστήματος που μελετάμε είναι ο μηχανισμός CUTER [39] που κατασκευάζει το DOM δέντρο της σελίδας. Τα στοιχεία που παρέχονται από την ανάλυση του δέντρου είναι πολύ σημαντικά αφού αυτά είναι που τελικά προσδιορίζουν τη θέση του κυρίως άρθρου μέσα στη σελίδα. Κατά τη διάρκεια της εξαγωγής του κειμένου, πραγματοποιείται ο χαρακτηρισμός των κόμβων του δέντρου, ενώ η τελική έξοδος αυτού του μηχανισμού είναι το σύνολο των χαρακτηρισμένων κόμβων, οι οποίοι ουσιαστικά αναπαριστούν τμήματα της σελίδας. Η τελική απόφαση του συστήματος σχετικά με το χρήσιμο κείμενο, γίνεται αφού επιλεγούν μόνο οι κόμβοι που έχουν αναφερθεί ότι περιλαμβάνουν χρήσιμη πληροφορία. Όταν αναλύουμε μια ιστοσελίδα στο DOM μοντέλο της, τότε κάθε στοιχείο του δέντρου ισοδυναμεί με ένα τμήμα του HTML κώδικα της. Έτσι, μετά την εκτέλεση του μηχανισμού εξαγωγής κειμένου, οι κόμβοι "χρήσιμης πληροφορίας" αναμένεται να περιέχουν κομμάτια του πραγματικού άρθρου. Ταυτόχρονα, με την ανάλυση του δέντρου, απαριθμούνται οι κόμβοι που το αποτελούν ακολουθώντας μια κατά βάθος αναζήτηση προκειμένου να γνωρίζουμε τη θέση τους στη σελίδα. Έχοντας τη θέση και το χαρακτηρισμό για τους κόμβους που μας ενδιαφέρουν, επιλέγουμε τον πρώτο και τον τελευταίο προκειμένου να καθοριστεί η θέση ολόκληρου του άρθρου.
Σε πολλές περιπτώσεις, τα άρθρα έτσι όπως εξάγονται από το μηχανισμό CUTER, παρόλο που περιέχουν το σώμα της είδησης, δεν περιλαμβάνουν τον τίτλο της. Αυτή η συμπεριφορά του μηχανισμού μπορεί να οδηγήσει σε απώλεια μιας εικόνας η οποία είναι τοποθετημένη μεταξύ του τίτλου και της αρχικής παραγράφου. Μια τέτοια διάταξη δεν πρέπει να θεωρηθεί ακραία περίπτωση αφού είναι πολύ συχνή στις ειδησεογραφικές σελίδες και μάλιστα παρατηρούμε ότι προτιμάται από τους περισσότερους συντάκτες. Η λύση σε αυτό το πρόβλημα που εμφανίζεται



Σχήμα 5.4: Διάγραμμα ροής εξαγωγής εικόνων

είναι η εύρεση εκείνου του κόμβου που περιέχει τον τίτλο και να τον θεωρήσουμε ως τον πρώτο κόμβο του άρθρου, ακόμα και αν δεν έχει χαρακτηριστεί ως κόμβος χρήσιμης πληροφορίας. Ωστόσο, ο τίτλος είναι εύλογο να εμφανίζεται παραπάνω από μια φορές μέσα στη σελίδα της είδησης, με χαρακτηριστικό παράδειγμα την εμφάνιση του στα μεταδεδομένα της HTML σελίδας. Το σύστημα χειρίζεται αυτή την περίπτωση με το να βρίσκει τον τελευταίο κόμβο που περιέχει τον τίτλο και ταυτόχρονα δεν βρίσκεται σε χαμηλότερη θέση από την αρχική παράγραφο του άρθρου.
Αφού ολοκληρωθεί η πρώτη φάση του αλγορίθμου και καθοριστεί το περίγραμμα του άρθρου, ο μηχανισμός εισάγεται στην κεντρική διαδικασία για την επιλογή των εικόνων που θα εξαχθούν. Η αρχική επιλογή γίνεται αφού εντοπιστούν οι εικόνες που περικλείονται στα όρια του άρθρου, χρησιμοποιώντας την αρίθμηση τους στο DOM δέντρο που στην ουσία εκφράζει την πραγματική θέση τους. Κάθε εικόνα που βρίσκεται, ελέγχεται βάσει του διαγράμματος ροής που φαίνεται στο σχήμα 5.4.

Η τελική επιλογή των εικόνων, βασίζεται στον έλεγχο των διαστάσεων τους. Ο πιο γρήγορος τρόπος να πραγματοποιηθεί αυτός ο έλεγχος χωρίς καν να ληφθεί η εικόνα, είναι η εξέταση των προκαθορισμένων χαρακτηριστικών που υποδεικνύουν πως θα φανεί η εικόνα στο φυλλομετρητή. Συγκεκριμένα, αναφερόμαστε στα χαρακτηριστικά ύψους (height) και πλάτους (width) της ετικέτας που χρησιμοποιεί η HTML για την εμφάνιση εικόνων. Ο μηχανισμός δέχε-

ται ως χρήσιμο περιεχόμενο την εικόνα όταν σε αυτό τον αρχικό έλεγχο προκύψει ότι τόσο το ύψος όσο και το πλάτος είναι μεταξύ των αποδεκτών ορίων και παράλληλα τα μεγέθη αυτά βρίσκονται σε σωστή αναλογία. Με τον όρο αναλογία εννοούμε το λόγο του πλάτους προς το ύψος της εικόνας (aspect ratio). Για τις ανάγκες του συστήματος που αναπτύσσουμε, χρησιμοποιούμε την εξίσωση 5.11.

$$ratio(w, h) = fracmax(w, h)min(w, h) \tag{5.11}$$

Έπειτα από αναλύσεις εικόνων που συνοδεύουν άρθρα ειδήσεων της βάσης δεδομένων του peRSSonal, παρατηρήσαμε ότι το μέσο πλάτος μιας εικόνας μπορεί να θεωρηθεί 300 pixels, ενώ το μέσο ύψος 270 pixels. Η πιο σημαντική παρατήρηση είναι ότι η συντριπτική πλειοψηφία των εικόνων ήταν πάνω από 150 pixels σε πλάτος και σε ύψος. Με βάση αυτή την παρατήρηση ο μηχανισμός επιλέχθηκε να λειτουργεί με όρια για το ελάχιστο αποδεκτό ύψος και πλάτος το οποίο ορίστηκε 150 pixels. Στόχος του καθορισμού των ελάχιστων ορίων είναι να απορρίπτονται οι πολύ μικρές εικόνες που ενδέχεται να είναι τοποθετημένες εντός της περιοχής του άρθρου. Τέτοιες εικόνες με μικρές διαστάσεις είναι συνήθως "εικονίδια" που αποτελούν μέρος του γραφιστικού σχεδιασμού της σελίδας. Ο έλεγχος της αναλογίας ύψους-πλάτους (aspect ratio) είναι αναγκαίος για την απόρριψη τόσο των διαφημίσεων, όσο και των λογότυπων των ίδιων των ιστοσελίδων. Αποτελεί κοινή παρατήρηση ότι οι περισσότερες διαφημίσεις έχουν αρκετά μεγάλη αναλογία ύψους-πλάτους συγκριτικά με τα πολυμέσα που σχετίζονται με την είδηση. Ο μηχανισμός στηρίζεται στο γεγονός ότι οι εικόνες που είναι ψηλές και στενές ή κοντές και πλατιές είναι πολύ πιθανόν να περιέχουν διαφημιστική πληροφορία. Είναι προφανές ότι αυτά τα όρια που τίθενται, επηρεάζουν τόσο την ακρίβεια όσο και την ανάκληση του συστήματος εξαγωγής. Για παράδειγμα, η μείωση των ορίων που τέθηκαν, θα οδηγούσε μεν σε μεγαλύτερο πλήθος εξαχθέντων εικόνων, αλλά θα ήταν αμφίβολο κατά πόσο αποτελούν χρήσιμη πληροφορία.

Η διαδικασία που παρουσιάστηκε, βασίζεται στο γεγονός ότι είναι δηλωμένες οι διαστάσεις της εικόνας μέσα στον HTML κώδικα και επομένως ήταν δυνατή η αποφυγή λήψης κάποιας εικόνας που θα είχε μη αποδεκτά χαρακτηριστικά. Ωστόσο, οι διαστάσεις μιας εικόνας δεν ορίζονται πάντα στον κώδικα της σελίδας ή ακόμα και αν ορίζονται μπορεί να είναι ελλιπείς, δηλαδή να δηλώνεται μόνο το ύψος ή μόνο το πλάτος. Ακόμα, η HTML επιτρέπει να γίνεται ο ορισμός των διαστάσεων βάσει των σχετικών διαστάσεων του τμήματος της σελίδας στο οποίο περιέχεται. Σε αυτές τις περιπτώσεις είναι επισφαλές να αποφανθούμε για το αν μια εικόνα πρέπει να εξαχθεί ή όχι, αφού δεν έχουμε τα πλήρη χαρακτηριστικά της. Έτσι, είναι αναπόφευκτη η λήψη της εικόνας με στόχο την ανάλυση των πραγματικών χαρακτηριστικών της. Για αυτή τη διαδικασία απαιτείται η χρήση κάποιας βιβλιοθήκης γραφικών που να μπορεί να αναλύσει τους διαφόρους τύπους φωτογραφιών που υπάρχουν στο διαδίκτυο.

Δεν πρέπει να ξεχνάμε ότι η ανάκτηση πολυμεσικών στοιχείων οδηγεί σε μεγάλη κατανάλωση πόρων και ότι ο μηχανισμός θα πρέπει να αποφεύγει περιττές λήψης δεδομένων. Η πιο διαδεδομένη τεχνική στα συστήματα του παγκόσμιου ιστού για αυτή την περίπτωση είναι η χρήση προσωρινής μνήμης προκειμένου να αποθηκεύεται πληροφορία για τις εικόνες που έχουν απορριφθεί στο παρελθόν. Η προσέγγιση που ακολουθούμε στηρίζεται στην ιδέα του πεπερασμένου συνόλου που αποδεικνύεται ότι μπορεί να λειτουργήσει πολύ ικανοποιητικά, χωρίς να απαιτείται μεγάλος αποθηκευτικός χώρος. Η υλοποίηση ενός τέτοιου συνόλου μπορεί να γίνει με οποιαδήποτε αφαιρετική δομή δεδομένων που μπορεί να αποθηκεύει δυναμικά συγκεκριμένες τιμές, οι οποίες να είναι αταξινόμητες και παράλληλα να μην επιτρέπεται η αποθήκευση της ίδιας τιμής δύο ή περισσότερες φορές. Συγκεκριμένα, οι τιμές που θα χρειαστεί να αποθηκευτούν θα πρέπει να μπορούν να χαρακτηρίσουν μοναδικά μια εικόνα και για αυτό το λόγο επιλέγουμε το URL της. Η διευθυνσιοδότηση της εικόνας σε HTML, γίνεται με το χαρακτηριστικό “src” της ετικέτας “img”. Σε αυτό το σημείο εκτελείται ο μηχανισμός που εξάγει το πλήρες URL της εικόνας, αφού γνωρίζουμε ότι η διευθυνσιοδότηση στον HTML κώδικα μπορεί να είναι είτε απόλυτη είτε σχετική.

Η μνήμη στην οποία αποθηκεύονται τα URLs, είναι αρχικά άδεια και επεκτείνεται με κάποιο νέο στοιχείο κάθε φορά που μια εικόνα έχει ανακτηθεί και κατόπιν απορριφθεί. Αξίζει να σημειώσουμε ότι οι εικόνες που ανήκουν σε αυτή την κατηγορία δεν χρησιμοποιούν διαφορετικές διευθύνσεις κάθε φορά. Όπως έχουμε αναφέρει τις περισσότερες φορές είναι μικρές στατικές εικόνες που αποτελούν μέρος της σχεδίασης της ιστοσελίδας με αποτέλεσμα να εμφανίζονται σε πολλά μέρη της. Οι ίδιοι οι σχεδιαστές για να επιταχύνουν τη φόρτωση της σελίδας, επιλέγουν να τοποθετούν τέτοιες εικόνες σε στατικές διευθύνσεις με στόχο την αποθήκευση αυτών των στοιχείων από τις μνήμες cache και τους διαμεσολαβητές (web proxy servers). Με την εισαγωγή κάποιας διεύθυνσης στο σύνολο των απορριφθέντων εικόνων, ο μηχανισμός μπορεί να αξιοποιήσει αυτή την πληροφορία στο μέλλον επιλέγοντας να αναζητήσει πρώτα την εικόνα στην προσωρινή μνήμη αντί να επαναλάβει τη λήψη της. Όλη η διαδικασία που περιγράφηκε καταναλώνει ένα μικρό μέρος αποθηκευτικού χώρου, προσφέροντας μακροπρόθεσμα χαμηλότερη κατανάλωση των πόρων του δικτύου και καλύτερο χρόνο εκτέλεσης. Ωστόσο, το κόστος της αναζήτησης στην προσωρινή μνήμη, δεν πρέπει να θεωρηθεί αμελητέο και για αυτό το λόγο αυτή η πρακτική εφαρμόζεται μόνο στο στάδιο που μια εικόνα έχει αποφασιστεί να ανακτηθεί.

Το επόμενο στάδιο εκτέλεσης αποτελεί ένα στάδιο επεξεργασίας για τις εικόνες που ανακτήθηκαν, αφού δεν ήταν δυνατό να βρεθούν στη μνήμη. Αρχικά ελέγχεται το μέγεθος σε bytes της εικόνας με στόχο να αποφύγουμε την ανάλυση αυτών που είναι πολύ μικρές σε μέγεθος. Για τον έλεγχο χρησιμοποιούμε ένα κατώφλι, κάτω από το οποίο θεωρούμε ότι είναι απίθανο να βρεθεί μια σχετική με το άρθρο εικόνα. Έπειτα από έρευνα στις φωτογραφίες των άρθρων ειδήσεων που έχουν συλλεχθεί, φαίνεται ότι το μέγεθος των εικόνων που μας ενδιαφέρουν είναι κατά μέσο όρο περίπου 25 kilobytes, ενώ η μικρότερη που βρέθηκε είναι 4 kilobytes. Λαμβάνοντας υπ΄ όψιν τα παραπάνω στοιχεία, θέτουμε το κατώφλι στα 2 kilobytes που συνήθως αντιστοιχεί σε εικόνα

πολύ μικρών διαστάσεων. Καταλήγοντας, τα αρχεία που λαμβάνονται και είναι μεγαλύτερα από αυτό το όριο, αναλύονται με τη βοήθεια της βιβλιοθήκης δυναμικού χειρισμού εικόνων GD. Αυτή τη στιγμή, το σύστημα είναι σε θέση να ανακτήσει το ύψος και το πλάτος και να αναλύσει εικόνες που ανήκουν στους παρακάτω τύπους αρχείων:

- GIF
- JPEG
- PNG

Αξίζει να αναφέρουμε ότι αυτοί οι τύποι εικόνας είναι οι πιο δημοφιλείς όσον αφορά την εφαρμογή τους σε σελίδες του παγκόσμιου ιστού. Έχοντας λοιπόν αναλύσει τις εικόνες, ο μηχανισμός είναι σε θέση να γνωρίζει τις ακριβείς διαστάσεις τους και να αποφανθεί για το αν πρέπει να εξαχθούν με παρόμοιο τρόπο όπως γίνεται στον αλγόριθμο που ακολουθεί.

```
images[] = Find_Images(startNode, endNode);

Foreach(images[] as image)
   If (Width_Height_Tags_Found(image)) Then
      If(image.width >= WIDTH_THRESHOLD &&
        image.height >= HEIGHT_THRESHOLD &&
        aspectRatio(image)<=RATIO_THRESHOLD)
      Then
         accept(image);
      Else
         reject(image);
      End If
   Else If (isInCache(image))
      Then
         reject(image);
      Else
         imageFile = download(image);
         findRealDimensions(imageFile);

         If (size(imageFile)>=SIZE_THRESHOLD &&
           image.width >= WIDTH_THRESHOLD &&
           image.height >= HEIGHT_THRESHOLD &&
           aspectRatio(image)<=RATIO_THRESHOLD)
         Then
```

```
            accept(image);
          Else
            addToCache(image);
            reject(image);
          End If
    End If
End For
```

5.3 Προεπεξεργασία κειμένου

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 λέξεων κλειδιών (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) [64].
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

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.

5.3.1 Μηχανισμός Προεπεξεργασίας για την Ελληνική γλώσσα

Στο πλαίσιο της εργασίας και προκειμένου να υπάρχει πλήρης υποστήριξη για το μηχανισμό, δημιουργήθηκε και ένας ελληνικός προεπεξεργαστής ο οποίος είναι σε θέση να πραγματοποιήσει την παραπάνω διαδικασία για προεπεξεργασία κειμένου αλλά σε ελληνικά κείμενα [38]. Το παρακάτω σχήμα δείχνει τη διαδικασία του συστήματος η οποία προσεγγίζει τον τρόπο με τον οποίο διεξάγεται η προεπεξεργασία από το σύστημά μας αλλά δεν ταυτίζεται.
Ο μηχανισμός stemming δημιουργήθηκε αρχικά προκειμένου να παρέχει υπηρεσίες stemming

για την ελληνική γλώσσα για το μηχανισμό perssonal. Στο επίπεδο που λειτουργεί μέσα στο σύστημα peRSSonal δέχεται σαν είσοδο XML αρχεία που περιέχουν τον τίτλο και το κείμενο άρθρων και σα σκοπό έχει να εφαρμόσει αλγορίθμους προεπεξεργασίας και να προσφέρει keywords στα συστήματα που ακολουθούν.
Στην ουσία ο ελληνικός προεπεξεργαστής, tagger και stemmer χρησιμοποιεί αλγορίθμους αφαίρεσης της κατάληξης προκειμένου να καταφέρει να εξάγει τη ρίζα μίας λέξης. Το σύστημα δέχεται σαν είσοδο ολόκληρες προτάσεις και χρησιμοποιεί το σύστημα tagging προκειμένου να εντοπίσει τα σωστά suffixes που πρέπει να αφαιρεθούν. Ο αλγόριθμος αφαίρεσης των καταλήξεων βασίζεται σε κανόνες και χρησιμοποιεί πίνακες με πιθανές καταλήξεις ανάλογα με το μέρος του λόγου. Η έξοδος του συστήματος είναι λίστα με τις λέξεις εισόδου, τη stemmed λέξη αλλά και το μέρος του λόγου στο οποίο ανήκει η λέξη (σύμφωνα πάντα με το σύστημα tagging).
Η πολυπλοκότητα του προτεινόμενου συστήματος είναι O(n) και φυσικά όπως ήδη αναφέρθηκε μπορεί να θεωρηθεί σαν συνδυασμός δύο διαφορετικών διαδικασιών: διαδικασίας tagging και διαδικασίας stemming. Παρά το γεγονός πως ο βασικός σκοπός είναι η δημιουργία ενός μηχανισμού stemming για την ελληνική γλώσσα εντούτοις δημιουργήθηκε παράλληλα και ο μηχανισμός tagging για δύο βασικούς λόγους. Ο πρώτος λόγος είναι γιατί μερικές λέξεις που ανήκουν σε συγκεκριμένα μέρη του λόγου δεν περιέχουν καθόλου πληροφορία και έτσι ο μηχανισμός σαν κομμάτι του peRSSonal μπορεί να τις απορρίπτει εντελώς χωρίς να τις θεωρεί κομμάτι της διαδικασίας. Χαρακτηριστικά, άρθρα, αντωνυμίες, πολλές φορές επίθετα ή και πολύ συνηθισμένα ρήματα δε μπορεί σε καμία περίπτωση να αποτελέσουν κομμάτι του νοήματος ενός κειμένου. Μάλιστα, αυτός ο διαχωρισμός μπορεί να λειτουργήσει ακόμα και πιο αποδοτικά από λίστες με stop-words και αφαίρεση μικρών σε μέγεθος λέξεων. Από την άλλη και ειδικά για την ελληνική γλώσσα, είναι πολύ εύκολο μερικές περιπτώσεις λέξεων και κυρίως λέξεις που κλίνονται ανώμαλα να μην είναι εφικτό να υποστούν stemming χωρίς πρώτα να εντοπίσουμε το μέρος του λόγου στο οποίο ανήκουν. Ο αλγόριθμος λειτουργεί σύμφωνα με τα παρακάτω βήματα:

1. Η είσοδος είναι ένα array από χαρακτήρες

2. Αρχικά όλοι οι χαρακτήρες πλην των αγγλικών και των ελληνικών γραμμάτων αφαιρούνται

3. Εν συνεχεία το κείμενο μετατρέπεται σε μικρά γράμματα ενώ τα σημεία στίξης μένουν άθικτα καθότι στην περίπτωση της ελληνικής γλώσσας παίζουν εξαιρετικά σημαντικό ρόλο για το tagging.

4. Η αρχική επεξεργασία κλείνει με δημιουργία συνδεδεμένων λιστών οι οποίες είναι έτοιμες να υποδεχθούν πληροφορία για κάθε λέξη.

Περνώντας στην αμιγώς αναλυτική διαδικασία tagging και stemming θα ξεκινήσουμε από το tagging του κειμένου το οποίο περιλαμβάνει δύο βασικά στάδια. Σε πρώτη φάση

κάθε λέξη ελέγχεται αν μπορεί απ' ευθείας να ενταχθεί σε κάποιο γραμματικό τύπο χρησιμοποιώντας λίστες με γνωστές καταλήξεις λέξεων από την Ελληνική Γραμματική [186]. Μερικές κατηγορίες γραμματικών τύπων που έχουν αυστηρά δικές του καταλήξεις σταματούν τη διαδικασία αν η λέξη βρεθεί σε αυτές ενώ άλλες περιμένουν να τελειώσει ο πρώτος γύρος. Σε περιπτώσεις που μία λέξη ταιριάζει σε περισσότερες από μία κατηγορίες τότε ελέγχεται η προηγούμενη λέξη καθότι ο γραμματικός τύπος των λέξεων καθορίζεται πλήρως από το χαρακτηρισμό της προηγούμενης. Παράλληλα με όλη την παραπάνω διαδικασία αποθηκεύεται και πληροφορία που σχετίζεται με τα προθέματα των λέξεων διότι είναι πολύ χρήσιμη για λέξεις που κλίνονται ανώμαλα. Ο δεύτερος γύρος του tagging έχει σα σκοπό να εντοπίσει όλες αυτές τις λέξεις που δεν κατηγοριοποιήθηκαν στην πρώτη προσπάθεια ενώ σαν ύστατη λύση δοκιμάζεται να βρεθεί κάποια κατηγοριοποίηση των λέξεων αφαιρώντας κάθε σημείο στίξης.
Έπειτα από την παραπάνω διαδικασία και αφού έχουμε καταφέρει να βρούμε (για όσες λέξεις έχουμε καταφέρει) το γραμματικό τύπο των λέξεων προχωρούμε στη διαδικασία του stemming. Σύμφωνα με αυτή τη διαδικασία από τις λέξεις αφαιρείται η κατάληξη σύμφωνα με του πίνακες καταλήξεων που υπάρχουν στον κάθε γραμματικό τύπο της εκάστοτε λέξης. Για τα ρήματα αλλά και γενικά για εξαιρέσεις λέξεων που γνωρίζουμε ότι κλίνονται ανώμαλα εφαρμόζεται επιπλέον διαδικασία εύρεσης σωστής ρίζας με τη βοήθεια της χρήσης προθέματος λέξης που έχει εντοπιστεί από προηγούμενη διαδικασία όπως ήδη αναφέρθηκε.

5.4 Εξαγωγή Περίληψης 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 [63] και [62].
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 5.12 για μια γενική βαθμολόγηση των προτάσεων.

$$S_i = \sum w_{k,i}(k_1 + k_2) \tag{5.12}$$

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 k2 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:

$$k_1 = 1 + 0.1x \tag{5.13}$$

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

$$k_2 = 1 + 1.2y \tag{5.14}$$

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:

$$y = fracnt * fracmt \tag{5.15}$$

Για να κανονικοποιήσουμε τις τιμές που προκύπτουν από την εξίσωση 5.12, 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 5.12, 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 5.12 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.

5.5 Μηχανισμός Κατηγοριοποίησης

Το υποσύστημα κατηγοριοποίησης βασίζεται στην μετρική ομοιότητας συνημιτόνου, σε εσωτερικά γινόμενα πινάκων και σε υπολογισμούς ζυγίσματος βαρών [44] και [64]. 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.

$$k_3 = \{ A \cdot cw_i 1 \tag{5.16}$$

Η παράμετρος A πρέπει να είναι μεγαλύτερη από το 1 και χρησιμοποιείται για να προσθέσει βάρος για την παράμετρο k3. Εάν θέλουμε η διαδικασία περίληψης να βασίζεται κυρίως στο k3, 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 k1 και k2. 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 5.12 μετατρέπεται ως εξής:

$$S_i = \sum w_{k,i}(k_1 + k_2) \cdot k_3 \tag{5.17}$$

5.6 Προσωποποίηση στο Χρήστη

Όπως έχουμε δει η πρώτη διαδικασία του συστήματος είναι η ανάκτηση άρθρων από το Διαδίκτυο με έναν περιοδικό και μεθοδικό τρόπο. Η επόμενη διαδικασία περιλαμβάνει βήματα ανάλυσης HTML σελίδων, εξαγωγή του χρήσιμου κειμένου, η κατηγοριοποίηση αλλά και η εξαγωγή περίληψης. Σε αυτό το σημείο έχουμε όλα τα προαπαιτούμενα, λοιπόν, για να παρουσιάσουμε την πληροφορία στο χρήστη. Για να το πραγματοποιήσουμε αυτό θα πρέπει να έχουμε πληροφορία για το προφίλ του χρήστη [58]. Η δημιουργία του προφίλ γίνεται σε ένα βήμα ενώ η διαμόρφωση του προφίλ είναι συνεχής. Με τη δημιουργία του προφίλ, τη διατήρησή του και τη συνεχή ανανέωση έχουμε τη δυνατότητα να προσφέρουμε ποιοτικές υπηρεσίες στους χρήστες οι οποίες είναι πλήρως προσαρμοσμένες στις ανάγκες τους. [65]

Για τις ανάγκες της προσωποποίησης έχουμε στην ουσία δύο βασικά βήματα. Σε πρώτο βήμα έχουμε την αρχική δημιουργία του προφίλ ενώ στην πορεία η ανανέωση του προφίλ είναι αέναη και εξαρτάται από τον τρόπο χρήσης του συστήματος από τον εκάστοτε χρήστη. Κάποια πρώτη πληροφορία αντλείται άμεσα από τους χρήστες όταν αυτοί εγγράφονται στο σύστημα. Αυτή η πληροφορία είναι προαιρετική ωστόσο αν δε δοθεί το προφίλ θα είναι πολύ γενικό και θα αργήσει να προσαρμοστεί στους χρήστες. Η διαδικασία αυτή πραγματοποιείται μέσα από το σύστημα εγγραφής (registration). Σε αυτή τη διαδικασία ο χρήστης καλείται να εισάγει τις προσωπικές του επιλογές αναφορικά με επτά βασικές κατηγορίες που διαθέτει το σύστημα. Οι κατηγορίες είναι: (a) business, (b) entertainment, (c) health, (d) politics, (e) technology, (f) education and (g) science. Η επιλογές του χρήστη αναφορικά με τις κατηγορίες είναι από -5 έως +5. Η επιλογή της αρνητικής διάθεσης προς μία κατηγορία αυτόματα σημαίνει και μηδενική προβολή άρθρων που σχετίζονται με αυτή την κατηγορία. Η κλίμακα χρησιμοποιείται για να δείξει το βαθμό εκκαθάρισης των εναπομεινάντων άρθρων από πληροφορία που σχετίζεται με επιλογές τις οποίες έχει αξιολογήσει αρνητικά ο χρήστης. Οι θετικές επιλογές δείχνουν τις κατηγορίες τις οποίες επιθυμεί ένας χρήστης να βλέπει και η κλίμακα δείχνει και εδώ το μέγεθος το οποίο ενδιαφέρει ένα χρήστη να παρακολουθεί.

Για κάθε μία κατηγορία υπάρχει ένας πίνακας στη ΒΔ ο οποίος είναι κατασκευασμένος με ζεύγη από λέξεις κλειδιά και τιμές που αντιπροσωπεύουν τις λέξεις κλειδιά που ανήκουν σε κάθε κατηγορία μαζί με μία τιμή η οποία δείχνει πόσο αντιπροσωπευτική είναι η λέξη γι αυτή την κατηγορία. Ο συσχετισμός προκύπτει από το TF-IDF βάρος κάθε λέξης κλειδί που προκύπτει από συλλογές κειμένων τα οποία κείμενα ανήκουν στις παραπάνω κατηγορίες. Η συλλογή των κειμένων έχει γίνει χειροκίνητα και πολλά από τα κείμενα έχουν απορριφθεί μετά από έλεγχο. Μάλιστα, στις τελικές λέξεις κλειδιά που είναι εκπρόσωποι κατηγοριών έχει πραγματοποιηθεί αναλυτικός έλεγχος για να δούμε το ενδεχόμενο να ανήκουν και σε άλλες κατηγορίες όπου ακολούθησε και δεύτερο κύκλος απόρριψης. έτσι, τελικά καταφέρνουμε να έχουμε λέξεις κλειδιά που αντιπροσωπεύουν σε μεγάλο βαθμό κάποια κατηγορία. Ο βαθμός αυτός εκφράζεται από την τιμή που ακολουθεί κάθε λέξη κλειδί της συλλογής μας. Η δημιουργία αυτών των κατηγοριών είναι ένα ξεχωριστό ζήτημα για την εργασία μας αλλά ωστόσο έχει καλυφθεί από την εργασία [2].

Το προφίλ του κάθε χρήστη όπως και οι κατηγορίες του συστήματος δεν είναι τίποτα περισσότερο από λέξεις κλειδιά ακολουθούμενα από έναν αριθμό. Ο αριθμός αυτός διαφέρει για το χρήστη απ' ότι για τις κατηγορίες στο ότι ο χρήστης μπορεί να έχει τόσο θετικά όσο και αρνητικά στοιχεία για τις λέξεις κλειδιά που διαθέτει στο προφίλ του. Όπως είναι φυσικό δεν είναι απλό έχουμε λέξεις κλειδιά με αρνητικό βάρος για μία κατηγορία για τον απλό λόγο ότι δεν υπάρχει μέτρο που να μας επιτρέπει να ορίσουμε το μέγεθος της μη συσχέτισης. Έτσι, θεωρούμε ότι οι λέξεις κλειδιά που δεν ανήκουν σε μία κατηγορία έχουν αρνητικό πρόσημο για τη συγκεκριμένη κατηγορία αλλά το μέγεθος αυτό δεν είμαστε σε θέση να το προσδιορίσουμε. Αντίθετα για το χρήστη ο προσδιορισμός της τιμής είναι σχετικά απλός γιατί ο κάθε χρήστης είναι μία μίξη κατηγοριών ίσως και μία ξεχωριστή κατηγορία όπου τόσο τα θετικά όσο και τα αρνητικά μπορούν

να εξαχθούν μέσα από τις διαδικασίες. Για αν φτιάξουμε το προφίλ ενός χρήστη, όπως ήδη αναφέραμε, εμφανίζουμε στο χρήστη έναν πίνακα αξιολόγησης των κατηγοριών. Ανάλογα με την επιλογή του χρήστη αρχικοποιείται το προφίλ του σύμφωνα με την εξίσωση 5.18.

$$\beta(x) = \sum_{\kappa=1}^{n} \beta\kappa_x(\kappa) * \epsilon(\kappa) \tag{5.18}$$

όπου β είναι η συσχέτιση για τη λέξη κλειδί x, $\beta\kappa_(x)$ είναι η συσχέτιση για τη λέξη κλειδί x στην κατηγορία κ και $\epsilon(\kappa)$ είναι η επιλογή του χρήστη για κάποια κατηγορία. ϵ ορίζεται ως:

$$\epsilon(\kappa) = \Delta * X_{\kappa}^{2} \tag{5.19}$$

Όπου Δ μπορεί να πάρει τιμές από 1 έως 5 σύμφωνα με το πόσο θέλουμε οι αρχικές επιλογές του χρήστη να επηρεάσουν τη δημιουργία του προφίλ του. Η τιμή 1 θεωρείται πολύ απαλή και συνήθως μας δίνει ένα γενικευμένο προφίλ ανεξάρτητα των επιλογών που κάνει ο χρήστη. Από την άλλη η τιμή 5 είναι πολύ μεγάλη ειδικά αν ο χρήστης έχει φτάσει τις επιλογές του στα άκρα (-5, +5) και αυτό μπορεί να έχει έντονα αρνητική επίδραση στο προφίλ του. Σε γενικές γραμμές φροντίζουμε να ελέγχουμε πριν θέσουμε τιμή για το Δ. Αν ο χρήστης έχει κάνει ακραίες επιλογές για τις κατηγορίες (πολλά -5 ή +5) τότε χρησιμοποιούμε ακόμα και την τιμή 1, ενώ αν ο χρήστης έχει επιλέξει γενικά μικρές τιμές (>-2 και < 2) τότε επιλέγουμε να χρησιμοποιήσουμε την τιμή 5. Γενικά, η τιμή που χρησιμοποιούμε στην πλειοψηφία των περιπτώσεων είναι η μέση τιμή, δηλαδή το 3.

Η συνάρτηση 5.18 στην ουσία επιλέγει τα πρώτα $\epsilon(\kappa)$ λέξεις κλειδιά από την κατηγορία (κ) τα οποία θεωρούνται και πιο αντιπροσωπευτικά για μία κατηγορία. με αυτό τον τρόπο καταφέρνουμε να δημιουργήσουμε το αρχικό προφίλ για κάθε χρήστη. Αυτό το αρχικό προφίλ είναι ένας πίνακας με λέξεις κλειδιά ακολουθούμενα τιμές που όπως ήδη αναφέραμε μπορεί να είναι είτε θετικές είτε αρνητικές. Οι θετικές τιμές χρησιμοποιούνται αρχικά από το σύστημα για να μπορέσουμε να ανακτήσουμε όλη αυτή την πληροφορία που ενδιαφέρει το χρήστη ενώ οι αρνητικές τιμές εκτός του ότι βοηθούν να αποφύγουμε πληροφορία που δεν ενδιαφέρει το χρήστη μας βοηθούν να φιλτράρουμε περαιτέρω την πληροφορία που του εμφανίζουμε με κατάλληλη βαθμολογία που πραγματοποιούμε για κάθε άρθρο που έχει επιλεγεί προς εμφάνιση. Ουσιωδώς αυτό που πραγματοποιούμε είναι ένας αλγόριθμος συσχέτισης κειμένων των άρθρων που υπάρχουν στη ΒΔ με το χρήστη του συστήματος.

Για να διατηρήσουμε το προφίλ ενός χρήστη χρησιμοποιούμε ένα σύνολο αλγορίθμων οι οποίοι προκύπτουν από τη χρήση των υπηρεσιών που προσφέρονται από το perssonal meta-portal. Ο σκοπός του κάθε αλγορίθμου σε αυτό το σημείο είναι η συνεχής ενημέρωση του προφίλ του χρήστη και ο τρόπος ενημέρωσης βασίζεται αποκλειστικά σε έμμεση ερμηνεία των ενεργειών που πραγματοποιεί ο χρήστης χωρίς να ζητείται ουσιωδώς άμεση είσοδος πληροφορίας. Αυτός

άλλωστε είναι και ο βασικό σκοπός μας να μη βάζουμε δηλαδή το χρήστη στη διαδικασία να διαμορφώσει με κάποιο τρόπο το προφίλ του αλλά να ερμηνεύουμε τις ενέργειες που πραγματοποιεί προκειμένου να διαμορφώνουμε ένα προφίλ.
Η πειραματική διαδικασία μας έχει δείξει πως ένα 10% το κειμένου μπορεί να περιέχει όλο το νόημά του κάτι το οποίο μεταφράζεται σε 10-15 λέξεις κλειδιά ανά κείμενο. Άλλωστε δεδομένου ότι ένα κείμενο περιέχει 100 περίπου λέξεις κλειδιά είναι επόμενο η σημασία του κειμένου να βρίσκεται στις 10 πιο συχνές λέξεις του κειμένου. Η πληροφορία που συλλέγεται προκειμένου να ανανεώσουμε το προφίλ ενός χρήστη είναι από τα άρθρα που επιλέγει να διαβάσει, από αυτά που δε διαβάζει, από αυτά που απορρίπτει (blacklist), ενώ σημαντική πληροφορία συλλέγεται και κατά τη διάρκεια ανάγνωσης ενός άρθρου. Όταν ο χρήστης συνδέεται με το σύστημα και ξεκινά μία συνεδρία, οι κινήσεις που πραγματοποιεί καταγράφονται πλήρως. Για κάθε λέξη κλειδί των άρθρων που εμφανίζονται στο χρήστη και αυτός επιλέγει ή απορρίπτει διατηρείται ένα βάρος. Όταν ο χρήστης διαβάζει ένα άρθρο μία αρχική εκτίμηση του βάρους δίνεται από τη συνάρτηση 5.20.

$$weight(keywords) = \frac{min(timereading(x), time2read(length(x)))}{time2read(length(x))} \cdot k \tag{5.20}$$

$$k = (1 + \frac{articleposition(x)^2}{\sqrt{articleposition(1)^2 + articleposition(2)^{+} articleposition(n)^2}}) \tag{5.21}$$

Αφού ο χρήστης ανοίξει ένα άρθρο για να το διαβάσει χρησιμοποιούμε έναν αλγόριθμο προκειμένου να αξιολογήσουμε τις λέξεις κλειδιά ανάλογα με τις ενέργειες που κάνει ο χρήστης. Αυτές οι ενέργειες ποικίλουν ανάλογα με τους χρήστες αλλά σε γενικές γραμμές είναι: (α) ανάγνωση όλου του κειμένου ενός άρθρο (όχι μόνο την περίληψη), (β) επίσκεψη στη σελίδα απ' όπου προκύπτει το άρθρο, (γ) επιλογή συναφών άρθρων, (δ) επιλογή tagging, (ε) επιλογή υπηρεσίες παρακολούθησης, κ.α. Κάθε φορά που ένας χρήστης χρησιμοποιεί το σύστημα και πραγματοποιεί οποιαδήποτε από τις παραπάνω ενέργειες αυτόματα μεταβάλλεται το προφίλ του.

```
Update_profile (a, b, c) {
Get_articles(a,b)
for each article {
if (full article)
if (time_viewed>Rar_ thr1 && time_viewed< Rar_thr2) {
Keywords_positive = article's top 5 frequent keywords
```

```
Update_list(Positive, Keywords_positive)}
else
//this is a summary
if(time_viewed > Rsum_thr1 && time_viewed < Rsum_thr2){
Keywords_positive = article's top 5 frequent keywords
Update_list(Positive, Keywords_positive)}
// also recover the negative articles
Get_articles(c)
for each article{
Keywords_negative = select top 5 frequent keywords
Update_list(Negative, Keywords_negative)
}
Get_article(lists){
//Recovers from the database browsed articles
//and the amount of time spent reading the full article or
//its summary (a,b) Recovers also the negative articles (c)
}
Update_list(list, keywords){
//either add the keyword to the list or increase its frequency
for each (keyword in keywords)
if (keyword not in list[])
list.add(keywords[keyword])
else
list.update_freq(keywords[keyword])
}
}
```

Κάθε μία από τις ενέργειες έχει και ένα ειδικό βάρος το οποίο φαίνεται στον πίνακα 5.1.

Πίνακας 5.1: Βάρη για αλλαγή του προφίλ χρήστη

Action and Weight (percentage)	
Read Similar Articles	10
Load Tagging	10
Track Article	20
"Star" Article	20
Load Original Page	10
Remove Article	-60

Τα βάρη προέρχονται από πειράματα που έγιναν σε χρήστες του συστήματος για το μέγεθος επιρροής που έχουν διάφορες ενέργειες των χρηστών και τη σημασία που έχει κάθε ενέργεια. Είναι σαφές πως το να θέλει κανείς να παρακολουθήσει ένα θέμα είναι πιο σημαντικό από την επιλογή ανάγνωσης παραπλήσιων άρθρων ενώ από την άλλη η απόρριψη ή η διαγραφή ενός άρθρου έχει πολύ αρνητική επιρροή γιατί ειδικά το δεύτερο θεωρείται ακραία ενέργεια δήλωσης προτίμησης ενός χρήστη.
Κάθε ενέργεια του χρήστη καταγράφεται και όταν μία σελίδα κλείνει ή αλλάζει τότε πραγματοποιείται μία πρώτη ανάλυση και ανανέωση σύμφωνα με την εξίσωση 5.20. Στο τέλος κάθε session κάθε λέξη κλειδί η οποία εμφανίστηκε σε ένα από τα έγγραφα με τα οποία ήρθε ο χρήστης σε επαφή θα έχει πάρει ένα βάρος, θετικό ή αρνητικό. Αν η λέξη κλειδί υπάρχει ήδη στο προφίλ του χρήστη, τότε γίνεται απλώς μία ανανέωση του βάρους αυτής της λέξης ενώ αν η λέξη κλειδί δεν υπήρχε στο προφίλ του χρήστη τότε αυτή προστίθεται. Προκειμένου να αποφύγουμε ακραίες τιμές για τις λέξεις κλειδιά που ανανεώνονται ή προστίθενται στο σύστημα δεν επιτρέπουμε για τις λέξεις που ανανεώνουμε να ξεπεράσουμε κατά τρεις φορές το υπάρχον βάρος σε κάθε ανανέωση ενώ για τις νέες λέξεις που εισάγονται στο προφίλ δεν επιτρέπουμε η τιμή τους να είναι διπλάσια από τη μέγιστη τιμή που υπάρχει στο προφίλ του χρήστη. Τα όρια υπάρχουν για να μη γίνονται ακραίες αλλαγές στο προφίλ του χρήστη από μία και μόνο συνεδρία αλλά να υπάρχει η δυνατότητα για ομαλή μετάβαση προς το σταθερό προφίλ του χρήστη.

5.7 Βοηθητικά Συστήματα

Κατά τη διάρκεια εκτέλεσης και ανάλυσης του μηχανισμού και ακριβώς επειδή ο μηχανισμός περιέχει πληθώρα συστημάτων ανακαλύψαμε πολλές αλλαγές και βελτιώσεις που μπορούν να γίνουν. Έτσι λοιπόν κατά τη διάρκεια εκτέλεσης της εργασίας πραγματοποιούσαμε παράλληλα και ανάπτυξη υποσυστημάτων με τρόπο κλιμακωτό που μπορούσαν να εκτελεσθούν χωρίς να επηρεάζουν τη λειτουργία του όλου συστήματος. Κάποια από τα υποσυστήματα αυτά που αξίζουν αναφοράς μιας και έχουν παρουσιαστεί σε διεθνή συνέδρια είναι το υποσύστημα online document grouping, το υποσύστημα trash article detection, και τα συστήματα search caching και personalized search και τέλος το σύστημα εξαγωγής περίληψης.

5.7.1 On-line document grouping

Ο μηχανισμός ενοποίησης άρθρων είναι μία διαδικασία η οποία λαμβάνει χώρα όταν παρουσιάζονται άρθρα σε κάποιο χρήστη και έχει σαν σκοπό να ενοποιήσει όλα αυτά τα άρθρα που παρουσιάζουν ακριβώς το ίδιο περιεχόμενο αλλά προκύπτουν από διαφορετικές πηγές [58]. Για να το επιτύχουμε αυτό θα πρέπει να αφαιρέσουμε διπλοεγγραφές από τη βάση αλλά

και να εφαρμόσουμε μία τεχνική εύκολης διασύνδεσης των άρθρων. Πέραν της παρουσίασης ενοποιημένων άρθρων για το ίδιο θέμα θα επιτύχουμε και μεγαλύτερη ταχύτητα γιατί πολλά άρθρα δε θα εμφανίζονται σαν ξεχωριστές οντότητες στο χρήστη αλλά σαν κομμάτι ενός συνόλου κειμένων. Ο αλγόριθμος που χρησιμοποιούμε προκειμένου να εντοπίσουμε όλα τα παραπλήσια άρθρα χρησιμοποιεί τη συσχέτιση συνημιτόνου και το σημαντικό είναι πως γίνεται δυναμικά και σε πραγματικό χρόνο. Προκειμένου να παρουσιάσουμε πως πραγματοποιείται η ενοποίηση των άρθρων θα δεχθούμε δύο βασικές αρχές: α. το σύστημα δεν έχει κάνει καμία ενοποίηση και άρα όλα τα άρθρα είναι μεμονωμένα και β. κάθε άρθρο μπορεί να ενοποιηθεί με ένα άλλο αν έχουν διαφορά 16 ωρών το πολύ και γ. το πιο παλιό άρθρο ενός πυρήνα ενιαίων άρθρων με το νεότερο άρθρο του ίδιου πυρήνα δεν πρέπει να έχουν διαφορά μεγαλύτερη των 16 ωρών. Αυτό πηγάζει από παλαιότερη εργασία μας όπου είχαμε αποδείξει πως ο χρόνος μέσα στον οποίο μπορεί να δημοσιευθεί ταυτόσημο άρθρο από διαφορετικές πηγές που συνηθίζουν να δημοσιεύουν πολύ συχνά είναι +-8 ώρες δηλαδή απόσταση 16 ωρών.

Όταν ένας χρήστης επιλέγει να δει ένα άρθρο (ή γενικά όταν εμφανίζεται ένα άρθρο σε ένα χρήστη), μια συνάρτηση αναλύει με ασύγχρονο τρόπο και λαμβάνει την ομάδα των άρθρων που σχετίζονται άμεσα με το άρθρο που διαβάζει ο χρήστη. Αν υπάρχει ήδη η ομάδα άρθρων τότε το σύνολο αυτών εμφανίζεται στο χρήστη χωρίς να χρειάζεται καμία ασύγχρονη κλήση. Παράλληλα, ακόμα κι αν υπάρχει η ομάδα, και επειδή το σύστημα προσθέτει νέα άρθρα κάθε 6-10 λεπτά αν το πιο νέο άρθρο στην ομάδα δεν είναι παλαιότερο από 16 ώρες τότε γίνεται έλεγχος μήπως υπάρχει και νέο άρθρο το οποίο μπορεί να εισαχθεί στον πυρήνα των ταυτόσημων άρθρων. Αν λοιπόν η ομάδα των άρθρων δεν υπάρχει, δημιουργείται την ώρα που ο χρήστη κάνει ανάγνωση. Ο χρήστης συμμετέχει και σε αυτή τη διαδικασία συστήματος καθότι ο έλεγχος για ταυτόσημα άρθρα δεν ελέγχει αποκλειστικά και μόνο τη συσχέτιση μεταξύ των άρθρων αλλά και τη συσχέτιση των ταυτόσημων άρθρων με τον ίδιο το χρήστη. Έτσι μπορεί κάποιο άρθρο να παρουσιάζει ομοιότητα με κάποιο που έχει παρουσιαστεί στο χρήστη αλλά αν η συσχέτιση με το χρήστη δεν είναι ικανοποιητική τότε το άρθρο μπορεί να απορριφθεί.

Αν το άρθρο που διαβάζει ο χρήστης έχει συσχέτιση Α με το συγκεκριμένο χρήστη τότε όσα άρθρα θεωρηθούν ταυτόσημα και άρα μπορούν να δημιουργήσουν ομάδα άρθρα θα πρέπει να έχουν $\pm \beta * A$ όπου β κυμαίνεται από 0.07 έως 0.1 σύμφωνα με την πειραματική μας διαδικασίας και γενικά εξαρτάται από την τιμή που έχει το Α. Αν η συσχέτιση ενός άρθρου με το χρήστη είναι μικρή (μικρότερη από 30%) τότε διαφαίνεται πως το όριο για το β θα πρέπει να είναι 0.1 ενώ αν η συσχέτιση ξεπερνά το 80% τότε θα πρέπει να είναι 0.07. Διαφαίνεται πως η χρήση του μέσου 0.085 είναι σε γενικές γραμμές ανεκτή καθότι συχνά έχουμε συσχετίσης της τάξης του 50%.

5.7.2 Εντοπισμός Άχρηστων Άρθρων

Ένα σημαντικό στοιχείο το οποίο πρέπει να λάβουμε υπόψη μας κατά τη λειτουργία του μηχανισμού είναι η ποιότητα των αποτελεσμάτων που δίνει σε κάθε επίπεδο. Αυτό που παρατηρήθηκε σε μεγάλο βαθμό κατά την εκτέλεση και λειτουργία του μηχανισμού είναι το γεγονός πως υπήρχαν περιπτώσεις όπου ο μηχανισμός εξαγωγής χρήσιμου κειμένου είχε μαντέψει λάθος. Δε μιλάμε για περιπτώσεις όπου υπάρχει χρήσιμο κείμενο αλλά και πολλά «σκουπίδια» αλλά για περιπτώσεις που δεν υπάρχει καθόλου χρήσιμο κείμενο στην έξοδο το οποίο φυσικά εντάσσεται στο περιθώριο σφάλματος του μηχανισμού. Σε αυτές τις περιπτώσεις λοιπόν διαπιστώσαμε πως μπορεί να δημιουργηθούν προβλήματα στη λειτουργία τους μηχανισμού και σκεφτήκαμε πως πρέπει να βρεθεί ένας τρόπος τόσο να εντοπίζονται αυτά τα άρθρα όσο και να «μαθαίνει» ο μηχανισμός από αυτά τα σφάλματα [61].
Ο εντοπισμός των «junk άρθρων» βασίζεται ιδιαίτερα στην κατηγορία η οποία αναφέρεται ρητά σε κάθε άρθρο και πρόκειται για μεταδεδομένο που προκύπτει από το RSS στο οποίο ανήκει το άρθρο. Αυτό το ονομάζουμε και προκατηγοριοποίηση. Στην ουσία για να εντοπίσουμε αυτού του τύπου τα άρθρα προσπαθούμε να συγκρίνουμε την προκατηγοριοποίηση που υπάρχει στα άρθρα με την κατηγοριοποίηση που εφαρμόζει το perssonal. Για το σκοπό αυτό χρησιμοποιούμε δύο ευρεστικές μεθόδους. Οι δύο μέθοδοι εφαρμόζονται σε κάθε άρθρο που εισάγεται στο σύστημα. Αν έστω και μία από αυτές επιτύχει τότε το άρθρο μαρκάρεται σαν junk καθότι ενδεχόμενα περιέχει «ανούσια» πληροφορία.
Η πρώτη μέθοδος είναι απλή. Αν ο μηχανισμός κατηγοριοποίησης του peRSSonal δώσει συσχέτιση με την κατηγορία προκατηγοριοποίησης σχεδόν μηδενική τότε θεωρούμε ότι υπάρχει πρόβλημα με το άρθρο. Φυσικά και δεν περιμένουμε η προκατηγοριοποίηση να είναι ταυτόσημη με την κατηγοριοποίηση που κάνει το σύστημά μας, εντούτοις όμως σε καμία περίπτωση δεν είναι δυνατόν το σύστημα να δώσει μηδενική συσχέτιση. Η συσχέτιση ενός άρθρου με κάθε κατηγορία είναι πληροφορία η οποία υπάρχει στη ΒΔ και προκύπτει από το μηχανισμό κατηγοριοποίησης. Έτσι μπορούμε να βρούμε όλη την παραπάνω πληροφορία. Σε γενικές γραμμές η κατηγοριοποίηση δεν είναι απόλυτη αλλά μας δίνει στοιχεία συσχέτισης ενός άρθρου με κάθε κατηγορία. Από εκεί και έπειτα ελέγχουμε κάποιες συνθήκες που σε γενικές γραμμές έχουμε διαπιστώσει ότι ισχύουν. Στη γενική περίπτωση αν MAX(a1cn) / MAX-1(a1cm) > 75% τότε θεωρούμε πως το άρθρο έχει αντιστοιχηθεί στην κατηγορία cn. Στατιστικά αν χωρίσουμε την κατηγοριοποίηση σε δύο group σύμφωνα με τη σχετική συσχέτικη τότε θα δούμε ότι έχουμε ένα πολύ μεγάλο κομμάτι άρθρων που παρουσιάζουν μεγάλες συσχετίσεις με μία ή περισσότερες κατηγορίες και ένα άλλο κομμάτι άρθρων που έχουν πολύ μικρές συσχετίσεις με τις κατηγορίες. Μετρώντας την τυπική απόκλιση σ εκάθε ομάδα μπορούμε να εντοπίσουμε πόσο «κοντά» είναι οι συσχετίσεις. Έτσι θα δούμε πως για το πρώτο γκρουπ έχουμε συχνά μία με τρεις συσχετίσεις με πολύ μεγάλες τιμές και όλες τις άλλες συσχετίσεις να έχουν πολύ μικρές τιμές (στην ουσία σαφής κατηγοριοποίηση, ή συσχέτιση με κατηγορίες). Αυτό που μας ενδιαφέρει είναι να εντο-

πίσουμε την απόκλιση των «μεγαλύτερων» τιμών και έτσι υπολογίζουμε πως Average([MAX – STDEV(HIGH VALUES)] / MAX) 0,8. Έτσι υπολογίζουμε πως το όριο για να θεωρήσουμε μία απόκλιση ανεκτή για να συμμετέχει στην κατηγοριοποίηση είναι το 0.75. Αυτό το σφιχτό κάτω όριο μας δίνει στοιχεία για σαφή κατηγοριοποίηση ή συσχέτιση με κατηγορίες και σε γενικές γραμμές όταν η n-οστή τιμή απέχει από την πρώτη τιμή λιγότερο από 75% τότε θεωρούμε αρκετά καλή πιθανότητα κατηγοριοποίησης και στη n-οστή κατηγορία.
Βασιζόμενοι στην παραπάνω ανάλυση θεωρούμε πως δημιουργείται πρόβλημα όταν οι n μεγαλύτερες κατηγορίες για τις οποίες ισχύει MAX-N/MAX > 0,75 και MAX – N – 1 / MAX < 0,75 δεν περιέχουν μέσα τους την κατηγορία της προκατηγοριοποίησης.

Ο παρακάτω αλγόριθμος δείχνει επακριβώς τη διαδικασία.

```
for_each article{
if (MAX-1/MAX < 0.75)
{
if( pre_category!=category_with_ MAX  mark;
}
else
{
for (i in 2:n)
{
if ( MAX-(i-1)/MAX to MAX-1/MAX > 0.75 AND MAX-(i) / MAX < 0.75)
{
if (pre_category! = category _with_MAX-(i-1) to category_with_MAX-1)
mark;
}
}
}
```

Επειδή το σύστημά μας έχει αρκετά διακριτές κατηγορίες έχουμε δει πως ένα άρθρο μπορεί να ανήκει το πολύ σε 3 κατηγορίες. Εδώ εντοπίζουμε και τη δεύτερη ευρεστική μέθοδο. Μάλιστα μετά από ανάλυση που κάναμε το 95% των περιπτώσεων αυτών είναι περιπτώσεις όπου η μέγιστη συσχέτιση είναι πάρα πολύ μικρή (<10%). Αυτό σημαίνει αυτόματα πως αν MAX-4/MAX>0,75 τότε αμέσως δημιουργείται και πάλι πρόβλημα και θα πρέπει να ελέγξουμε το άρθρο.
Καταλήγοντας θα πρέπει να πούμε πως ο εντοπισμός των προβληματικών άρθρων με τις παραπάνω μεθόδους αποδίδει σε πολύ σημαντικό βαθμό και βοηθά στο να απαλλαγούμε από άρθρα τα οποία δημιουργούν προβλήματα στο μηχανισμό μας. Η σκέψη για τους αλγορίθμους που χρησιμοποιούμε είναι απλή καθότι η κατηγοριοποίηση βασίζεται στις λέξεις κλειδιά που εξάγονται

από κάθε κείμενο συνεπώς μία κακή κατηγοριοποίηση, δεδομένου ότι έχουμε αρκετά καλή πληροφορία για τις λίστες λέξεων κλειδιών των κατηγοριών, μπορεί να σημαίνει δύο πράγματα: α. δεν έχουμε αρκετή πληροφορία για να κατηγοριοποιήσουμε επαρκώς ένα κείμενο το οποίο περιέχει άγνωστες προς το σύστημα λέξεις ή το κείμενο είναι τέτοιο ώστε να δημιουργεί πρόβλημα στην κατηγοριοποίηση, δεν περιέχει δηλαδή σωστή πληροφορία. Σε συνέχεια του συστήματος αυτού θα προχωρήσουμε και στη δημιουργία μιας κατηγορίας junk από όσα άρθρα εντοπίζουμε και επιβεβαιώνουμε ότι είναι junk προκειμένου να πετύχουμε ακόμα καλύτερα αποτελέσματα στο μηχανισμό.

5.7.3 Pre-fetching άρθρων στο perssonal

Για το prediction module του prefetcher, δηλαδή το τμήμα εκείνο που αναλαμβάνει ποια άρθρα θα προ-ανακτηθούν κάθε φορά, χρησιμοποιούμε τον μηχανισμό προσωποποίησης του perSSonal.
Η προσαρμογή του μηχανισμού στις ανάγκες του χρήστη γίνεται με τον καθορισμό των πεδίων ενδιαφέροντος κατά τη διαδικασία εγγραφής του χρήστη στο σύστημα. Συγκεκριμένα, του δίνεται η επιλογή να επιλέξει μέσω μιας κλίμακας το βαθμό ενδιαφέροντος για κάθε μία από τις 7 κύριες κατηγορίες όπως Business, Entertainment, Health κτλ. Επίσης, ο μηχανισμός έχει τη δυνατότητα να διαμορφώνει το προφίλ του χρήστη κατά την περιήγηση του στην ιστοσελίδα, χρησιμοποιώντας το ιστορικό των άρθρων που έχει επισκεφθεί. Έτσι, στην προσωποποιημένη προβολή, τα άρθρα θα εμφανίζονται ταξινομημένα σύμφωνα με τα διαμορφωμένα ενδιαφέροντα του χρήστη.
Έτσι, όταν ο χρήστης κάνει login στο σύστημα, βλέπει στο user homepage που ανήκει στο πρώτο κανάλι του perSSonal και συγκεκριμένα στα personalized νέα, ένα σύνολο από άρθρα που τον ενδιαφέρουν και έχουν προκύψει ως αποτέλεσμα του μηχανισμού προσωποποίησης. Όταν ο χρήστης επισκεφτεί το user homepage, τότε σε τακτά χρονικά διαστήματα αρχίζει να τρέχει ο μηχανισμός του prefetching και αρχίζει να φέρνει τα XML των άρθρων που υπάρχουν στην τρέχουσα σελίδα, μαζί με τα XML των related similar και identical άρθρων του. Η συνάρτηση prefetch λαμβάνει υπόψιν 2 παραμέτρους, ένα πίνακα με τα αναγνωριστικά των άρθρων που υπάρχουν στη σελίδα που βρίσκεται ο χρήστης, καθώς και τη βαθμολογία του κάθε άρθρου που προκύπτει από τον μηχανισμό προσωποποίησης και εκφράζει το πόσο ενδιαφέρον ήταν το άρθρο στον χρήστη. Τα άρθρα παρουσιάζονται στη σελίδα κατά φθίνουσα σε σειρά αξίας για τον χρήστη, με το άρθρο που βρίσκεται στο πάνω μέρος της σελίδας να έχει την μεγαλύτερη βαθμολογία, το αμέσως επόμενο της αμέσως μικρότερη βαθμολογία κ.ο.κ. Συνεπώς το prefetching τρέχει σε τακτά χρονικά διαστήματα και φέρνει άρθρα από την αρχή της σελίδας προς το κάτω μέρος. Τα άρθρα που γίνονται prefetch μαζί με τα related και identical τους αποθηκεύονται σε μια μεταβλητή τύπου SESSION σε μια σελίδα του server

που καλείτε με AJAX. Αυτό γίνεται επειδή το PHP session μπορεί να κρατήσει πληροφορίες στον server για μελλοντική χρήση, και αυτή την πληροφορία μπορεί να την περάσει σε άλλες σελίδες. Σαν αποτέλεσμα, υπάρχει μια κοινή session μεταβλητή για όλες τις προσωποποιημένες σελίδες που θα επισκεφτεί ο χρήστης και κρατάει τα prefetched άρθρα από όλες τις σελίδες που επισκέφθηκε.
Οπότε, όταν ο χρήστης θέλει να βρει ένα άρθρο καθώς και τα related/identical/similar άρθρα ελέγχει αν τα XML τους υπάρχουν αποθηκευμένα στην μεταβλητή session και αν ναι, τότε δεν χρειάζεται να εκτελεστούν τα queries στην βάση για την ανάκτηση των αντίστοιχων XML.
Όταν ο χρήστης αλλάξει σελίδα τότε η διαδικασία του prefetching στην προηγούμενη σελίδα σταματά και αρχίζει νέο prefetching στην καινούρια σελίδα. Παρόλα αυτά όλα τα prefetched άρθρα από τις προηγούμενες σελίδες, διατηρούνται στην μεταβλητή session, οπότε αν ο χρήστης να επιλέξει να γυρίσει σε σελίδα που είχε προσπελαστεί παλιότερα θα κάνει prefetch μόνο τα άρθρα που δεν είχαν γίνει prefetched παλιότερα.
Για να μην υπερφορτωθεί το bandwidth του χρήστη, ο μηχανισμός δεν τρέχει συνέχεια αλλά κάθε ένα prefetch interval και φέρνει k άρθρα κάθε φορά. Στη συνέχεια υποθέτουμε ότι ο χρήστης διαβάζει τα άρθρα από το πάνω μέρος της σελίδας προς τα κάτω. Αυτό όπως αναφέρθηκε παραπάνω έχει βάση γιατί ο μηχανισμός προσωποποίησης του Perssonal παρουσιάζει τα νέα με σειρά ενδιαφέροντος για τον χρήστη. Για να είναι αποτελεσματικό το prefetching πρέπει ενόσω όταν ο χρήστης τελειώσει το διάβασμα πχ των k πρώτων άρθρων, και αρχίσει να διαβάζει το k+1 άρθρο, να έχει τελειώσει το prefetching των k επόμενων άρθρων ώστε να αντιληφθεί μηδενική καθυστέρηση στις πράξεις του. Δηλαδή το prefetching θα τρέχει κάθε Inerval k με :

Prefetch_interval (k) = k * (μέσος χρόνος προσπέλασης κάθε άρθρου από τον χρήστη – μέσος χρόνος για prefetching).

Ο μηχανισμός του prefetching τρέχει όταν ο χρήστης επισκεφτεί την σελίδα με τα προσωποποιημένα νέα, και εκμεταλλεύεται το χρόνο φορτώματος της σελίδας καθώς και το χρόνο που "σπαταλά" ο χρήστης να περιηγηθεί στο site, και να επιλέξει ποιο άρθρο θα διαβάσει, ώστε να φέρει τα k πρώτα άρθρα. Στη συνέχεια ενώ ο χρήστης διαβάζει κάποια από τα k άρθρα, ο μηχανισμός τρέχει και φέρνει τα k επόμενα κ.ο.κ. Τέλος, ο μηχανισμός σταματάει να τρέχει αφού φέρει τα 100 πρώτα προσωποποιημένα άρθρα, το οποίο και είναι ένα λογικό όριο καθώς ο μέσος χρήστης συνήθως δεν διαβάζει ακόμα και σε ακραίες περιπτώσεις περισσότερα από 100 άρθρα.
Συγκεντρωτικά σε ψευδοκώδικα ο μηχανισμός τρέχει ως εξής:

1. Articles[] : all the article id's of the page
2. Rank[] : the rank of every article id

```
3. Session[]  : session  holds all the previously prefetched articles
4. for_each_page_the_user_visits
5. {
6.       Gather the Articles[] ;
7.     Sort Articles[] by Rank in descending order;
8.       Remove those article id's that belong to session and to     Articles[] as well;
9.       for (i=0; i<Articles[].length; times=times+k)
10.       {
11.      articles_to_be_prefetched = [Articles[i+0], Articles [i+1], Articles[i+2],…, Articles[i+k]];
12. prefetch (articles_to_be_prefetched);
13. sleep (Prefetch_interval(k));
14.      }
          15.  }
```

5.7.4 Προσωποποιημένη Αναζήτηση με υποστήριξη Caching

Στο επόμενο σχήμα παρουσιάζονται τα βασικά στάδια της εκτέλεσης του αλγόριθμου προσωποποιημένης αναζήτησης [59] και [60]. Στο πρώτο στάδιο, ο χρήστης υποβάλλει την επερώτηση του. Το σύστημα αντιστοιχίζει τις λέξεις κλειδιά που έδωσε ο χρήστης σε λέξεις που υπάρχουν ήδη αποθηκευμένες στη ΒΔ από τη διαδικασία κατηγοριοποίησης των άρθρων που φτάνουν καθημερινά στο σύστημα. Στο επόμενο στάδιο και προτού γίνει εκκίνηση της διαδικασίας αναζήτησης άρθρων στη ΒΔ, ελέγχεται η cache μνήμη του συστήματος που περιέχει αποτελέσματα από παρελθοντικές επερωτήσεις που υπέβαλλαν χρήστες του συστήματος. Εάν εντοπιστεί επερώτηση στην cache που να έχει τις ίδιες παραμέτρους με την τρέχουσα επερώτηση και η οποία να έχει υποβληθεί από τον ίδιο χρήστη στο πρόσφατο παρελθόν τότε τα άρθρα-αποτελέσματα ανασύρονται γρήγορα από την cache χωρίς να έχει γίνει καμία αναζήτηση. Ορισμένες φορές, είναι απαραίτητο να εκτελεστεί μια περιορισμένη αναζήτηση αν το σύστημα διαπιστώσει ότι τα cached αποτελέσματα δεν επαρκούν για να καλύψουν όλο το χρονικό διάστημα για το οποίο ο χρήστης υπόβαλε την επερώτηση. Αφού ανακτηθούν όλα τα άρθρα που απαντούν στην επερώτηση του χρήστη ξεκινάει η φάση της ταξινόμησης και επιλογής τους έτσι ώστε η διαδικασία να είναι προσωποποιημένη στο προφίλ και στις προτιμήσεις του χρήστη. Από τις λέξεις-κλειδιά της επερώτησης με μια πολύπλοκη διαδικασία που θα παρουσιασθεί και θα αναλυθεί στη συνέχεια του κεφαλαίου, βρίσκονται άλλες λέξεις-κλειδιά που ήδη υπάρχουν στο σύστημα και που χαρακτηρίζονται ως σχετικές των λέξεων που έδωσε αρχικά ο χρήστης. Οι νέες αυτές λέξεις δεν χρησιμοποιούνται καθόλου στην διαδικασία ανάσυρσης άρθρων από τη βάση δεδομένων του συστήματος αλλά ο σκοπός τους είναι να βελτιώσουν την σειρά με την οποία θα εμφανιστούν τα άρθρα στον ίδιο το χρήστη. Αυτό για μας αποτελεί μια αναβάθμιση

της ποιότητας αναζήτησης, μιας και όσο πιο ψηλά στη λίστα των άρθρων του αποτελέσματος βρίσκονται τα άρθρα που ουσιαστικά επιθυμούσε ο χρήστης τότε μεγαλύτερη είναι και η επιτυχία της αναζήτησης καθώς και η αποτελεσματικότητα του μηχανισμού μας. Με βάση αυτήν την «εμπλουτισμένη» επερώτηση που προκύπτει γίνεται η ταξινόμηση των άρθρων και η παρουσίασή τους στον τελικό χρήστη. Μετά από κάθε αναζήτηση ακολουθεί ενημέρωση της cache είτε υπό μορφή ενημέρωσης των cached αποτελεσμάτων για τις ήδη υπάρχουσες επερωτήσεις ή υπό μορφή προσθήκης των αποτελεσμάτων για νέες επερωτήσεις που δεν υπήρχαν στην cache.

Στη συνέχεια θα εξετασθεί λεπτομερειακά κάθε βήμα της διαδικασίας που περιγράφηκε με το βάρος να δίνεται στην διαδικασία της αναζήτησης καθώς και την προσωποποίησης μέσω της αναδιάταξης των άρθρων του αποτελέσματος. Για το λόγο αυτό, όπου κρίνεται αναγκαίο θα υπάρχουν διαφωτιστικά σχήματα, διαγράμματα ροής, κομμάτια κώδικα απευθείας μέσα από το σύστημα καθώς και ψευδοκώδικα για την απλούστευση της παρουσίασης ορισμένων σημείων.

ΚΕΦΑΛΑΙΟ 6

ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Μια κοινωνία ανθίζει όταν οι γηραιότεροι φυτεύουν δένδρα των οποίων δε θα απολαύσουν τη σκιά

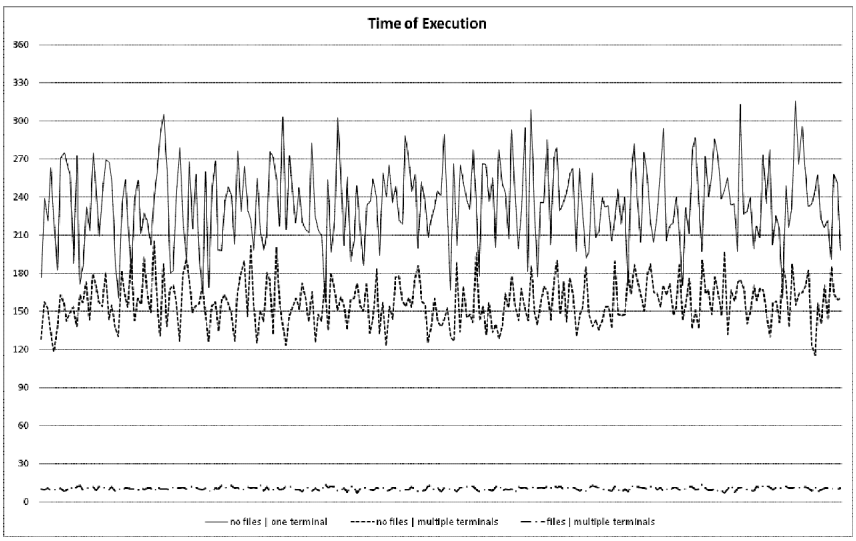
(Ανώνυμος)

Στο κεφάλαιο παρουσιάζουμε την πειραματική διαδικασία που πραγματοποιήσαμε με το μηχανισμό μας σε πλήρη λειτουργία. Κάθε ένα σύστημα ελέγχεται ξεχωριστά ενώ στο τέλος εμφανίζονται τα συγκεντρωτικά αποτελέσματα από το peRSSonal meta-portal

Ο μηχανισμός ο οποίος έχουμε αναπτύξει πραγματοποιεί μία σειρά από διαδικασίες προκειμένου να επιτύχει το επιθυμητό αποτέλεσμα που είναι η προσωποποιημένη προβολή πληροφορίας που έχει συλλεχθεί από πηγές του Διαδικτύου. Όπως έχουμε δει ήδη έως τώρα στην εργασία τα βασικά επίπεδα του συστήματος είναι τρία και οι αυτόνομοι βασικοί μηχανισμοί είναι 6: (α) ανάκτηση πληροφορίας από το Διαδίκτυο, (β) εξαγωγή χρήσιμης πληροφορίας από σελίδες του Διαδικτύου, (γ) προ-επεξεργασία πληροφορίας, (δ) κατηγοριοποίηση πληροφορίας, (ε) αυτοματοποιημένη περίληψη πληροφορίας και τέλος (στ) μηχανισμός παρουσίασης πληροφορίας. Όπως έχουμε ήδη δει, κάθε ένα από τα παραπάνω συστήματα μπορεί να αναλυθεί σε τεράστιο βάθος ωστόσο η βασική μας έννοια είναι να καταφέρουμε να δημιουργήσουμε ένα μηχανισμό που θα μπορεί να προσφέρει εύκολα και γρήγορα, ποιοτικά αποτέλεσματα σε όλους τους χρήστες του συστήματος και πάνω απ' όλα να μπορέσει να ενσωματώσει την υποκειμενικότητα του χρήστη σε κάθε διαδικασία. Παράλληλα θα πρέπει να τονίσουμε πως ο μηχανισμός παρουσίασης πληροφορίας μπορεί να χωριστεί σε πληθώρα υποσυστημάτων, όμως, η λειτουργία αυτού του μηχανισμού είναι τόσο συμπαγής και ενιαία που τα αποτελέσματά της θα παρουσιαστούν συνολικά.
Σε αυτή την ενότητα θα μελετήσουμε πειραματικές διαδικασίες που σα σκοπό έχουν να δείξουν ποιοτικά και ποσοτικά στοιχεία για κάθε μηχανισμό του συστήματος ενώ η ενότητα περιλαμβάνει και παραδείγματα της λειτουργίας του συστήματος από πραγματικούς χρήστες, τον ίδιο μας τον εαυτό δηλαδή κατά τη διάρκεια χρήσης του συστήματος. Τέλος, για πρώτη φορά θα έρθει στο φως το σύστημα υποστήριξης Ελληνικών το οποίο λειτουργεί από τον Απρίλιο του 2010 ωστόσο δεν υπάρχει σε καμία online έκδοση του μηχανισμού. Μέσα, λοιπόν, από αυτή την εργασία θα κάνουμε και παρουσίαση του συστήματος σε λειτουργία για την ελληνική γλώσσα.

6.1 Μηχανισμός advaRSS

Ο μηχανισμός crawling που αναπτύξαμε ονομάζεται advaRSS [41] και όπως έχουμε αναφέρει είναι ένας mixed crawler. Όπως κάθε τέτοιος μηχανισμός θεωρούμε πως θα πρέπει να είναι σε θέση να αναλύει μεγάλο αριθμό στοιχείων εισόδου αλλά παράλληλα θα πρέπει να διατηρεί και μία ισορροπία τόσο στην κατανάλωση ιδίων πόρων όσο και πόρων του δικτύου. Επιπρόσθετα, ο μηχανισμός αποδίδει καλύτερα εφόσον εφαρμόζει παραλληλία στις διαδικασίες πράγματα που επιτυγχάνεται με τον έλεγχο από ένα κεντρικοποιημένο μηχανισμό ελέγχου. Είναι λοιπόν αναμενόμενο ο μηχανισμός μας λειτουργώντας παράλληλα να μπορεί να πετύχει μεγάλες ταχύτητες στις διαδικασίες του, ωστόσο σε αυτό το σημείο τίθεται ένα σημαντικό ζήτημα. Το γεγονός ότι ο μηχανισμός απαιτεί κεντρικό έλεγχο αλλά και διαθέτει κεντρικοποιημένη βάση δεδομένων ενδεχόμενα να επηρεάζει την εκτέλεση του συστήματος. Επιπλέον, υπάρχει πιθανότητα αργοπορίας του μηχανισμού με δεδομένο ότι χρησιμοποιούμε πολλούς αλγορίθμους αλλά και μεγάλο αριθμό στοιχείων εισόδου.



Σχήμα 6.1: Χρονο Εκτέλεσης του συστηματος advaRSS σε διαφορετικα set-ups

Για όλους τους παραπάνω λόγους και προκειμένου να είμαστε σίγουροι πως το σύστημα θα μπορεί να ανταποκριθεί σε πολύ μεγάλες εισόδους χωρίς πρόβλημα και άρα είναι ένα σύστημα που μπορεί να λειτουργήσει με δεδομένα πολύ μεγάλης κλίμακας πραγματοποιήσαμε τρία διαφορετικά πειράματα χρησιμοποιώντας στην ουσία τρία διαφορετικά setups για το σύστημα. Αυτό πραγματοποιήθηκε ώστε να επιβεβαιώσουμε πως η αρχιτεκτονική και λειτουργία του συστήματος είναι τέτοια που να μπορεί να ανταποκριθεί σε δύσκολες συνθήκες. Έτσι δημιουργήσαμε τα τρία παρακάτω συστήματα:

1. ένας απλός υπολογιστής ο οποίος δεν έχει παραλληλία, δεν έχει τερματικά αλλά είναι ο ίδιος το σύστημα που πραγματοποιεί όλες τις διαδικασίες του crawler σειριακά. Παράλληλα σε αυτό το σύστημα αφαιρέσαμε κάθε ίχνος αλγορίθμου crawling adaptation on time.

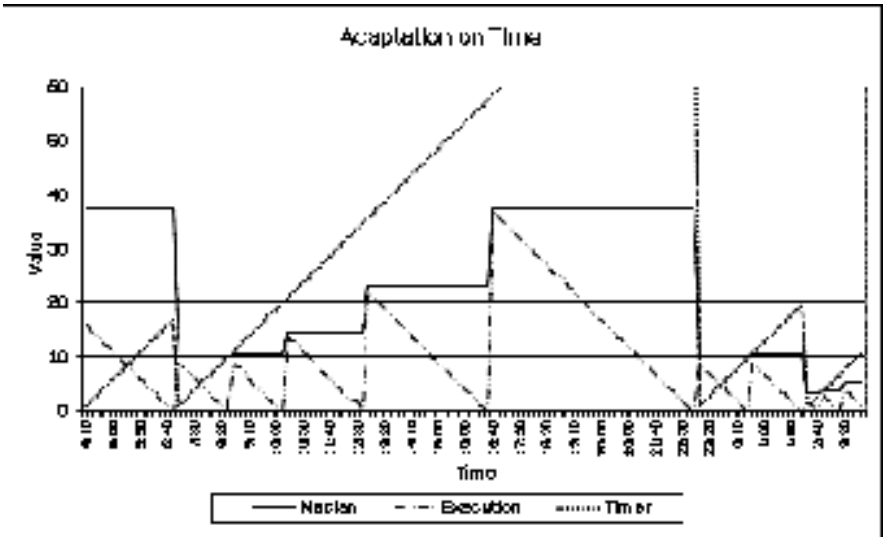
2. Το δεύτερο σύστημα έχει τα χαρακτηριστικά του crawler μας αλλά δεν έχει καμία διαδικασία κανενός αλγορίθμου adaptation

3. το τρίτο σύστημα, προφανώς, είναι ο advaRSS ακριβώς όπως λειτουργεί και εκτελείται στο σύστημά μας.

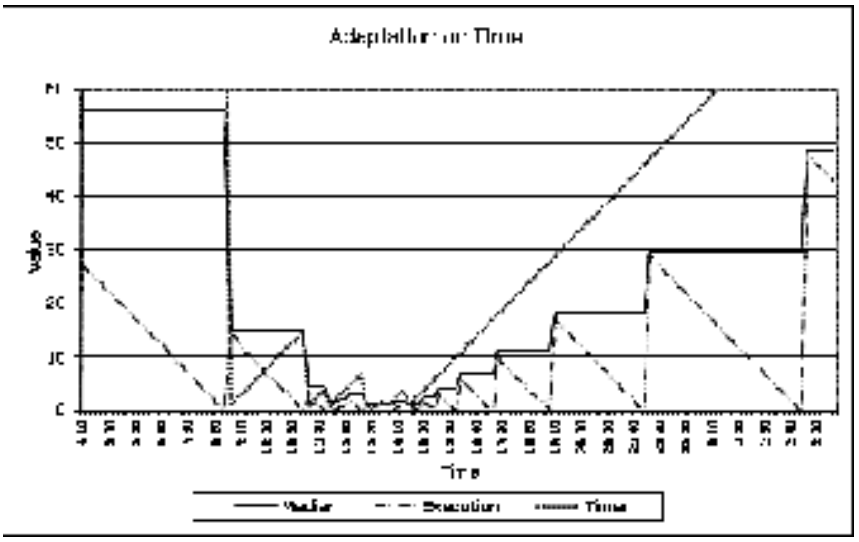
Όπως είναι εμφανές και από το σχήμα, το τρίτο σύστημα, που είναι και αυτό που χρησιμοποιούμε, έχει σαφώς πολλαπλάσια καλύτερη συμπεριφορά από τα άλλα δύο συστήματα που περιγράψαμε παραπάνω. Αυτό μας δείχνει πως ο αλγόριθμός μας και το system setup που έχουμε για το σύστημα μπορεί να ανταποκριθεί στις απαιτήσεις που έχουμε αυτή τη στιγμή

από το σύστημα. Μάλιστα, στην προσπάθειά μας να δυσκολέψουμε λίγο περισσότερο την κατάσταση για το real system setup δοκιμάσαμε να βάλουμε σαν είσοδο δεκαπλάσια στοιχεία απ' ότι στα άλλα setups. Η απόδοση που λάβαμε σαν έξοδος είχε ελάχιστη διαφορά από το παραπάνω σχήμα. Αυτό δικαιολογείται διότι αφενώς ο μηχανισμός μας είναι πλήρως κατανεμημένος, οι διαδικασίες ελέγχονται κεντρικά, ελαχιστοποιείται η επικοινωνία με τη βάση δεδομένων και το πιο σημαντικό οι αλγόριθμοι adaptation που χρησιμοποιούνται παρέχουν στο σύστημα προστασία τόσο από το να υπερφορτωθεί το ίδιο και άρα οι πόροι του αλλά και να υπερφορτώσει τους πόρους του δικτύου. Ενημερωτικά τα παραπάνω νούμερα για την εκτέλεση του μηχανισμού προκύπτουν από είσοδο 1000 περίπου RSS feed (τόσα είχαν επιλεγεί περίπου για είσοδος τη στιγμή της εκτέλεσης), ενώ για το πείραμα αυξημένης δυσκολίας για το μηχανισμό μας η είσοδος ήταν κάποιες χιλιάδες RSS feeds κάποια από τα οποία φυσικά απορρίφθηκαν από το μηχανισμό adaptation.
Έχοντας εξασφαλίσει κάτι πολύ βασικό για το μηχανισμό μας, την αποδεκτή και ποιοτική λειτουργία κάτω από δύσκολες συνθήκες προχωρήσαμε σε πειράματα προκειμένου να εξασφαλίσουμε ποιοτικά αποτελέσματα από τη χρήση των αλγορίθμων προσαρμογής του μηχανισμού στη χρονική συμπεριφορα (temporal behavior) των RSS feeds. Αυτή τη στιγμή το σύστημα διαθέτει έναν μεγάλο αριθμό από RSS feed αλλά αναμένουμε ο αριθμός αυτός να γίνει πραγματικά τεράστιος όταν το σύστημα θα βρεθεί σε κανονική λειτουργία. Σε αυτή την περίπτωση αναμένουμε το σύστημα να είναι σε θέση να πραγματοποιεί ελέγχους σε κάποιες χιλιάδες RSS feeds ανά X λεπτά (Μ.Ο.(X) = 8 λεπτά), όπου X η περίοδος εκτέλεσης του μηχανισμού. Η προσαρμογή στο ρυθμό αλλαγής ενός RSS feed την οποία παρουσιάζουμε επιτρέπει στο μηχανισμό να αποφεύγει να ελέγχει όλα τα feeds που διαθέτει αλλά μόνο αυτά για τα οποία έχει γίνει πρόβλεψη ότι μπορεί να έχουν υποστεί κάποια αλλαγή από την τελευταία φορά που εκτελέστηκε ο μηχανισμός. Στα σχήματα 6.2, 6.3 και 6.4 παρατηρούμε τον τρόπο προσαρμογής του μηχανισμού σε τρία ενδεικτικά RSS feeds.

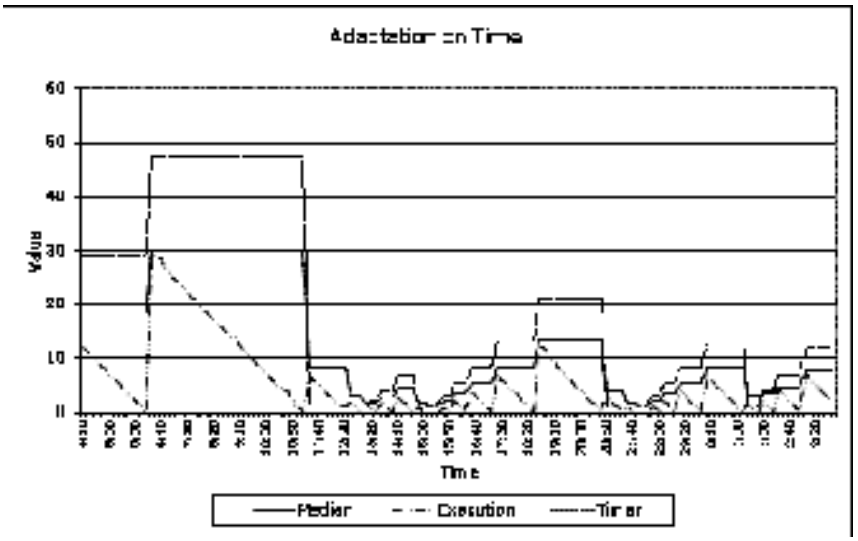
Όπως είναι εμφανές από την πειραματική διαδικασία που ακολουθούμε και παρατηρούμε στα σχήματα 6.2, 6.3 και 6.4, ο μηχανισμός δεν ελέγχει κάθε RSS feed κάθε φορά που εκτελείται ο advaRSS αλλά τα ελέγχει κάθε στιγμή που πιστεύει ότι τα συγκεκριμένα RSS έχουν δημοσιεύσει κάποιο άρθρο. Στα τρία σχήματα βλέπουμε τρία αρκετά χαρακτηριστικά παραδείγματα. Το σχήμα 6.2 μας δείχνει ένα RSS feed που συνηθίζει να δημοσιεύει πολλά άρθρα μαζί 2-3 φορές ημερησίως. Ο έλεγχος μας δείχνει πως πρόκειται για ένα RSS από δικτυακό τόπο αμερικάνικης εφημερίδας (έτσι εξηγούνται και οι ώρες δημοσίευσης). Στη δεύτερη περίπτωση 6.3 ερχόμαστε αντιμέτωποι με ένα RSS feed το οποίο δημοσιεύει για ένα μεγάλο διάστημα της ημέρας (εργάσιμες ώρες) και παραμένει ανενεργό για ένα εξίσου μεγάλο διάστημα (βράδι). Το τρίτο παράδειγμα 6.4 είναι χαρακτηριστικό παράδειγμα δικτυακού τόπου που μένει ελάχιστες ώρες ανενεργό. Πρόκειται για χαρακτηριστικό pattern ενός ενημερωτικού / ειδησεογραφικού blog πολλαπλών χρηστών. Πέραν των παραπάνω patterns που δείχνουν την απόκριση του μηχα-



Σχήμα 6.2: Προσαρμογή Στην Περίοδο Δημοσίευσης του RSS - 1



Σχήμα 6.3: Προσαρμογή Στην Περίοδο Δημοσίευσης του RSS - 2



Σχήμα 6.4: Προσαρμογή Στην Περίοδο Δημοσίευσης του RSS - 3

νισμού στις αλλαγές που προκύπτουν στα feeds που ελέγχονται είναι πολύ σημαντικό να δούμε δύο μετρικές που θα εκφράσουν την ποιότητα των αλγορίθμων. Η πρώτη μετρική σχετίζεται με το ποσοστό μείωσης της χρήσης των resources, πάντοτε σε σύγκριση με ένα μηχανισμό που θα έκανε έλεγχο σε όλα τα RSS ανά Χ λεπτά. Παρατηρούμε πως με τον αλγόριθμο που χρησιμοποιούμε μπορούμε να μειώσουμε την περίοδο ελέγχου ενός RSS feed περίπου 5 φορές (αναφέρεται στη χειρότερη περίπτωση των RSS που αλλάζουν πολύ συχνά, διότι για τις περιπτώσεις που έχουμε σπάνια δημοσίευση από μία πηγή μπορούμε να πετύχουμε πολύ μεγαλύτερη μείωση). Από που προκύπτει όμως αυτός ο αριθμός για το μέγεθος μείωσης των ελέγχων. Αν ελέγχαμε ένα RSS feed κατά μέσο όρο ανά 8 λεπτά τότε σε μία μέρα θα το ελέγχαμε 180 φορές περίπου. Ο αλγόριθμος που φαίνεται στο τρίτο σχήμα κάνει έλεγχο στο feed που αλλάζει συχνά 34 φορές σε μία μέρα. Η μείωση λοιπόν είναι της τάξης των 180 / 34 5 μονάδων. Αν ωστόσο ελέγχαμε την πρώτη περίπτωση της σπάνιας δημοσίευσης στο RSS feed ο μηχανισμός μας κάνει έλεγχο 10 φορές που είναι 18 φορές λιγότερες απ' ότι ένας μηχανισμός χωρίς time adaptation. Αυτή η μείωση δε θα μας έλεγε τίποτα αν δεν ελέγχαμε μία άλλη σημαντική παράμετρο. Ένας μηχανισμός ο οποίος ελέγχει 180 φορές τη μέρα, ανά οκτώ λεπτά δηλαδή, κάποια πηγή για δημοσιευμένα νέα, αναμένεται να "καθυστερήσει" το πολύ 8 λεπτά να ανακτήσει ένα νέο και αυτό στη χειρότερη περίπτωση που το νέο θα δημοσιευθεί αμέσως μόλις ο μηχανισμός κάνει την Κ-οστή εκτέλεσή του, οπότε αναμένεται να λάβει αυτό το άρθρο στην Κ+1 εκτέλεση. Όταν όμως ο μηχανισμός βασίζεται σε βάσεις γνώσης για να προβλέψει πότε θα είναι το επόμενο δημοσίευμα σε ένα RSS τότε μπορεί ένα λάθος να κοστίσει σε μη επικαιροποιημένο set άρθρων στο μηχανισμό. Σύμφωνα, λοιπόν, με τα πειράματά μας ο μηχανισμός πετυχαίνει καθυστέρηση κατά μέσο όρο περίπου 17 λεπτά, λαμβάνει δηλαδή κατά μέσο όρο τα άρθρα που δημοσιεύονται 17 λεπτά αργό-

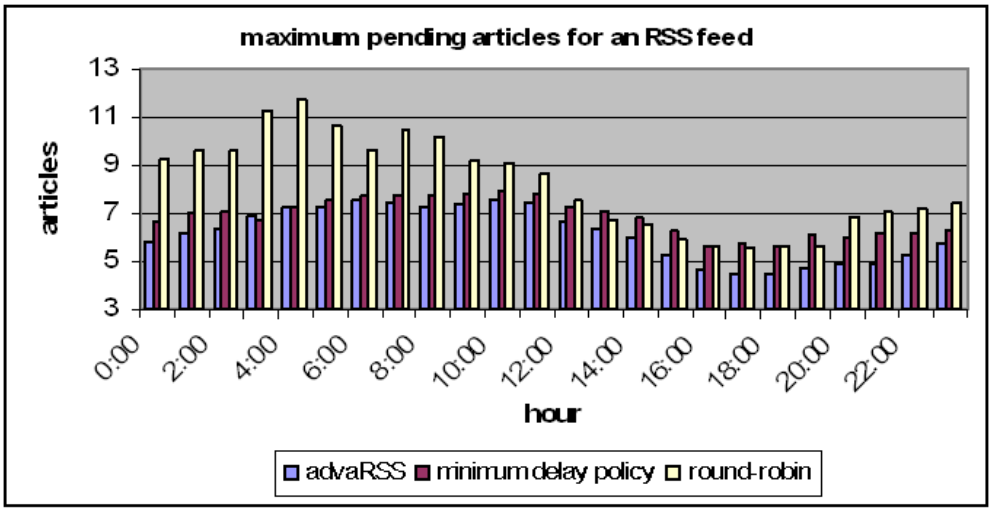
τερα συγκριτικά με την ώρα δημοσίευσης. Ο τρόπος υπολογισμού έγινε συγκρίνοντας το χρόνο που ανέφερε ένα άρθρο με το χρόνο στον οποίο το ανακτήσαμε. Βέβαια, θα πρέπει να τονίσουμε το απόλυτα παράδοξο που παρατηρήθηκε: όταν ο μηχανισμός έτρεχε ανά 8 λεπτά χωρίς κανέναν αλγόριθμο προσαρμογής ο Μ.Ο. καθυστέρησης ανάκτησης ήταν σχεδόν 10 λεπτά. Ωστόσο, γνωρίζουμε ότι ο μηχανισμός εκτελείται ακαριαία και από την άλλη ακόμα και κάθε φορά που γινόταν εκτέλεση να είχαμε καταφέρει να είμαστε στη χειρότερη περίπτωση και πάλι θα είχαμε μία καθυστέρηση της τάξης των 8 λεπτών όχι όμως και μεγαλύτερη. Τα προβλήματα αυτά μας οδηγησαν σε δύο συμπεράσματα. Από τη μία οι διαφορές λεπτών μεταξύ των δικών μας server και πολλών servers του διαδικτύου προσέγγιζαν ακόμα και τα 25 λεπτά. Την ώρα που το δικό μας ρολόι έδειχνε 10:00 ώρα αγγλίας, αγγλικό ειδησεογραφικό πρακτορείο δημοσίευε μόλις άρθρο (συνεχή refresh στη σελίδα) με ώρα 9:40, κάτι το οποίο μπορεί ενδεχόμενα να είναι και πονηρό (πρωτεία ίσως στην ενημέρωση αλλά ας μην είμαστε καχύποπτοι). Από την άλλη, υπήρχαν περιπτώσεις που προφανώς τα RSS ανανεώνονταν σε διαφορετικό χρόνο από το χρόνο δημοσίευσης (ανά 10 λεπτά μέσω κάποιας χρονοπρογραμματισμένης διαδικασίας). Έτσι, λοιπόν, και θεωρώντας ανεκτά τα δύο νούμερα παρατηρήσαμε πως η διαφορά μας από ένα μηχανισμό συνεχούς ανίχνευσης είναι σχεδόν 2 μονάδες (17/10).

Η εξέταση του μηχανισμού δεν τελειώνει εδώ καθότι η απλή προσαρμογή είναι ένας αλγόριθμος που μεταβάλλει απλώς το χρόνο ελέγχου με θετικά όπως είδαμε αποτελέσματα. Στη συνέχεια θα συγκρίνουμε το μηχανισμό ελέγχου που προτείνουμε και βασίζεται στο posting history συγκριτικά με άλλες πολιτικές ανάκτησης πληροφορίας. Η διαδικασία αυτή βασίστηκε σε δεδομένα που συλλέχθηκαν για 90 μέρες χρησιμοποιώντας ενδεικτικά RSS feeds από τη συλλογή μας. Σε πρώτη φάση αυτό που ελέγχουμε είναι ο μέγιστος αριθμός άρθρων που βρίσκονται στην ουρά κάποιας πηγής. Ως τέτοια χαρακτηρίζουμε τα άρθρα που ο μηχανισμός δεν έχει συλλέξει ακόμα από μία πηγή. Αυτή η μέτρηση βοηθά να κατανοήσουμε τη χρησιμότητα της υλοποίησης που παρουσιάστηκε αφού επικεντρωνόμαστε στη μέγιστη ποσότητα των άρθρων που μπορεί να έχει δημοσιεύσει κάποια πηγή και δεν έχουν ακόμα συλλεχθεί. Η σύγκριση γίνεται μεταξύ της πολιτικής που ακολουθούμε εμείς για το μηχανισμό και δύο άλλων πολιτικών:

1. round-robin πολιτική η οποία τοποθετεί τα RSS feeds σε μία ουρά και τα ελέγχει με τη μέθοδο FIFO (που σημαίνει ότι θα επισκευθεί ξανά μία σελίδα όταν έχει ελέγξει πρώτα όλες τις υπόλοιπες)

2. πολιτική που χρησιμοποιεί posting patterns για να ελαχιστοποιήσει τη συνολική καθυστέρηση των ανακτημένων άρθρων. Ως καθυστέρηση ορίζεται η χρονική περίοδος μεταξύ της δημοσίευσης και της ανάκτησης κάποιου άρθρου.

Όπως είναι φανερό από το γράφημα 6.5, η πολιτική που ελαχιστοποιεί τη συνολική καθυστέρηση αυξάνει κατά μέσο όρο περίπου 11,2% το πλήθος των άρθρων που δεν έχουν ακόμα συλλεχθεί (pending articles). Με τη round-robin μέθοδο παρατηρούμε ότι η αύξηση αγγίζει το 33,4%.

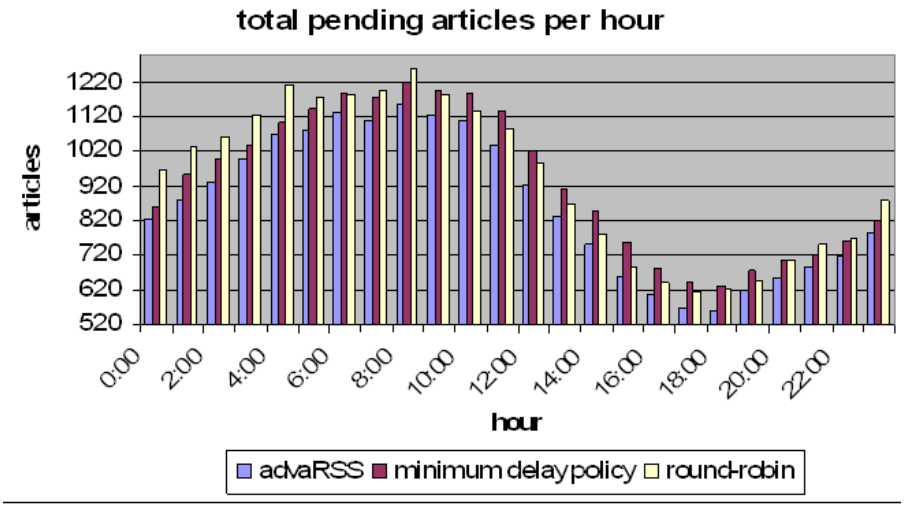


Σχήμα 6.5: Μέσος όρος του μέγιστου αριθμού άρθρων μιας πηγής

Εκτός από την παραπάνω μέτρηση που ουσιαστικά αναφέρεται στην πηγή με τον μέγιστο αριθμό μη ανακτηθέντων άρθρων, είναι χρήσιμο να υπολογίσουμε και το συνολικό αριθμό αυτών των άρθρων στο σύστημα. Στο επόμενο πείραμα χρησιμοποιούμε ένα σύνολο από 460 RSS feeds που αντιστοιχούν σε δικτυακούς τόπους ενημέρωσης με διάφορους ρυθμούς ανανέωσης. Το ζητούμενο είναι να μετρηθεί το άθροισμα των μη ανακτηθέντων άρθρων όλων των πηγών. Αξίζει να σημειωθεί ότι και τα δυο πειράματα που περιγράφονται, έγιναν αφού πρώτα συλλέχθηκε η απαραίτητη πληροφορία για ένα διάστημα 90 ημερών και ύστερα εφαρμόστηκε η κάθε πολιτική ξεχωριστά. Δηλαδή, τα αποτελέσματα μπορούν να θεωρηθούν αντικειμενικά αφού έγιναν στο ίδιο σύνολο δεδομένων. Από το γράφημα 6.6 φαίνεται το πλήθος των άρθρων που δεν έχουν συλλεχθεί ακόμα από το σύστημα, για κάθε ώρα της ημέρας. Μπορούμε να διακρίνουμε ότι με την πολιτική ελαχιστοποίησης της συνολικής καθυστέρησης, ο αριθμός των άρθρων είναι 7,5% περισσότερο από την προτεινόμενη προσέγγιση. Τέλος, τα αποτελέσματα του round-robin προγραμματισμού δείχνουν 8,5% περισσότερα άρθρα. Οι μετρήσεις έγιναν ορίζοντας κατάλληλα το ρυθμό ανάκτησης ώστε να προσπελαύνεται το 15% των RSS feeds της βάσης σε κάθε ώρα.

6.2 Εξαγωγή Χρήσιμου Κειμενου

Η διαφοροποίηση του εργαλείου CUTER με τα υπάρχοντα εργαλεία εξαγωγής χρήσιμου περιεχομένου, δεν κατέστησε δυνατή την εκτέλεση συγκριτικών πειραμάτων. Ωστόσο, η πειραματική διαδικασία βασίστηκε στην μέτρηση του ποσοστού του χρήσιμου κειμένου που έχει εξαχθεί και στο ποσοστό εισχώρησης ανεπιθύμητων δεδομένων. Η βέλτιστη υλοποίηση θα πε-



Σχήμα 6.6: Συνολικός αριθμός μη ανακτηθέντων άρθρων

τύχαινε 100% εξαγωγή χρήσιμου κειμένου και 0% εισχώρηση άχρηστου περιεχομένου. Οι ιστοσελίδες που περιέχουν τα άρθρα χωρίστηκαν σε τρεις κατηγορίες:

1. σελίδες που περιέχουν την είδηση σε ενιαίο άρθρο,
2. σε διασπασμένο άρθρο,
3. σε άρθρο ακολουθούμενο από σχόλια χρηστών.

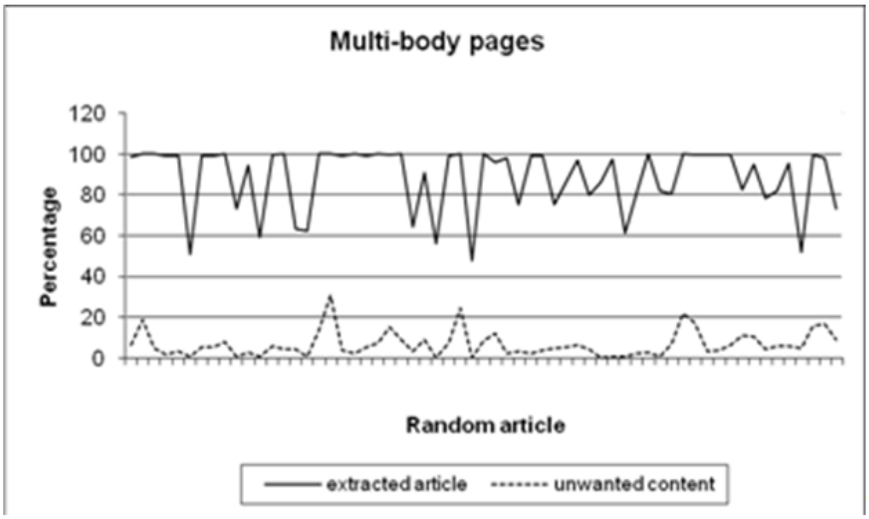
Στην κατηγορία (1) ανήκουν οι ειδήσεις, των οποίων το κείμενο δεν διακόπτεται από μη χρήσιμο περιεχόμενο και αποτελούν την ευκολότερη περίπτωση για τον μηχανισμό εξαγωγής.

Στην κατηγορία (2), στο σώμα του άρθρου παρεμβάλλονται σύνδεσμοι, εικόνες, διαφημίσεις κτλ. κάνοντας τη διαδικασία της εξαγωγής λιγότερο αποδοτική, ενώ παρατηρούμε ότι συχνά εξάγεται μόνο ένα μέρος του πραγματικού άρθρου. Το κείμενο του είδησης που χάνεται είναι συνήθως το τελευταίο μέρος του άρθρου.

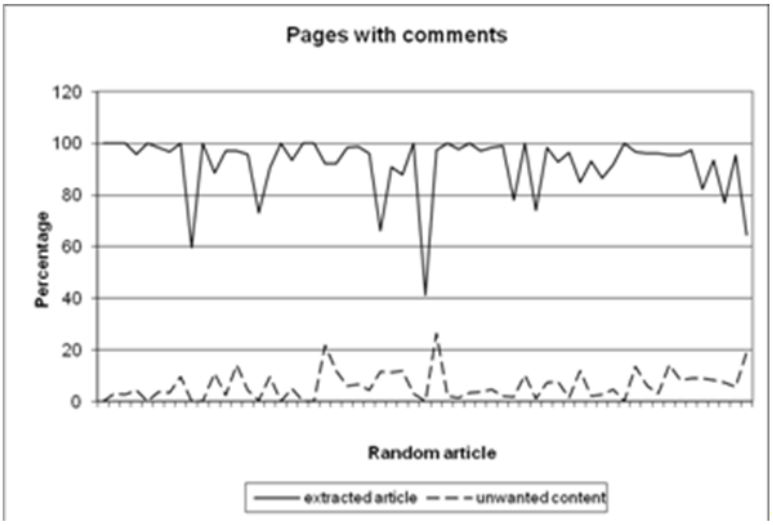
Τέλος, στην κατηγορία (3) ανήκουν οι ιστοσελίδες που διαθέτουν τη δυνατότητα σχολιασμού των άρθρων, όπως τα blogs. Τα σχόλια των χρηστών είναι σχετικά με το άρθρο και σε αρκετές περιπτώσεις βρίσκονται κοντά σε αυτό, δυσκολεύοντας την αναγνώριση τους.

Από τα πειράματα των σχημάτων 6.7, 6.8 και 6.9 φαίνεται ότι το σύστημα είναι αρκετά αποδοτικό για ενιαία άρθρα, ενώ εισάγει ανεπιθύμητο περιεχόμενο στις άλλες μορφές δόμησης. Ο επόμενος πίνακας παρουσιάζει την απόδοση του εργαλείου για τις 3 προαναφερθείσες μορφές των άρθρων. Ως ανάκληση ορίζουμε το ποσοστό του αρχικού άρθρου που έχει

Σχήμα 6.7: Ενιαία άρθρα



Σχήμα 6.8: Κατακερματισμένα άρθρα



Σχήμα 6.9: Άρθρα με σχόλια

εξαχθεί. Αντίστοιχα, ως ποσοστό μη χρήσιμης πληροφορίας, θεωρούμε το κλάσμα του μεγέθους του ανεπιθύμητου κειμένου προς το μέγεθος του κειμένου που εξάγει το σύστημα.
Μπορούμε να πούμε ότι το σύστημα επιτυγχάνει ικανοποιητική απόδοση σε όλες τις περιπτώσεις. Στην πρώτη περίπτωση εξάγονται περισσότερο από το 95% των πραγματικών δεδομένων από την πηγή, ενώ το ποσοστό της μη χρήσιμης πληροφορίας διατηρείται μικρό. Στις άλλες δύο περιπτώσεις μειώνεται η απόδοση, εξάγοντας όμως πάνω από το 90% των δεδομένων του άρθρου. Μεταφράζοντας τα ποσοστά του παραπάνω πίνακα σε πραγματικό μήκος κειμένου, παρατηρούμε ότι για ένα συνηθισμένο άρθρο 250 λέξεων (περίπου 1500 χαρακτήρες), ο μηχανισμός μπορεί να εξάγει με επιτυχία περίπου 220 λέξεις. Οι υπόλοιπες 30 λέξεις θα αποτελούν θόρυβο. Για μεγαλύτερα άρθρα που έχουν περίπου 800-1000 λέξεις, το σύστημα μπορεί να εξάγει 720-900 λέξεις αντίστοιχα.
Το CUTER είναι μέρος ενός μεγαλύτερου συστήματος, το οποίο εκτός των άλλων, πραγματοποιεί περίληψη στα εξαχθέντα άρθρα. Είναι ενδιαφέρον να τονίσουμε ότι με τις διαδικασίες περίληψης φαίνεται ότι αφαιρείται αυτόματα ο θόρυβος (μη χρήσιμη πληροφορία), αφού είναι δυνατόν να ανιχνευτεί ότι δεν πρόκειται για μη σχετικό περιεχόμενο, συγκριτικά με την είδηση.

6.2.1 Εξαγωγή Εικόνων

Όπως αναφέρθηκε, το σύστημα μπορεί να θεωρηθεί ως ένας συνδυασμός δύο επιμέρους υποσυστημάτων, τα οποία υπό προϋποθέσεις είναι δυνατόν να λειτουργούν και ανεξάρτητα, αφού η είσοδος του ενός δεν εξαρτάται άμεσα από την έξοδο του άλλου. Το πρώτο υποσύστημα

εκτελεί την αυτόματη εξαγωγή εικόνων από ειδησεογραφικά άρθρα, ύστερα από την εφαρμογή ενός συνόλου από ελέγχους. Όπως και στα περισσότερα συστήματα ανάκτησης πληροφορίας η κύρια αξιολόγηση βασίζεται στο κατά πόσο είναι ποιοτικά τα αποτελέσματα που παράγονται. Με τον όρο ποιοτικά αποτελέσματα εννοούμε ότι τα αποτελέσματα θα πρέπει να περιέχουν όσο το δυνατόν πιο πολλές εικόνες εκ των οποίων οι περισσότερες να μην είναι άσχετες με το άρθρο. Για αυτό το λόγο, χρησιμοποιούμε τις μετρικές της ακρίβειας και της ανάκλησης για να εκφραστούν όσο το δυνατόν πιο αντικειμενικά τα πειραματικά αποτελέσματα. Ως ακρίβεια, ορίζουμε το κλάσμα του πλήθους των εικόνων που ανακτήθηκαν σωστά δια το πλήθος των συνολικών εικόνων που ανακτήθηκαν.

$$precision = \frac{useful_{images} \cap retrieved_{images}}{retrieved_{images}} \tag{6.1}$$

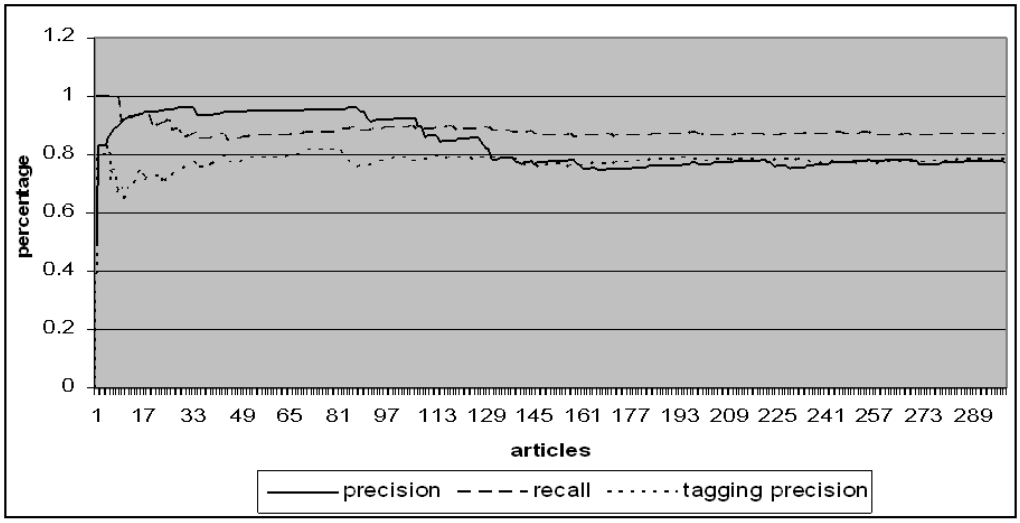
Η ανάκληση ορίζεται ως το κλάσμα του πλήθους των εικόνων που σωστά ανακτήθηκαν δια του συνόλου των εικόνων που έπρεπε να ανακτηθούν.

$$precision = \frac{useful_{images} \cap retrieved_{images}}{useful_{images}} \tag{6.2}$$

Για το δεύτερο υποσύστημα, που περιλαμβάνει τη διαδικασία της απόδοσης μιας περιγραφής σε μια εικόνα που έχει εξαχθεί, είναι χρήσιμο να μετρήσουμε τη συχνότητα αντιστοίχισης μιας αντιπροσωπευτικής ετικέτας. Στη συντριπτική πλειοψηφία των άρθρων, ο συγγραφέας έχει ήδη συμπεριλάβει την περιγραφή της εικόνας, οπότε αυτή θεωρείται ως η αντιπροσωπευτική ετικέτα. Ο παραπάνω συλλογισμός μπορεί να εκφραστεί στην εξίσωση 6.3.

$$tagging = 1, retrieved_{tag} = original_{tag} tagging = 0, else \tag{6.3}$$

Αξίζει να σημειωθεί ότι στα πειράματα που πραγματοποιήθηκαν, η ορθότητα του υποσυστήματος χαρακτηρισμού εικόνων ελέγχθηκε μόνο στις περιπτώσεις επιτυχούς εξαγωγής, δηλαδή στις περιπτώσεις που ανακτήθηκε σωστά μια εικόνα του άρθρου.
Όπως έχει γίνει σαφές, ο μηχανισμός σχεδιάστηκε για να λαμβάνει ως είσοδο σελίδες ειδησεογραφικού περιεχομένου. Συνεπώς, για τα πειράματα επιλέχθηκε ως είσοδος ένα σύνολο από ειδησεογραφικές ιστοσελίδες που είναι πιθανό να δίνονται και στην πραγματική εκτέλεση του. Επιλέχθηκαν τυχαία 300 τυχαίες σελίδες από τη βάση δεδομένων του peRSSonal, με άρθρα διαφόρων κατηγοριών. Αυτό σημαίνει ότι ο μηχανισμός εφαρμόστηκε σε διαφορετικούς ιστότοπους, με τον καθένα να έχει διαφορετική σχεδίαση. Στόχος είναι η μέτρηση της ακρίβειας και της ανάκλησης που πετυχαίνει ο μηχανισμός στο σύνολο αυτών των σελίδων, τόσο για τη



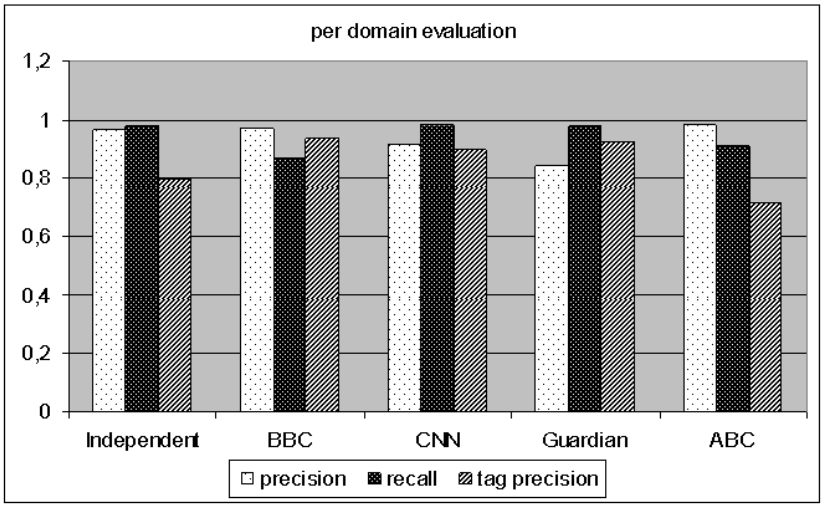
Σχήμα 6.10: Μέση ακρίβεια, ανάκληση και ακρίβεια εξαγωγής

διαδικασία της εξαγωγής, όσο και για τη διαδικασία αυτόματης απόδοσης ετικέτας. Με την τυχαία επιλογή των ιστοσελίδων, όπως είναι αναμενόμενο, περιλαμβάνονται άρθρα που περιέχουν καμία, μία ή περισσότερες εικόνες. Στο σχήμα που ακολουθεί φαίνεται η διαμόρφωση των μετρικών ακρίβειας και ανάκλησης που χρησιμοποιήθηκαν, καθ' όλη τη διάρκεια εκτέλεσης του μηχανισμού.

Στο παραπάνω διάγραμμα φαίνεται ότι οι μετρικές ακρίβειας και ανάκλησης συγκλίνουν σχετικά γρήγορα και σταθεροποιούνται μετά από την εξέταση περίπου 170 άρθρων. Αυτή η παρατήρηση ενισχύει τον ισχυρισμό μας ότι τα 300 άρθρα είναι ένα ικανό δείγμα για να αναδείξει την απόδοση του μηχανισμού που αναπτύχθηκε. Οι τελικές τιμές για τη διαδικασία της εξαγωγής, διαμορφώνονται σε:

- 77% ακρίβεια
- 87% ανάκληση
- 81,7% F1-value

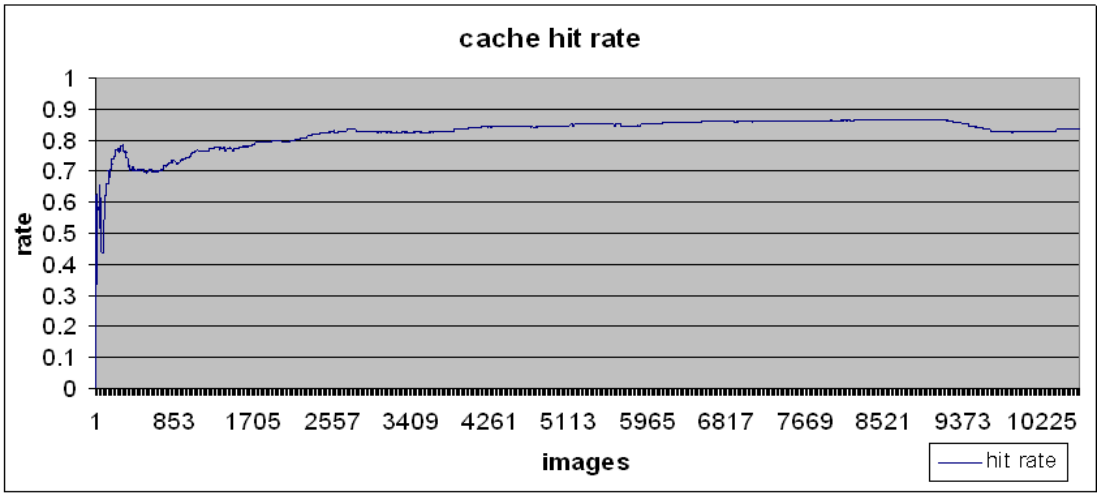
Παρατηρούμε ότι η ακρίβεια επηρεάζεται από διαφημίσεις που είναι ενσωματωμένες στο σώμα του άρθρου, οι οποίες λόγω των χαρακτηριστικών τους και συγκεκριμένα λόγω των διαστάσεων τους, λανθασμένα θεωρούνται ως εικόνες σχετικές με το άρθρο. Άλλος ένας λόγος που συμβάλλει στη δυσκολία αναγνώρισης των διαφημίσεων, είναι και η μοναδικότητα τους καθώς πολλές φορές παράγονται αυτόματα σε κάθε επίσκεψη του χρήστη, έχοντας μοναδικό URL για κάθε άρθρο, γεγονός που καθιστά άσκοπη τη χρήση της προσωρινής μνήμης (cache) που διατηρεί ο μηχανισμός. Από την ανάλυση των αποτελεσμάτων, παρατηρούμε επίσης ότι η



Σχήμα 6.11: μέση ακρίβεια, ανάκληση και ακρίβεια εξαγωγής ανά ιστότοπο

ανάκληση μειώνεται περισσότερο από τη χρήση CSS και Javascript χαρακτηριστικών για την εμφάνιση εικόνων στον εκάστοτε ιστότοπο. Τέτοια χαρακτηριστικά ενεργοποιούν τμήματα κώδικα τα οποία εκτελούνται στον απομακρυσμένο υπολογιστή από τον browser του επισκέπτη. Αυτό έχει ως συνέπεια, η ανάλυση του DOM μοντέλου της σελίδας να μην είναι αρκετή για να εκφράσει την εγγύτητα του άρθρου με την εικόνα, με αποτέλεσμα να απορρίπτονται εικόνες γιατί βρέθηκαν ότι δεν είναι "κοντά" στο κυρίως σώμα του άρθρου. Σχετικά με την απόδοση του υποσυστήματος χαρακτηρισμού των εικόνων, βλέπουμε ότι πετυχαίνει να εντοπίσει την ετικέτα της εικόνας στο 78% των περιπτώσεων. Ωστόσο, ακόμα και όταν δεν επιτυγχάνεται η σωστή εξαγωγή, παρατηρήσαμε ότι συνήθως λαμβάνεται ένα τμήμα του άρθρου ως περιγραφή της εικόνας. Αν και δε θεωρούμε αυτό το αποτέλεσμα ως σωστή ετικέτα για την εικόνα, είναι ενδιαφέρον να αναφερθεί ότι πολλές φορές είναι σχετικό με την εικόνα και θα μπορούσε να χρησιμοποιηθεί σε αλγορίθμους κατηγοριοποίησης της εικόνας βάσει της περιγραφής τους. Όπως έχει ήδη αναφερθεί στα προηγούμενα κεφάλαια, ο μηχανισμός στηρίζεται στην ανάλυση του DOM δέντρου της σελίδας για να εξάγει το τελικό περιεχόμενο. Η DOM αναπαράσταση είναι άρρηκτα συνδεδεμένη με την εμφάνιση της σελίδας έτσι όπως θα τη δει ο επισκέπτης από τον φυλλομετρητή του. Αυτό το γεγονός οδηγεί στο εύλογο συμπέρασμα ότι η απόδοση του συστήματος μεταβάλλεται ανάλογα με το σχεδιασμό και τη δομή της κάθε ιστοσελίδας που ανακτάται. Για να ποσοτικοποιηθεί η μεταβολή της απόδοσης, είναι αναγκαίο να παραθέσουμε το πως διαμορφώνονται οι μετρικές ακρίβειας και ανάκλησης, όταν αναλύονται σελίδες συγκεκριμένων ιστότοπων. Στο διάγραμμα 6.11, παραθέτουμε τη μέση τιμή που λαμβάνουν οι μετρικές κατά την εφαρμογή του μηχανισμού σε πέντε δημοφιλείς πύλες ενημέρωσης.

Άλλο ένα μέρος του συστήματος που πρέπει να αξιολογηθεί για την αποδοτικό-



Σχήμα 6.12: Συχνότητα επιτυχούς αναζήτησης στην cache

τητα του είναι και εκείνο που αποθηκεύει το ιστορικό των ανακτήσεων που έγιναν στο παρελθόν (caching). Σε συστήματα προσωρινής αποθήκευσης, η μετρική που χρησιμοποιούνται κατά κόρον είναι εκείνη που εκφράζει τη συχνότητα με την οποία το αντικείμενο που αναζητείται μπορεί ή όχι να βρεθεί στη μνήμη cache (hit και miss συχνότητες). Στο πείραμα που ακολουθεί δείχνουμε το ποσοστό επιτυχίας εύρεσης του στοιχείου που αναζητούμε (hit rate), ξεκινώντας από μια αρχικά άδεια μνήμη. Τα αποτελέσματα προέκυψαν ύστερα από την επεξεργασία περισσοτέρων από 10630 εικόνων διαφόρων σελίδων και μπορούν να αποτυπωθούν στο παρακάτω σχήμα.

Όπως φαίνεται από το διάγραμμα 6.12, η απόδοση της χρήσης αυτής της μνήμης είναι αυξημένη, αφού το ποσοστό επιτυχούς αναζήτησης είναι περίπου 84%. Ο κύριος στόχος αυτής της μνήμης είναι η μείωση της κατανάλωσης των πόρων του συστήματος, δηλαδή η αποφυγή της λήψης της ίδιας εικόνας παραπάνω από μία φορά. Έτσι το 84% ποσοστό επιτυχίας μεταφράζεται σε 84% λιγότερες ανακτήσεις εικόνων, γεγονός που οδηγεί σε σημαντικά χαμηλότερη χρησιμοποίηση των δικτυακών πόρων του συστήματος.

6.3 Προεπεξεργασία Κειμένου

Για το μηχανισμό προεπεξεργασίας κειμένου καταγράφουμε αρχικά τις αναλύσεις που έχουν γίνει για τη λειτουργία του μηχανισμού και στην πορεία ελέγχουμε διάφορες μετρικές. Παράλληλα ελέγχουμε την εκτέλεση του μηχανισμού που πραγματοποιεί προ-επεξεργασία ελληνικών κειμένων.

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, διεξάγομε πειραματική διαδικασία ώστε να εντοπιστεί ποιο είναι το ελάχιστο δυνατό μήκος από keywords του αρχικού κειμένου που πρέπει να διατηρηθεί, 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.

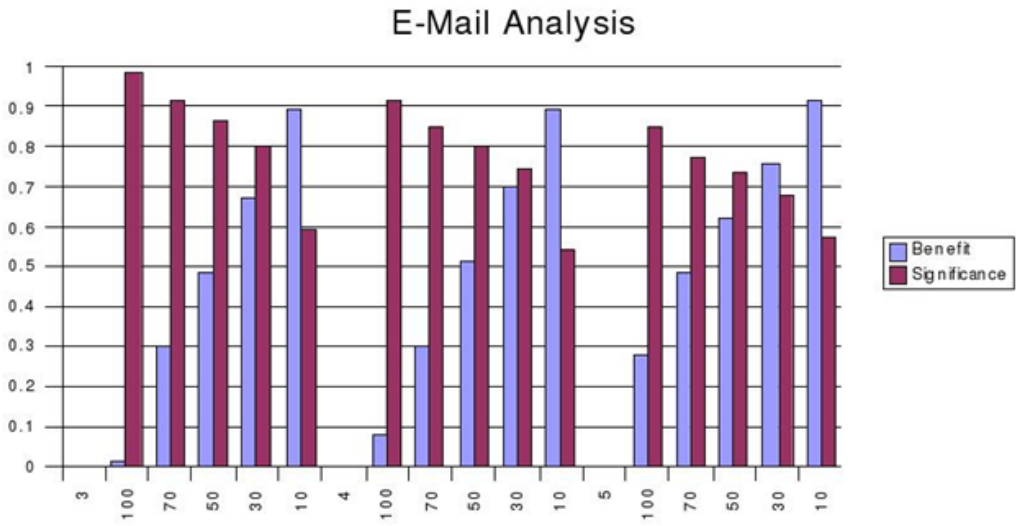
Για να «μετρηθεί» η διαφορά του νοήματος μεταξύ δύο κειμένων (δηλ. εκείνου στο οποίο έχουμε ελάχιστο μήκος λέξεων 4 και εκείνου που έχουμε ελάχιστο μήκος λέξεων 6), χρησιμοποιήθηκε μια απλή έκδοση του SVM αλγορίθμου.
Αν υποθέσουμε ότι έχουμε έναν πίνακα με όλα τα keywords και τις συχνότητές τους για το κείμενο A, και έναν πίνακα του κειμένου B, τότε μπορούμε να υπολογίσουμε τη συσχέτιση μεταξύ των δύο κειμένων όπως φαίνεται στην εξίσωση 6.4.

$$x = a \cdot by = |a| \cdot |b| z = \frac{x}{y} r = sin(z) \tag{6.4}$$

όπου x είναι το εσωτερικό γινόμενο των πινάκων a και b και y το γινόμενο των νορμών (2) του A και του B. 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.

6.3.1 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 πειρα-

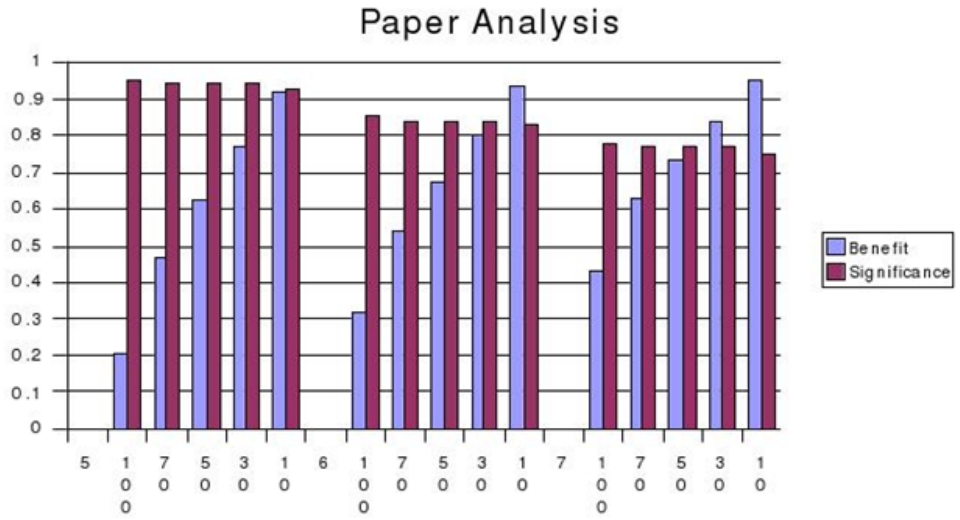


Σχήμα 6.13: Ανάλυση κειμένων ηλεκτρονικού ταχυδρομείου

ματικής διαδικασίας χρησιμοποιήθηκε ελάχιστο μήκος λέξεων τριών, τεσσάρων και πέντε γραμμάτων. Τα αποτελέσματα συνοψίζονται στην γραφική απεικόνιση του παρακάτω σχήματος.

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 λέξεις κλειδιά 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.

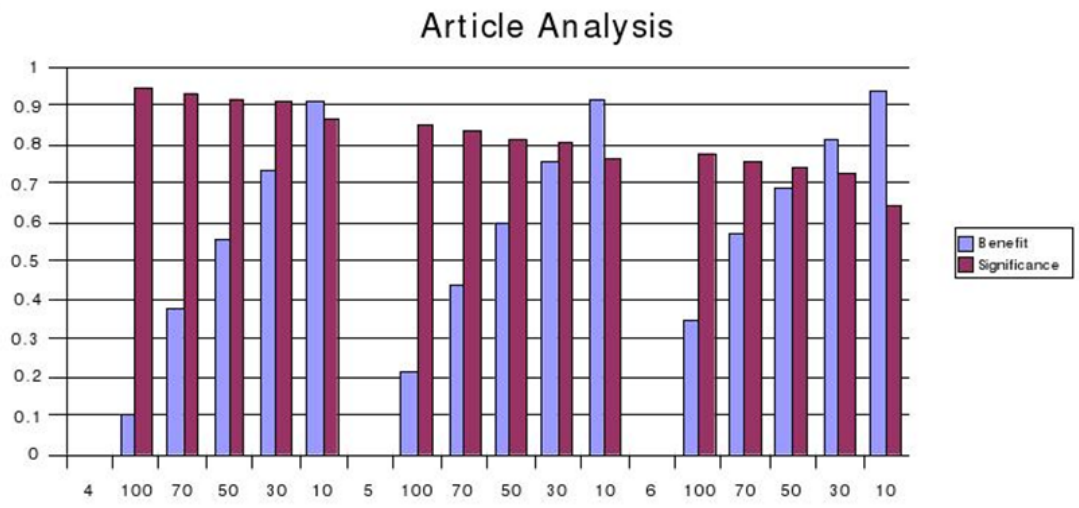


Σχήμα 6.14: Ανάλυση κειμένων ερευνητικών δημοσιεύσεων

6.3.2 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.

Όπως μπορούμε να δούμε από τη γραφική παράσταση του σχήματος 6.14, 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 επανάληψη 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.



Σχήμα 6.15: Ανάλυση κειμένων άρθρων

6.3.3 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 λέξεων κλειδιών 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.

6.3.4 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.

6.4 Μηχανικός προεπεξεργασίας Ελληνικών Κειμένων

Σε αντίθεση με άλλους αλγορίθμους που πραγματοποιούν stemming, ο μηχανισμός δεν προσπαθεί να εξάγει απόλυτα γραμματικά σωστά τις ρίζες των λέξεων, αν και στην πλειονότητα των περιπτώσεων το καταφέρνει, αρκεί να μπορεί να δώσει ίδια απάντηση για λέξεις που πηγάζουν από το ίδιο θέμα. Στην ουσία αυτό που επιθυμούν να επιτύχουν είναι να μπορέσουμε να εντοπίσουμε ίδια ρίζα για λέξεις που πράγματι έχουν την ίδια ρίζα, άσχετα αν αυτή η ρίζα είναι η σωστή απάντηση και αρκεί με αυτή τη ρίζα να μη «μπερδεύουμε» άλλες λέξεις του συστήματος. Για παράδειγμα, όταν έχουμε να κάνουμε με ρήματα θεωρούμε τη διαδικασία του stemming ως σωστή όταν η απάντηση είναι η ρίζα του πρώτου ενικού σε όποια κλίση και αν βρούμε το ρήμα. Αυτή η μέθοδος προσεγγίζει αυτό που λέγεται lemmatization

ωστόσο είναι ξεκάθαρο για εμάς πως αυτός είναι ο ουσιαστικός τρόπος λειτουργίας του συστήματός μας. Αντίστοιχα για τα επίθετα μετατρέπουμε τα στοιχεία για να ταιριάξουν στο πρώτο ενικό του αρσενικού, ενώ αν δεν υπάρχει αρσενικό τότε προσπαθούμε να ταιριάξουμε με το πρώτο που υπάρχει, αν δεν υπάρχει ενικός με το πρώτο πρόσωπο του πληθυντικού αριθμού, κ.ο.κ. Επιπρόσθετα παρά το γεγονός ότι δεν είναι αναγκαίο για να θεωρήσουμε τη διαδικασία του stemming σωστή προσπαθούμε να εντοπίσούμε ίδια ρίζα τόσο για ουσιαστικά όσο και για ρήματα που πράγματι έχουν την ίδια ρίζα. Σε γενικές γραμμές σκοπός μας είναι να δημιουργήσουμε ένα stemmer, ωστόσο τον προσαρμόζουμε πλήρως στις ανάγκες του perssonal οι οποίες δεν επιβάλουν σωστό stemming αλλά επιβάλεται η σωστή αντιστοίχηση. Αντίστοιχα, για τα επιρρήματα προσπαθούμε και γι αυτά να βρούμε τα αντίστοιχα ρήματα από τα οποία προκύπτουν για να προσπαθήσουμε το stemming. Θεωρητικά για τις υπόλοιπες λέξεις η διαδικασία του stemming είναι τετριμένη και γι αυτό αξιολογούμε αν το tagging επιτυγχάνει. Για να αξιολογήσουμε τον αλγόριθμό μας χρησιμοποιήσαμε δύο διαφορετικά σετ κειμένων. Το πρώτο ήταν κατασκευασμένο από άρθρα και ειδήσεις όπως αυτά χρησιμοποιούνται στο perssonal τα οποία είχαν κάποιες χιλιάδες λέξεις. Από την άλλη μεριά το δεύτερο set δημιουργήθηκε από την ένωση mail που ανταλάσσαμε μεταξύ μας και τα οποία εν πολλοίς συχνά έχουν ορολογία, greeklish, λάθος σύνταξη και γενικά πολλά χαρακτηριστικά που θα μπέρδευαν ένα σύστημα.

Ο χρόνος εκτέλεσης του συστήματος πάνω στα κείμενα εισόδου ήταν μη μετρήσιμος καθώς το σύστημα στο server που χρησιμοποιούμε μπορεί να αναλύσει περίπου 163.000 (εκατόν εξήντα τρεις χιλιάδες) χαρακτήρες το δευτερόλεπτο. Αρχικά αξιολογήσαμε το σύστημα tagging του μηχανισμού σε κάθε λέξη του κειμένου εισόδου και στη συνέχεια αξιολογήσαμε το μηχανισμό stemming πάνω στο κείμενο. Η σύγκριση που κάναμε στο stemmer έγινε με τον μοναδικό ελληνικό που έχουμε βρει και είναι αυτός του Γεώργιου Νταή [158]. Ακριβώς επειδή είναι τέτοια και τα δικά του πειράματα, αγνοήσαμε κάθε λέξη εκτός από ουσιαστικά, επίθετα, ρήματα και επιρήματα. Άλλωστε λέξεις που ανήκουν σε άλλες κατηγορίες πέραν αυτών δύσκολα έχουν κάποια σημασία για το νόημα του κειμένου, κάτι πολύ βασικό για το σύστημά μας. Παράλληλα πρόκειται για λέξεις που είναι πολύ συχνές και πολυχρησιμοποιημένες και η ρίζα τους είναι γνωστή κάτι που θα ανέβαζε την ποιότητα των αποτελεσμάτων μας χωρίς αυτό να αντικατοπτρίζει την πραγματικότητα.

Από τις λέξεις που είχαμε σαν είσοδο διαπιστώσαμε ότι καταφέραμε να επιτύχουμε 98,6% ακρίβεια του μηχανισμού tagging και 96,7% επιτυχία στο μηχανισμό stemming πάνω στις προαναφερθείσες κατηγορίες λέξεων. Από τα σφάλματα που έκανε ο μηχανισμός περίπου 13% ήταν αποτέλεσμα over-stemming το οποίο σημαίνει πως αφαιρέσαμε περισσότερα γράμματα απ’ τι έπρεπε. Το υπόλοιπο 87% περίπου των λέξεων είτε ήταν under-stemmed, δηλαδή δεν αφαιρέσαμε όσα γράμματα έπρεπε είτε ήταν ανώμαλες λέξεις και ο μηχανισμός δε γνώριζε πως ακριβώς να συμπεριφερθεί. Συγκεκριμένα το 75% των λαθών προέκυψαν από ανώμαλα ρήματα, γεγονός το οποίο μπορεί να αποφευχθεί γενικά και όχι μόνο από το μηχανισμό μας με ένα και μόνο τρόπο, με την προσθήκη εξαιρέσεων. Παράλληλα και για το ίδιο σετ λέξεων

Πίνακας 6.1: Συγκριτικά αποτελέσματα για μετρικές μεταξύ του Ntais stemmer και το G.I.C.S stemmer

	Ntais stemmer	G.I.C.S.
Index compression factor	76.8%	80.9%
Mean modified Hamming Distance	1.95	2.73
Median modified Hamming Distance	2	2

εκτελέσαμε τον αλγόριθμο του Νταή [158] ώστε να μπορέσουμε να κάνουμε μία σύγκριση των δύο μηχανισμών.
Η σύγκριση μας έδειξε πως ο αλγόριθμος του Νταή [158] μπορεί να επιτύχει 91,1% για το πρώτο σετ κειμένων (άρθρα).
Αναφορικά με το δεύτερο σετ το οποίο περιλάμβανε κείμενο από email ο μηχανισμός μας κατάφερε να φτάσει στο 96% περίπου αναφορικά με το stemming ενώ η απόδοσή του στο tagging μειώθηκε δραματικά, καθότι όπως αναφέρθηκε τα email δεν έχουν σωστή σύνταξη. Από τα λάθη που εντοπίσαμε ότι έκανε ο μηχανισμός και πάλι ένα ποσοστό 7% ήταν λάθη του μηχανισμού ενώ το 80% των υπολοίπων λαθών ήταν ανώμαλα ρήματα που δεν υπήρχαν σαν εξαιρέσεις στο μηχανισμό μας. Συγκρικά και πάλι με τον αλγόριθμο του Νταή, πετυχαίνουμε και πάλι καλύτερη απόδοση καθότι ο μηχανισμός του νταή φτάνει την ακρίβεια του 88,15%.

Σαν επόμενο πείραμα χρησιμοποιούμε τρείς διαφορετικές μετρικές σύγκρισης μηχανισμών stemming [92]. Η πρώτη μετρική ονομάζεται index compression factor (ICF) και μας δείχνει το ποσοστο μείωσης του κειμένου μετά το stemming.

$$ICF = \frac{n - s}{n} \tag{6.5}$$

Όπου n είναι ο αριθμός των λέξεων της συλλογής κειμένων και s ο αριθμός των παραγόμενων stems. Οι άλλες δύο μετρικές σχετίζονται με το Hamming Distance και είναι η σταθμισμένη μέση και μέση απόσταση Hamming. Η απόσταση Hamming ορίζεται σαν ο αριθμός των χαρακτήρων δύο string που είναι διαφορετική στην ίδια θέση. Η είσοδος που δίνουμε στο σύσ-τημα είναι πάνω από 10.000 λέξεις και τα αποτελέσματα φαίνονται στον πίνακα 6.1.

6.5 Μηχανισμοί Κατηγοριοποίησης και Εξαγωγής Περίληψης

Κάθε μια από τις εξισώσεις 5.12 και 5.17 που είδαμε σε προηγούμενο κεφάλαιο για την βαθμολόγηση των προτάσεων ελέγχθηκε σε κάποια προκατηγοριοποιημένα (από ανθρώ-

Πίνακας 6.2: Συγκριτικά αποτελέσματα για μετρικές μεταξύ του MSWord Summarizer και του προτεινόμενου μηχανισμού

	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

πους) 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 5.12 και 5.17 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.

6.5.1 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. 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

Πίνακας 6.3: 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	Similar Articles	Precision	Recall
10 min	0	0,5	0,66
8 hours	8	0,5	0,66
12 hours	31	0,66	0,5
18 hours	43	0,66	0,66
24 hours	59	0,66	0,66
30 hours	88	0,75	0,75
36 hours	103	0,75	0,8

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.

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. 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

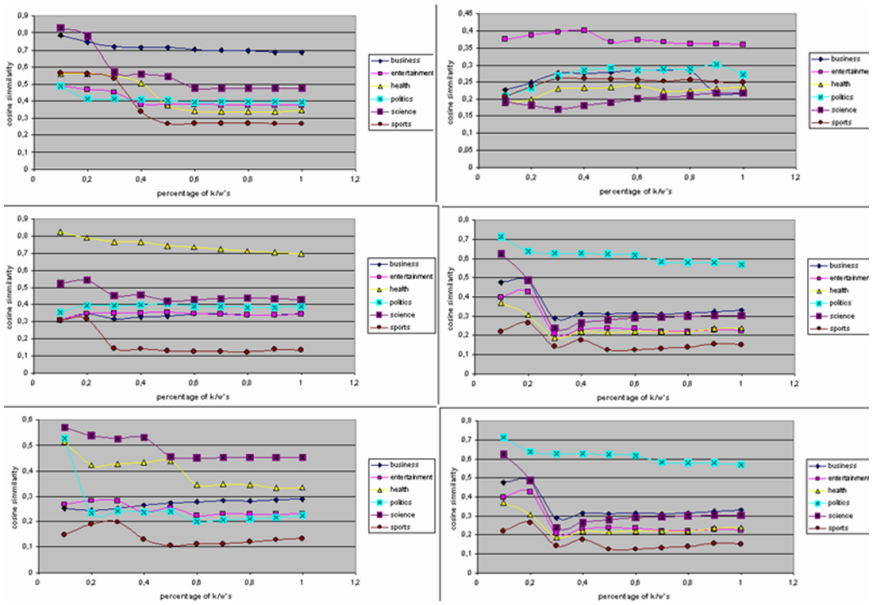
το πλήρες κείμενο του άρθρου, (β) δώσε στους χρήστες τις περιλήψεις που προέκυψαν τόσο από την εξίσωση (6), όσο και από την εξίσωση (8), και (γ) άφησε τους χρήστες να επιλέξουν ποια περίληψη θεωρούν ως περισσότερο αντιπροσωπευτική για το άρθρο που διάβασαν. Η αντίστροφη διαδικασία εξετάστηκε επίσης, 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 k4 (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.

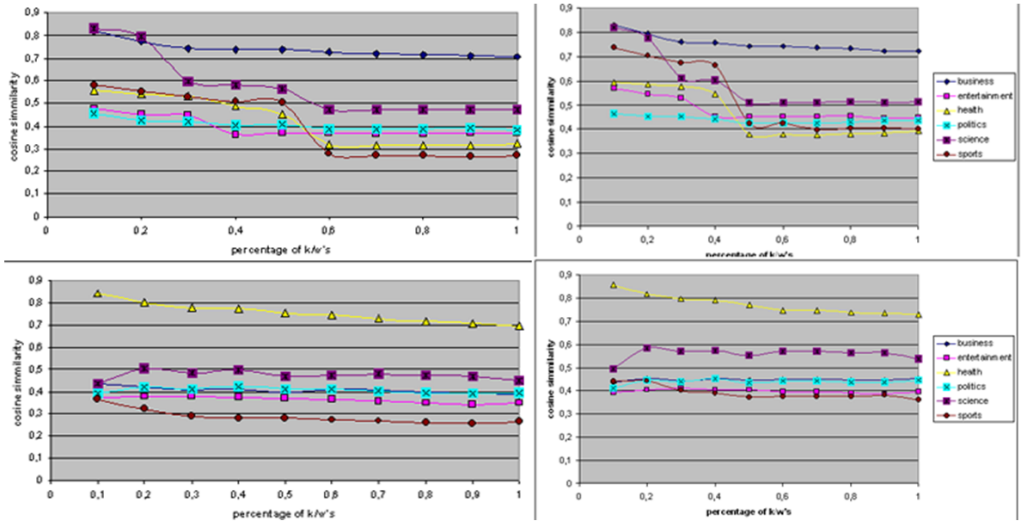
6.5.2 Αλληλεπίδραση μεταξύ της διαδικασίας περίληψης και κατηγοριοποίησης

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 ούτως ώστε 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.

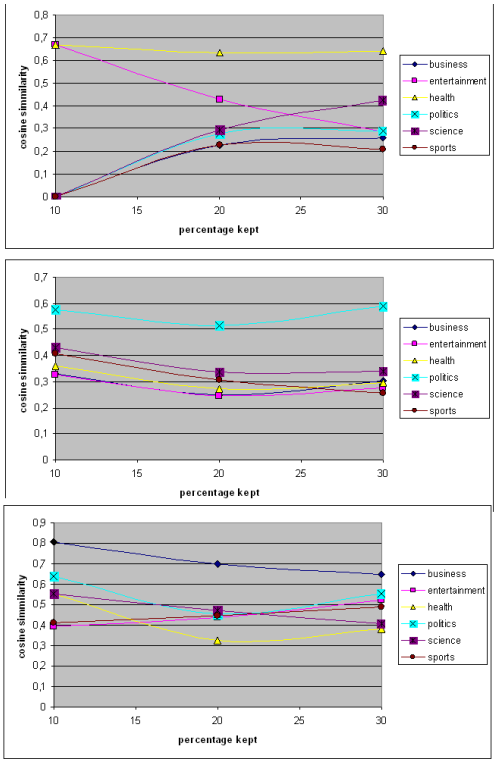
Από την εικόνα 6.16 (αποτελέσματα διαδικασίας κατηγοριοποίησης), προκύπτει ότι ένα ποσοστό 30% των keywords του κειμένου πρέπει να κρατηθούν από την διαδικασία κατηγοριοποίησης ώστε αυτή να είναι βέλτιστη. Αν και ένα μικρότερο ποσοστό μπορεί να είναι επαρκές ώστε να αποφασιστεί η κατηγορία του κειμένου, 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 εικόνα 6.17, ένα κείμενο μπορεί να επιτύχει καλύτερο σκορ χρησιμοποιώντας ένα ελάχιστο μήκος 5 γραμμάτων για τα keywords και κρατώντας 50% των keyowrds



Σχήμα 6.16: Ομοιότητα συνημιτόνου των κειμένων σε σχέση με τις κατηγορίες. Το training set κατασκευάζεται με χρήση του 50% των keywords (διαδικασία προεπεξεργασίας)



Σχήμα 6.17: Σύγκριση Ομοιότητας Συνημιτόνου - Πρώτη στήλη στο 50% των keywords. Δεύτερη στήλη στο 100% των keywords του training set.



Σχήμα 6.18: Ομοιότητα συνημιτόνου που μετρήθηκε για την κατηγοριοποίηση περιλήψεων χρησιμοποιώντας διάφορα ποσοστά για την δημιουργία των περιλήψεων

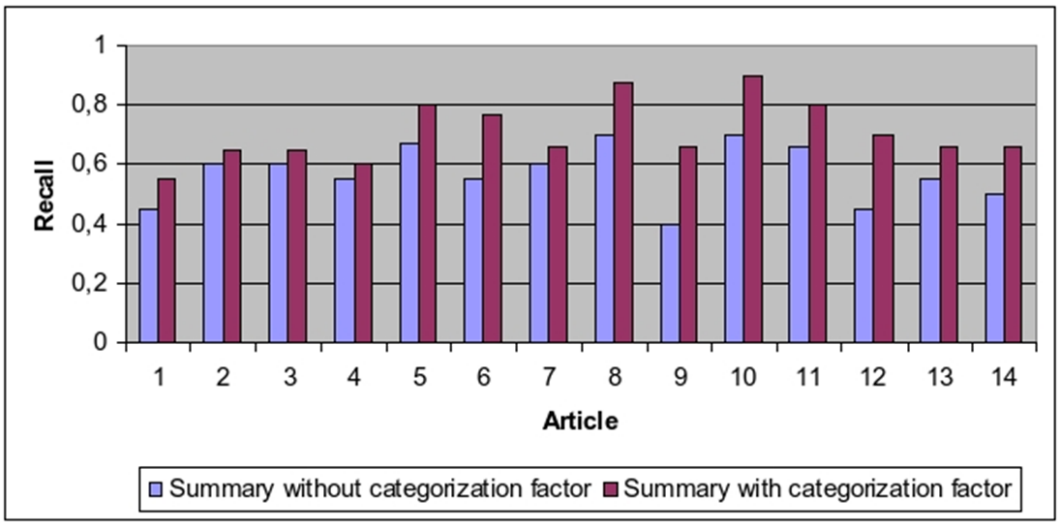
που προκύπτουν. Με αυτό τον τρόπο, η βάση γνώσης είναι πιο φιλτραρισμένη, ενώ δεν μένουν έξω από τη διαδικασία keywords σημαντικά για κάποια/ες κατηγορία/ες. Στο επόμενο βήμα της πειραματικής διαδικασίας, θέλουμε να εξεταστεί η επιρροή που έχει η διαδικασία περίληψης στο στάδιο της κατηγοριοποίησης. 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.

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%,

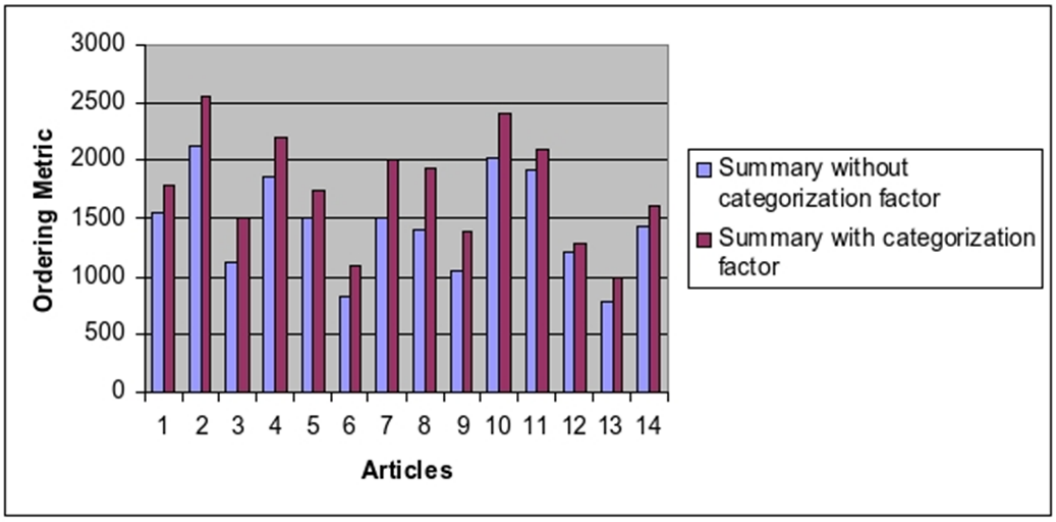
για την παραγωγή της περίληψης του κειμένου, μπορούμε να κατηγοριοποιήσουμε την περίληψη σωστά στην κατηγορία του κειμένου. Με αυτό τον τρόπο γλιτώνουμε ένα τεράστιο ποσοστό της δουλειάς που πρέπει να γίνει στην πλευρά της κατηγοριοποίησης, αφού η περίληψη είναι μόνο ένα μικρό μέρος του κειμένου. Αυτό 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». 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.

6.6 Προσωποποιημένη Προβολη Περιεχομένου

Το σημείο στο οποίο ίσως πρέπει να δώσουμε μεγαλύτερη βαρύτητα είναι αυτό που σχετίζεται με το perssonal meta-portal. Αυτό γίνεται διότι αυτό είναι το κομμάτι το οποίο είναι εμφανές στους χρήστες. Μάλιστα, για πρώτη φορά μέσα από αυτή την εργασία θα δούμε



Σχήμα 6.19: Σύγκριση της ανάκλησης των περιλήψεων οι οποίες εξήχθηκαν με και χωρίς την χρήση του παράγοντα κατηγοριοποίησης



Σχήμα 6.20: Σύγκριση της μετρικής σειράς από περιλήψεις που εξήχθηκαν με και χωρίς τον παράγοντα κατηγοριοποίησης

την ελληνική έκδοση του μηχανισμού που πλέον μπορεί να λειτουργεί με μεγάλη επιτυχία δίνοντάς μας χρήσιμα αποτελέσματα για να μπο-ρέσουμε να αναβαθμίσουμε ακόμα περισσότερο το μηχανισμό που κατασκευάζουμε. Το peRSSonal meta-portal είναι κατασκευασμένο από σειρά στοιχείων τα οποία είναι διαθέσιμα σαν ενσωματωμένα micro-applications και τα οποία μπορούν να λειτουργήσουν επικουρικά στις διαδικασίες που πραγματοποιεί ο χρήστης κατά την ανάγνωση άρθρων και ειδήσεων. Μιας και το meta portal που έχουμε φτιάξει αποτελείται από μία σειρά υποσυστημάτων θα ήταν καλό να επικεντρώσουμε την προσοχή μας σε συγκεκριμένα στοιχεία που αξίζει να αναλύσουμε στην πειραματική διαδικασία. Ένα από τα θέματα που θα δούμε αναλυτικά κατά τη διάρκεια της ανάλυσης και πειραματικής διαδικασίας είναι ο τρόπος λειτουργίας του δικτυακού τόπου τον οποίο έχουμε δημιουργήσει. Ο τρόπος λειτουργίας σχετίζεται βασικά και κύρια από τα δομικά στοιχεία τα οποία απαρτίζουν το δικτυακό τόπο αλλά και τον τρόπο με τον οποίο παρουσιάζεται η πληροφορία αλλά και οι παρεχόμενες υπηρεσίες στους χρήστης. Η ανάλυση έχει σαν σκοπό να δούμε αναλυτικά τι υπηρεσίες παρέχονται στους χρήστες για να προχωρήσουμε στη συνέχεια σε επιπλέον ανάλυση του τρόπου με τον οποίο δημιουργείται και αλλάζει συνεχώς το προφίλ ενός χρήστη καθώς και τα πειράματα που έγιναν προς αυτή την κατεύθυνση. Επιπλέον, θα δούμε στοιχεία τα οποία σχετίζονται με τον τρόπο πρόσβασης στην πληροφορία αλλά και θέματα που έχουν να κάνουν με τις επιλογές του χρήστη και πως αυτό μπορεί να αλλάξει στοιχεία που τον αφορούν.

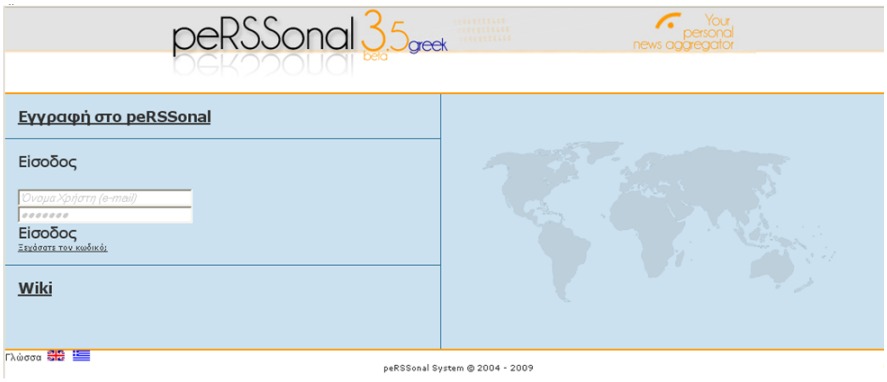
Αφού δούμε διεξοδικά πως μπορούμε να πειραματιστούμε με τις σελίδες του peRSSonal θα προχωρήσουμε στον αναλυτικό έλεγχο των πειραματικών διαδικασιών που οδηγούν στη δια-μόρφωση του προφίλ που πραγματοποιείται για τους χρήστες του συστήματος. Η δια-μόρφωση του προφίλ των χρηστών είναι μεγάλης σημασίας για το σύστημά μας διότι περιέχει όλη αυτή την πληροφορία που αποτελεί το δομικό λίθο της προσωποποίησης. Έτσι λοιπόν, στην πειραματική διαδικασία που πραγματοποιούμε θα ασχοληθούμε με στοιχεία που σχετίζονται με τη διαμόρφωση του προφίλ του χρήστη και την ποιότητα που παρέχεται στο χρήστη μετά τη διαμόρφωση του προφίλ. Τα στοιχεία αυτά δείχνουν τον τρόπο με τον οποίο επηρεάζεται το προφίλ του χρήστη με την πάροδο του χρόνου και τι σημαίνει αυτό στην πληροφορία που του παρουσιάζεται μέσα από το peRSSonal.

Επιπλέον των παραπάνω θα δούμε και πειραματικές διαδικασίες που σχετίζονται με συστήματα που λειτουργούν μέσα στο σύστημα per-sonal. Τα συστήματα αυτά σχετίζονται με υπηρεσίες όπως είναι το article tagging, αλλά και το πολύ σημαντικό κομμάτι της ομαδοποίησης άρθρων. Το τελευταίο αποτελεί ένα πολύ σημαντικό κομμάτι του μηχανισμού το οποίο έχουμε καταφέρει να λειτουργούμε σε πραγματικό χρόνο, όπως ακριβώς μας επιτάσσει η ταχύτητα του διαδικτύου.

Το σημείο στο οποίο ίσως πρέπει να δώσουμε μεγαλύτερη βαρύτητα είναι αυτό που σχετίζεται με το perssonal meta-portal. Αυτό γίνεται διότι αυτό είναι το κομμάτι το οποίο είναι εμφανές στους χρήστες. Μάλιστα, για πρώτη φορά μέσα από αυτή την εργασία θα δούμε την ελληνική έκδοση του μηχανισμού που πλέον μπορεί να λειτουργεί με μεγάλη επιτυχία

δίνοντάς μας χρήσιμα αποτελέσματα για να μπορέσουμε να αναβαθμίσουμε ακόμα περισσότερο το μηχανισμό που κατασκευάζουμε. Το peRSSonal meta-portal είναι κατασκευασμένο από σειρά στοιχείων τα οποία είναι διαθέσιμα σαν ενσωματωμένα micro-applications και τα οποία μπορούν να λειτουργήσουν επικουρικά στις διαδικασίες που πραγματοποιεί ο χρήστης κατά την ανάγνωση άρθρων και ειδήσεων. Μιας και το meta portal που έχουμε φτιάξει αποτελείται από μία σειρά υποσυστημάτων θα ήταν καλό να επικεντρώσουμε την προσοχή μας σε συγκεκριμένα στοιχεία που αξίζει να αναλύσουμε στην πειραματική διαδικασία. Ένα από τα θέματα που θα δούμε αναλυτικά κατά τη διάρκεια της ανάλυσης και πειραματικής διαδικασίας είναι ο τρόπος λειτουργίας του δικτυακού τόπου τον οποίο έχουμε δημιουργήσει. Ο τρόπος λειτουργίας σχετίζεται βασικά και κύρια από τα δομικά στοιχεία τα οποία απαρτίζουν το δικτυακό τόπο αλλά και τον τρόπο με τον οποίο παρουσιάζεται η πληροφορία αλλά και οι παρεχόμενες υπηρεσίες στους χρήστης. Η ανάλυση έχει σαν σκοπό να δούμε αναλυτικά τι υπηρεσίες παρέχονται στους χρήστες για να προχωρήσουμε στη συνέχεια σε επιπλέον ανάλυση του τρόπου με τον οποίο δημιουργείται και αλλάζει συνεχώς το προφίλ ενός χρήστη καθώς και τα πειράματα που έγιναν προς αυτή την κατεύθυνση. Επιπλέον, θα δούμε στοιχεία τα οποία σχετίζονται με τον τρόπο πρόσβασης στην πληροφορία αλλά και θέματα που έχουν να κάνουν με τις επιλογές του χρήστη και πως αυτό μπορεί να αλλάξει στοιχεία που τον αφορούν.
Αφού δούμε διεξοδικά πως μπορούμε να πειραματιστούμε με τις σελίδες του peRSSonal θα προχωρήσουμε στον αναλυτικό έλεγχο των πειραματικών διαδικασιών που οδηγούν στη δια-μόρφωση του προφίλ που πραγματοποιείται για τους χρήστες του συστήματος. Η δια-μόρφωση του προφίλ των χρηστών είναι μεγάλης σημασίας για το σύστημά μας διότι περιέχει όλη αυτή την πληροφορία που αποτελεί το δομικό λίθο της προσωποποίησης. Έτσι λοιπόν, στην πειραματική διαδικασία που πραγματοποιούμε θα ασχοληθούμε με στοιχεία που σχετίζονται με τη διαμόρφωση του προφίλ του χρήστη και την ποιότητα που παρέχεται στο χρήστη μετά τη διαμόρφωση του προφίλ. Τα στοιχεία αυτά δείχνουν τον τρόπο με τον οποίο επηρεάζεται το προφίλ του χρήστη με την πάροδο του χρόνου και τι σημαίνει αυτό στην πληροφορία που του παρουσιάζεται μέσα από το peRSSonal.
Επιπλέον των παραπάνω θα δούμε και πειραματικές διαδικασίες που σχετίζονται με συστήματα που λειτουργούν μέσα στο σύστημα per-sonal. Τα συστήματα αυτά σχετίζονται με υπηρεσίες όπως είναι το article tagging, αλλά και το πολύ σημαντικό κομμάτι της ομαδοποίησης άρθρων. Το τελευταίο αποτελεί ένα πολύ σημαντικό κομμάτι του μηχανισμού το οποίο έχουμε καταφέρει να λειτουργούμε σε πραγματικό χρόνο, όπως ακριβώς μας επιτάσσει η ταχύτητα του διαδικτύου.

Παρά το γεγονός πως το σύστημα συνήθιζε να είναι ανοιχτό προς μη εγγεγραμμένους χρήστης, η χρήση αναβαθμισμένων αλγορίθμων σε κάθε στοιχείο του μηχανισμού καθιστά στην ουσία ανούσια αυτή τη λειτουργία και έτσι όπως βλέπουμε και από την αρχική σελίδα 6.21 οι χρήστες πρέπει να εγγραφούν στο σύστημα για να μπορέσουν να το χρησιμοποιήσουν.



Σχήμα 6.21: peRSSonal meta-portal

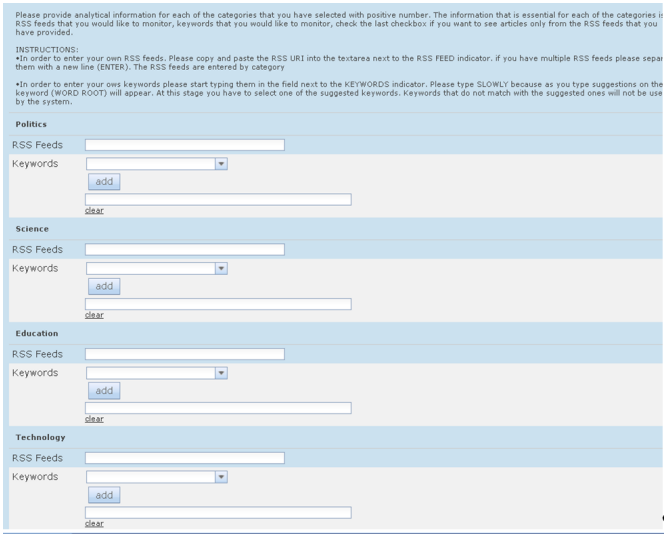
Σχήμα 6.22: Εγγραφή στο peRSSonal meta-portal

Η εγγραφή στο σύστημα είναι απλή και απαιτεί απλώς μία σειρά από βήματα που πρέπει να πραγματοποιήσει ένας χρήστης. Σε πρώτο επίπεδο ένας χρήστης καλείται να δώσει κάποια προσωπικά στοιχεία για να γίνει η εγγραφή. Τα στοιχεία αυτά δεν είναι τίποτα περισσότερο από ένα e-mail και ένας κωδικός πρόσβασης στις σελίδες του peRSSonal 6.22.

Η εγγραφή όπως μπορούμε να δούμε χρειάζεται ένα e-mail απλά και μόνο για να έχουμε ένα αναγνωριστικό για το χρήστη και να μπορούμε να έρθουμε σε κάποια υποτυπώδη επικοινωνία για θέματα που αφορούν το χρήστη. Επίσης, υπάρχει και η σκέψη για τη δημιουργία ημερήσιου personalized newsletter για κάθε χρήστη και άρα ένα στοιχείο επικοινωνίας είναι απαραίτητο για να ενημερώνεται ο χρήστης.
Στη συνέχεια περνάμε σε στοιχεία που αφορούν το προφίλ του χρήστη και ουσιαστικά ζητάμε

Σχήμα 6.23: Επιλογή κατηγοριών στο peRSSonal meta-portal

πολύ βασικά στοιχεία από τους χρήστες. Δεδομένου ότι το σύστημά μας χρησιμοποιεί επτά διαφορετικές κατηγορίες, καλούμε τους χρήστες να δηλώσουν τις προτιμήσεις τους προς αυτές τις κατηγορίες. Ο χρήστης που θα φτιάξουμε για να δούμε τα πειράματά μας θα έχει θετικά στοιχεία για κάποιες κατηγορίες ενώ για άλλες θα έχει αρνητικά, όπως δηλαδή αναμένουμε να κάνει κάθε χρήστης ο οποίος σκοπεύει να χρησιμοποιήσει το σύστημά μας. Αυτό θα μας βοηθήσει και στην πειραματική διαδικασία που θα κάνουμε πάνω στο χρήστη μας 6.23. Όπως βλέπουμε στο σχήμα 6.23 ο χρήστης έχει τη δυνατότητα να καθορίσει για κάθε κατηγορία του συστήματος στοιχεία που σχετίζονται με την προτίμηση που έχει αναφορικά με την κατηγορία. Αυτή είναι η σελίδα που δίνει την είσοδο για τη δημιουργία του πρώτου προφίλ χρήστη. Σύμφωνα με τους συνδυασμούς που μπορούμε να κάνουμε και ακριβώς επειδή οι κατηγορίες μεταβάλλονται δυναμικά με τη λειτουργία του συστήματος θεωρητικά δε μπορεί να προκύψουν από την αρχή δύο ταυτόσημα προφίλ χρηστών, εκτός κι αν αυτοί δηλώσουν επακριβώς τα ίδια στοιχεία στη συγκεκριμένη σελίδα και κάνουν εγγραφή με διαφορά λίγων ωρών, μέσα στις οποίες δεν έχει συντελεστεί καμία απολύτως αλλαγή στις κατηγορίες του συστήματος. Με τις επιλογές που κάνει ο χρήστης σε αυτή τη σελίδα δημιουργείται όπως είπαμε και το προφίλ και είναι πολύ σημαντικό ο χρήστης σε αυτό το επίπεδο να κατανοήσει πως οι επιλογές του μπορεί να παίξουν σημαντικό ρόλο στη διαμόρφωση του προφίλ του. Όπως είδαμε μέσα από την έρευνα που κάναμε, όσο πιο συνειδητοποιημένος είναι ο χρήστης με τις επιλογές του σε αυτή τη σελίδα τόσο πιο εύκολα μπορούμε να δημιουργήσουμε ένα σταθερό προφίλ γι αυτόν. Φυσικά το σύστημα δεν απαγορεύει στους χρήστες να δημιουργήσουν ένα πολύ γενικό προφίλ δίνοντας μηδενικά στοιχεία στο σύστημα (π.χ. να δώσουν επιλογή 0 σε κάθε κατηγορία σε αυτό το βήμα). Σε αυτή την περίπτωση, και πάλι το σύστημα θα μπορέσει να προχωρήσει στη διαμόρφωση του προφίλ, ωστόσο αυτό θα



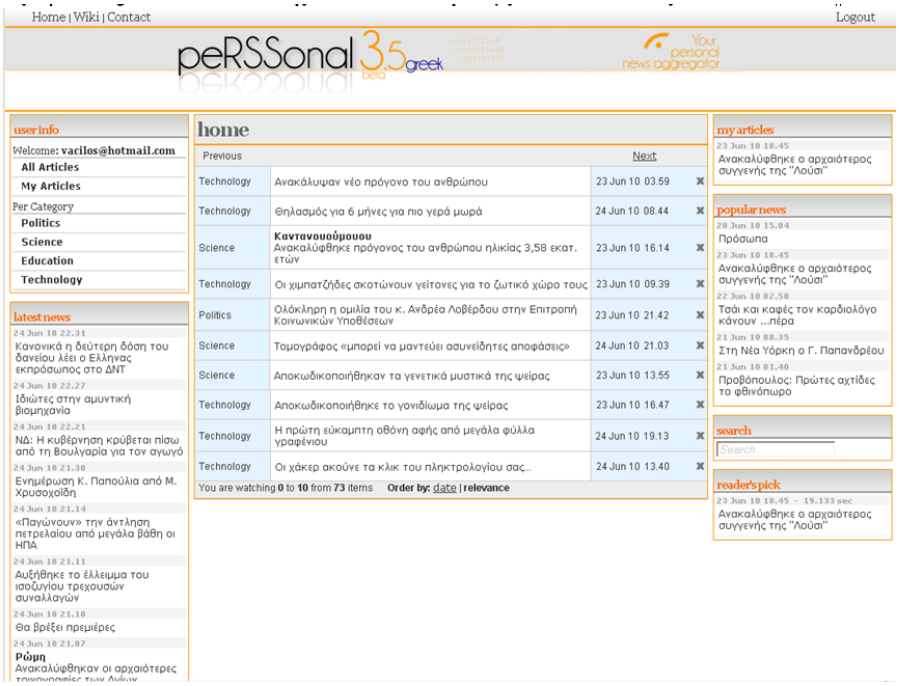
Σχήμα 6.24: Επιλογή λέξεων κλειδιών και RSS feeds στο peRSSonal meta-portal

γίνει με μεγαλύτερη δυσκολία αλλά και μεγαλύτερη προσπάθεια από τον ίδιο το χρήστη.
Ένα επόμενο βήμα αλλά εξαιρετικά σημαντικό το οποίο απευθύνεται κυρίως στους εξοικειωμένους χρήστες είναι το βήμα επιλογής προσωπικών λέξεων κλειδιών αλλά και προσωπικών RSS feed για παρακολούθηση μέσα από το σύστημα peRSSonal
reffig:perssonal4. Ο χρήστης του συστήματος μπορεί να δώσει στοιχεία που αφορούν κάθε μία κατηγορία από αυτές στις οποίες έχει δείξει ενδιαφέρον (θετικό ή μηδέν) και τα στοιχεία αυτά έχουν να κάνουν με λέξεις κλειδιά ή με RSS feeds. Μάλιστα επιτρέπεται στους χρήστες να επιλέξουν αν επιθυμούν να παρακολουθούν αποκλειστικά και μόνο άρθρα από τα RSS feeds που έχουν εισάγει οι ίδιοι κάτι το οποίο γενικά δεν προτείνεται από το σύστημα καθότι είναι σχεδιασμένο με τέτοιο τρόπο ώστε να μπορεί να ξεχωρίσει τις ειδήσεις που αφορούν το χρήστη.

6.6.1 Η πρώτη σελίδα του χρήστη

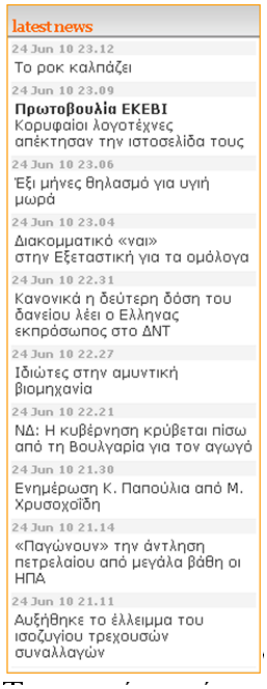
Μετά την εγγραφή μας στο σύστημα αξίζει να ρίξουμε μία ματιά στην πρώτη σελίδα που μας εμφανίζει το σύστημα όταν πραγματοποιούμε είσοδο 6.25.

Το σύστημα είναι σχετικά απλό και προσπαθεί να παρέχει όσο το δυνατόν λιγότερη πληροφορία στους χρήστες. Έτσι μπορούμε να διακρίνουμε ένα βασικό μενού με όλα τα απαραίτητα στοιχεία που χρειάζεται ο χρήστης για να πραγματοποιήσει περιαγωγή σε όλα τα στοιχεία που τον αφορούν και είναι: Τα άρθρα που έχει διαβάσει, όλα τα τελευταία άρθρα, και τα άρθρα ανά κατηγορία6.26.



Σχήμα 6.25: peRSSonal meta-portal - Αρχική Σελίδα Χρήστη

Σχήμα 6.26: peRSSonal meta-portal - Βασικό Μενού Χρήστη



Σχήμα 6.27: peRSSonal meta-portal - Δευτερεύον αριστερό μενού

Ο χρήστης μπορεί να επιλέξει να δει όλα τα άρθρα που τον ενδιαφέρουν σύμφωνα με τις επιλογές που έχει κάνει το σύστημα, μπορεί να επιλέξει να δει όλα τα άρθρα που έχει επιλέξει προς ανάγνωση και ακόμα να δει μεμονωμένα κάθε κατηγορία από αυτές που έχει επιλέξει να παρακολουθεί. Είναι σημαντικό σε αυτό το σημείο να τονίσουμε πως αρχικά το να βλέπει ο χρήστης όλα τα άρθρα μπορεί να είναι λίγο επικίνδυνο ειδικά αν αυτά είναι πολλά και ειδικά αν έχει κάνει ακραίες επιλογές κατά τη διάρκεια δημιουργίας του προφίλ. Αυτό συμβαίνει διότι οι πρώτες σελίδες με άρθρα θα περιέχουν αποκλειστικά και μόνο άρθρα τα οποία αφορούν βασικά τις κατηγορίες τις οποίες έχει βαθμολογήσει ο χρήστης με πολύ μεγάλο βαθμό.
Στη συνέχεια και συνεχίζοντας στο αριστερό μενού ο χρήστης έρχεται σε επαφή με πληροφορία η οποία αλλάζει σε πραγματικό χρόνο συνολικά για το δικτυακό τόπο και είναι ένα menu το οποίο είναι κοινό για όλους τους χρήστες του συστήματος. Πρόκειται για τη ροή εισροής ειδήσεων στο σύστημα και υπάρχει για να προσφέρει πληρότητα. Μάλιστα, υπάρχει εκεί γιατί πολλές φορές, αρκετοί χρήστες δείχνουν ιδιαίτερο ενδιαφέρον στη ροή των ειδήσεων έτσι όπως πραγματοποιείται σε ένα δικτυακό τόπο 6.27. Το μενού αυτό προσφέρει απλώς άμεση πρόσβαση σε άμεση πληροφορία. Βλέποντας τα μενού που έχει η αρχική σελίδα θα περάσουμε να εξετάσουμε τα μενού που βρίσκονται στο δεξί μέρος της σελίδας που προβάλλεται στο χρήστη. Το μενού αυτό παρέχει πληροφορίες για τα πιο πρόσφατα άρθρα που έχει αναγνώσει ένας χρήστης προκειμένου να επανέλθει σε αυτά άμεσα χωρίς να χρειαστεί να περάσει από τη σελίδα “My

Σχήμα 6.28: peRSSonal meta-portal - Κεντρική σελίδα | Στο σχήμα είναι σημειωμένες ενότητες για καλύτερη κατανόηση

Articles". Εν συνεχεία και αμέσως κάτω από τις επιλογές που έχει κάνει ο χρήστης μπορούμε να δούμε μία λίστα με τις επιλογές όλων των χρηστών. Πρόκειται στην ουσία για τα πιο πρόσφατα άρθρα που έχουν τα περισσότερα hit στο σύστημα. Ακολουθεί ένα μενού το οποίο παρέχει την υπηρεσία αναζήτησης και η δεξιά πλευρά της σελίδας κλείνει με μια υπηρεσία που ονομάζεται reader's pick και παρουσιάζει τα άρθρα που έχουν αναγνωσθεί περισσότερο υπό την έννοια του χρόνου που έχουν ξοδέψει οι χρήστες σε κάθε άρθρο. Μάλιστα, προκειμένου να γίνεται πιο σωστός υπολογισμός αυτής της μετρικής λαμβάνουμε υπόψη μας το μέγεθος κάθε άρθρου το οποίο βλέπει ο εκάστοτε χρήστης.

Έχοντας ολοκληρώσει από τα διαφορετικά μενού τα οποία βλέπει ο χρήστης αξίζει να περάσουμε στο κεντρικό μέρος της αρχικής σελίδας του χρήστη. Το κεντρικό μέρος χωρίζεται σε τρία διαφορετικά οριζόντια μέρη και τρία κάθετα 6.28. Τα τρία οριζόντια κομμάτια του μηχανισμού χωρίζουν το κεντρικό μέρος στο πάνω μέρος (A) το οποίο περιέχει τη σελιδοποίηση, το

κεντρικό κομμάτι (B) που έχει τις ουσιαστικές πληροφορίες και το κάτω κομμάτι (C) το οποίο έχει πληροφορίες που σχετίζονται με τον αριθμό των άρθρων που έχει φέρει το σύστημα για να προβληθούν στο χρήστη καθώς και η επιλογή για το χρήστη αν επιθυμεί να βλέπει τα άρθρα βάσει συσχέτισης ή βάση χρόνου.

Το κεντρικό κομμάτι είναι χωρισμένο σαφώς σε τρεις ενότητες οι οποίες είναι και χρωματισμένες κατάλληλα. Η πρώτη στήλη (1) δείχνει την κατηγορία στην οποία ανήκει το άρθρο. Η κατηγορία αυτή δεν έχει καμία απολύτως σχέση με την κατηγοριοποίηση που κάνει ο μηχανισμός αλλά προέρχεται απ' ευθείας από την κατηγορία στην οποία ανήκει το άρθρο από το δικτυακό τόπο από τον οποίο προέρχεται. Στη συνέχεια, η στήλη (2) μας παρέχει τον τίτλο του άρθρου έτσι όπως αυτός έχει εξαχθεί από τα RSS feeds και έχει εγγραφεί στη βάση δεδομένων. Τέλος, η στήλη (3) μας παρέχει στοιχεία για την ημερομηνία και ώρα του άρθρου και μας παρέχει την πρώτη βασική δυνατότητα αλληλεπίδρασης με το σύστημα και διαμόρφωσης του προφίλ. Το X που φαίνεται στο δεξί μέρος της στήλης χρησιμοποιείται από το χρήστη εφόσον αυτός δεν ενδιαφέρεται για ένα άρθρο και θέλει με λίγα λόγια να το ξεφορτωθεί. Στην ουσία, ο χρήστης μπορεί να επιλέξει να διαβάσει ένα άρθρο εφόσον του αρέσει ο τίτλος ή να προσθέσει ένα άρθρο στη μαύρη λίστα.

Ας ξεκινήσουμε τη διαδικασία προσπαθώντας να ανακαλύψουμε τι ακριβώς συμβαίνει όταν ένας χρήστης προσπαθεί να διαβάσει ένα άρθρο. Όταν ο χρήστης εντοπίσει ένα άρθρο το οποίο τον ενδιαφέρει αρκεί να πατήσει πάνω στον τίτλο του για να το διαβάσει. Ας δοκιμάσουμε λοιπόν όπως δείχνει και η παραπάνω εικόνα να ανοίξουμε το πρώτο άρθρο με τίτλο: «ανακάλυψαν νέο πρόγονο του ανθρώπου» της κατηγορίας Τεχνολογία 6.29. Όπως βλέπουμε η σελίδα μοιάζει αρκετά με αυτή της αρχικής σελίδας, διαθέτουμε δηλαδή ένα αριστερό και ένα δεξί μενού και ένα κεντρικό κομμάτι όπου βρίσκεται η είδηση. Ωστόσο, αξίζει να δούμε τα νέα χαρακτηριστικά που έχουμε σε αυτή τη σελίδα.

Σε αυτό το σημείο αξίζει να ξεκινήσουμε τη διαδικασία ανάλυσης από το κεντρικό μέρος, το οποίο χωρίζεται και πάλι σε οριζόντια κομμάτια. Στο πάνω μέρος βλέπουμε τον τίτλο του άρθρου, στη συνέχεια ακολουθούν τρεις ενδείξεις για το αν θέλουμε να δούμε την περίληψη του άρθρου, το πλήρες σώμα του άρθρου και τις εικόνες που σχετίζονται με το άρθρο, ακολουθεί το σώμα του άρθρου με ότι εικόνες έχουν εξαχθεί γι αυτό και το κάτω μέρος της σελίδας περιλαμβάνει το link από τη σελίδα από την οποία προέρχεται το άρθρο. Εξαιρετικά απλό στη χρήση και περιεκτικό όπως θα το ήθελε κάθε χρήστης. Σε αυτή τη σελίδα όμως παρατηρούμε διαφορές και στα μενού τα οποία βρίσκονται εκατέρωθεν του άρθρου.

Το αριστερό μέρος διαθέτει ένα διαφορετικό δεύτερο μενού που σχετίζεται με το tagging που έχει γίνει στο άρθρο. Για την ακρίβεια πρόκειται για interactive tagging καθότι καλούμε το χρήστη να επιλέξει, εφόσον επιθυμεί, από τα tags που βλέπει αυτά που του φαίνονται ενδιαφέροντα. Η επιλογή μπορεί να γίνει με τις ενδείξεις – και + που υπάρχουν όπως φαίνεται και στην εικόνα. Όπως μπορούμε να δούμε το tagging που γίνεται περιλαμβάνει stemmed λέξεις κλειδιά, χωρίς αυτό να εμποδίζει κάποιον να επιλέξει κάποια λέξη και να τη χρησιμοποιήσει. Έτσι, αν ένας χρήστης θεωρήσει πως η λέξη «απολίθωμ» που φυσικά είναι η stemmed λέξη κλειδί για τα απο-

Σχήμα 6.29: peRSSonal meta-portal - Σελίδα Ανάγνωσης Άρθρου

Σχήμα 6.30: peRSSonal meta-portal - Tagging Άρθρου

Σχήμα 6.31: peRSSonal meta-portal - Συναφή Άρθρα ενός Άρθρου

λιθώματα, το απολιθωμένος, κ.α. αρκεί να πατήσει το σημείο + που βρίσκεται δίπλα στη λέξη. Περνώντας στη δεξιά μεριά βρίσκεται ίσως ένα από τα πιο σημαντικά κομμάτια του συστήματος και δεν είναι άλλο από το μηχανισμό εύρεσης παραπλήσιων άρθρων και δημιουργίας ομάδων άρθρων. Το μενού αυτό πραγματοποιεί σε πραγματικό χρόνο ανάλυση του άρθρου που διαβάζει ο χρήστης και προσπαθεί αν εντοπίσει άρθρα που μπορεί να σχετίζονται με αυτό. Όπως μπορούμε να δούμε για το συγκεκριμένο άρθρο ο μηχανισμός έχει να μας προτείνει άρθρα τα οποία όπως μπορούμε να καταλάβουμε από τους τίτλους έχουν άμεση συσχέτιση με το άρθρο που διαβάζουμε. Προφανώς, εκτός από παραπλήσια άρθρα, θα αποτελέσουν κομμάτι της ομάδας άρθρων που θα δημιουργηθεί 6.31.

Τα άρθρα που ήδη έχουμε διαβάσει φεύγουν από την αρχική σελίδα κάθε φορά που την επισκεπτόμαστε και βρίσκονται στις σελίδες με τα άρθρα του χρήστη (My Articles), ενώ όλες οι ενέργειες που πραγματοποιεί ο χρήστης κατά τη διάρκεια μίας συνεδρίας καταγράφονται και μπορούν ακόμα και σε πραγματικό χρόνο να αλλάξουν τις προτιμήσεις που έχει ένας χρήστης.
Είδαμε τις πιο βασικές διαδικασίες του συστήματος peRSSonal. Όπως είδαμε, προσπαθούμε να παρέχουμε στους χρήστες όσο το δυνατόν πιο απλές υπηρεσίες γίνεται προκειμένου οι χρήστες να επικεντρώνεται στην ανάγνωση ειδήσεων και να σπαταλούν το λιγότερο δυνατό χρόνο. Στη συνέχεια θα προχωρήσουμε σε ανάλυση της λειτουργίας του προφίλ του χρήστη και στα πειραματικά αποτελέσματα που διαθέτουμε από την ανάλυση του προφίλ του χρήστη.

6.6.2 Προσαρμογή στο προφίλ του χρήστη

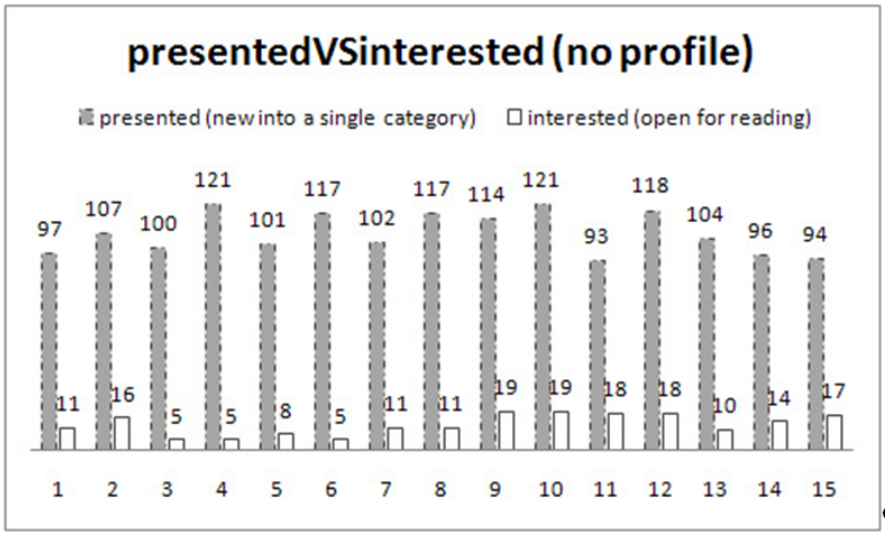
Ένα πολύ σημαντικό πείραμα που πρέπει να γίνει για την ανάλυση του μηχανισμού είναι η δυνατότητα του μηχανισμού να προσαρμόζεται στο προφίλ του χρήστη και να το προσεγγίζει με όσο το δυνατόν μεγαλύτερη ταχύτητα και ακρίβεια. Στην ουσία αυτό που θέλουμε να ελέγξουμε είναι δύο στοιχεία που αφορούν την προσωποποίηση στο χρήστη. Αφενός έχει αξία να εντοπίσουμε το compression που πετυχαίνουμε στην πληροφορία και από την άλλη να δούμε

αν με αυτή τη συμπύκνωση πληροφορίας μπορούμε να έχουμε ακόμα θετικά αποτελέσματα για το χρήστη. Ας δούμε όμως λίγο πιο πρακτικά το συγκεκριμένο θέμα. Όταν το σύστημα λειτουργεί χωρίς στοιχεία προσωποποίησης τότε περιμένουμε να εμφανίζει έναν αριθμό από Χ άρθρα στο χρήστη. Ο χρήστης κοιτώντας αυτά τα άρθρα ένα προς ένα (τους τίτλους) θα ενδιαφερθεί για Ψ από αυτά τα οποία και θα διαβάσει. Στην ιδανική περίπτωση θα θέλαμε αυτά τα άρθρα να είναι και τα Ψ πρώτα που του εμφανίζονται. Σε ακόμα πιο ιδανική περίπτωση θα θέλαμε να έχουμε εμφανίσει μόνο Ψ. Με το σύστημα προσωποποίησης που χρησιμοποιούμε εμφανίζουμε στο χρήστη Χ' άρθρα, όπου Χ'<=Χ και Χ' υποσύνολο του Χ. Ο χρήστης από αυτά τα άρθρα επιλέγει να διαβάσει τα Ψ', όπου όπως αποδεικνύεται από την πειραματική διαδικασία δε μπορεί να δοθεί μία σχέση ανάμεσα σε Ψ' και Ψ. Αυτό που θα θέλαμε στην ιδανική περίπτωση είναι η ταύτιση των Χ', Ψ' και Ψ. Αυτό φυσικά ισχύει στην περίπτωση που υπάρχει απόλυτη συσχέτιση των Ψ' και Ψ. Χρησιμοποιώντας αμιγώς μαθηματικούς τύπους θα μπορούσαμε να πούμε πως το Ψ' είναι υποσύνολο του Ψ (ή και το ίδιο το Ψ). ωστόσο, η πρακτική μας έδειξε πως η συμπεριφορά του χρήστη αλλάζει ανάλογα με το περιβάλλον. Έτσι, εμφανίζονται και περιπτώσεις όπου το Ψ' είναι διαφορετικό του Ψ.

Ας περάσουμε όμως στην πράξη για να δούμε τι μας δείχνουν τα πειραματικά αποτελέσματα από τη λειτουργία του συστήματός μας. Αρχικά μετράμε για μία σειρά από άρθρα και για συγκεκριμένους χρήστες τα άρθρα τα οποία τους παρουσιάζονται χωρίς προσωποποίηση και αυτά που επιλέγουν να διαβάζουν. Στη συνέχεια για τους ίδιους χρήστες βλέπουμε το σύνολο των άρθρων που τους παρουσιάζει το σύστημα όταν εφαρμόζει προσωποποίηση και ποια από αυτά επιλέγουν να διαβάσουν.

Στο σχήμα 6.32 βλέπουμε τα άρθρα που διαβάζουν οι χρήστες συγκριτικά με αυτά που τους παρουσιάζονται όταν δεν εφαρμόζουμε κανένα αλγόριθμο προσωποποίησης.

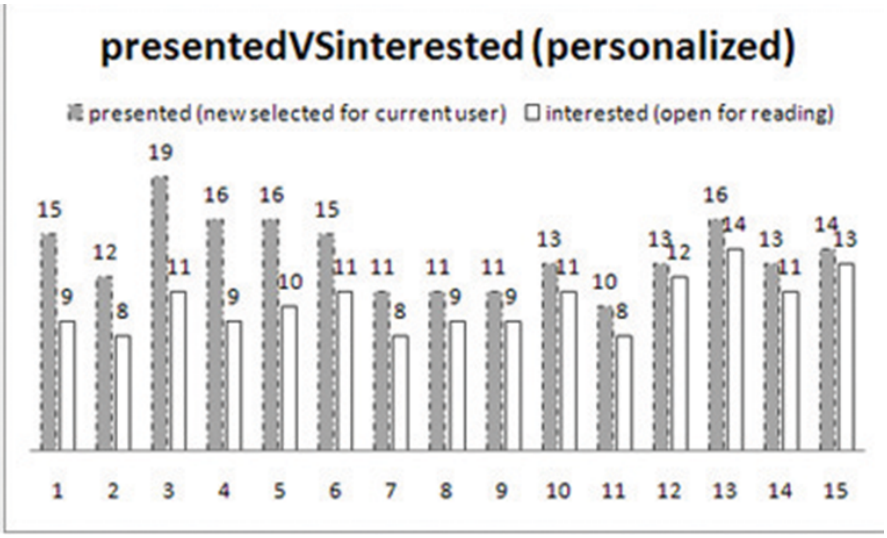
Για να γίνει πιο σαφές, το σχήμα 6.32 παρουσιάζει το σύνολο των άρθρων μίας κατηγορίας που προέκυψαν μία ημέρα και παρουσιάστηκαν στους χρήστες και παράλληλα βλέπουμε πόσα από αυτά τα άρθρα επέλεξε ο χρήστης να διαβάσει. Συνεπώς βλέπουμε τις τιμές του Χ και Ψ για επιλεγμένες μέρες και όπως μπορούμε να δούμε ο Μ.Ο.(Χ) = 103 ενώ ο Μ.Ο.(Ψ) = 11. Στην ουσία, βλέπουμε πως ο χρήστης επιλέγει για να διαβάσει περίπου 10% των άρθρων που του εμφανίζονται. Το σημαντικό σε αυτό το σημείο είναι το γεγονός που μας προξενεί πολύ μεγάλη εντύπωση και είναι το εξής: ο ίδιος χρήστης όταν του εμφανίσαμε τα ίδια άρθρα έπειτα από κάποια ώρα επέλεξε να διαβάσει διαφορετικό αριθμό άρθρων. Ένα πρώτο συμπέρασμα στο οποίο καταλήγουμε σε αυτό το σημείο είναι πως ο χρήστης αδυνατεί να διακρίνει πλήρως την πληροφορία που τον ενδιαφέρει όταν αυτή βρίσκεται ανάμεσα σε πληθώρα άλλων δεδομένων που τον ενδιαφέρουν. Η απάντηση των χρηστών στην ερώτηση γιατί την δεύτερη ή τρίτη φορά δεν επέλεξαν το ίδια άρθρα ήταν πως πολύ απλά δεν τα είχαν προσέξει ανάμεσα στα τόσα πολλά που είχαν να διαβάσουν. Αυτό το συμπέρασμα θα μας φανεί πολύ χρήσιμο όταν στη συνέχεια θα εντοπίσουμε αυτό που ήδη αναφέραμε και τη διαφορά που παρουσιάζουν τα Ψ και Ψ'. όπως



Σχήμα 6.32: Σύγκριση άρθρων που παρουσιάστηκαν με τα άρθρα για τα οποία παρουσιάστηκε ενδιαφέρον - χωρίς προφίλ χρήστη

είναι προφανές από αυτά που έχουμε αναφέρει πρόκειται για μία διαφορά η οποίο οφείλεται αποκλειστικά και μόνο στο γεγονός πως στην πρώτη περίπτωση η πληθώρα πληροφορίας απο-προσανατολίζει τους χρήστες. Ας περάσουμε όμως να δούμε τι ακριβώς αλλάζει από τη μεριά του χρήστη όταν έχουμε προσωποποίηση της πληροφορίας. Το σχήμα 6.33 μας δίνει πληροφορίες για τον αριθμό των άρθρων που παρουσιάζονται σε ένα χρήστη σε ημερήσια βάση και για μία και μόνο κατηγορία.

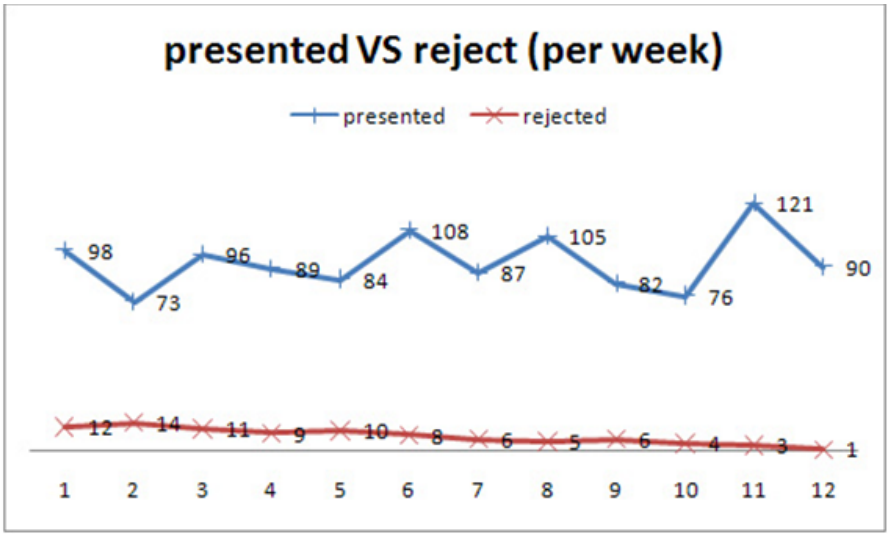
Όπως μπορούμε να διακρίνουμε σε αυτή την περίπτωση έχουμε πολύ λιγότερα άρθρα να παρουσιάζονται στους χρήστες και φυσικά αξίζει να αναφέρουμε πως τα άρθρα στα οποία έγινε η προσωποποίηση ταυτίζονται ένα προς ένα με το προηγούμενο σχήμα. Βλέπουμε λοιπόν πως σε αυτή την περίπτωση παρουσιάζονται στο χρήστη πολύ λιγότερα άρθρα Χ' από τα οποία επιλέγει να αναγνώσει Ψ'. Όπως είναι σαφές και από το σχήμα, ο Μ.Ο.(Χ') είναι περίπου 11 ενώ ο Μ.Ο.(Ψ') είναι περίπου 10. Σε αυτή την περίπτωση είναι ενδιαφέρον να διακρίνουμε δύο σημαντικά στοιχεία. Αρχικά αξίζει να δούμε τι ποσοστό compression επιτύχαμε, δηλαδή σε τι ποσοστό καταφέραμε να μειώσουμε την πληροφορία που παρουσιάζουμε στο χρήστη. Είναι σαφές πως η μείωση αγγίζει το επίπεδο του 90%, ένα νούμερα πραγματικά εντυπωσιακό. Αν το δούμε όμως σε επίπεδο portal θα καταλάβουμε πως αυτός ο αριθμός αντικατοπτρίζει μία πραγματικότητα του διαδικτύου. Όταν περιηγούμαστε σε μία ιστοσελίδα και ειδικά σε ένα portal μπορούμε εύκολα να διαπιστώσουμε πως ένα τεράστιο κομμάτι της πληροφορίας που μας εμφανίζεται μας είναι πραγματικά αδιάφορο έως και άχρηστο. Γι' αυτό και θεωρούμε αυτό το ποσοστό μείωσης της πληροφορίας εξαιρετικά εντυπωσιακό. Τι επίπτωση έχει όμως η αλλαγή αυτή στην προβολή πληροφορίας για ένα χρήστη του συστήματος; Η μείωση δε σημαίνει απαραίτητα πως η ζωή του



Σχήμα 6.33: Σύγκριση άρθρων που παρουσιάστηκαν με τα άρθρα για τα οποία παρουσιάστηκε ενδιαφέρον - με προφίλ χρήστη

χρήστη έγινε και πιο εύκολη. Ωστόσο, όπως μπορούμε να διακρίνουμε η τιμή του Ψ' κινείται στα ίδια επίπεδα με προηγούμενα. Σε αυτό το σημείο είναι σημαντικό να δούμε το ποσοστό επικάλυψης του Ψ' με το Ψ. Στην ιδανική περίπτωση όπως έχουμε ήδη αναφέρει αναμένουμε το Ψ να ταυτίζεται με το Ψ' και ακόμα πιο ιδανικά να ταυτίζεται και με το Χ'. ωστόσο, όπως ήδη αναφέρθηκε βλέπουμε ένα πραγματικά παράδοξο φαινόμενο. Όσο κι αν απαιτούμε από το χρήστη στην πρώτη περίπτωση να προσπαθήσει να επιλέξει πραγματικά όλα τα αρθρα τα οποία θα διάβαζε σε περίπτωση που επισκεπτόταν πραγματικά έναν ειδησεογραφικό δικτυακό τόπο παρατηρούμε πως συχνά το Ψ' περιέχει άρθρα που ενώ υπάρχουν στο Χ δεν υπάρχουν μέσα στο σύνολο Ψ. Αρκεί λοιπόν να τονίσουμε πως παρατηρούμε επικάλυψη του Ψ' με το Ψ μεγαλύτερη από 95% αλλά παρατηρούμε επιπλέον τη συμπεριφορά του χρήστη να μεταβάλλεται όταν του παρουσιάζουμε λίγα σε αριθμό άρθρα. Δε θα έλεγε κανείς πως αυτό είναι εντελώς παράδοξο ωστόσο δε μας επιτρέπει να κάνουμε απόλυτη σύγκριση μεταξύ των συνόλων Ψ' και Ψ που θα μας έδινε σαφή αριθμητική απάντηση για την προσωποποίηση. Παρά το γεγονός λοιπόν ότι δεν έχουμε σαφή απάντηση πάνω σε αυτό, οι ποιοτικές μετρήσεις μας δείχνουν θεαματικά αποτελέσματα για τον τρόπο λειτουργίας της κατηγοριοποίησης και την επιλογή των άρθρων που εμφανίζονται στον τελικό χρήστη.

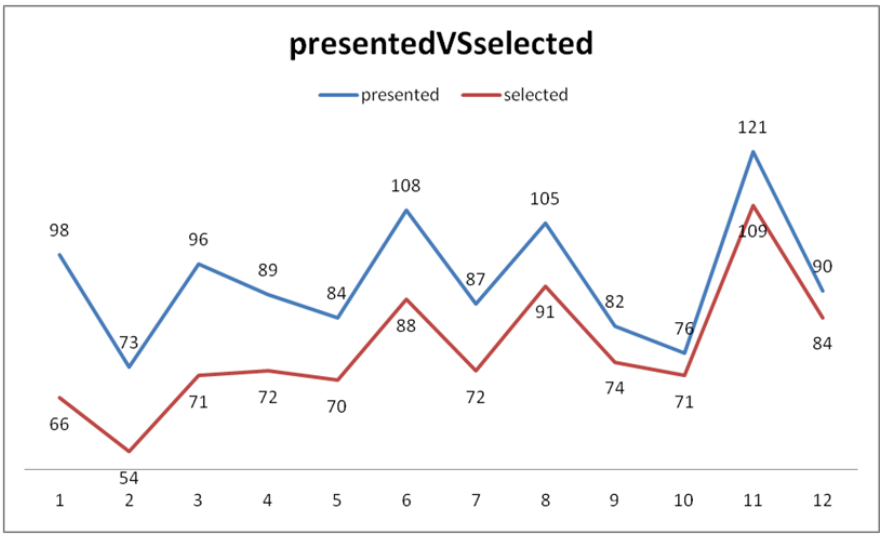
Επιπλέον μετρήσεις έχουν γίνει στην προσπάθεια να δούμε τον ακριβή τρόπο λειτουργίας της κατηγοριοποίησης και η αλήθεια είναι πως οι παραπάνω μετρήσεις έγιναν αφού είχαμε εξασφαλίσει πως το προφίλ του χρήστη έχει διαμορφωθεί επαρκώς. Τι σημαίνει όμως διαμόρφωση του προφίλ του χρήστη και πως συντελείται αυτό. Όπως έχουμε ήδη δει, ο χρήστης έχει τη δυνατότητα να τοποθετήσει κάποια άρθρα σε αυτό που ονομάζουμε blacklist, στην ουσία δηλαδή να



Σχήμα 6.34: Σύγκριση άρθρων που παρουσιάστηκαν με τα άρθρα τα οποία ο χρήστης απέρριψε (εβδομαδιαία)

απορρίψει ένα άρθρο το οποίο δε θεωρεί καθόλου σχετικό με το προσωπικό του προφίλ. Αυτή η ενέργεια των χρηστών μπορεί να μας δώσει στοιχεία για τη διαμόρφωση του προφίλ του χρήστη μέσα στο χρόνο. Όπως είναι λογικό, όσο περισσότερα άρθρα απορρίπτει ένας χρήστης τόσο πιο ατελές θεωρείται το προφίλ του. Η μετρική που παρουσιάζουμε στο επόμενο σχήμα μας δίνει πληροφορία σχετικά με αυτό. Το σχήμα παρουσιάζει τον αριθμό των άρθρων που παρουσιάζονται σε ένα χρήστη μέσα σε μία βδομάδα και αφορούν μία κατηγορία και τον αριθμό από αυτά που ο χρήστης αποφάσισε να απορρίψει. Για το συγκεκριμένο χρήστη έχουμε θεωρήσει απλή χρήση του δικτυακού τόπου, δηλαδή, δύο επισκέψεις στο δικτυακό τόπο τυχαίες ώρες της ημέρας οι οποίες περιλαμβάνουν πλήρη ανάγνωση του περιεχομένου που υπάρχει τη στιγμή που ο χρήστης εισέρχεται 6.34.

Όπως μπορούμε να δούμε από το σχήμα 6.34 οι μετρήσεις μας έγιναν για ένα διάστημα 12 εβδομάδων. Σε αυτές τις 12 εβδομάδες ο χρήστης εκτός των άλλων ενεργειών που έκανε στο δικτυακό τόπο παράλληλα προχωρούσε σε απόρριψη των άρθρων τα οποία δεν τον ενδιέφεραν. Όπως μπορούμε να δούμε κατά το πέρασμα των εβδομάδων ο αριθμός των άρθρων τα οποία ο χρήστης απέρριπτε μειωνόταν. Αυτό σημαίνει πως το σύστημα σταματά να του παρουσιάζει άρθρα τα οποία δεν τον ενδιαφέρουν χωρίς παράλληλα να μειώνεται ο συνολικός αριθμός των άρθρων που παρουσιάζεται στο χρήστη. Το παραπάνω σχήμα είναι μία πολύ σημαντική ένδειξη για το σύστημα προσωποποίησης. Αξίζει όμως να προχωρήσουμε και σε μία επιπλέον ανάλυση για να δούμε πάνω στα ίδια άρθρα και για τον ίδιο χρονικό ορίζοντα τι επιλογές έκανε ο χρήστης, πόσα άρθρα δηλαδή επέλεξε για να διαβάσει.

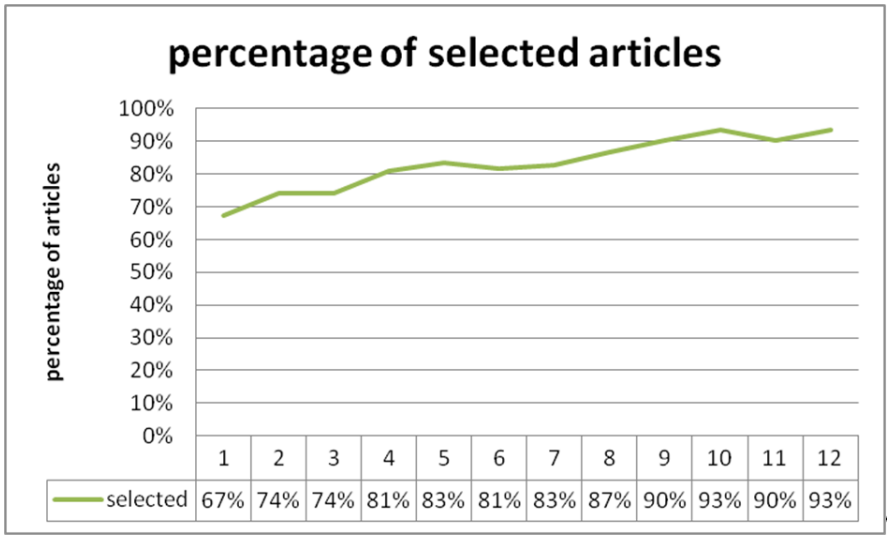


Σχήμα 6.35: Σύγκριση άρθρων που παρουσιάστηκαν με τα άρθρα τα οποία επιλέχθηκαν (εβδομαδιαία)

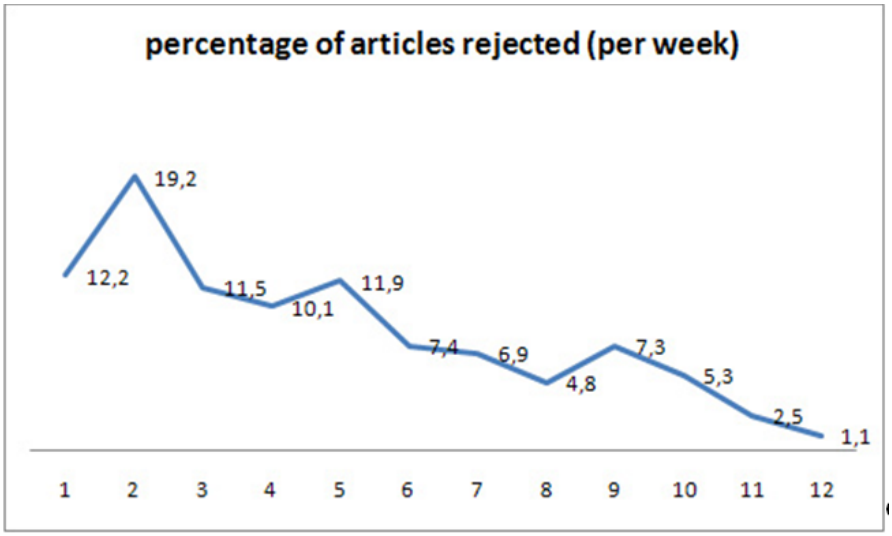
Όπως μπορούμε να δούμε και από το σχήμα 6.35 είναι πολύ σημαντικό το γεγονός ότι ο χρήστης ενδιαφέρεται για όλο και περισσότερα άρθρα από αυτά που του εμφανίζονται και στο επόμενο σχήμα φαίνεται πιο καθαρά καθότι απεικονίζεται το ποσοστό των άρθρων που του εμφανίζονται και ο χρήστης επιλέγει να διαβάσει.

Από το σχήμα 6.36 γίνεται σαφές πως το ποσοστό είναι από την αρχή σε πολύ υψηλά επίπεδα καθότι μπορεί άμεσα σε λίγες μέρες λειτουργίας να δημιουργήσει ένα σεβαστό προφίλ, ωστόσο βλέπουμε πως με την πάροδο του χρόνου μπορεί να φτάσει σε επίπεδα πάνω από 90%. Είναι εξίσου σημαντικό να δούμε παράλληλα και το ποσοστό των άρθρων που απορρίπτονται από το χρήστη, τα άρθρα δηλαδή τα οποία επιλέγει να τοποθετήσει στη μαύρη λίστα.

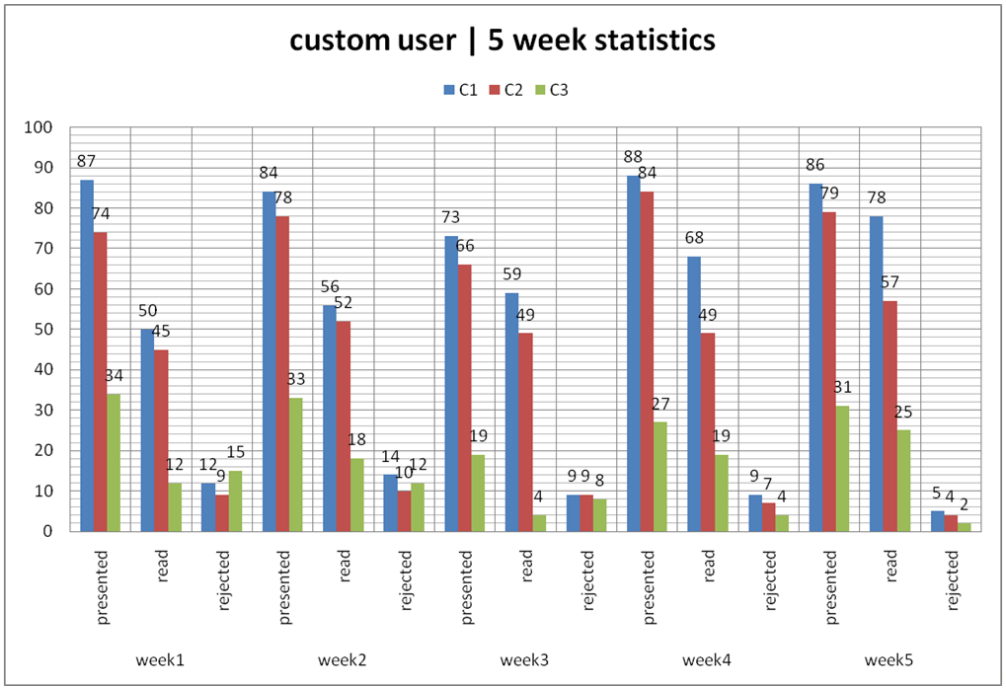
Παρατηρούμε στο σχήμα 6.37 πως ενώ αρχικά ο χρήστης μπορεί να φτάσει σε σημείο να απορρίπτει ακόμα και 20% των άρθρων τα οποία του εμφανίζονται μετά από κάποιο διάστημα δεν έχει πλέον στις σελίδες του σχεδόν καθόλου ανεπιθύμητα άρθρα ή καλύτερα άρθρα τα οποία δε σχετίζονται με το προφίλ του. Φυσικά, αυτό το ποσοστό λάθους της τάξης του 1% το οποίο παρατηρούμε για το μηχανισμό μας θεωρείται εξαιρετικά μικρό ειδικά αν το συγκρίνουμε με το ποσοστό επιλεγμένων άρθρων που μπορεί να ξεπεράσει το 90%.
Σε αυτή την ενότητα είδαμε τα πειράματα που έγιναν πάνω στο μηχανισμό και σχετίζονται με την προσωποποίηση που πραγματοποιεί το σύστημα. Διαπιστώσαμε πως το σύστημα αρχικά μπορεί να δώσει αρκετά μεγάλο όγκο πληροφορίας η οποία δε σχετίζεται με το προφίλ του χρήστη. Είναι αναμενόμενο αν δεν έχουμε πληροφορίες για το χρήστη να μην είμαστε σε θέση να εντοπίσουμε όλη εκείνη την πληροφορία που ενδεχόμενο τον ενδιαφέρει σε απόλυτο βαθμό. Ωστόσο,



Σχήμα 6.36: Ποσοστό επιλεγμένων άρθρων (συγκριτικά με τον αριθμό όσων εμφανίστηκαν)



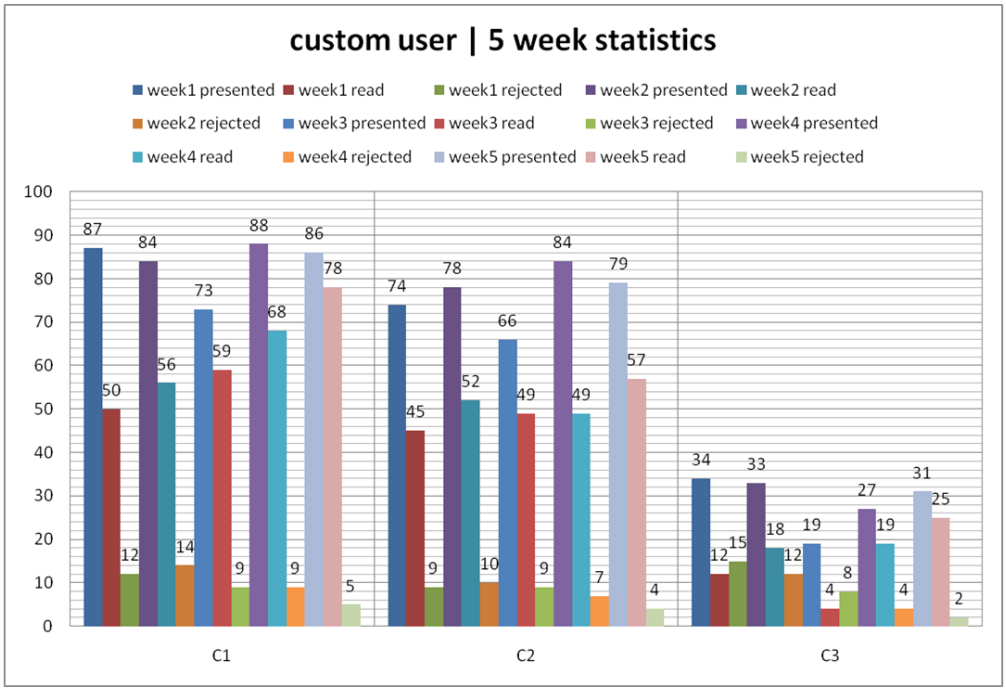
Σχήμα 6.37: Ποσοστό blacklisted άρθρων (συγκριτικά με τον αριθμό όσων εμφανίστηκαν)



Σχήμα 6.38: Στατιστικά Χρήστη για λειτουργία 5 εβδομάδων

διαπιστώσαμε πως σε μικρό μόνο χρονικό διάστημα ο μηχανισμός μπορεί να λειτουργήσει με τέτοιο τρόπο που να προσφέρει αποκλειστικά και μόνο πληροφορία που ενδιαφέρει το χρήστη. Για να δούμε αυτό το πράγμα σε απόλυτους αριθμούς αρκεί να σκεφτούμε πως αν ένας χρήστης παρακολουθεί τη σελίδα με τα νέα που το εμφανίζονται σε καθημερινή βάση, έχει επιλέξει να παρακολουθεί 3 κατηγορίες από τις 7 του συστήματος και έχει φτάσει σε ένα σεβαστό σημείο διαμόρφωσης προφίλ τότε μέσα σε μία βδομάδα του παρουσιάζονται:
Περίπου 300 άρθρα από όλες τις κατηγορίες. 270 άρθρα αποφασίζει να διαβάσει. 3 άρθρα θεωρεί πως δεν ταιριάζουν στο προφίλ του. Τα νούμερα είναι κάτι περισσότερο από πειστικά. Για την ακρίβεια το επόμενο σχήμα 6.38 παρουσιάζει αναλυτικά στατιστικά στοιχεία τα οποία έχουμε συγκεντρώσει από την προσωπική χρήση που έχει κάνει ο χρήστης που εμείς έχουμε δημιουργήσει για το μηχανισμό.

Από το διάγραμμα 6.38 είναι σαφές πως μέσα σε 5 εβδομάδες έχουμε καταφέρει να οριστικοποιήσουμε το προφίλ και να προσφέρουμε αρκετά καλά δεδομένα στο χρήστη. Ενδιαφέρον παρουσιάζει και η σύγκλιση που έχουν τόσο τα κείμενα που διαβάζει ο χρήστης όσο και αυτά που απορρίπτει όπως είναι εμφανές στο ακόλουθο σχήμα 6.39 που είναι ακριβώς ίδιο με το παραπάνω αλλά ανεστραμμένο. Έτσι λοιπόν βλέπουμε σε κάθε κατηγορία πως τα άρθρα που διαβάζει ο χρήστες τείνουν προς τα άρθρα που εμφανίζονται στο χρήστη ενώ τα άρθρα που απορρίπτει ο χρήστης σε κάθε κατηγορία τείνουν στο 0.



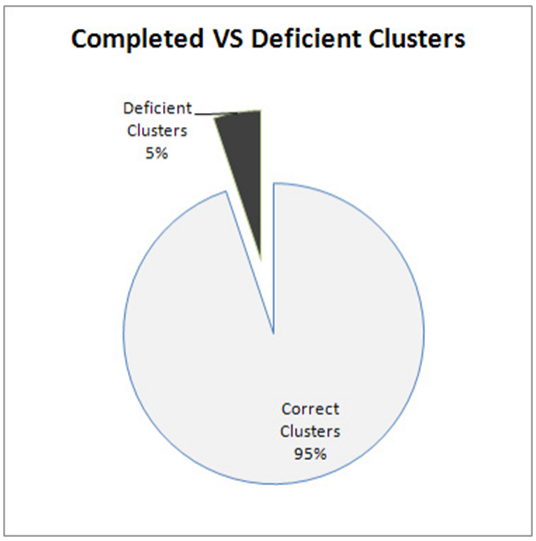
Σχήμα 6.39: Στατιστικά Χρήστη για λειτουργία 5 εβδομάδων (ανεστραμμένοι άξονες)

Ένα πολύ σημαντικό στοιχείο του μηχανισμού το οποίο παρουσιάζει εξίσου μεγάλο ενδιαφέρον με την προσωποποίηση στο χρήστη αποτελεί και η παρουσίαση των συναφών άρθρων και η δημιουργία ομάδων άρθρων μέσα από αυτή τη διαδικασία. Στην ουσία πρόκειται για ένα μηχανισμό ο οποίος είναι ζωτικής σημασίας για το σύστημα γιατί με τη βοήθειά του αποφεύγονται οι διπλές εμφανίσεις άρθρων προς το χρήστη και έτσι γίνεται καλύτερη η εμπειρία περιήγησης στο δικτυακό τόπο. Παράλληλα η διαδικασία βοηθά πολύ τους χρήστες να μπορούν να εντοπίσουν πολλές πηγές του ίδιου άρθρου και με αυτό τον τρόπο αφενός να μπορούν να βλέπουν πολλές διαφορετικές πτυχές του ίδιου άρθρου αλλά και πολλές οπτικές. Η διαδικασία αυτή γίνεται τόσο για κάθε χρήστη ξεχωριστά όσο και συνολικά για το σύστημα. Ο κάθε χρήστης που διαβάζει άρθρα έχει τη δυνατότητα να παράγει τις δικές του ομάδες άρθρων αλλά παράλληλα οι ομάδες δημιουργούνται και συνολικά για το δικτυακό τόπο για να μπορούν να είναι προσπελάσιμες από όλους τους χρήστες.

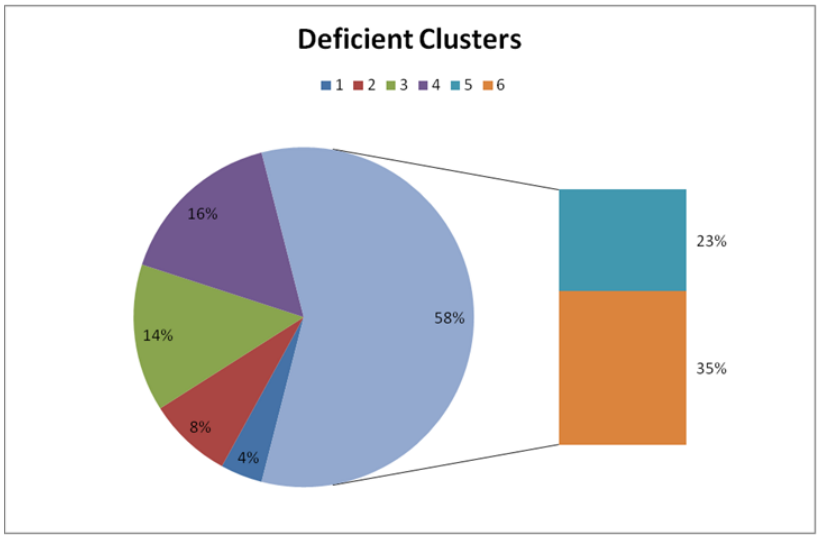
Η ομαδοποίηση των άρθρων του δικτυακού τόπου βασίζεται στη συσχέτιση συνημιτόνου και είναι μία εργασία η οποία πραγματοποιείται σε πραγματικό χρόνο. Είναι πολύ σημαντικό για το σύστημα να είναι σε θέση να πραγματοποιήσει αυτή τη διαδικασία real time upon demand διότι με αυτό τον τρόπο εξασφαλίζει ελάχιστη κατανάλωση πόρων. Αυτό επιτυγχάνεται διότι από τη μία δεν είναι σίγουρο ότι όλα τα άρθρα του συστήματος θα προσπελαστούν από τους χρήστες, ή έστω οι ίδιες ομάδες άρθρων. Αυτό σημαίνει πως αν κάνουμε τη διαδικασία από πριν μπορεί

να καταναλώσουμε πόρους χωρίς κανένα λόγο. Από την άλλη ελέγχεται το ενδεχόμενο η ομαδοποίηση άρθρων να δίνει καλύτερα αποτελέσματα αν εμπλέκει στις διαδικασίες της τον ίδιο το χρήστη. Στο επίπεδο παρουσίασης πληροφορίας είναι εύκολο να εμπλέξουμε το χρήστη στη διαδικασία γιατί γνωρίζουμε ποιος είναι ο χρήστης που παρακολουθεί το άρθρο ενώ είναι αδύνατο σε ένα προηγούμενο επίπεδο να προβλέψουμε με απόλυτη επιτυχία τα άρθρα που βλέπει ο χρήστης. Έτσι λοιπόν κρίνεται σκόπιμο να κάνουμε αυτή τη διαδικασία στο στάδιο της προβολής πληροφορίας για να είναι εφικτή και η εμπλοκή του χρήστη η οποία γίνεται με την άμεση συσχέτιση με την ομάδα άρθρων. Κάθε άρθρο που θεωρείται ότι έχει συνάφεια με το άρθρο που διαβάζει ο χρήστης και άρα μπορεί να αποτελέσει κομμάτι την ομάδας άρθρων ελέγχεται παράλληλα για συνάφεια με το χρήστη. Σε αυτό το σημείο κάνουμε τη θεώρηση πως ο χρήστης θα διαβάσει μόνο άρθρα που τον ενδιαφέρουν και συνεπώς αν κάποιο άρθρο δεν ενδιαφέρει το χρήστη δε μπορεί να βρεθεί στην ίδια θέση με άρθρα που τον ενδιαφέρουν.

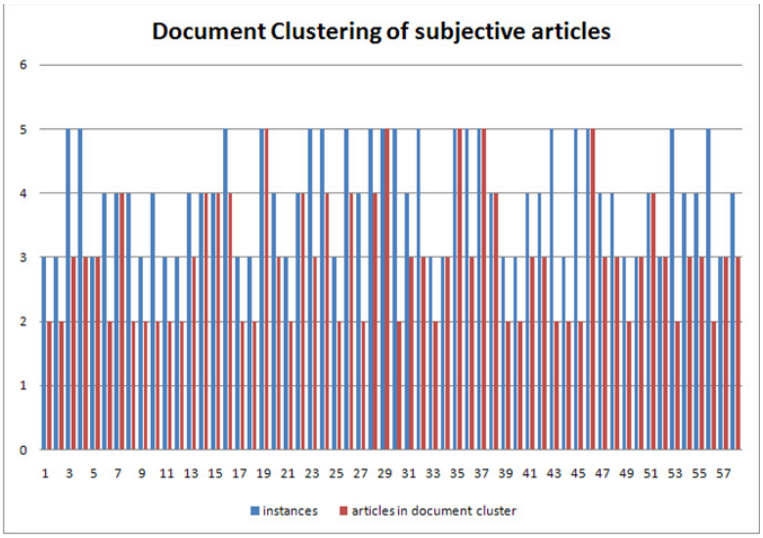
Για τις πειραματικές μας διαδικασίες πάνω στο θέμα των συναφών άρθρων και των ομάδων άρθρων κάναμε ένα απλό αλλά αποτελεσματικό πείραμα. Επιλέξαμε 7 διαφορετικούς δικτυακούς τόπους μεγάλων ειδησεογραφικών πρακτορείων και τους παρατηρούσαμε τακτικά προκειμένου να εντοπίσουμε ίδια άρθρα που δημοσιεύονται και στους 7. Μάλιστα προσπαθήσαμε να βρίσκουμε άρθρα που δεν αναφέρονται σε μία ανακοίνωση κάποιας είδησης αλλά ακόμα περισσότερο παρουσιάζουν με κριτική άποψη κάποιο συγκεκριμένο θέμα. Φυσικά, φροντίσαμε όλα τα άρθρα να έχουν χαρακτηριστικά για να θεωρούνται ακριβώς ίδια αλλά και να καλύπτουν τους βασικούς «χρονικούς» περιορισμούς που έχει το peRSSonal. Στη συνέχεια και αφού καταγράφομε τα άρθρα στους διαφορετικούς δικτυακούς τόπους και αφού ελέγχαμε ότι το σύστημά μας έχει εντοπίσει και κατεβάσει όλα τα άρθρα ανοίγαμε ένα από αυτά επιλεγμένο τυχαία και ελέγχαμε τα μέλη της ομάδας άρθρων που δημιουργούσε το peRSSonal. Το παρακάτω διάγραμμα δείχνει το ποσοστό επιτυχίας του μηχανισμού στον εντοπισμό των ίδιων στην ουσία άρθρων από διαφορετικές πηγές 6.40. Όπως μπορούμε να δούμε στο σχήμα 6.40 ο μηχανισμός καταφέρνει να εντοπίσει επιτυχώς ένα πολύ μεγάλο κομμάτι των συναφών άρθρων που δημοσιεύονται σε διαφορετικές πηγές. Σε αυτό το σημείο αξίζει να τονίσουμε πως η παρουσία σε 7 διαφορετικούς μεγάλους δικτυακούς τόπους δεν είναι πολύ συχνό φαινόμενο εκτός κι αν πρόκειται για κάποια πολύ ιδιαίτερη και μεγάλη είδηση. Παράλληλα θα πρέπει να δούμε και το φαινόμενο της δημοσίευσης συναφούς είδησης στο ίδιο domain το οποίο επίσης είναι πολύ συνηθισμένο και πραγματοποιείται πολύ συχνά κυρίως με δικτυακούς τόπους που αντί για να ενημερώνουν (update) ένα άρθρο δημοσιεύουν μία καινούρια είδηση απόλυτα σχετική με το αρχικό άρθρο. Σε αυτό το σημείο αξίζει να κάνουμε μία επιπλέον εξέταση στο σύστημά μας γιατί αν και το ποσοστό 95% θεωρείται αρκετά μεγάλο θα πρέπει να δούμε τι ακριβώς πηγαίνει λάθος με αυτό το 5% που είναι προβληματικές ομάδες άρθρων. Το επόμενο γράφημα 6.41 μας δείχνει πληροφορίες για τον αριθμό των άρθρων που βρέθηκαν στις ομάδες άρθρων, πόσα δηλαδή από τα 7 μπόρεσε να εντοπίσει ο μηχανισμός. Όπως μας δείχνει το σχήμα 6.41 η μεγάλη πλειοψηφία των ομάδων άρθρων περιέχουν 5 ή 6 άρθρα ενώ ένα 40% των ελλιπών ομάδων περιέχουν 4 ή λιγότερα άρθρα.



Σχήμα 6.40: Ολοκληρωμένες και ελλιπείς ομάδες άρθρων



Σχήμα 6.41: Ανάλυση Αριθμού Άρθρων ελλιπών ομάδων



Σχήμα 6.42: Ανάλυση Αριθμού Άρθρων ελλιπών ομάδων

Μάλιστα, ένα πολύ μικρό ποσοστό δεν περιέχει καθόλου παραπλήσια άρθρα παρά το γεγονός ότι υπάρχουν. Σε αυτό το σημείο σίγουρο μπορούμε να δούμε διάφορες αιτίες. Η πιο σημαντική είναι σίγουρα σφάλμα του μηχανισμού, ωστόσο μπορούμε να εντοπίσουμε και μία σειρά από άλλα στοιχεία τα οποία δημιουργούν προβλήματα στο μηχανισμό. Τέτοια μπορεί να είναι: το γεγονός ότι ο χρήστης δεν έχει σαφώς διαμορφωμένο προφίλ με αποτέλεσμα να επηρεάζει αρκετά το μηχανισμό όταν επιχειρεί τους αλγορίθμους συνάφειας με άρθρα. Μία άλλη αιτία είναι το γεγονός ότι πολλά άρθρα έχουν την ίδια θεματολογία αλλά εντελώς διαφορετική οπτική για το θέμα. Για να δούμε πόσο πολύ μπορεί να επηρεαστεί ο μηχανισμός από άρθρα που έχουν το ίδιο θέμα αλλά το παρουσιάζουν με διαφορετική οπτική προχωρήσαμε στο παρακάτω πείραμα. Εντοπίσαμε χειροκίνητα διάφορα άρθρα που δημοσιεύονταν σε δικτυακούς τόπους, είχαν κοινή θεματολογία, ικανοποιούσαν τα κριτήρια να ενταχθούν από το σύστημα στην ίδια ομάδα άρθρων αλλά στο περιεχόμενο ήταν σαφές ότι παρουσίαζαν διαφορετική οπτική για κάποιο θέμα. Αφού χρησιμοποιήσαμε για το σύστημα ένα χρήστη με διαμορφωμένο προφίλ προσπαθήσαμε να δούμε πως ακριβώς κατασκευάζονται οι ομάδες άρθρων γι αυτά τα άρθρα. Το σχήμα 6.42 παρουσιάζει πραγματικά πολύ μεγάλο ενδιαφέρον για τον τρόπο με τον οποίο το σύστημα αντιλαμβάνεται την έννοια των συναφών άρθρων διαφορετικής οπτικής. Αρχικά όπως είναι σαφές, αν και το σύστημα διαθέτει πάνω από μισό εκατομμύριο άρθρο δεν είναι εύκολη διαδικασία να εντοπιστούν άρθρα με τέτοιες ιδιαιτερότητες, να έχουν δηλαδή ίδιο θέμα αλλά να το παρουσιάζουν με άλλη οπτική. Στα λίγα τέτοια άρθρα που εντοπίσαμε ανακαλύψαμε πως πραγματικά αυτά αποτελούν ένα σημαντικό κομμάτι άρθρων για τα οποία ο μηχανισμό αποτυγχάνει να τα ομαδοποιήσει, ωστόσο είναι σημαντικό να δούμε αν εντοπίζει τη συνάφεια και δεν επιτρέπει την ομαδοποίηση ή αποτυγχάνει και στα δύο. Η αποτυχία εύρεσης της συνάφειας σίγουρα εγεί-

ρει ένα σημαντικό θέμα αλλά η ομαδοποίηση είναι κάτι διαφορετικό. Στην ουσία και τα δύο προκύπτουν από την ίδια διαδικασία αλλά η ομαδοποίηση έχει πιο σφιχτά όρια. Στην ουσία αυτό που πραγματοποιείται όταν έχουμε τέτοιου είδους άρθρα είναι το εξής. Το σύστημα μπορεί να διακρίνει τη συνάφεια με το αρχικό άρθρο και στην πλειοψηφία των περιπτώσεων αυτά τα άρθρα επιτυγχάνουν να χαρακτηριστούν σα συναφή. Ωστόσο, όπως δείχνει και το σχήμα δεν καταφέρουν όλα να περάσουν το «σκληρό όριο» που τίθεται για την ομαδοποίηση των άρθρων. Η διαδικασία είναι αντίστοιχη της διαδικασίας που χρησιμοποιούμε για να τοποθετήσουμε ένα άρθρο στο training set μίας κατηγορίας. Ένα μεγαλό ποσοστό αρκεί για να χαρακτηρίσει ένα άρθρο σε μία κατηγορία ωστόσο για να θεωρηθεί ένα άρθρο ως πιθανό κομμάτι του training set μίας κατηγορίας θα πρέπει η συνάφεια να ξεπερνά ένα σκληρό όριο (μεταβλητό με την πάροδο του χρόνου).

Είδαμε μέσα από την πειραματική διαδικασία του μηχανισμού πως το σύστημα βρίσκεται σε πλήρη λειτουργία σύμφωνα με τις προδιαγραφές που έχουμε θέσει και φυσικά είναι σημαντικό να ελεγχθεί η λειτουργία του σε μεγάλη κλίμακα προκειμένου να δούμε τον τρόπο λειτουργίας κάτω από αυτές τις συνθήκες. Η πειραματική διαδικασία μας έδωσε πολύ καλά αποτελέσματα βάσει των προδιαγραφών που έχουμε και για τους χρήστες που εμείς δημιουργήσαμε για το σύστημα διαπιστώσαμε πως λειτουργεί περισσότερο από ικανοποιητικά. Κάθε σύστημα του μηχανισμού μας έδινε μεγάλη ακρίβεια στα αποτελέσματα κάτι που γίνεται χαρακτηριστικά εμφανές από τη διεπαφή με το χρήστη.

ΚΕΦΑΛΑΙΟ 7

ΣΥΜΠΕΡΑΣΜΑΤΑ

Άκουε πολλά, λάλει καίρια

(Βίας ο Πρινεύς)

Το παρόν κεφάλαιο περιέχει συμπεράσματα για την εργασία που συντελέστηκε στο πλαίσιο της διδακτορικής διατριβής και σχετίζεται με τους τομείς στους οποίους πραγματοποιήθηκε μελέτη, έρευνα και ανάπτυξη.

Ο μηχανισμός peRSSonal αποτελεί ένα ολοκληρωμένο σύστημα το οποίο περιέχει διαδικασίες ανάκτησης, ανάλυσης, επεξεργασίας και μεταφόρτωσης πληροφορίας με μεθοδικό και προσωποποιημένο τρόπο με υποστήριξη της ελληνικής γλώσσας σε κάθε βήμα, κάτι πρωτοφανές για τα ελληνικά δεδομένα. Το σύστημα ξεκίνησε την ανάπτυξή του το 2004 και από τότε έχει αλλάξει πολλές μορφές σε κάθε επίπεδό του ενώ παράλληλα συνεχίζει να αλλάζει μέρα με τη μέρα κυρίως μέσα από την πληθώρα εργασιών οι οποίες πραγματοποιούνται και συντελούν τα μέγιστα στην ενίσχυση του συστήματος.

Μέσα από την εργασία που πραγματοποιήσαμε είχαμε την ευκαιρία να ασχοληθούμε με πολλά και διαφορετικά ερευνητικά πεδία, και να κοιτάξουμε άλλα επιφανειακά καθότι αυτό ήταν και το βάθος στο οποίο θέλαμε να προσεγγίσουμε και άλλους διεξοδικά και με πολλή λεπτομέρεια. Η εργασία που πραγματοποιήσαμε είχε σαν αποτέλεσμα αρκετές δημοσιεύσεις σε διεθνή συνέδρια και περιοδικά τόσο σε επίπεδο υποσυστήματος όσο και σε συνολικό επίπεδο σα μηχανισμός. Ο μηχανισμός αυτός επισήμως έχει εφαρμοστεί σε ένα και μόνο σύστημα με μεγάλη επιτυχία και παράλληλα έχουν γίνει πολλές προσεγγίσεις για εφαρμογή του μηχανισμού σε συστήματα μεγάλης κλίμακας χωρίς ωστόσο να έχουμε ακόμα κάποιο σύστημα σε επίπεδο παραγωγής μεγάλης κλίμακας. Φυσικά κάτι τέτοιο θα μας έδινε τη δυνατότητα να ελέγξουμε τις λειτουργίες του μηχανισμού σε μεγάλη κλίμακα αλλά και να συλλέξουμε πληροφορίες που σχετίζονται με τη λειτουργία του συστήματος από πολλούς διαφορετικούς χρήστες. Όπως γίνεται βέβαια σαφές ένα τέτοιο σύστημα θέλει καθημερινό έλεγχο και υποστήριξη διότι πολλοί χρήστες σημαίνουν και πολλές ευθύνες αλλά και πολλά προβλήματα τα οποία θα πρέπει να βρίσκουν μία άμεση λύση. Η εργασία αυτή μας βοήθησε, όπως αναφέρθηκε ήδη, να μελετήσουμε πληθώρα ερευνητικών πεδίων και να συλλέξουμε χρήσιμες πληροφορίες οι οποίες μας οδήγησαν στη δημιουργία ενός κάθετου συστήματος με απεριόριστες δυνατότητες. Αυτό που καταλάβαμε μέσα από την εργασία που πραγματοποιήσαμε είναι πως συγκεκριμένα στάδια του μηχανισμού θα πρέπει να είναι εξαιρετικά προσεγμένα ενώ άλλα αρκεί να προσφέρουν απλώς ποιοτική πληροφορία και ανεκτά αποτελέσματα για να οδηγήσουν το μηχανισμό συνολικά σε πολύ καλή λειτουργία. Αναφορικά με το μηχανισμό ανάκτησης πληροφορίας, μέσα από την εργασία μας οδηγηθήκαμε στη δημιουργία ενός ολοκληρωμένου συστήματος το οποίο είναι σε θέση να πραγματοποιεί ανάκτηση πληροφορίας με τρόπο μεθοδικό και απρόσκοπτο και μάλιστα χρησιμοποιώντας τεχνικές πρόβλεψης για το χρόνο στον οποίο κάποιο RSS feed θα ανανεωθεί με πληροφορία. Σε αυτό τον τομέα εισάγαμε την έννοια της ευγένειας και σε επίπεδο RSS crawling διότι όσο μικρό κι αν είναι το κόστος σε ένα σύστημα πολύ μεγάλης κλίμακας η κατανάλωση πόρων για τον έλεγχο των RSS feeds μπορεί να είναι τεράστια. Παράλληλα, μελετήσαμε και αναπτύξαμε αλγορίθμους οι οποίοι έχουν σα βασικό σκοπό να μπορούν να προβλέπουν τον τρόπο με τον οποίο ανανεώνεται πληροφορία σε ένα feed. Αυτό το πραγματοποιήσαμε χρησιμοποιώντας δύο στάδια, ένα στο οποίο γίνεται ανάγνωση και συχνός έλεγχος του RSS feed για να μπορέσουμε να συλλέξουμε ένα πολύ αναλυτικό posting history και ένα δεύτερο στάδιο στο οποίο χρησιμοποιούμε τα patterns που έχουμε δημιουργήσει για να μπορούμε να προβλέψουμε το χρόνο στον οποίο

θα δημοσιευθεί νέα πληροφορία. Παράλληλα, ο ίδιος μηχανισμός εφαρμόζει και intra-domain ελέγχους προκειμένου να αποφύγει τη συγκέντρωση διπλών άρθρων από διαφορετικά RSS του ίδιου domain.

Στη συνέχεια πραγματοποιήσαμε μελέτη για ένα μηχανισμό εξαγωγής χρήσιμου κειμένου από HTML σελίδες του διαδικτύου. Ο μηχανισμός αυτός βασίζεται τόσο σε ποιοτικά όσο και σε ποσοτικά στοιχεία που διαθέτουμε από την ανάλυση που πραγματοποιούμε στη σελίδα. Αποδείξαμε πως με τον αλγόριθμο που χρησιμοποιούμε για το συγκεκριμένο σύστημα μπορούμε να επιτύχουμε εξαγωγή χρήσιμου κειμένου με ακρίβεια πάνω από 90% σε κάθε περίπτωση. Οι διαφορετικές περιπτώσεις που υπάρχουν για τις σελίδες με άρθρα όπως έχουμε δει είναι άρθρα ενιαίου σώματος, άρθρα κατακερματισμένου σώματος και άρθρα τα οποία περιλαμβάνουν και σχόλια χρηστών. Η τελευταία περίπτωση είναι πάρα πολύ συνηθισμένη στα blogs τα οποία όπως είναι πλέον γνωστό τείνουν να αποτελέσουν σημαντικότερη πηγή ενημέρωσης (από άποψη επισκεψιμότητας) ακόμα και από τις σελίδες ειδησεογραφικών πρακτορείων. Αποδείξαμε, λοιπόν, πως ο αλγόριθμος που χρησιμοποιούμε μπορεί να προχωρήσει σε ανάλυση των σελίδων και να εντοπίσει το χρήσιμο κείμενο και να το εξάγει. φυσικά, είδαμε ότι υπάρχουν περιπτώσεις όπου μπορεί να γίνουν σφάλματα, ωστόσο αυτές είναι ελάχιστες και σχετίζονται κυρίως με περίεργο τρόπο στη δομή της σελίδας. Επιπρόσθετα και για το συγκεκριμένο μηχανισμό μελετήσαμε στοιχεία που σχετίζονται και με την εξαγωγή πολυμέσων από τις σελίδες που έχουν τα άρθρα και πιο συγκεκριμένα επικεντρωθήκαμε στην εξαγωγή εικόνων σχετικών με τα άρθρα. Σε αυτή την περίπτωση είδαμε πως ο μηχανισμός ο οποίος βασίζεται στη δομή της HTML σελίδας μπορεί να πετύχει πάρα πολύ υψηλά ποσοστά στην εξαγωγή πολυμέσων από σελίδα. Αυτό συμβαίνει διότι για κάθε σελίδα διαθέτουμε το DOM μοντέλο από το οποίο κατασκευάζεται και φυσικά τους κόμβους που περιέχουν το χρήσιμο κείμενο. Ο τρόπος κατασκευής μίας HTML σελίδας καθιστά σαφές στην πλειοψηφία των περιπτώσεων πως η εικόνα είναι ένας επιπλέον κόμβος ανάμεσα στο σώμα του κειμένου. Αυτή η διαπίστωση σε συνδυασμό με στοιχεία που έχουν να κάνουν με τη μορφολογία των εικόνων που χρησιμοποιούνται στα άρθρα του διαδικτύου κάνουν όπως είδαμε εύκολη την εξαγωγή εικόνων σχετικών με τα άρθρα. Ο μηχανισμός εξαγωγής χρήσιμου κειμένου αλλά και πολυμέσων είναι από τα πολύ βασικά συστήματα του peRSSonal καθότι αυτός συγκεντρώνει όλη την πληροφορία που θα εισαχθεί στο σύστημα προεπεξεργασίας και είναι κρίσιμο να πετυχαίνουμε μεγάλη ακρίβεια αλλά και πολύ ποιοτικά αποτελέσματα. Όπως είδαμε, ο μηχανισμός αυτός μπορεί να επιτύχει και τα δύο παραπάνω και μάλιστα να έχει καλύτερα αποτελέσματα από αντίστοιχους μηχανισμούς.

Στη συνέχεια μελετήσαμε συστήματα που σχετίζονται με προεπεξεργασία κειμένου. Στην ουσία μελετήσαμε ένα μηχανισμό ο οποίος έχει σα βασικές προδιαγραφές τόσο τη λεξικολογική ανάλυση κειμένου και την εξαγωγή λέξεων κλειδιών αλλά επιπρόσθετα ο μηχανισμός αυτός χρησιμοποιείται και για την εξαγωγή πληροφορίας που σχετίζεται με τη θέση κάθε λέξης στο κείμενο, τη συχνότητα με την οποία εμφανίζεται και σε περιπτώσεις που έχουμε γλώσσες που χρειάζονται τέτοια πληροφορία (όπως τα ελληνικά) και στοιχεία που αφορούν το μέρος του λόγου που

είναι κάθε λέξη του κειμένου. Στη μελέτη που κάναμε διαπιστώσαμε πως για την πλειονότητα των γλωσσών τα βήματα τα οποία πρέπει να ακολουθούνται είναι συγκεκριμένα και έχουν να κάνουν με ομογενοποίηση του κειμένου, αφαίρεση των σημείων στίξης, αφαίρεση λέξεων πολύ μικρού μεγέθους, αφαίρεση κοινών λέξεων, χρήση λεξικού για διόρθωση ή διαγραφή λέξεων και τέλος εφαρμογή αλγορίθμου stemming. Φυσικά όπως ήδη είπαμε στην μελέτη που κάναμε για την ανάπτυξη ελληνικού stemmer για την υποστήριξη ελληνικών από το σύστημα έγινε σαφές πως υπάρχουν περιπτώσεις που τα παραπάνω βήματα αλλάζουν. Έτσι, για την υποστήριξη της γαλλικής γλώσσας χρειάζεται να ορίσουμε τις λίστες με τα stopwords, και τον αλγόριθμο stemming και αμέσως το σύστημα θα αποκτήσει πλήρη υποστήριξη προεπεξεργασίας κειμένου για τη συγκεκριμένη γλώσσα. Σε άλλες περιπτώσεις, όπως διαπιστώσαμε στην υλοποίηση που κάναμε για την ελληνική γλώσσα είδαμε πως τα παραπάνω δεν είναι αρκετά. Η υλοποίηση που κάναμε για τον ελληνικό stemmer βασιζόταν σε μεγάλο βαθμό σε έναν ελληνικό tagger που επίσης αναπτύξαμε, με αποτέλεσμα τα παραπάνω βήματα να αλλάζουν τελείως στην υλοποίηση που τελικά εφαρμόσαμε. Μέσα από την έρευνα που κάναμε διαπιστώσαμε πως ο μηχανισμός tagging και stemming που φτιάξαμε για την ελληνική γλώσσα πετυχαίνει πολύ μεγάλη απόδοση και μάλιστα καλύτερη από τους αντίστοιχους αλγορίθμους που υπάρχουν για την ελληνική γλώσσα.
Προχωρόντας στους μηχανισμούς κατηγοριοποίησης και εξαγωγής περίληψης κειμένου, όπως διαφαίνεται είναι και τα συστήματα που μελετήσαμε λιγότερο στην εργασία μας αναφορικά με την ανάλυση και ανάπτυξη που έγινε συγκριτικά με τις εκδόσεις που είχε παλαιότερα το σύστημα peRSSonal. Φυσικά η μελέτη των συστημάτων ήταν εκτενής και οι συγκρίσεις με υπάρχοντα συστήματα απαραίτητες προκειμένου να καταφέρουμε να πετύχουμε βέλτιστο αποτέλεσμα αλλά για αυτό που θέλουμε να επιτύχουμε συνολικά στο σύστημά μας. Ο μηχανισμός κατηγοριοποίησης βασίζεται σε ευρετικές 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. Το training set που χρησιμοποιήσαμε για το σύστημα είναι δικής μας κατασκευής καθότι το σύστημα έχει συνολικά συγκεκριμένες προδιαγραφές που δεν ταιριάζουν στα έτοιμα set που διατίθενται στο Διαδίκτυο. Ο μηχανισμός κατηγοριοποίησης όπως έχουμε δει χρησιμοποιείται ουσιωδώς για να αποκαλύψει συσχέτιση με περισσότερες της μίας κατηγορίας για ένα κείμενο. Αυτό μας βοηθά να εντάξουμε ένα κείμενο σε πολλές κατηγορίες και με αυτό τον τρόπο να παρουσιάσουμε μεγαλύτερη ακρίβεια στην παρουσίαση πληροφορίας σε ένα χρήστη.
Ο μηχανισμός αυτόματης εξαγωγής περίληψης βασίζεται επίσης σε ευρετικές μεθόδους και η λειτουργία του εντοπίζεται στη βαθμοδότηση των προτάσεων προκειμένου να κρατηθούν αυτές με τη μεγαλύτερη βαθμολογία. Ο αλγόριθμος βαθμοδότησης των προτάσεων βασίζεται σε φυσικά χαρακτηριστικά που μπορεί να έχει ένα κείμενο όπως η συχνότητα των λέξεων κλειδιών αλλά και η επανεμφάνιση στην ίδια πρόταση ή στον τίτλο του κειμένου, το μέρος του λόγου κάθε λέξης και γενικά στοιχεία που προκύπτουν από την προεπεξεργασία. Είδαμε πως επικου-

ρικά στην εξαγωγή περίληψης μπορεί να λειτουργήσει και η κατηγοριοποίηση κειμένου και η βαθμολόγηση των προτάσεων να αλλάξει σύμφωνα με την κατηγορία που ανήκει το κείμενο και οι λέξεις κλειδιά που εξάγονται από αυτό. Είδαμε παράλληλα πως δε χρειάζεται να κρατάμε μόνο ένα κομμάτι του κειμένου σαν περίληψη αλλά ολόκληρο το κείμενο με τις προτάσεις τοποθετημένες σύμφωνα με τη βαθμολογία που αυτές λαμβάνουν. Με αυτό τον τρόπο διατηρούμε το σύνολο της πληροφορίας και έχουμε τη δυνατότητα εμφάνισης τόσου μεγέθους όσο κρίνεται σκόπιμο για να μπορεί ένας χρήστης να πραγματοποιεί ανάγνωση της πληροφορίας όποια κι αν είναι η συσκευή από την οποία έρχεται σε επαφή με την πληροφορία. Είδαμε πως μόνο ένα μικρό κομμάτι της περίληψης και μάλιστα συνήθως οι 4-5 πρώτες προτάσεις μπορούν να περιέχουν αρκετά στοιχεία για να κατανοήσουμε το γενικό νόημα του κειμένου και σε γενικές γραμμές μπορούν να οδηγήσουν σε ποιοτικότερη κατηγοριοποίηση αν συνδυάσουμε τους μηχανισμούς κατηγοριοποίησης και εξαγωγής περίληψης.

Τέλος, είδαμε πως μεγάλο βάρος έχει δοθεί στη μελέτη του μηχανισμού παρουσίασης πληροφορίας αλλά και βασικότερα προσωποποίησης της πληροφορίας στον τελικό χρήστη. Πρόκειται για ένα πολύ σημαντικό κομμάτι του μηχανισμού καθότι είναι αυτό που φέρνει σε επαφή την πληροφορία που έχουμε συλλέξει με τους εκάστοτε χρήστες του συστήματος. Για το σύστημα αυτό είδαμε κάποια πολύ σημαντικά χαρακτηριστικά που διαθέτει, όπως τη δυνατότητα της απόρριψης άρθρων, τη δυνατότητα της παρουσίασης σχετικών άρθρων και είδαμε πως μπορεί να λειτουργήσει η μονάδα δημιουργίας ομάδας άρθρων που ανήκουν σε διαφορετικά domain. Πρόκειται στην ουσία για την ίδια είδηση δημοσιευμένη ή αναδημοσιευμένη σε πολλούς δικτυακούς τόπους. Σε αυτή την περίπτωση ο μηχανισμός είναι σε θέση να μπορέσει να εντοπίσει τις ταυτόσημες ειδήσεις και να δημιουργήσει ομάδες ίδιων άρθρων που έχουν χαρακτηριστικές ιδιότητες τις οποίες και μελετήσαμε. Αποδείξαμε παράλληλα μέσα από τις πειραματικές διαδικασίες που κάναμε πως ο μηχανισμός αυτός μπορεί να επιτύχει πολύ μεγάλα ποσοστά επιτυχίας στην ομαδοποίηση άρθρων χρησιμοποιώντας χαρακτηριστικά όπως τη συσχέτιση κειμένων αλλά και άλλες μετρικές οι οποίες μελετήσαμε και ισχύουν για άρθρα που έχουν ίδιο περιεχόμενο. Στην ουσία ο μηχανισμός αυτός προσφέρει στους τελικούς χρήστες μία υπηρεσία περιήγησης σε άρθρα του διαδικτύου δίνοντας τη δυνατότητα για πλήρη προσωποποίηση στο προφίλ του εκάστοτε χρήστη. Μιλώντας για προσωποποίηση αξίζει να αναφέρουμε πως ο μηχανισμός προσωποποίησης στο χρήστη μπορεί να προσφέρει μία ποιοτική υπηρεσία που επιτρέπει στους χρήστες να έχουν πρόσβαση μόνο σε άρθρα τα οποία είναι μέσα στα ενδιαφέροντά τους και να μη βομβαρδίζονται με πληροφορία που τους είναι αδιάφορη ή και άχρηστη. Είδαμε πως ο μηχανισμός είναι σε θέση μέσα σε διάστημα τεσσάρων εβδομάδων με μία λογική ημερήσια χρήση του συστήματος από τους χρήστες να εντοπίσει ένα αρκετά αναλυτικό προφίλ για κάθε χρήστη και να πετύχει πάνω από 75% ακρίβεια στη σχετικότητα των άρθρων που επιθυμούν να βλέπουν οι χρήστες. Παράλληλα, είδαμε πως μπορεί να περιοριστεί κατά 90% ο αριθμός των άρθρων που εμφανίζονται συνολικά στο χρήστη.

Συμπερασματικά είδαμε ένα μηχανισμό ο οποίος μπορεί να λειτουργήσει συνολικά για να προσφέρει στους χρήστες μία εναλλακτική εμπειρία όταν αυτοί πραγματοποιούν καθημερινές εργασίες στο διαδίκτυο όπως είναι αυτή της ανάγνωσης ειδήσεων. Ο μηχανισμός μπορεί να προσφέρει πολλές σημαντικές υπηρεσίες στους χρήστες, ακόμα και σε αυτούς που επιθυμούν να έχουν μία απλή σφαιρική ενημέρωση αλλά και σε αυτούς που θέλουν να λάβουν πολύ αναλυτική πληροφόρηση. Τα δυνατά σημεία του μηχανισμού εντοπίζονται στο γεγονός ότι ο χρήστης δε χρειάζεται να επισκέπτεται κάθε σελίδα ειδησεογραφικού περιεχομένου προκειμένου να λάβει ενημέρωση ενώ παράλληλα το σύστημα μπορεί να λειτουργήσει αποκλειστικά πάνω σε στοιχεία που επιθυμεί ο χρήστης (δικά του RSS feeds και λέξεις κλειδιά). Πολύ σημαντικό είναι και το γεγονός πως εκτός από την εγγραφή του χρήστη σε κανένα άλλο σημείο δε ζητείται από το χρήστη να δώσει ρητά κάποια εντολή στο σύστημα για διαμόρφωση του προφίλ του. Ακόμα και η υπηρεσία tag rating κατά την οποία ένας χρήστης δηλώνει ρητά προτίμηση για λέξεις κλειδιά είναι ένα παιχνίδι όπου ο χρήστης δηλώνει αν του αρέσει ή δεν του αρέσει κάποιο keyword του κειμένου, πράγμα που πλέον συμβαίνει σε κάθε δικτυακό τόπο κοινωνικής δικτύωσης. Τέλος, ο μηχανισμός μπορεί να δώσει για κάθε χρήστη έξοδο σε μορφή XML (RSS) και άρα είναι εφικτή η πρόσβαση στην πληροφορία ανεξαρτήτως του μέσου. Φυσικά, όπως και για κάθε RSS feed, είναι απαραίτητη η «ανακάλυψη» και η εγγραφή (δηλαδή θα πρέπει ο χρήστης και να εντοπίσει το peRSSonal αλλά και να εγγραφεί σε αυτό.

Τέλος, όπως κάθε μηχανισμός που προκύπτει από ερευνητική εργασία, για να ολοκληρωθεί και να μπορέσει να λειτουργήσει στον κόσμο του Διαδικτύου, είναι απαραίτητο να περάσει στο στάδιο της παραγωγής. Ο μηχανισμός έχει ανάγκη να βρεθεί σε αυτό το επίπεδο τόσο για να βελτιωθεί όσο και για να δοκιμαστεί εκτενώς. Είναι σαφές πως είναι ο μοναδικός τρόπος απόλυτου ελέγχου κάθε λειτουργίας, διόρθωσης, βελτίωσης και τελειοποίησης του μηχανισμού σα σύνολο αλλά και κάθε επιμέρους συστήματος.

ΚΕΦΑΛΑΙΟ 8

ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Εις το σκαλί για να πατήσεις τούτο
πρέπει με το δικαίωμά σου νάσαι
πολίτης εις των ιδεών την πόλι.
Και δύσκολο στην πόλι εκείνην
είναι
και σπάνιο να σε πολιτογραφήσουν.
Στην αγορά της βρίσκεις
Νομοθέτας
που δεν γελά κανένας τυχοδιώκτης.
Εδώ που έφθασες, λίγο δεν είναι
τόσο που έκαμες, μεγάλη δόξα

(Κ. Π. Καβάφης)

Το παρόν κεφαλαιο αναλύει τη μελλοντικη εργασια που μπορει να πραγματοποιηθεί στο πλαίσιο βελτίωσης του μηχανισμού peRSSonal παρέχοντας αναλυτικα στοιχεία για τη βελτίωση που μπορει να γινει σε καθε κομματι του μηχανισμού.

Θα μπορούσαμε να αναλωθούμε με μία ολόκληρη εργασία προκειμένου να αναλύσουμε τη μελλοντική εργασία που μπορεί να γίνει πάνω στο σύστημα perssonal. Όπως έχει γίνει ήδη κατανοητό, το σύστημα κατασκευάζεται από πληθώρα υποσυστημάτων και έτσι κάθε ένα από αυτά, που εκ των πραγμάτων αποτελεί μία ολόκληρη ερευνητική περιοχή, θα μπορούσε να αναπτυχθεί σε τεράστιο βαθμό. Ως εκ τούτου στην περιγραφή της μελλοντικής εργασίας θα δούμε τόσο μεμονωμένα όσο και συνολικά τις βελτιώσεις που θα μπορούσαν να πραγματοποιηθούν. Αυτό διότι τόσο κάθε μηχανισμός σαν αυτόνομη οντότητα μπορεί να αναπτυχθεί, να βελτιωθεί και να αλλάξει ριζικά χωρίς να επηρεάζεται η λειτουργία των υπολοίπων μηχανισμών αλλά και όλοι μαζί οι μηχανισμοί μπορούν να αλλάξουν προς μία κατεύθυνση προκειμένου να γίνει συνολικά βελτίωση του peRSSonal.

8.1 Μηχανισμός Ανάκτησης

Ο μηχανισμός ανάκτησης πληροφορίας χρησιμοποιεί αλγορίθμους προκειμένου να μπορεί να ελέγχει τη ροή των ειδήσεων και να προσπαθεί να προσαρμόζεται χρονικά στις αλλαγές που πραγματοποιούνται στα RSS feeds. Σε κάθε περίπτωση η απόδοση του μηχανισμού αυτού μπορεί να αυξηθεί εφόσον εφαρμόσουμε ακόμα πιο αναλυτικούς αλγορίθμους. Ωστόσο, σε κάθε περίπτωση θα πρέπει να σκεφτόμαστε το trade-off που υπάρχει από κάθε ενέργεια που κάνουμε. Αυτή τη στιγμή ο μηχανισμός μπορεί να εφαρμόζει μία περίοδο εκμάθησης πάνω στα RSS feeds και από τη στιγμή που θα αναλύσει τις χρονικές αλλαγές ενός feed εφαρμόζει έναν αλγόριθμο επιλογής κάθε RSS feed ανάλογα με το ρυθμό δημοσίευσης στο πέρασμα της ώρας. Πρόκειται για ένα σύστημα που στην ουσία βασίζεται σε στατικά patterns τα οποία δημιουργούνται άπαξ και στην πορεία χρησιμοποιούνται σαν σχέδιο για το μηχανισμό επιλογής και πρόβλεψης. Στην ουσία θα μπορούσε να χρησιμοποιηθεί δυναμικό pattern για κάθε RSS έτσι ώστε το αρχικό pattern να μεταβάλλεται στην πορεία του χρόνου καθότι δεν είναι απίθανο ένα RSS feed να αλλάζει συμπεριφορά με την πάροδο του χρόνου.
Ένας άλλος σημαντικός παράγοντας για τη χρονική προσαρμογή και πρόβλεψη είναι να μπορέσουμε να αξιολογήσουμε το μέγεθος που μπορεί να έχει μία είδηση σύμφωνα με τον τρόπο που αυτή δημοσιεύεται. Είναι αδιαμφισβήτητο πως σε περίοδο εκλογών έχουμε διαφορετικό pattern δημοσίευσης πολιτικών άρθρων απ' ότι έχουμε για άλλες περιόδους. Επιπρόσθετα όταν συμβαίνει ένα έκτακτο γεγονός, όπως μία φυσική καταστροφή, παρατηρούμε πως για τις επόμενες ώρες ή και μέρες αλλάζει και πάλι ο ρυθμός με τον οποίο προκύπτουν οι ειδήσεις. Αν θα μπορούσε ο μηχανισμός να προβλέψει το μέγεθος μίας τυχαίας είδησης η οποία δεν εντοπίστηκε από το μηχανισμό και άρα είτε είναι τυχαία και δε θα πρέπει να δοθεί σημασία είτε όμως είναι σημαντική και άρα θα πρέπει να αλλάξει ο ρυθμός ανανέωσης και πρόβλεψης για το συγκεκριμένο feed. Αναφορικά με τη σημαντικότητα μίας είδησης σίγουρα μία μελλοντική εργασία θα μπορούσε να είναι ο εντοπισμός σημαντικών ειδήσεων. Μάλιστα και εφόσον η ερευνητική περιοχή της

ανάκτησης τέτοιου είδους πληροφορίας ωθείται προς αυτή την κατεύθυνση θα μπορούσαμε να μελετήσουμε τρόπους εντοπισμού τέτοιων ειδήσεων και μάλιστα εφαρμογή σε focused crawler ο οποίος θα έχει σαν σκοπό τον εντοπισμό αυτών των ειδήσεων. Σε κάθε περίπτωση ο μηχανισμός θα μπορούσε να ενισχυθεί με επικουρικά συστήματα. Ο τρόπος λειτουργίας αυτή τη στιγμή προσφέρει ποιοτικά αποτελέσματα για τις προδιαγραφές που θέτουμε στο μηχανισμό μας. Από την άλλη, τίποτα δε μας απαγορεύει να προσθέσουμε επιπλέον υποσυστήματα τα οποία θα λειτουργούν επικουρικά του αρχικού μηχανισμού και τα οποία θα έχουν στην ευθύνη τους τη δυναμική αλλαγή των patterns ενός RSS feed αλλά και τη λειτουργία focused crawling για τον εντοπισμό σημαντικών γεγονότων. Προφανώς τα δύο παραπάνω θα μπορούσαν να συνδυαστούν για να προκύψουν τα patterns δημοσίευσης σημαντικών άρθρων ανάλογα με τη σημαντικότητά τους.

8.2 Μηχανισμός Εξαγωγής Χρήσιμων κειμένων και multimedia

Ένα από τα ιδιαίτερα κομμάτια του μηχανισμού μας είναι και αυτό που έχει σα σκοπό να απομονώσει το κείμενο του άρθρου από την HTML σελίδα στην οποία έχει δημοσιευθεί. Ουσιωδώς και αν τηρούνταν οι προδιαγραφές λειτουργίας των RSS feeds η πληροφορία αυτή θα υπήρχε ατόφια μέσα στο XML αρχείο του RSS και δε θα υπήρχε καμία απολύτως ανάγκη για ανάλυση και ανάκτηση των σελίδων που περιέχουν τα άρθρα. Ωστόσο, όπως είδαμε αυτό κρίνεται απαραίτητο και εκτός από την ανάκτηση αυτό που πρέπει να γίνει είναι εντοπισμός και εξαγωγή του χρήσιμου κειμένου από τη σελίδα. Όπως είδαμε και στην ανάλυση των αλγορίθμων αυτό επιτυγχάνεται με μετρικές οι οποίες βασίζονται στην αποδόμηση της σελίδας και τον εντοπισμό των στοιχείων της σελίδας που έχουν το κείμενο του άρθρο. Για να εξαχθεί αυτή η πληροφορία χρησιμοποιούμε μετρικές οι οποίες βασίζονται αφενός στον εντοπισμό των κόμβων που έχουν μεγάλο κομμάτι καθαρού κειμένου και από την άλλη στη γειτνίαση αυτών των κόμβων προκειμένου να εξασφαλίσουμε ότι το κείμενο είναι ενιαίο. Μία αλλαγή που πραγματοποιήσαμε στο μηχανισμό σχετιζόταν με το σημείο έναρξης της αναζήτησης για κείμενο που δεν είναι άλλο από τον κόμβο με τον τίτλο του άρθρου. Τον τίτλο του άρθρου δεν πρέπει να ξεχνάμε ότι τον έχουμε ανακτήσει από το RSS feed καθότι αυτός βρίσκεται εκεί.
Κάτι που θα μπορούσε να μας βοηθήσει επίσης είναι το γεγονός πως κάποιες φορές το RSS feed περιέχει και κάποιο κομμάτι του κειμένου. Αυτό σημαίνει πως ενδεχόμενα να μπορούμε να εντοπίσουμε και που ακριβώς βρίσκεται κομμάτι του κειμένου αν όχι ολόκληρο και με αυτό τον τρόπο να είμαστε περισσότερο σίγουροι για το κείμενο που εξάγουμε.
Ένα σύστημα που θα μπορούσε να βοηθήσει επιπλέον το μηχανισμό εξαγωγής χρήσιμου κειμένου είναι λίστες με λέξεις οι οποίες αποτελούν χαρακτηριστικές απαγορευμένες λέξεις για να περιληφθούν σε καθαρό κείμενο. Η έρευνά μας έχει δείξει πως τέτοιες λέξεις μπορεί να είναι : “written by”, “copyright”, “last updated” και άλλες οι οποίες είναι χαρακτηριστικές λέξεις και

φράσεις που τοποθετούνται στο τέλος ενός άρθρου. Αυτός ο μηχανισμός θα μπορούσε επιπρόσθετα να περιέχει και μηχανισμό συσχέτισης κειμένων. Με αυτό τον τρόπο θα μπορούσαμε να έχουμε πίνακες με λέξεις κλειδιά οι οποίες εξάγονται από κείμενο που δε θέλουμε να συλλέξουμε και κάθε κομμάτι κειμένου που εξάγεται από τους κόμβυος που εντοπίζουμε να συγκρίνεται σημασιολογικά με αυτές τις λέξεις κλειδιά και οποιαδήποτε ομοιότητα να οδηγεί σε απόρριψη του κόμβου. Όπως και σε κάθε μηχανισμό θα πρέπει σε αυτό το επίπεδο να τονίσουμε πως αυτό που μας ενδιαφέρει είναι να έχουμε αρκούντως ποιοτικά αποτελέσματα σε ικανοποιητικό χρόνο διατηρώντας παράλληλα την κατανάλωση πόρων σε χαμηλά επίπεδα.

Περνώντας στο κομμάτι του CUTER το οποίο ασχολείται με την εξαγωγή πολυμέσων θα πρέπει να τονίσουμε πως έχουμε καταφέρει να φτιάξουμε ένα μηχανισμό ο οποίος είναι σε θέση να εξάγει με πολύ μεγάλη επιτυχία οποιεσδήποτε εικόνες σχετίζονται με το άρθρο που διαβάζουμε. Αυτό είναι εμφανές και μέσω πειραμάτων που πραγματοποιήσαμε αλλά και από το γεγονός πως η online έκδοση του μηχανισμού είναι πλέον γεμάτη με πληθώρα άρθρων που περιέχουν τουλάχιστον μία σχετική εικόνα. Ο μηχανισμός αυτός θα μπορούσε να εξελιχθεί ακόμα περισσότερο ώστε να μπορεί να εξάγει και άλλα πολυμέσα πλην των εικόνων, όπως τα πλέον δημοφιλή βίντεο αλλά και ηχητικά ντοκουμέντα. Βέβαια για την πραγματοποίηση αυτού θα πρέπει να λάβουμε υπόψη μας και τον τρόπο αποθήκευσης αυτής της πληροφορίας, δηλαδή των πολυμέσων. Ήδη με το ρυθμό με τον οποίο ανακτούμε άρθρα από το διαδίκτυο και χωρίς να έχουμε καθόλου πολυμέσα αποθηκευμένα σε κάποια βάση δεδομένων ή σε κάποιο σύστημα αρχείων καταφέρνουμε να χρειαζόμαστε κάποια GBs εβδομαδιαίως. Αυτό σημαίνει πως αν αποφασίζαμε να εξάγουμε και να αποθηκεύουμε πολυμέσα θα χρειαζόμαστε οπωσδήποτε τεράστιους αποθηκευτικούς χώρους και ανεξάντλητο bandwidth για να προσφέρουμε τις υπηρεσίες μας. Από την άλλη η αποθήκευση της πηγής του πολυμέσου σημαίνει πως αν πρόκειται για εικόνα μπορούμε να την προβάλλουμε στο χρήστη μέσα από τις σελίδες μας αλλά χρησιμοποιώντας το original URL και άρα bandwidth από τον πραγματικό server. Δε γίνεται όμως το ίδιο και για τα embedded videos στις ιστοσελίδες καθότι πέραν των υπηρεσιών που έχουν φτιαχτεί με αυτό το σκοπό και άρα προσφέρουν remote embedded video οι υπηρεσίες των portals συνήθως δεν επιτρέπουν την απομακρυσμένη εκτέλεση του βίντεο και συνεπώς θα πρέπει κανείς να επισκεφθεί την πραγματική σελίδα για να δει το βίντεο.

Τέλος, επειδή έχουμε πληθώρα interdomain RSS feeds θα μπορούσαμε να αξιοποιήσουμε το γεγονός ότι σε γενικές γραμμές οι σελίδες έχουν ενιαίο layout και συνεπώς μπορούμε να εντοπίσουμε το Pattern μίας σελίδας κατά τη διάρκεια της ανάλυσης. Αυτό θα ήταν πολύ χρήσιμο ώστε να μπορέσουμε να εντοπίσουμε ακριβώς που τοποθετείται το χρήσιμο κείμενο στο HTML structure ενός domain και συνεπώς να μπορούμε εύκολα με δυναμικούς wrappers να εξάγουμε με μεγάλη επιτυχία το κείμενο. Σε αυτή την περίπτωση υπάρχει πάντα ο κίνδυνος μικρές αλλαγές στο layout να καταστρέφουν συνολικά τη λειτουργία του μηχανισμού και συνεπώς φαίνεται να είναι μία λύση που ενδεχόμενα να μπορεί να δώσει πιο ποιοτικά αποτελέσματα αλλά σε κάθε περίπτωση η αποτυχία του μπορεί να είναι πολύ καταστροφική.

8.3 Μηχανισμός Προεπεξεργασίας Κειμένου

Ο μηχανισμός προ-επεξεργασίας κειμένου είναι ένα σύστημα το οποίο έχει σα βασικό σκοπό να εξάγει λέξεις κλειδιά από τα κείμενα που δέχεται σαν είσοδο. Ωστόσο, ο μηχανισμός αυτός πραγματοποιεί πολλές παράλληλες εργασίες προκειμένου να υποστηρίξει πλήρως όλους τους αλγορίθμους των συστημάτων που έπονται στο σύστημα peRSSonal. Έτσι, λοιπόν, άλλες εργασίες που πραγματοποιούνται σχετίζονται με εύρεση για κάθε λέξη κλειδί της συχνότητας εμφάνισης μέσα στο κείμενο, εύρεση των προτάσεων μέσα στις οποίες βρίσκεται η λέξη κλειδί αλλά και πληροφορίας σχετικά με το αν βρέθηκε μία λέξη στον τίτλο.
Όλα τα μεταδεδομένα που εξάγονται από το μηχανισμό είναι απαραίτητα για την άρτια λειτουργία του συστήματος. Ωστόσο, επειδή ο μηχανισμός έχει σα σκοπό να μπορεί να υποστηρίζει κάθε γλώσσα θα πρέπει να δοθεί μεγάλη προσοχή στα σημεία τα οποία εξαρτώνται από τη γλώσσα ώστε να είναι εύκολη η υποστήριξη και υλοποίηση του μηχανισμού σε κάθε γλώσσα. Το σύστημα έχει δημιουργηθεί έτσι ώστε με τον προσδιορισμό ενός συστήματος stemming για κάθε γλώσσα ο μηχανισμός μπορεί να πραγματοποιήσει άμεσα την υποστήριξη πολλών γλωσσών. Πρόσφατα, και με την προσπάθεια προσθήκης της ελληνικής γλώσσας διαπιστώσαμε πως σε κάποιες γλώσσες που έχουν ιδιαιτερότητες που ή που χρειαστήκαμε κάτι περισσότερο από έναν απλό stemmer, η διαδικασία δεν ήταν τόσο απλή όσο αναμέναμε. Η ύπαρξη ενός tagger ο οποίος διατηρούσε πληροφορία για κάθε λέξη κλειδί που ανέλυε δημιουργούσε προβλήματα καθότι πρόσθετε ένα επιπλέον βήμα στο μηχανισμό το οποίο δεν υπήρχε προηγούμενως. Σε αυτή την περίπτωση και μελλοντικά θα πρέπει να δοθεί προσοχή στο γεγονός πως μπορεί κάποιες ανωμαλίες σε κάποιες γλώσσες να προκαλούν την ανάγκη για περισσότερα βήματα στο μηχανισμό προεπεξεργασίας. Έτσι, αντί για να έχουμε αυτόνομο σύστημα το οποίο με την προσθήκη αλγορίθμου stemming να μπορεί να υποστηρίξει κάθε γλώσσα θα πρέπει να αλλάξουμε το σύστημα ώστε να έχει απλώς ενιαία είσοδο και ενιαία έξοδο και τα ενδιάμεσα βήματα να είναι απλά ένα μαύρο κουτί. Στην παρούσα φάση έχουμε πάει ένα βήμα πιο μακριά προσπαθώντας να επιτύχουμε μέγιστη κλιμάκωση, ωστόσο καθώς φάνηκε η μέγιστη κλιμάκωση σε αυτό το σημείο δεν είναι η καλύτερη λύση. Θα πρέπει να μεταφερθεί η κλιμάκωση ένα βήμα πιο πάνω και απλά να χρειαζόμαστε μία υλοποίηση η οποία να μπορεί να δεχθεί σαν είσοδο ένα κείμενο και να μπορεί να δώσει σαν έξοδο λίστες με τις λέξεις κλειδιά, τη συχνότητά τους μέσα στο κείμενο αλλά και τις προτάσεις στις οποίες εμφανίζονται. Παράλληλα, θα πρέπει να προβλέπεται χώρος για μεταδεδομένα όπως το μέρος του λόγου κάθε λέξης στο σημείο που εντοπίστηκε αλλά και τη συχνότητα εμφάνισης της λέξης στο τίτλο του κειμένου ή σε σημαντικά κομμάτια του κειμένου γενικότερα (π.χ. στη λεζάντα μίας εικόνας). Μελλοντικά θα μπορούσαμε επίσης να εντάξουμε στη λειτουργία του συστήματος αυτού του επιπέδου και στοιχεία όπως ο έλεγχος των λέξεων μέσω εργαλείων που ήδη υπάρχουν και διατίθενται ελεύθερα στο διαδίκτυο και θα μπορούσαν να βελτιώσουν την ποιότητα των αποτελεσμάτων αλλά και τη λειτουργία του συστήματος. Για παράδειγμα είναι πολύ σημαντικό το γεγονός ότι θα μπορούσαμε να κάνουμε χρήση μηχανισμών εντοπισμού συνώνυ-

μων κάποιων λέξεων αλλά και αντώνυμων. Γενικότερα, σε τέτοια συστήματα είναι σημαντική η χρήση ενός θησαυρού λέξεων που θα βελτίωνε την ποιότητα των εξαγόμενων αποτελεσμάτων.

8.4 Μηχανισμός Κατηγοριοποίησης Κειμένου

Ο μηχανισμός κατηγοριοποίησης κειμένου που έχουμε στο σύστημα peRSSonal έχει δημιουργηθεί για να λειτουργεί βοηθητικά σε κάποιες βασικές λειτουργίες του συστήματος. Για παράδειγμα, μέσα από τη διαδικασία κατηγοριοποίησης ο μηχανισμός καταφέρνει να αναλύσει το training set που έχει και να δημιουργήσει τις αρχικές κατηγορίες που αποτελούν και τη ρίζα του peRSSonal. Στη συνέχεια και ακόμα και αν ο μηχανισμός αυτός θεωρείται βασικός και πυρήνας του συστήματός μας έχουμε να παρατηρήσουμε πως η κατηγοριοποίηση λειτουργεί κυρίως επικουρικά. Η κατηγοριοποίηση που πραγματοποιείται βασίζεται στη συσχέτιση συνημίτονου των άρθρων τα οποία εισάγονται συγκριτικά με τις ήδη υπάρχουσες κατηγορίες του συστήματος. Ο μηχανισμός όπως είδαμε μπορεί να χρησιμοποιηθεί επιπλέον για να βοηθήσει συστήματα όπως η εξαγωγή περίληψης ή ο μηχανισμός εύρεσης trash articles να λειτουργήσουν πιο ποιοτικά. Συνολικά, αν θέλαμε να δούμε το μηχανισμό κατηγοριοποίηση κειμένου θα πρέπει να αλλάξουμε εντελώς την οπτική από την οποία τον βλέπουμε και το λόγο για τον οποίο έχει κατασκευαστεί στην ουσία. Αυτό σημαίνει αυτομάτως ένα μηχανισμό που θα έχει εντελώς διαφορετικό ρόλο στο σύστημα και όχι απλώς επικουρικό. Από τη στιγμή που αυτός ο μηχανισμός μπορεί να υποστηρίξει εύκολα και άμεσα άλλα υποσυστήματα του peRSSonal τότε θα πρέπει να έχει και διαφορετική προσέγγιση και συνεπώς και διαφορετική κατασκευή. Ο μηχανισμός αυτός έχει όλα τα απαραίτητα στοιχεία εισόδου για να λειτουργήσει χρησιμοποιώντας πληθώρα αλγορίθμων από αυτούς που είναι διαθέσιμοι στη βιβλιογραφία. Ωστόσο, δε θα σταθούμε στον τρόπο λειτουργίας αλλά θα πρέπει να κοιτάξουμε και άλλα στοιχεία του μηχανισμού.
Ο μηχανισμός κατηγοριοποίησης θα μπορούσε σε πρώτη φάση με τη βοήθεια έτοιμων εργαλείων να μετατραπεί αυτόνομα σε πολύγλωσσο σύστημα. Για να γίνει αυτό χρειάζεται να υπάρχει λεξικό αλλά και μηχανισμός εύρεσης συνωνύμων και αντωνύμων. Αυτό θα άλλαζε όλη τη διαδικασία λειτουργίας του μηχανισμού. Κατά τη διάρκεια κατασκευής των πρότυπων λιστών από λέξεις κλειδιά που αντιπροσωπεύουν κάθε κατηγορία, ο μηχανισμός θα μπορούσε να το κάνει αυτό για κάθε γλώσσα για την οποία διαθέτει λεξικό και αυτό που λέμε θησαυρό λέξεων. Παράλληλα, ο μηχανισμός θα μπορούσε να χρησιμοποιήσει επιπλέον στοιχεία εισόδου για να κατασκευάζει τόσο θετικά όσο και αρνητικά βάρη για τις λέξεις κλειδιά που διαθέτει κάθε κατηγορία. Με αυτό τον τρόπο θα είχαμε διαφορετικά αποτελέσματα κατά τη διάρκεια κατηγοριοποίησης η οποία θα μπορούσε να περάσει σε διαφορετικά επίπεδα και στη διαδικασία σχεδιασμού αλλά και στη διαδικασία μάθησης.
Τέλος ο μηχανισμός κατηγοριοποίησης πληροφορίας όπως αναφέρθηκε θα μπορούσε να λάβει γενικότερες διαστάσεις για να χρησιμοποιηθεί σε μεγαλύτερο βάθος. Δεδομένου ότι στις μέρες

μας γίνεται πολύς λόγος τόσο για σημαντικότητα που έχει μία είδησης και από την άλλη, για ανάλυση περιεχομένου κάθε είδησης θα μπορούσε στο σύστημά μας το ρόλο αυτό να τον έχει ο μηχανισμός κατηγοριοποίησης. Αναφορικά με τη σημαντικότητα κάθε είδησης ο μηχανισμός μπορεί να μεταβάλλει τη λειτουργία του, της απλής δηλαδή κατηγοριοποίησης, και να μπορεί να εντοπίζει ή να δίνει στοιχεία για την πιθανότητα μία είδηση να είναι σημαντική. Παράλληλα, και φυσικά πάντα με τον κατάλληλο σχεδιασμό, είναι εφικτό να μπορεί ο μηχανισμός να αξιολογήσει το περιεχόμενο μίας είδησης. Αναφορικά με το τελευταίο και επειδή και αυτό είναι αντικείμενο σχολιασμού τα τελευταία χρόνια μπορεί ο μηχανισμός να εντοπίζει αν το άρθρο είναι αντικειμενικό, αν μεροληπτεί ή αν κατακρίνει σύμφωνα με το περιεχόμενό του.

8.5 Μηχανισμός Εξαγωγής Περίληψης

Αναφορικά με το μηχανισμό περίληψης του συστήματος, θα πρέπει να τονιστεί πως ο μηχανισμός βρίσκεται ήδη σε στάδιο αλλαγής. Μία πρώτη βασική αλλαγή που πραγματοποιήθηκε τελευταία στο μηχανισμό είναι πως επιστρέφει το πλήρες μέγεθος του κειμένου με βαθμολογημένες τις προτάσεις και όχι μία περίληψη. Το στάδιο αυτό μεταφέρθηκε σαν εργασία για το μηχανισμό παρουσίασης πληροφορίας. Αυτό συνέβη διότι εφόσον έχουμε έτοιμο το πλήρες κείμενο βαθμολογημένο μπορούμε να δείξουμε κομμάτι αυτού στο χρήστη ανάλογα με το μέγεθος της συσκευής που διαθέτει. Από την άλλη και προκειμένου να παρέχουμε προσωποποιημένες περιλήψεις προς τους χρήστες αρκεί να διαλέξουμε το μέγεθος κειμένου που θέλουμε να παρουσιάσουμε.

Από την άλλη ο μηχανισμός αυτός είναι ένας μηχανισμός που δε χρησιμοποιείται για κανέναν άλλο λόγο εκτός του να παρουσιάσει σημαντικό κομμάτι του κειμένου προς τους χρήστες αντί για ολόκληρο το κείμενο. Δεδομένης της κατάστασης του διαδικτύου και δεδομένου ότι πλέον πολλά άρθρα έχουν πολύ μικρό μέγεθος ενώ όποιος διαβάζει αναλύσεις άρθρων γνωρίσει ότι θα διαβάσει ένα μακροσκελές κείμενο δεν κρίνεται σκόπιμο να συνεχίσει η λειτουργία αυτού του μηχανισμού. Όπως αναφέρθηκε από τη μία έχουμε άρθρα μικρού μεγέθους, πολύ περιεκτικά που συχνά απλώς κάνουν αναφορά σε μία είδηση και συνήθως παρουσιάζονται σε συνέχειες όπως προκύπτουν τα γεγονότα και από την άλλη έχουμε τα πολύ μεγάλα άρθρα που έχουν αναλύσεις, τις οποίες λίγοι διαβάζουν και συνήθως μία περίληψη αυτών δεν είναι αντιπροσωπευτική του περιεχομένου ειδικά αν το άρθρα που αναλύει δείχνει μία συγκεκριμένη οπτική των πραγμάτων. Ως εκ τούτου ο μηχανισμός εξαγωγής περίληψης είτε θα πρέπει να αλλάξει τελείως κατεύθυνση λειτουργίας, είτε θα πρέπει να καταργηθεί.

8.6 Μηχανισμός Προσωποποίησης στο Χρήστη

Το σημαντικό κομμάτι της παρουσίασης πληροφορίας και προσωποποίησης της πληροφορίας στο χρήστη είναι ίσως το σύστημα που αγγίζει περισσότερο από κάθε άλλο του συστήματος την εξέλιξη της τεχνολογίας διαδικτύου όπως συμβαίνει με ραγδαίους ρυθμούς τα τελευταία χρόνια. Έχει ήδη παρατηρηθεί στο σύστημα που αναπτύσουμε πως κάτι που σχεδιάζεται σήμερα και υλοποιείται μέσα σε ένα μήνα και αφορά την παρουσίαση άρθρων αλλά και τα δεδομένα προσωποποίησης στο χρήστη μπορεί να θεωρείται σε πολύ μικρό διάστημα παρωχημένο. Ως εκ τούτου, ο σχεδιασμός που έγινε για την παρούσα εργασία προσπαθεί να εξασφαλίσει το ελάχιστο. Ότι δηλαδή ο τρόπος με τον οποίο συντελείται η δόμηση της πληροφορίας αλλά και τα δεδομένα που εισέρχονται στο μηχανισμό προσωποποίησης γίνονται με τέτοιο τρόπο ώστε να αποτελούν μία ενιαία αναλοίωτη βάση παρά τις αλλαγές που θα πραγματοποιούνται στο περιτύλιγμα που αφορά την παρουσίαση ή τη συλλογή δεδομένων.

Οι επεκτάσεις που μπορεί να γίνουν τόσο στο σύστημα παρουσίασης πληροφορίας όσο και στο σύστημα προσωποποίησης στο χρήστη είναι πολλές και αφορούν τόσο την παρουσίαση της πληροφορίας όσο και τον τρόπο προσωποποίησης στο χρήστη. Από τη μία η παρουσίαση της πληροφορίας στο χρήστη είναι σίγουρα ζήτημα αντίληψης του κάθε χρήστη αλλά και του κάθε σχεδιαστή. Στην πιο πρόσφατη έκδοση του συστήματος έγινε μία προσπάθεια να απεμπλακούν τελείως τα δεδομένα από τη δομή του δικτυακού τόπου. Ωστόσο, επειδή για πρώτη φορά έγινε πλήρης χρήση από το σύστημα εργαλείων web2.0 ώστε το περιβάλλον να προσωμοιάζει περισσότερο ένα application παρά ένα δικτυακό τόπο δεν επιτύχαμε να πετύχουμε 100% ανεξαρτητοποίηση ανάμεσα στα δύο. Ως εκ τούτου, οπωσδήποτε μία πρώτη πρόταση που έχουμε για το σύστημα είναι η πλήρης απεμπλοκή των συστημάτων συγκέντρωσης και δόμησης πληροφολρίας και του συστήματος παρουσίασης πληροφορίας. Με αυτό τον τρόπο θα γίνει εφικτό ο χρήστης να μπορεί να χτίζει το δικό του layout αλλά και να προσφέρεται πληθώρα από skins ανάλογα τι αρέσει στον κάθε χρήστη. Σίγουρα, το θέμα της παρουσίασης πληροφορίας μπορεί να αποτελέσει από μόνο του μία ολόκληρη εργασία που οπωσδήποτε περιλαμβάνει τόσο ερευνητικές περιοχές όπως αυτή της αλληλεπίδρασης ανθρώπου υπολογιστή όσο και αυτή που έχει να κάνει αμιγώς με το σχεδιασμό ιστοσελίδων, τη γραφιστική και την αντίληψη που έχουν οι web designers. Αυτό το κομμάτι είναι το κομμάτι του customization και φυσικά δεν προσφέρει τίποτα περισσότερο από απλά διαφορετική οπτική εμπειρία στον κάθε χρήστη.

Σίγουρα ένα κομμάτι του peRSSonal το οποίο παρουσιάζει τεράστιο ενδιαφέρον σχετίζεται απόλυτα με την προσωποποίηση πληροφορίας στο χρήστη. Όπως είδαμε το σύστημα εφαρμόζει μία σειρά από τεχνικές για να μπορεί να εξασφαλίσει ποιοτική προσαρμογή στο προφίλ ενός χρήστη και με αυτό τον τρόπο να μπορεί να παρουσιάζει στον κάθε χρήστη αποκλειστικά και μόνο πληροφορία που τον ενδιαφέρει. Τα στοιχεία που θα μπορούσε ο καθένας να αναφέρει ως αυτά που μπορεί να αφορούν το προφίλ ενός χρήστη σίγουρα μπορεί να αποτελέσουν πληροφορία εισόδου για το σύστημα προσωποποίησης. Όπως έχουμε ήδη αναφέρει, μέχρι αυτή τη στιγμή

αναλύουμε στοιχεία που έχουν να κάνουν τόσο με τον τρόπο περιαγωγής του χρήστη όσο και με μετρικές που σχετίζονται απόλυτα με την αίσθηση που έχουν οι δοκιμαστικοί χρήστες του συστήματός μας ο οποίοι περιορίζονται στα άτομα της ομάδας εργασίας. Πέρα, λοιπόν, από τις μετρικές που βρήκαμε στη βιβλιογραφία και αυτές που εφαρμόσαμε βάσει της λειτουργίας που κάνουμε στο σύστημα θα μπορούσε κανείς να πει ότι υπάρχει πληθώρα στοιχείων τα οποία μπορούν να συνεισφέρουν στη διαμόρφωση του προφίλ ενός χρήστης και συνεπώς στη συνολική προσωποποίηση που πραγματοποιείται στο σύστημα. Κάποια βασικά πράγματα έχουν ωστόσο και σε αυτό το στάδιο να κάνουν με εμπλουτισμό των λέξεων κλειδιών με συνώνυμες ή αντίθετες και μάλιστα αυτό να συμβαίνει σε όλη τη διάρκεια της ανανέωσης του προφίλ και όχι μόνο στη φάση της δημιουργίας αυτού. Είναι προφανές φυσικά πως ο κάθε χρήστης έχει και τα δικά του προσωπικά χαρακτηριστικά όχι μόνο όταν επιλέγει να κάνει κάτι αλλά και για το λόγο για τον οποίο ενεργεί. Αυτό σημαίνει πως ενώ εμείς αξιολογούμε σαν κάτι πολύ σημαντικό το γεγονός ότι ένας χρήστης που διαβάζει ένα αρχείο από τις σελίδες του peRSSonal μπορεί να αποφασίσει να μεταβεί στην πραγματική σελίδα του άρθρου μπορεί για τον ίδιο το χρήστη να μη σημαίνει τίποτα. Αυτό πραγματοποιείται διότι το δείγμα χρηστών από το οποίο εξάγαμε την πληροφορία για τη σημαντικότητα κάθε ενέργειας δεν ήταν μεγάλο. Πραγματικά, κάτι το οποίο αξίζει να γίνει είναι αξιολόγηση του μηχανισμού από μεγάλο αριθμό χρηστών και όχι μόνο από μερικούς χρήστες πολλοί από τους οποίους δημιουργούνται συχνά από το ίδιο φυσικό πρόσωπο για δοκιμές των βασικών συστατικών του συστήματος.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Face-to-face vs. cyberspace: Finding the middle ground author=. Technical report.

[2] Βασίλειος Πουλόπουλος. Προσωποποιημένη Προβολή Περιεχομένου του Διαδικτύου με τεχνικές Προεπεξεργασίας, Αυτόματης Κατηγοριοποίησης και Αυτόματης Εξαγωγής Περίληψης. Master's thesis, 2007.

[3] Proposal for an open profiling standard. http://www.w3.org/TR/NOTE-OPS-FrameWork, June 1997.

[4] Pidl - personalized information description language. http://www.w3.org/TR/1999/NOTE-PIDL-19990209, February 1999.

[5] Learning user interest dynamics with a three-descriptor representation. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):212--225, 2001.

[6] The platform for privacy preferences 1.0 (p3p1.0) specification. http://www.w3.org/TR/P3P/, April 2002.

[7] Extensible markup language (xml) 1.0. http://www.w3.org/TR/REC-xml/, November 2008.

[8] Amazon.com. http://www.amazon.com, June 2010.

[9] The associated press - the essential global news network. http://www.ap.org/, June 2010.

[10] Bbc. http://www.bbc.co.uk, June 2010.

[11] Bloomberg - business & financial news, breaking news headlines. http://www.bloomberg.com/, June 2010.

[12] Cnn.com international - breaking, world, business, sports, entertainment and video news. http://www.cnn.com, June 2010.

[13] Composite capability/preference profiles (cc/pp): Structure and vocabularies 2.0. http://www.w3.org/TR/2010/NOTE-CCPP-struct-vocab2-20100629/, June 2010.

[14] Dom, document object model w3c standard. http://www.w3.org/DOM/, June 2010.

[15] Extensible markup language (xml). http://www.w3.org/XML/, June 2010.

[16] Facebook. http://www.facebook.com, June 2010.

[17] Foxnews.com - breaking news, latest news, current news. http://www.foxnews.com/, June 2010.

[18] Gnu wget - gnu project - free software foundation. http://www.gnu.org/software/wget/, June 2010.

[19] Google news. http://news.google.gr/, June 2010.

[20] Google news. http://www.google.com, June 2010.

[21] Heritrix - internet archive's open-source, extensible, web-scale, archival-quality web crawler project. http://crawler.archive.org/, June 2010.

[22] ht://dig - internet search engine software. http://www.htdig.org/, June 2010.

[23] Html, hyper-text markup language, html 4.01 specification, w3c standard. http://www.w3.org/TR/REC-html40/, June 2010.

[24] Httrack website copier - offline browser. http://www.httrack.com/, June 2010.

[25] Larbin web crawler. http://larbin.sourceforge.net/index-eng.html, June 2010.

[26] Linkedin | relationships matter. http://www.linkedin.com/, June 2010.

[27] Methabot web crawler. http://bithack.se/methabot/, June 2010.

[28] Normal distribution by wolfram mathworld. http://mathworld.wolfram.com/NormalDistribution.html, June 2010.

[29] Nutch open source web search engine. http://lucene.apache.org/nutch/, June 2010.

[30] Passport.net. http://www.passport.net, June 2010.

[31] Rdf site summary (rss) 1.0. http://web.resource.org/rss/1.0/spec, June 2010.

[32] Reuters - business and financial news, breaking us & international news | reuters.com. http://www.reuters.com/, June 2010.

[33] The top news headlines on current events from yahoo! news - yahoo! news. http://news.yahoo.com/, June 2010.

[34] Twitter. http://www.twitter.com/, June 2010.

[35] Web information retrieval environment (wire). http://www.cwr.cl/projects/WIRE/, June 2010.

[36] Websphinx: A personal, customizable web crawler. http://www.cs.cmu.edu/~rcm/websphinx/, June 2010.

[37] Wikipedia: News aggregator. http://en.wikipedia.org/wiki/Feed_aggregator, June 2010.

[38] George Adam, Kostas Asimakis, Christos Bouras, and Vassilis Poulopoulos. An efficient mechanism for stemming and tagging: the case of greek language. In *Advanced Knowledge - based Systems, Invited Session of the 14th International Conference on Knowledge - based and Intelligent Information & Engineering Systems*, Cardiff Wales, UK, September 10 2010.

[39] George Adam, Christos Bouras, and Vassilis Poulopoulos. Cuter: An efficient useful text extraction mechanism. In *The 2009 IEEE International Symposium on Mining and Web(WAM09)*, pages 125 -- 130, Bradford, UK, 26 - 29 May 2009.

[40] George Adam, Christos Bouras, and Vassilis Poulopoulos. Monitoring rss feeds. In *International Conference on Knowledge Management and Knowledge Technologies (I-KNOW 09)*, Gratz, Austria, 2 - 4 September 2009.

[41] George Adam, Christos Bouras, and Vassilis Poulopoulos. Utilizing rss feeds for crawling the web. In *IADIS European Conference on Data Mining*, Algavre, Portugal, June 18 - 20 2009.

[42] George Adam, Christos Bouras, and Vassilis Poulopoulos. Image extraction from online text streams. In *The 2010 IEEE International Symposium on Mining and Web (MAW10)*, pages 609--614, Perth, Australia, April 20 - 23 2010.

[43] 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, 1999.

[44] Ioannis Antonellis, Christos Bouras, and Vassilis Poulopoulos. Personalized news categorization through scalable text classification. In *Proceedings of the 8th Asia-Pacific Web Conf*, pages 391--401. Springer, 2006.

[45] Chinatsu Aone, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen. *A trainable summarizer with knowledge acquired from robust NLP techniques*, pages 71--80. 1999.

[46] Chidanand Apté, Fred Damerau, and Sholom M. Weiss. Towards language independent automated learning of text categorization models. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 23--30, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[47] Hiroki Arimura, Atsushi Wataki, Ryoichi Fujino, and Setsuo Arikawa. A fast algorithm for discovering optimal string patterns in large text databases. In *ALT '98: Proceedings of the 9th International Conference on Algorithmic Learning Theory*, pages 247--261, London, UK, 1998. Springer-Verlag.

[48] Alejandro Pena Ayala. Student modelling based on ontologies. In *ACIIDS '09: Proceedings of the 2009 First Asian Conference on Intelligent Information and Database Systems*, pages 392--397, Washington, DC, USA, 2009. IEEE Computer Society.

[49] Ricardo A. Baeza-Yates. Introduction to data structures and algorithms related to information retrieval. pages 13--27, 1992.

[50] R. Barzilay and L. Lee. Catching the drift: probabilistic content models, with applications to generation and summarization. In *In HLT-NAACL 2004: Proceedings of the Main Conference*, pages 113--120, 2004.

[51] Sotiris Batsakis, Euripides G. M. Petrakis, and Evangelos Milios. Improving the performance of focused web crawlers. *Data Knowl. Eng.*, 68(10):1001--1013, 2009.

[52] Adam L. Berger and Vibhu O. Mittal. Ocelot: a system for summarizing web pages. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 144--151, New York, NY, USA, 2000. ACM.

[53] Helmut Berger and Dieter Merkl. A comparison of text-categorization methods applied to n-gram frequency statistics. In *Australian Conference on Artificial Intelligence*, pages 998--1003, 2004.

[54] Michael K. Bergman. The deep web: Surfacing hidden value. *Journal of Electronic Publishing In TAKING LICENSE: Recognizing a Need to Change*, 7(1).

[55] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American Magazine*, may 2001.

[56] H. W. Beyer. Crc standard mathematical tables. *Boca Raton, FL: CRC Press*, pages 533--534, 1987.

[57] Paolo Boldi, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. Ubicrawler: a scalable fully distributed web crawler. *Softw. Pract. Exper.*, 34(8):711--726, 2004.

[58] Christos Bouras and Vassilis Poulopoulos. Dynamic user context web personalization in meta - portals. In *The Fifteenth IEEE Symposium on Computers and Communications (ISCC'10)*, Riccione, Italy, June 22 - 25 2010. best paper award.

[59] Christos Bouras, Vassilis Poulopoulos, and Panagiotis Silintziris. Date - based dynamic caching mechanism. In *IADIS European Conference on Data Mining*, Algavre, Portugal, June 18 - 20 2009.

[60] Christos Bouras, Vassilis Poulopoulos, and Panagiotis Silintziris. Personalized news search in www: Adapting on user's behavior. In *The Fourth International Conference on Internet and Web ASpplications and Services - ICIW 2009*, pages 125 -- 130, Venice, Italy, June 18 - 20 2009.

[61] Christos Bouras, Vassilis Poulopoulos, and George Tsichritzis. Trash article detection using categorization techniques. In *IADIS International Conference Applied Computing*, Rome, Italy, November 19 - 21 2009.

[62] Christos Bouras, Vassilis Poulopoulos, and Vassilis Tsogkas. Efficient summarization based on categorized keywords. In *DMIN*, pages 285--291, 2007.

[63] Christos Bouras, Vassilis Poulopoulos, and Vassilis Tsogkas. Creating dynamic personalized rss summaries. In *8th Industrial Conference on Data Mining β€“ ICDM 2008*, pages 1 -- 15, Leipzig, Germany, 16 - 18 July 2008.

[64] Christos Bouras, Vassilis Poulopoulos, and Vassilis Tsogkas. Perssonal's core functionality evaluation: Enhancing text labeling through personalized summaries. *Data & Knowledge Engineering*, 64(1):330 -- 345, 2008.

[65] Christos Bouras, Vassilis Poulopoulos, and Vassilis Tsogkas. Adaptation of rss feeds based on the user profile and on the end device. *J. Netw. Comput. Appl.*, 33(4):410--421, 2010.

[66] J.P. Bowen and S. Filippini-Fantoni. Personalization and the web from a museum perspective.

[67] danah boyd. Identity production in a networked culture: Why youth heart myspace. In *AAAS 2006*, St. Louis, Missouri, February 19 2006.

[68] B. Bredeweg and R.G.F. Winkels. Student modelling through qualitative reasoning. In J. Greer, editor, *Student Modelling: The Key to Individualized Knowledge--Based Instruction*, pages 63--98. Springer Verlag, Berlin, 1994. NATO ASI--Series F: Computer and Systems Sciences, Vol. 125.

[69] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107--117, 1998.

[70] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335--336, New York, NY, USA, 1998. ACM.

[71] Maria Fernanda Caropreso, Stan Matwin, and Fabrizio Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. pages 78--102, 2001.

[72] James Caverlee and David Buttler. Discovering objects in dynamically-generated web pages, 2003.

[73] Philip K. Chan. Constructing web user profiles: A non-invasive learning approach. In *In Web Usage Analysis and User Profiling, LNAI 1836*, pages 39--55. Springer-Verlag, 2000.

[74] Chia-Hui Chang and Shao-Chen Lui. Iepad: information extraction based on pattern discovery. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 681--688, New York, NY, USA, 2001. ACM.

[75] David Maxwell Chickering, David Heckerman, and Christopher Meek. A bayesian approach to learning bayesian networks with local structure. In *In Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 80--89. Morgan Kaufmann, 1997.

[76] Roger Clarke. Web 2.0 as syndication. *J. Theor. Appl. Electron. Commer. Res.*, 3(2):30--43, 2008.

[77] Cyril Cleverdon. The cranfield tests on index language devices. pages 47--59, 1997.

[78] William W. Cohen. Text categorization and relational learning. In *In Proceedings of the Twelfth International Conference on Machine Learning*, pages 124--132. Morgan Kaufman, 1995.

[79] Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of the 18th conference on Computational linguistics*, pages 201--207, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[80] 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.

[81] Vanessa Paz Dennen. Constructing academic alter-egos: identity issues in a blog-based community. *Identity in the Information Society*, 2(1), DECEMBER 2009.

[82] Hai Dong, Farookh Khadeer Hussain, and Elizabeth Chang. State of the art in semantic focused crawlers. In *ICCSA '09: Proceedings of the International Conference on Computational Science and Its Applications*, pages 910--924, Berlin, Heidelberg, 2009. Springer-Verlag.

[83] Lauren B. Doyle. Semantic road maps for literature searchers. *J. ACM*, 8(4):553--578, 1961.

[84] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pages 148--155, New York, NY, USA, 1998. ACM.

[85] H. P. Edmundson. New methods in automatic extracting. *J. ACM*, 16(2):264--285, 1969.

[86] Jenny Edwards, Kevin McCurley, and John Tomlin. An adaptive model for optimizing performance of an incremental web crawler. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 106--113, New York, NY, USA, 2001. ACM.

[87] Robert Epstein. The truth about online dating, February 2007.

[88] 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.

[89] W. Feller. An introduction to probability theory and its applications. *New York: Wiley*, 1968.

[90] Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Fact or fiction: Content classification for digital libraries, 2001.

[91] Andrew T. Fiore, Lindsay Shaw Taylor, G.A. Mendelsohn, and Marti Hearst. Assessing attractiveness in online dating profiles. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 797--806, New York, NY, USA, 2008. ACM.

[92] William B. Frakes and Christopher J. Fox. Strength and similarity of affix removal stemming algorithms. *SIGIR Forum*, 37(1):26--30, 2003.

[93] William J. Frawley, Gregory Piatetsky-shapiro, and Christopher J. Matheus. Knowledge discovery in databases: an overview, 1992.

[94] James C. French, Allison L. Powell, Jamie Callan, Charles L. Viles, Travis Emmitt, and Kevin J. Prey. Comparing the performance of database selection algorithms. Technical report, Charlottesville, VA, USA, 1999.

[95] Johannes F?rnkranz, Tom Mitchell, and Ellen Riloff. A case study in using linguistic phrases for text categorization on the www. In *In Working Notes of the AAAI/ICML Workshop on Learning for Text Categorization*, pages 5--12. AAAI Press, 1998.

[96] Danilo Fum, Giovanni Guida, and Carlo Tasso. Forward and backward reasoning in automatic abstracting. In *Proceedings of the 9th conference on Computational linguistics*, pages 83--88, , Czechoslovakia, 1982. Academia Praha.

[97] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 482--490, New York, NY, USA, 2004. ACM.

[98] Jeremy Goecks and Jude Shavlik. Automatically labeling web pages based on normal user actions, 1999.

[99] R. P. Goldman and E. Cherniak. A language for construction of belief networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(3):196--208, 1993.

[100] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop on Automatic Summarization*, pages 40--48, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[101] John Grohol. Anonymity and online community: Identity matters. http://www.alistapart.com/articles/identitymatters, April 4 2006.

[102] David A. Grossman and Ophir Frieder. *Information Retrieval: Algorithms and Heuristics*. Springer, 2nd edition, 2004.

[103] Suhit Gupta, Gail E. Kaiser, Peter Grimm, Michael F. Chiang, and Justin Starren. Automating content extraction of html documents. *WORLD WIDE WEB - INTERNET AND INFORMATION SYSTEMS*, pages 179--224, 2005.

[104] U. Hahn and U. Reimer. Semantic parsing and summarizing of technical texts in the topic system. *Informations linguistik*, pages 153--183, 1986.

[105] Jeffrey T. Hancock, Catalina Toma, and Nicole Ellison. The truth about lying in online dating profiles. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 449--452, New York, NY, USA, 2007. ACM.

[106] Jeffrey T. Hancock, Catalina Toma, and Nicole Ellison. The truth about lying in online dating profiles. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 449--452, New York, NY, USA, 2007. ACM.

[107] Dick Hardt. How sxip works. https://sxip.org/docs/specs/how-sxip-works.pdf2004. whitepaper.

[108] Allan Heydon and Marc Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219--229, 1999.

[109] Eduard Hovy and Chin-Yew Lin. Automated text summarization and the summarist system. In *Proceedings of a workshop on held at Baltimore, Maryland*, pages 197--214, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[110] Qinghua Hu, Daren Yu, and Zongxia Xie. Neighborhood classifiers. *Expert Syst. Appl.*, 34(2):866--876, 2008.

[111] P. S. Jacobs and Lisa F. Rau. Scisor: extracting information from on-line news. *Commun. ACM*, 33(11):88--97, 1990.

[112] T. Joachims, D. Freitag, and Tom Mitchell. Webwatcher: A tour guide for the world wide web. In *Proceedings of the 1997 IJCAI*, August 1997.

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

[114] Ioannis Katakis, Grigorios Tsoumakas, Evangelos Banos, Nick Bassiliades, and Ioannis Vlahavas. An adaptive personalized news dissemination system. *J. Intell. Inf. Syst.*, 32(2):191--212, 2009.

[115] Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun'ichi Tsujii. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 1--8, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[116] Hoang Kiem and Do Phuc. Discovering motiv based association rules in a set of dna sequences. In *RSCTC '00: Revised Papers from the Second International Conference on Rough Sets and Current Trends in Computing*, pages 386--390, London, UK, 2001. Springer-Verlag.

[117] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression, 2002.

[118] Robert Krovetz. Viewing morphology as an inference process. Technical report, Amherst, MA, USA, 1993.

[119] Robert Krovetz. Viewing morphology as an inference process. Technical report, Amherst, MA, USA, 1993.

[120] Chandan Kumar, Prasad Pingali, and Vasudeva Varma. Generating personalized summaries using publicly available web documents. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 103--106, Washington, DC, USA, 2008. IEEE Computer Society.

[121] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68--73, New York, NY, USA, 1995. ACM.

[122] F.W. Lancaster and E.G. Fayen. Information retrieval on-lin. 1973.

[123] Steve Laurence and C. Lee Giles. Accessibility of information on the web. *Nature*, (400), 1999.

[124] Edward Lazowska, David Notkin, Brian Pinkerton, and Brian Pinkerton. Webcrawler: Finding what people want, 2000.

[125] Michael Learmonth. Want 5,000 more facebook friends? that'll be 654.30 dolars. Technical report, September 2009.

[126] Barry M. Leiner, Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff. A brief history of the internet. *SIGCOMM Comput. Commun. Rev.*, 39(5):22--31, 2009.

[127] M. Lennon, D. Pierce, and P Tarry, B..and Willett. An evaluation of the stemming algorithms. Technical report, 1981.

[128] Mark Levene. An introduction to search engines and web navigation. *Pearson*, 2005.

[129] David D. Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In *In Third Annual Symposium on Document Analysis and Information Retrieval*, pages 81--93, 1994.

[130] G. Linden. Personalized search primer -- and google's approach, August 2007.

[131] Bing Liu. Web data mining: Exploring hyperlinks, contents and usage data. *Springer*, 2007.

[132] Ling Liu, Calton Pu, and Wei Tang. Webcq - detecting and delivering information changes on the web. In *In Proc. Int. Conf. on Information and Knowledge Management (CIKM*, pages 512--519. ACM Press, 2000.

[133] J.B. Lovins. Development of a stemming algorithm. Technical report, 1968.

[134] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159--165, 1958.

[135] Jayant Madhavan, David Ko, Lucja Kot, Vignesh Ganapathy, Alex Rasmussen, and Alon Halevy. Google's deep web crawl. *Proc. VLDB Endow.*, 1(2):1241--1252, 2008.

[136] Inderjeet Mani. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA, 1999.

[137] Inderjeet Mani and Eric Bloedorn. Summarizing similarities and differences among related documents. *Inf. Retr.*, 1(1-2):35--67, 1999.

[138] Inderjeet Mani and George Wilson. Robust temporal processing of news. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 69--76, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[139] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[140] Massimo Marchiori. The quest for correct information on the web: hyper search engines. *Computer Networks and ISDN Systems*, 29(8-13):1225 -- 1235, 1997. Papers from the Sixth International World Wide Web Conference.

[141] Daniel Marcu. The rhetorical parsing of natural language texts. In *ACL-35: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 96--103, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[142] Bernd Marcus, Franz Machilek, and Astrid Schutz. Personality in cyberspace: Personal web sites as media for personality expressions and impressions. *Journal of Personality and Social Psychology*, 90(6):1014--1031, 2006.

[143] Brij Masand, Gordon Linoff, and David Waltz. Classifying news stories using memory based reasoning. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 59--65, New York, NY, USA, 1992. ACM.

[144] N. Mathe and J.R. Chen. User-centered indexing for adaptive information access. *International Journal of User Modeling and User Adapted Interaction*, 6(2--3):225--261, 1996.

[145] L.A. Mather and J. Note. Discovering encyclopedic structureand topics in text laura a. matherbritannica.com, inc.

[146] Kathleen McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74--82, New York, NY, USA, 1995. ACM.

[147] Kathleen R. McKeown, Vasileios Hatzivassiloglou, Regina Barzilay, Barry Schiffman, David Evans, and Simone Teufel. Columbia multi-document summarization: Approach and evaluation. In *In Proceedings of the Document Understanding Conference (DUC01*, 2001.

[148] Stefano Mizzaro and Carlo Tasso. Ephemeral and persistent personalization in adaptive information access to scholarly publications on the web. In *AH '02: Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 306--316, London, UK, 2002. Springer-Verlag.

[149] Dunja Mladenic and Marko Grobelnik. Word sequences as features in text-learning. In *In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98*, pages 145--148, 1998.

[150] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142--151, 2000.

[151] Calvin N. Mooers. Information retrieval viewed as temporal signaling. In *Proceedings of the International Congress of Mathematicians*, 1952.

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

[153] Thierry Nabeth. Understanding the identity concept in the context of digital social environments. 2005.

[154] Marc Najork and Janet L. Wiener. Breadth-first crawling yields high-quality pages. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 114--118, New York, NY, USA, 2001. ACM.

[155] Gonzalo Navarro and Ricardo Baeza-Yates. Proximal nodes: a model to query document databases by content and structure. *ACM Trans. Inf. Syst.*, 15(4):400--435, 1997.

[156] H.T. Ng, W.B. Goh, and K.L. Low. Feature selection, perception learning, and a usability case study for text categorization. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 67--73, 1997.

[157] Chikashi Nobata, Nigel Collier, and Jun ichi Tsujii. Automatic term identification and classification in biology texts. In *In Proc. of the 5th NLPRS*, pages 369--374, 1999.

[158] G. Ntais. Development of a stemmer for the greek language. Master's thesis, 2007.

[159] José A. Olivas. Fuzzy sets and web meta-search engines. In *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*, pages 537--552. 2008.

[160] C.D. Paice. Another stemmer. *ACM SIGIR Forum*, 23(1):56--61, 1990.

[161] Odile Paliès, Michel Caillot, Evelyne Cauzinille-Marmèche, Jean-Louis Laurière, and Jaques Mathieu. Student modelling by a knowledge-based system. *Computational Intelligence*, 2:99--107, 1986.

[162] Michael Pazzani, Jack Muramatsu, and Daniel Billsus. Syskill & webert: Identifying interesting web sites. In *In Proc. 13th Natl. Conf. on Artificial Intelligence*, pages 54--61, 1998.

[163] Joseph J. Pollock and Antonio Zamora. Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15(4):226--232, 1975.

[164] M. Porter. The porter stemming algorithm. Technical report.

[165] Dragomir R. Radev, Hongyan Jing, Malgorzata Sty, and Daniel Tam. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919--938, 2004.

[166] A. F. R. Rahman, H. Alam, and year = 2001 R. Hartono title = Content Extraction from HTML Documents, journal = In 1st Int. Workshop on Web Document Analysis (WDA2001).

[167] PC Reghu Raj and S. Raman. Content identification and semantic indexing of text documents. In *Proceedings Of the Indo European Conference on Multilingual Communication Technologies (IEMCT-02)*, pages 203--217, 2002.

[168] Ellen Riloff and Jessica Shepherd. A corpus-based approach for building semantic lexicons. In *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117--124, 1997.

[169] Stephen E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. pages 143--160, 1988.

[170] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613--620, 1975.

[171] Gerard. Salton. *Automatic Information Organization and Retrieval.* McGraw Hill Text, 1968.

[172] Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. pages 413--418, 1999.

[173] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. pages 355--364, 1997.

[174] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Inf. Process. Manage.*, 33(2):193--207, 1997.

[175] M. Saravanan, Pc Reghu Raj, and S. Raman. Summarization and categorization of text data in high-level data cleaning for information retrieval, 2003.

[176] M. Saravanan and S. Raman. The term distribution model for summarization of multiple documents. In *Proceedings of the Indo European Conference on Multilingual Communication Technologies (IEMCT 2002)*, pages 182--192, 2002.

[177] R.C. Schank. *Reading and Understanding: Teaching from the Perspective of Artificial Intelligence*. Lawrence Erlbaum Associates, 1982.

[178] Jude Shavlik, Susan Calcari, Tina Eliassi-Rad, and Jack Solock. An instructable, adaptive interface for discovering and monitoring information on the world-wide web. In *IUI '99: Proceedings of the 4th international conference on Intelligent user interfaces*, pages 157--160, New York, NY, USA, 1999. ACM.

[179] Ben Shneiderman, Donald Byrd, and W. Bruce Croft. Sorting out searching: a user-interface framework for text searches. *Commun. ACM*, 41(4):95--98, 1998.

[180] A. Siibak. Casanovas of the virtual world. how boys present themselves on dating websites. In *Young People at the Crossroads: 5th International Conference on Youth Research*, pages 83 -- 91, Petrozavodsk, Republic of Karelia, Russian Federation, September 1-5 2007.

[181] Manfred Stede. Lexical paraphrases in multilingual sentence generation. In *Machine Translation*, pages 75--107, 1996.

[182] D Sullivan. Google ramps up personalized search, February 2007.

[183] Russell Swan and James Allan. Automatic generation of overview timelines. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49--56, New York, NY, USA, 2000. ACM.

[184] John I. Tait. Spotting and discovering terms through natural language processing by christian jacquemin. the mit press. *Nat. Lang. Eng.*, 10(2):195--196, 2004.

[185] Nabeth Thierry. D2.3: Models, fidis deliverable. Technical report, October 2005. FIDIS project.

[186] M. Triantafillidis. *Modern Greek Grammar (Dimotiki) (in Greek). Reprint with corrections*. Institute of Modern Greek Studies, Thessaloniki (1941), 1978.

[187] Zeynep Tufekci. Can you see me now? audience and disclosure regulation in online social network sites, 2008.

[188] H. Turtle and W. B. Croft. Inference networks for document retrieval. In *SIGIR '90: Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1--24, New York, NY, USA, 1990. ACM.

[189] Kostas Tzeras and Stephan Hartmann. Automatic indexing based on bayesian inference networks. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 22--35, New York, NY, USA, 1993. ACM.

[190] Johan van Wamelen and Dennis de Kool. Web 2.0: a basis for the second society? In *ICEGOV '08: Proceedings of the 2nd International Conference on Theory and Practice of Electronic Governance*, pages 349--354, New York, NY, USA, 2008. ACM.

[191] Asimina Vasalou and Adam N. Joinson. Me, myself and i: The role of interactional context on self-presentation through avatars. *Comput. Hum. Behav.*, 25(2):510--520, 2009.

[192] Jean-Noël Vittaut and Patrick Gallinari. Machine learning ranking for structured information retrieval. In *Proc. ECIR*, pages 338--349, 2006.

[193] Nina Wacholder, Dvid K. Evans, and Judith L. Klavans. Automatic identification and organization of index terms for interactive browsing. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 126--134, New York, NY, USA, 2001. ACM.

[194] Chirayu Wongchokprasitti and Peter Brusilovsky. Newsme: A case study for adaptive news systems with open user model. In *ICAS '07: Proceedings of the Third International Conference on Autonomic and Autonomous Systems*, page 69, Washington, DC, USA, 2007. IEEE Computer Society.

[195] WA Woods and JG Schmolze. The kl-one family. semantic networks in artificial intelligence. pages 133--178, 1992.

[196] Manuel Montes y Go'mez, Alexander Gelbukh, and Aurelio Lo'pez-Lo'pez. Mining the news: Trends, associations, and deviations. *COMPUTATIONS AND SYSTEMS*, 5(1):14--24, 2001.

[197] Yiming Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *SIGIR*, pages 13--22, 1994.

[198] Yiming Yang and Christopher G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Trans. Inf. Syst.*, 12(3):252--277, 1994.

[199] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412--420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[200] Demetrios Zeinalipour-Yazti and Marios D. Dikaiakos. Design and implementation of a distributed crawler and filtering processor. In *NGITS '02: Proceedings of the 5th International Workshop on Next Generation Information Technologies and Systems*, pages 58--74, London, UK, 2002. Springer-Verlag.

ΓΛΩΣΣΑΡΙ

ad-hoc γι' αυτό το σκοπό. 52

Association Patterns Πρότυπα Συσχέτισης. 36

blog web log. 3, 4, 41

Computational Linguistics Υπολογιστική Γλωσσολογία. 54

corpus Συλλογή Κειμένων με κοινά χαρακτηριστικά. 50

crawler Μηχανισμός που χρησιμοποιείται για τη δημιουργία ενός offline αντιγράφου του Διαδικτύου. 6, 20

CSS Cascading Style Sheets. 38

Data and Knowledge Management Εξόρυξη Δεδομένων και Γνώσης. 31

Data Mining Εξόρυξη Δεδομένων. 31

Ephemeral Associations Εφήμερες Συσχετίσεις. 37

Focused Crawler Εφαρμογή που λειτουργεί όπως ο crawler αλλά κάνει στοχευμένη και ποιοτική ανάκτηση δεδομένων βάσει στοιχείων που τίθενται ακόμα και ως ερωτήματα παράλληλα με το feed url του.. 38

gadgets Χαρακτηρισμός για όλες τις Ηλεκτρονικές Συσκευές Τελευταίας Τεχνολογίας. 11

HMM Hidden Markov Model. 44

HTML hypertext markup language. Η βασική γλώσσα των σελίδων διαδικτύου. 6, 38, 41, 42

Information Retrieval Ανάκτηση Πληροφορίας: όρος που χρησιμοποιείται για να περιγράψει κάθε διαδικασία λήψης μέρους δεδομένων από κάποια συλλογή δεδομένων. 25

Internet Το Διαδίκτυο, ή αλλιώς Παγκόσμιος Ιστός, www, κ.α.. 5

keywords Λέξι Κλειδί. Στο σύστημα που θα αναλύσουμε λέξεις κλειδιά είναι οι ρίζες των λέξεων που εξάγουμε από τα χρήσιμα κείμενα. 16

Knowledge Discovery in Databases Ανάκτηση Γνώσης από Βάσεις Δεδομένων. 37

large scale Ο όρος στην πληροφορική αναφέρεται κυρίως σε συστήματα τα οποία αποτελούνται από πολλούς μηχανισμούς που παράγουν αποτέλεσμα με τη μεταξύ τους συνεργασία. Επίσης τα συστήματα αυτά έχουν μεγάλο όγκο πληροφορίας, μεγάλο αριθμό χρηστών, κλπ.. 19

link Σύνδεσμος του Διαδικτύου - βλ. URL. 28

LSI Latent Semantic Analysis: Λανθάνουσα Σημασιολογικη Δεικτοδότηση. 26

microapplications Εφαρμογή, συνήθως web based, η οποία είναι μικρής κλίμακας και εξυπηρετεί συγκεκριμένο σκοπό. 6

Microsite Πρόκειται για εσωτερικό Δικτυακό τόπο, ένα ολόκληρο site που βρίσκεται εμφωλευμένο σε ένα μεγαλύτερο. 5

NLP Natural Language Processing. 45--47

online Αναφέρεται συνήθως στη σύνδεση στο Διαδίκτυο. 4

PDF Portable Document Format. 38

portal Δικτυακός Τόπος μεγάλης κλίμακας που συνήθως χαρακτηρίζεται σαν πληροφοριακός κόμβος. 3

proxy server Ενδιάμεσος Εξυπηρετητής του Διαδικτύου. 39

RSS Really Simple Syndication. 3

RSS feeds Ονομασία των αρχείων που βασίζονται στο πρωτόκολλο RSS. 5, 19

RSS Reader Πρόγραμμα για ανάγνωση RSS. 3

script Κομμάτι Κώδικα. 27

search engines Μηχανή Αναζήτησης. Υπηρεσία που προσφέρει τη δυνατότητα υποβολής ερωτημάτων και λήψης απαντήσεων από σελίδες του διαδικτύου που σχετίζονται με το ερώτημα. 16

Semantic Web Σημασιολογικός Ιστός. Όρος που χρησιμοποιείται τα τελευταία χρόνια για να περιγράψει το νόημα που έχει η πληροφορία και όχι την ίδια την πληροφορία. 23

set-top box Συσκευές μεσαίου μεγέθους με αρκετή υπολογιστική ισχύ ώστε να προσφέρουν υπηρεσίες συνήθως μέσω cable ή broadband και χρησιμοποιούνται σε συνδυασμό με τηλεόραση. 19

stemmed Πρόκειται για όρο που χρησιμοποιείται στη λεξικολογική ανάλυση και υποδηλώνει τη ρίζα μίας λέξης. 16

stemming Πρόκειται για όρο που χρησιμοποιείται στη λεξικολογική ανάλυση και υποδηλώνει τη ρίζα μίας λέξης. 45

Subtopic Regions Περιοχές Δευτερεύουσας Σημασίας. 54

SVM Support Vector Machines. 44

Text Summarization Διαδικασία Εξαγωγής Περίληψης από Κείμενο. 46

TF-IDF Term Frequency - Inverse Document Frequency. 45

Topic Identification Αναγνώριση Θεμάτων. 54

URL Uniform Resource Locator: Αναφέρεται στη διεύθυνση μίας ιστοσελίδας. 20

web clipping Διαδικασία εξαγωγής μέρους ιστοσελίδας που διαθέτει χρήσιμο κομμάτι πληροφορίας. 41

WWW World Wide Web. 38

XML Extensible Markup Language. 13

Κοινωνία της Πληροφορίας Όρος που αναφέρεται σε όλο τον τομέα της πληροφορικής. 23

ενημερωτικές πύλες Δικτυακοί Τόποι που σαν πρωταρχικό σκοπό έχουν την ενημέρωση για την επικαιρότητα. 12

χρήσιμο κείμενο Χρήσιμο Κείμενο ονομάζουμε το κομμάτι μίας ιστοσελίδας που περιέχει το κείμενο για το οποίο ενδιαφέρεται κυρίως ο μέσος χρήστης. 6, 18