Πανεπιστημιο Πατρων
Τμημα Μηχανικων Η/Υ και Πληροφορικης

Διπλωματικη Εργασια

Στα Πλαισια του Μεταπτυχιακου Διπλωματος Ειδικευσης (ΜΔΕ) 'Επιστημη και Τεχνολογια Υπολογιστων' του Τμηματος Μηχανικων Η/Υ & Πληροφορικης

Προσωποποιημενη Προβολη Περιεχομενου του Διαδικτυου σε Desktop Εφαρμογη με Τεχνικες ανακτησης δεδομενων, προεπεξεργασιας κειμενου, αυτοματης κατηγοριοποιησης και εξαγωγης περιληψης

Τσόγκας Βασίλειος

Α.Μ. 558

Επιβλέπων Καθηγητής:
Χρήστος Μπούρας,
Καθηγητής

Τριμελής Επιτροπή:
1. Χρήστος Μπούρας, Καθηγητής
2. Ευστράτιος Γαλλόπουλος, Καθηγητής
3. Χρήστος Μακρής, Επίκουρος Καθηγητής

Νοεμβριος 2008

'αφιερωμένη στους ανθρώπους που βρήκαν τον τρόπο να με αγαπούν'

Πρόλογος

Ζούμε σε μια κοινωνία αλλαγής και προόδου. Σε μια κοινωνία που χαρακτηρίζεται από τον τεράστιο όγκο της πληροφορίας που διακινείται μέσα στις τάξεις της. Κυρίως όμως διανύουμε την εποχή της κατάργησης των συνόρων και της αδιάλειπτης επικοινωνίας μεταξύ των ανθρώπων. Το διαδίκτυο αποτελεί τον τροχό γι' αυτές τις αλλαγές· η ποσότητα όμως των δεδομένων που υπάρχουν και διακινούνται μέσω αυτού είναι τόσο τεράστια, ώστε να αποσπά τους πολίτες της κοινωνίας αυτής στην προσπάθειά τους να βρουν χρήσιμη πληροφορία και επομένως να μετατρέπεται σε τροχοπέδη της αλλαγής.

Με την πραγματικότητα των υπέρογκων και ολοένα αυξανόμενων πηγών κειμένου στο διαδίκτυο, καθίστανται αναγκαία η ύπαρξη μηχανισμών οι οποίοι βοηθούν τους χρήστες ώστε να λάβουν γρήγορες απαντήσεις στα ερωτήματά τους. Η παρουσίαση προσωποποιημένου, συνοψισμένου και προκατηγοριοποιημένου περιεχομένου στους χρήστες, κρίνεται απαραίτητη σύμφωνα με τις επιταγές της συνδυαστικής έκρηξης της πληροφορίας που είναι ορατή σε κάθε 'γωνία' του διαδικτύου. Ζητούνται άμεσες και αποτελεσματικές λύσεις ώστε να 'τιθασευτεί' αυτό το χάος πληροφορίας που υπάρχει στον παγκόσμιο ιστό, λύσεις που είναι εφικτές μόνο μέσα από ανάλυση των προβλημάτων και εφαρμογή σύγχρονων μαθηματικών και υπολογιστικών μεθόδων για την αντιμετώπισή τους.

Η συγκεκριμένη εργασία αποτελεί κομμάτι μίας προσπάθειας ετών εργασίας και έρευνας μέσα στο χώρο της επιστήμης των υπολογιστών. Είναι τιμή μου που μου δόθηκε αυτή η ευκαιρία να περιδιαβώ τα ενδιαφέροντα μονοπάτια της έρευνας μέσα στο χώρο του Πανεπιστημίου που μετά από 7 χρόνια σχεδόν σπουδών και εργασιών αποτελεί τμήμα του εαυτού μου. Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της εργασίας μου καθηγητή κ. Χρήστο Μπούρα και τον καλό μου φίλο Β. Πουλόπουλο για την ευκαιρία αυτή.

Ομοίως, θέλω να ευχαριστήσω τους καθηγητές του ΤΜΗΥΠ για την τιμή που μου έκαναν να είναι μέλη της τριμελούς επιτροπής, τον καθηγητή κ. Ευστράτιο Γαλλόπουλο και τον επίκουρο καθηγητή κ. Χρήστο Μακρή.

Ευχαριστώ επίσης την Αντιγόνη και την οικογένειά μου για την στήριξη που μου δίνουν τόσα χρόνια σε οτιδήποτε χρειαστώ, πριν το χρειαστώ. Η παρουσία τους και η υπομονή τους είναι το πιο σημαντικό για μένα.

Κλείνοντας θα ήθελα να εκφράζω την ελπίδα μου ότι η διατριβή που έγινε κατά την διάρκεια εκπόνησης της εργασίας πρόσθεσε έστω και ένα μικρό λιθαράκι στο αχανές πεδίο της επιστήμης των υπολογιστών. Ελπίζω η ανάγνωσή της να είναι τόσο ευχάριστη, ενδιαφέρουσα και δημιουργική όσο ήταν η συγγραφή της.

Τσόγκας Βασίλης, Πάτρα, Νοέμβριος 2008

Επιτελική σύνοψη

Σκοπός της Μεταπτυχιακής εργασίας είναι η επέκταση και η βελτίωση του μηχανισμού που δημιουργήθηκε στα πλαίσια της πτυπτυχιακής διπλωματικής εργασίας που εκπόνησα με τίτλο 'Αλγόριθμοι και Τεχνικές Δημιουργίας Περίληψης Κειμένου και Εφαρμογή σε Συσκευές Μικρού Μεγέθους'.

Στα πλαίσια της παραπάνω διπλωματικής εργασίας, δημιουργήθηκε ένας ολοκληρωμένος μηχανισμός ο οποίος μπορεί αυτόματα να αναλύει κείμενα του διαδικτύου προκειμένου να εξάγει λέξεις-κλειδιά. Μέσα από αυτή την ανάλυση προκύπτουν οι σημαντικότερες προτάσεις του κειμένου που το χαρακτηρίζουν και οι οποίες μπορούν, αν συνενωθούν, να αποτελέσουν μια σύντομη περίληψη του κειμένου. Ο μηχανισμός αξιοποιεί γνώσεις για την κατηγορία του κειμένου καθώς και για τις προτιμήσεις που παρουσιάζουν οι χρήστες του προκειμένου να βελτιώσει και να φιλτράρει τα αποτελέσματα που παρουσιάζονται. Το σύστημα που κατασκευάστηκε έχει τα εξής βασικά υποσυστήματα: μηχανισμός ανάκτησης δεδομένων και εξαγωγής χρήσιμου κειμένου από τον παγκόσμιο ιστό, μηχανισμός εξαγωγής λέξεων-κλειδιών από το πηγαίο κείμενο, μηχανισμός κατηγοριοποίησης κειμένου, ο οποίος μπορεί να συμμετάσχει στη διαδικασία εξαγωγής περίληψης και να ενδυναμώσει τα αποτελέσματά της, μηχανισμοί προσωποποίησης περιεχομένου στο χρήστη και φυσικά, μηχανισμός εξαγωγής περίληψης. Οι παραπάνω μηχανισμοί είναι ενσωματωμένοι σε ένα σύστημα αποδελτίωσης, το PeRSSonal, το οποίο χρησιμοποιείται για την ανάκτηση / προεπεξεργασία / κατηγοριοποίηση / προσωποποίηση και περίληψη άρθρων από ειδησεογραφικούς τόπους του διαδικτύου.

Σκοπός της παρούσας εργασίας είναι η ενίσχυση των υπαρχόντων διαδικασιών του μηχανισμού που δημιουργήθηκε με καλύτερες και αποτελεσματικότερες μεθόδους και αλγορίθμους, καθώς και η δημιουργία μιας desktop εφαρμογής που θα αξιοποιεί στο έπακρο τις δυνατότητες παρουσίασης του συστήματος μέσω του κλασικού client-server μοντέλου.

Πιο συγκεκριμένα, αναβαθμίζονται όλα τα στάδια λειτουργίας του μηχανισμού. Έτσι, το στάδιο ανάκτησης δεδομένων από τον ιστό ενισχύεται με έναν νέο, πιο αποτελεσματικό crawler. Ο αλγόριθμος που υλοποιείται σε αυτό το στάδιο λαμβάνει υπ' όψιν του, μεταξύ άλλων, και τον ρυθμό μεταβολής των RSS Feeds που αναλύει προκειμένου να αποφανθεί αν θα επισκεφθεί τη σελίδα του νέου. Αποφεύγονται έτσι άσκοπες εκτελέσεις της διαδικασίας του crawling και ουσιαστικά εξοικονομούνται πόροι του συστήματος. Παράλληλα, οι αλγόριθμοι αναγνώρισης και εξαγωγής χρήσιμου κειμένου έχουν ενισχυθεί και βελτιστοποιηθεί ώστε να εκτελούνται ταχύτερα και να επιστρέφουν με υψηλότερη ακρίβεια το περιεχόμενο που ανταποκρίνεται στο ωφέλιμο κείμενο μιας ιστοσελίδας.

Η διαδικασία προεπεξεργασίας του κειμένου και εξαγωγής των λέξεων-κλειδιών από αυτό, έχει επίσης βελτιωθεί σημαντικά. Οι αλγόριθμοι πλέον δέχονται ρύθμιση μέσω παραμέτρων που μετα-

βάλλονται ανάλογα το κείμενο και την πηγή του. Επιπλέον, το σύστημα μπορεί να αναγνωρίσει κείμενα όλων των βασικών γλωσσών με μια αρθρωτή (modular) αρχιτεκτονική. Παράλληλα, η διαδικασία εύρεσης λέξεων-κλειδιών έχει ενισχυθεί με την δυνατότητα εξαγωγής των ουσιαστικών του κειμένου, που συνήθως φέρουν το μεγαλύτερο ποσοστό 'νοήματος' μιας πρότασης, και γενικότερα δυνατότητα αναγνώρισης των μερών του λόγου των προτάσεων. Το υποσύστημα αυτό τέλος έχει σχεδιαστεί για να υποστηρίζει (μελλοντικά) και πολυμεσικό περιεχόμενο το οποίο μπορεί να αντιστοιχίσει με λέξεις κλειδιά.

Ακολουθώντας, βρίσκονται οι μηχανισμοί κατηγοριοποίησης κειμένου και εξαγωγής της περίληψης αυτού οι οποίοι επίσης έχουν ενισχυθεί και παρουσιάζουν καλύτερα αποτελέσματα σε σχέση με την αρχική έκδοση του συστήματος. Η διαδικασία περίληψης έχει βελτιωθεί σημαντικά με τεχνικές που αξιοποιούν τη γνώση του συστήματος τόσο για το ίδιο το κείμενο όσο και για τον χρήστη που ζητάει την περίληψη. Η διαδικασία κατηγοριοποίησης επίσης επωφελείται από την περίληψη του κειμένου αξιοποιώντας τη, ως μικρότερη και συνοπτικότερη έκδοση του αρχικού κειμένου, προκειμένου να αποφανθεί σε περιπτώσεις που δεν είναι εντελώς ξεκάθαρο σε ποια κατηγορία ανήκει το κείμενο.

Η διαδικασία ολοκληρώνεται με την προσωποποιημένη παρουσίαση των αποτελεσμάτων στη μεριά του χρήστη. Ο αλγόριθμος προσωποποίησης λαμβάνει υπ' όψιν του πολλές παραμέτρους, μεταξύ των οποίων το ιστορικό περιήγησης, οι χρόνοι που μένει ο χρήστης σε κάποιο άρθρο και οι επιλογές του στην εφαρμογή για να παράγει το προφίλ του. Ο αλγόριθμος προσωποποίησης που προτείνεται ουσιαστικά 'μαθαίνει' από τις επιλογές του χρήστη και προσαρμόζεται στις πραγματικές προτιμήσεις του με το πέρασμα του χρόνου. Έτσι το σύστημα μπορεί να ανταποκρίνεται στις διαρκώς μεταβαλλόμενες προτιμήσεις των χρηστών.

Στην τελική φάση της ροής της πληροφορίας, τα αποτελέσματα επιστρέφονται στην εφαρμογή που τρέχει ο χρήστης στην επιφάνεια εργασίας του και που αποτελεί μέρος της παρούσας εργασίας. Ο σκοπός της client-side εφαρμογής είναι να αξιοποιήσει και να παρουσιάσει την πληροφορία που εκτιμάται ότι ενδιαφέρει τον χρήστη, μορφοποιώντας την κατάλληλα ώστε να είναι πραγματικά χρήσιμη και ευανάγνωστη. Σκοπός δεν είναι να 'πλημμυριστεί' ο χρήστης με ακόμη περισσότερη πληροφορία από αυτή που μπορεί να βρει μόνος του στο διαδίκτυο, αλλά να φιλτραριστεί αυτή ώστε να αντιπροσωπεύει πραγματικά τα ενδιαφέροντα του χρήστη. Η εφαρμογή που αναπτύχθηκε στηρίζεται σε standard πρωτόκολλα τόσο μετάδοσης όσο και μορφοποίησης της πληροφορίας και είναι εύκολα παραμετροποιήσιμη από τον χρήστη, ενώ παράλληλα προσφέρει πλήθος λειτουργιών που την καθιστούν ικανή να αντικαταστήσει τις κοινές μεθόδους καθημερινής ενημέρωσης που χρησιμοποιούν οι χρήστες του διαδικτύου.

Executive Summary

The aim of this Master of Science thesis is the expansion and improvement of the mechanism that was constructed within the scope of my bachelor thesis entitled: "Algorithms and techniques for generating text summaries and applications in small screen devices".

In the aforementioned thesis, a complete mechanism was constructed which can automatically analyze internet texts in order to extract key-words. Through this analysis, the most important text sentences which characterize it are identified and which, if joined, can form a short summary of the whole text. The mechanism takes advantage of the text's category knowledge as well as the user's preferences in order to improve and filter more accurately the presented results. The constructed system consists of the following basic subsystems: data retrieval mechanism and useful text extraction mechanism for texts deriving from the web, keyword extraction mechanism, categorization mechanism, which can take part in the process of summary extraction and enhance its results, mechanisms for personalizing content to the end user and of course, text summarization mechanism. The aforementioned mechanisms are incorporated to an internet news indexing system, PeRSSonal, which is used for the retrieval / preprocessing / categorization / personalization and the summarization of articles that derive from news portal of the web.

The aim of the current thesis is the amendment of the existing procedures of the mechanism that was constructed with better and more effective methods and algorithms, as well as the development of a desktop application which shall exploit to the maximum the presentation capabilities of the system though the classic client-server model.

More specifically, all the operation stages of the mechanism are upgraded. Thus, the data retrieval stage is improved with a new, more effective web crawler. The implemented algorithm at this stage takes into consideration, among others, the modification rate of the RSS Feeds that are analyzed in order to decide if the article's page should be fetched. In this manner, unneeded crawling executions are bypassed and system resources are conserved. Furthermore, the recognition and useful text extraction algorithms are enhanced in order to run faster and return with higher precision the content which responds to the useful text of an article's page.

The text preprocessing keyword extraction unneeded are also significantly improved. The algorithms now are parametrized and are adjusted according to the text and its origin. Moreover, the system can recognize the texts language through a modular architecture. In addition, the keyword extraction procedure is enhanced with noun retrieval capabilities. Nouns usually baring the most semantic meaning of the text are now identified and can be weighted accordingly. This subsystem is also designed to support multimedia content which will be correlated with

keywords.

One step more, the categorization and summarization mechanism are improved with heuristics that deliver better results than the initial version of the system. The summarization procedure has improved significantly with techniques that utilize the system's knowledge not only for the text itself, but also for the user requesting the summary. The categorization procedure is also benefitted by the text's summary using it as a shorter, more meaningful version of the initial text, in order to decide in occasions that the categorization of the full text does not give clear results.

The procedure concludes with the personalized presentation of the results on the user's side. The personalization algorithm takes into consideration many parameters, along which the browsing history, the times spent by the user at a text's summary or full body, etc. The algorithm is also "leaning" by the user choices and adjusts itself to the real user preferences as time passes. Thus the system can actually respond positively to the continually changing user preferences.

In the final stage of the flow of information, the results are returned to the application that the user is running on his/her desktop and the development of which is part of this thesis. The aim of the client side application is to utilize and properly present the information that the system has decided to be user-interesting. This information is suitably formatted so as to be really useful and readable on the desktop application. We are not targetting to the "information flooding" of the user, but contrary, to the filtering of information in order to truly represent the user's interests. The developed application is based on standard protocols for the transmission and formatting of information and is easily adjustable by the user, while it also offers many functions which make it able to replace the common methods for the user's everyday internet news reading needs.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Συντομογραφίες

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

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

> Above all things, reverence yourself.
>
> Pythagoras, Greek Mathematician, 497 BC

Σκοπός της Μεταπτυχιακής εργασίας είναι η επέκταση και η βελτίωση του μηχανισμού που δημιουργήθηκε στα πλαίσια της Διπλωματικής εργασίας που εκπόνησα με τίτλο 'Αλγόριθμοι και Τεχνικές Δημιουργίας Περίληψης Κειμένου και Εφαρμογή σε Συσκευές Μικρού Μεγέθους'.

Στα πλαίσια της παραπάνω διπλωματικής εργασίας, δημιουργήθηκε ένας ολοκληρωμένος μηχανισμός ο οποίος μπορεί αυτόματα να κάνει αναλύει κείμενα του διαδικτύου προκειμένου να εξάγει λέξεις-κλειδιά. Μέσα από αυτή την ανάλυση προκύπτουν οι σημαντικότερες προτάσεις του κειμένου που το χαρακτηρίζουν και οι οποίες μπορούν, αν συνενωθούν, να αποτελέσουν μια σύντομη περίληψη του κειμένου. Ο μηχανισμός αξιοποιεί γνώσεις για την κατηγορία του κειμένου καθώς και για τις προτιμήσεις που παρουσιάζουν οι χρήστες του προκειμένου να βελτιώσει και να φιλτράρει τα αποτελέσματα που παρουσιάζονται. Το σύστημα που κατασκευάστηκε έχει τα εξής βασικά υποσυστήματα: μηχανισμός ανάκτησης δεδομένων και εξαγωγής χρήσιμου κειμένου από τον παγκόσμιο ιστό, μηχανισμός εξαγωγής λέξεων-κλειδιών από το πηγαίο κείμενο, μηχανισμός κατηγοριοποίησης κειμένου, ο οποίος μπορεί να συμμετάσχει στη διαδικασία εξαγωγής περίληψης και να ενδυναμώσει τα αποτελέσματά της, μηχανισμοί προσωποποίησης περιεχομένου στο χρήστη και φυσικά, μηχανισμός εξαγωγής περίληψης. Οι παραπάνω μηχανισμοί είναι ενσωματωμένοι σε ένα σύστημα αποδελτίωσης, το PeRSSonal, το οποίο χρησιμοποιείται για την ανάκτηση / προεπεξεργασία / κατηγοριοποίηση / προσωποποίηση και περίληψη άρθρων από ειδησεογραφικούς τόπους του διαδικτύου.

Σκοπός της παρούσας εργασίας είναι η ενίσχυση των υπαρχόντων διαδικασιών του μηχανισμού που δημιουργήθηκε με καλύτερες και αποτελεσματικότερες μεθόδους και αλγορίθμους, καθώς και η δημιουργία μιας desktop εφαρμογής που θα αξιοποιεί στο έπακρο τις δυνατότητες παρουσίασης του συστήματος μέσω του κλασικού client-server μοντέλου.

Πιο συγκεκριμένα, αναβαθμίζονται όλα τα στάδια λειτουργίας του μηχανισμού. Έτσι, το στάδιο ανάκτησης δεδομένων από τον ιστό ενισχύεται με έναν νέο, πιο αποτελεσματικό crawler. Ο αλγόριθμος που υλοποιείται σε αυτό το στάδιο λαμβάνει υπ' όψιν του, μεταξύ άλλων, και τον ρυθμό

μεταβολής των RSS Feeds που αναλύει προκειμένου να αποφανθεί αν θα επισκεφθεί τη σελίδα του νέου. Αποφεύγονται έτσι άσκοπες εκτελέσεις της διαδικασίας του crawling και ουσιαστικά εξοικονομούνται πόροι του συστήματος. Παράλληλα, οι αλγόριθμοι αναγνώρισης και εξαγωγής χρήσιμου κειμένου έχουν ενισχυθεί και βελτιστοποιηθεί ώστε να εκτελούνται ταχύτερα και να επιστρέφουν με υψηλότερη ακρίβεια το περιεχόμενο που ανταποκρίνεται στο ωφέλιμο κείμενο μιας ιστοσελίδας.

Η διαδικασία προεπεξεργασίας του κειμένου και εξαγωγής των λέξεων-κλειδιών από αυτό, έχει επίσης βελτιωθεί σημαντικά. Οι αλγόριθμοι πλέον δέχονται ρύθμιση μέσω παραμέτρων που μεταβάλλονται ανάλογα το κείμενο και την πηγή του. Επιπλέον, το σύστημα μπορεί να αναγνωρίσει κείμενα όλων των βασικών γλωσσών με μια αρθρωτή (modular) αρχιτεκτονική. Παράλληλα, η διαδικασία εύρεσης λέξεων-κλειδιών έχει ενισχυθεί με την δυνατότητα εξαγωγής των ουσιαστικών του κειμένου, που συνήθως φέρουν το μεγαλύτερο ποσοστό 'νοήματος' μιας πρότασης, και γενικότερα δυνατότητα αναγνώρισης των μερών του λόγου των προτάσεων. Το υποσύστημα αυτό τέλος έχει σχεδιαστεί για να υποστηρίζει (μελλοντικά) και πολυμεσικό περιεχόμενο το οποίο μπορεί να αντιστοιχίσει με λέξεις κλειδιά.

Ακολουθώντας, βρίσκονται οι μηχανισμοί κατηγοριοποίησης κειμένου και εξαγωγής της περίληψης αυτού οι οποίοι επίσης έχουν ενισχυθεί και παρουσιάζουν καλύτερα αποτελέσματα σε σχέση με την αρχική έκδοση του συστήματος. Η διαδικασία περίληψης έχει βελτιωθεί σημαντικά με τεχνικές που αξιοποιούν τη γνώση του συστήματος τόσο για το ίδιο το κείμενο όσο και για τον χρήστη που ζητάει την περίληψη. Η διαδικασία κατηγοριοποίησης επίσης επωφελείται από την περίληψη του κειμένου αξιοποιώντας τη, ως μικρότερη και συνοπτικότερη έκδοση του αρχικού κειμένου, προκειμένου να αποφανθεί σε περιπτώσεις που δεν είναι εντελώς ξεκάθαρο σε ποια κατηγορία ανήκει το κείμενο.

Η διαδικασία ολοκληρώνεται με την προσωποποιημένη παρουσίαση των αποτελεσμάτων στη μεριά του χρήστη. Ο αλγόριθμος προσωποποίησης λαμβάνει υπ' όψιν του πολλές παραμέτρους, μεταξύ των οποίων το ιστορικό περιήγησης, οι χρόνοι που μένει ο χρήστης σε κάποιο άρθρο και οι επιλογές του στην εφαρμογή για να παράγει το προφίλ του. Ο αλγόριθμος προσωποποίησης που προτείνεται ουσιαστικά 'μαθαίνει' από τις επιλογές του χρήστη και προσαρμόζεται στις πραγματικές προτιμήσεις του με το πέρασμα του χρόνου. Έτσι το σύστημα μπορεί να ανταποκρίνεται στις διαρκώς μεταβαλλόμενες προτιμήσεις των χρηστών.

Στην τελική φάση της ροής της πληροφορίας, τα αποτελέσματα επιστρέφονται στην εφαρμογή που τρέχει ο χρήστης στην επιφάνεια εργασίας του και που αποτελεί μέρος της παρούσας εργασίας. Ο σκοπός της client-side εφαρμογής είναι να αξιοποιήσει και να παρουσιάσει κατάλληλα την πληροφορία που εκτιμάται ότι ενδιαφέρει τον χρήστη, μορφοποιώντας την κατάλληλα ώστε είναι πραγματικά χρήσιμη και ευανάγνωστη. Σκοπός δεν είναι να 'πλημμυριστεί' ο χρήστης με ακόμη περισσότερη πληροφορία από αυτή που μπορεί να βρει μόνος του στο διαδίκτυο, αλλά να φιλτραριστεί αυτή ώστε να αντιπροσωπεύει πραγματικά τα ενδιαφέροντα του χρήστη. Η εφαρμογή που αναπτύχθηκε στηρίζεται σε standard πρωτόκολλα τόσο μετάδοσης όσο και μορφοποίησης της πληροφορίας και είναι εύκολα παραμετροποιήσιμη από τον χρήστη, ενώ παράλληλα προσφέρει πλήθος λειτουργιών που την καθιστούν ικανή να αντικαταστήσει τις κοινές μεθόδους καθημερινής ενημέρωσης που χρησιμοποιούν οι χρήστες του διαδικτύου.

Μέσα από την εργασία προέκυψαν αποτελέσματα που έχουν να κάνουν με σύγκριση αλγορίθμων σε όλα τα παραπάνω στάδια του μηχανισμού αλλά και ανταπόκριση του μηχανισμού στις ανάγκες του χρήστη. Τα αποτελέσματα αυτά είναι ιδιαίτερα ενθαρρυντικά και μας παρακινούν για περαιτέρω έρευνα στα πεδία με τα οποία καταπιάνεται, καθώς και επέκταση του συστήματος PeRSSonal.

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

Κεφάλαια σε Βιβλία

- PeRSSonal, the Automatic Summarization, Text Categorization, Personalized Syndication, System. Handbook of Research on Social Interaction Technologies and Collaboration Software: Concepts and Trends, IGI Global, C. Bouras, V. Poulopoulos, V. Tsogkas, 2008, (to appear) [68]

  Abstract Η τεχνολογική πρόοδος και η ευκολία στην πρόσβαση της πληροφορίας έχουν αλλάξει δραματικά την κατάσταση του παγκόσμιου ιστού τα τελευταία χρόνια. Αυτή η αλλαγή έχει επίσης επηρεάσει τον τρόπο και τη συχνότητα στην οποία τα άρθρα δημιουργούνται και δημοσιεύονται στο διαδίκτυο. Κάθε μέρα, χιλιάδες άρθρων παράγονται από την πληθώρα των news portals, μικρά και μεγάλα που υπάρχουν στον παγκόσμιο ιστό. Αυτή η αίσθηση της ελευθερίας την οποία το διαδίκτυο δημιουργεί, προσελκύει όλο και πιο πολλούς χρήστες, όχι μόνο να διαβάσουν σε καθημερινή βάση τη 'διαδικτυακή τους εφημερίδα', αλλά και να παράγουν τα δικά τους άρθρα ή πηγές πληροφοριών. Εξάλλου, η τελευταία 'μόδα' του blogging ξεφεύγει από την απλή τήρηση ενός προσωπικού ιστολογίου αλλά δρα ως ένα μέσω ανταλλαγής πληροφοριών.

Διεθνή περιοδικά

- PeRSSonal's core functionality evaluation: enchancing text labeling through personalized summaries. Data and Knowledge Engineering Journal, Elsevier Science, 2008, Vol. 64, Issue 1 , C. Bouras, V. Poulopoulos, V. Tsogkas, 2008, pp. 330 - 345 [69]

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

Διεθνή συνέδρια

- Improving text summarization using noun retrieval techniques. Advanced Knowledge – based Systems, Invited Session of the 12nd International Conference on Knowledge – based and Intelligent Information & Engineering Systems (KES 2008), Zagreb, Croatia, C. Bouras, V. Tsogkas, 3 - 5 September 2008, pp. 593 - 600 [70]

  Abstract Η αυτόματη περίληψη και η κατηγοριοποίηση κειμένου είναι εδώ και καιρό δύο από τις πιο σημαντικές διεργασίες της ανάκτησης πληροφορίας. Η δημιουργία ενός γενικού, πολυχρηστικού μηχανισμού που παράγει καλά αποτελέσματα και για τις δύο διεργασίες μοιάζει να είναι μία καλή λύση για τις περισσότερες ανάγκες ανάκτησης πληροφορίας, τουλάχιστον σε ότι έχει να κάνει με κειμενική πληροφορία. Σε αυτή τη δημοσίευση, παρουσιάζουμε τις τεχνικές εξαγωγής κωδικολέξεων, ερευνώντας τα αποτελέσματα που έχει η αναγνώριση των μερών του λόγου ενός κειμένου πάνω στην διαδικασία περίληψης ενός υπάρχοντος συστήματος.

- Creating dymanic personalized RSS summaries. 8th Industrial Conference on Data Mining – 8th Industrial Conference on Data Mining – ICDM 2008, , Leipzig, Germany, C. Bouras, V. Poulopoulos, V. Tsogkas, 16 - 18 July 2008 [67]

  Abstract Τα συστήματα παραγωγής αυτόματης, υψηλής ποιότητας περίληψη κειμένου είναι δύσκολα τόσο στην κατασκευή όσο και στην αξιολόγηση, εν΄ μέρει διότι τα κείμενα διαφέρουν σε διάφορες διαστάσεις: μήκος, στυλ γραψίματος και χρήση λεξιλογίου. Σε αυτή τη δημοσίευση προτείνουμε ένα περιβάλλον το οποίο, αξιοποιώντας RSS Feeds, είναι ικανό να προσωποποιήσει τα αποτελέσματα στις ανάγκες του χρήστη και τις δυνατότητες της τελικής συσκευής ώστε να παρουσιάζει στον τελικό χρήστη μόνο ένα κλάσμα των άρθρων, καλύπτοντας μόνο την χρήσιμη πληροφορία που υπάρχει σε αυτά. Οι περιλήψεις που παράγονται αξιοποιούν έναν ζυγισμένο συνδυασμό από στατιστικά και γλωσσολογικά χαρακτηριστικά που οδηγεί στην βαθμολόγηση και την επιλογή των προτάσεων. Η διαδικασία υποβοηθείται από τα αποτελέσματα κατηγοριοποίησης καθώς και από αλγορίθμους προσωποποίησης. Ο μηχανισμός αξιολογείται χρησιμοποιώντας κλασικές μετρικές ακρίβειας - ανάκλησης μαζί με στατιστικά αποτελέσματα που προέκυψαν από τη χρήση τους. Σε αυτό το περιβάλλον εργασίας, δημιουργήσαμε τον σύστημα PeRSSonal το οποίο είναι ικανό να παράγει προσωποποιημένα, προ-κατηγοριοποιημένα και δυναμικά RSS Feeds που στοχεύονται για χρήση σε συσκευές μικρού μεγέθους.

- Efficient Summarization Based On Categorized Keywords. The 2007 International Conference on Data Mining (DMIN'07), Las Vegas, Nevada, USA, C. Bouras, V. Poulopoulos, V. Tsogkas, 25 - 28 June 2007 [65]

  Abstract Η πληροφορία που υπάρχει στο διαδίκτυο είναι αρκετά μεγάλη ώστε να εκτρέπει τους χρήστες στην προσπάθεια αναζήτησης πληροφορίας. Προκειμένου να αποφευχθούν τα προβλήματα που δημιουργούνται από την πληθώρα δεδομένων του διαδικτύου πολλοί μηχανισμοί προσωποποίησης 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.

- Personalizing text summarization based on sentence weighting. IADIS European First International Conference Data Mining (ECDM 2007), Lisbon, Portugal, C. Bouras, V. Poulopoulos, V. Tsogkas, 3 - 8 July 2007 [66]

  Abstract Η πληροφορία που υπάρχει στο διαδίκτυο είναι τόσο μεγάλη ώστε να εμποδίζει τους χρήστες στην προσπάθεια εύρεσης χρήσιμης πληροφορίας. Παράλληλα, η μεγάλη ανάπτυξη της τεχνολογίας όσον αφορά τις συσκευές μικρού μεγέθους και η δυνατότητα αυτών

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

- The importance of the difference in text types to keyword extraction: Evaluating a mechanism. 7th International Conference on Internet Computing 2006 (ICOMP 2006), Las Vegas, Nevada, USA, C. Bouras, C. Dimitriou, V. Poulopoulos, V. Tsogkas, 26 - 29 June 2006 [64]

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

1.1 Περιγραφή του προβλήματος

Το διαδίκτυο είναι πλέον παντού: σε κάθε συσκευή, σε κάθε μεριά του σπιτιού στην κοινωνία ολόκληρη. Η χρήση του διαδικτύου την οκταετία 2000-2008 έχει αυξηθεί κατά το ασύλληπτο ποσοστό του 305,5% [22] και το μέγεθός του, το 2005 τουλάχιστον, ήταν περί τα 5 εκατομμύρια Terrabytes. Είναι χαρακτηριστικό ότι η μηχανή αναζήτησης Google [16] που δεικτοδοτεί την πληροφορία του διαδικτύου τα τελευταία 10 χρόνια, με την ταχύτητα που έχουν οι αλγόριθμοι δεικτοδότησής της, χρειάζεται χοντρικά 300 χρόνια για να δεικτοδοτήσει όλο αυτό το περιεχόμενο (μη λαμβάνοντας προφανώς υπ' όψιν το νέο περιεχόμενο που έχει προστεθεί στην πορεία). Παράλληλα, μια συνδυαστική έκρηξη φαίνεται να έχει λάβει χώρα όσον αφορά στις τεχνολογίες που χρησιμοποιούνται στο διαδίκτυο και κατ' επέκταση στις νέες υπηρεσίες. Όλα αυτά τα στοιχεία μας οδηγούν στο συμπέρασμα ότι η διαδικασία αναζήτησης και η επιτυχής εύρεση πληροφορίας που μας ενδιαφέρει στο διαδίκτυο είναι αν μη τι άλλο μια υπόθεση δύσκολη.

Θα μπορούσε εύκολα να ειπωθεί ότι όπως κάθε κοινωνία, έτσι και το Διαδίκτυο, έχει τα δικά

του προβλήματα. Πηγή αυτών των προβλημάτων μπορεί να θεωρηθεί η 'άναρχη δόμησή του', η έλλειψη σαφούς νομοθεσίας αλλά και η αίσθηση ελευθερίας που αφήνει τους '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.

Η ελευθερία της έκφρασης και του λόγου παγκοσμίως διασφαλίζεται πλέον από τον τρόπο με τον οποίο διακινείται το περιεχόμενο στο Διαδίκτυο. Η διάχυση γνώσης και εμπειρίας θα μπορούσαν επίσης να χαρακτηριστούν σαν θετικά επακόλουθα από την ύπαρξη μεγάλου όγκου πληροφορίας στον παγκόσμιο ιστό. Θα πρέπει όμως κανείς να αναλογιστεί κατά πόσο όλος αυτός ο όγκος πληροφορίας και όλες οι πηγές ενημέρωσης του Διαδικτύου είναι έγκυρες. Δεν υπάρχει απολύτως κανένας μηχανισμός που να μπορεί να διασφαλίσει σε κάθε επισκέπτη του Διαδικτύου πως οι σελίδες που παρακολουθεί και το περιεχόμενο που συλλέγει είναι αξιόπιστο και ποιοτικό. Πλέον, ακόμα και ο μέσος χρήστης, γνωρίζει μηχανισμούς μέσα από τους οποίους μπορεί να βρει στοιχεία για οποιοδήποτε θέμα. Κανείς όμως δε μπορεί να του εγγυηθεί επιτυχία και ταχύτητα στη διαδικασία ανεύρεσης αλλά πάνω απ' όλα, ποιότητα στα αποτελέσματα της εκάστοτε αναζήτησής του. Απαιτούνται καινοτόμες τεχνικές, νέες ιδέες και νέες προσεγγίσεις για να αντιμετωπιστεί το πρόβλημα. Οι χρήστες δεν θέλουν απλά πληροφορία, θέλουν να μπορούν να εντοπίζουν εύκολα και γρήγορα ποιοτική πληροφορία, πληροφορία που τους ενδιαφέρει και ταιριάζει με το ύφος τους. Ακόμα περισσότερο, επιθυμούν αυτή η πληροφορία να τους προσφέρετε μέσα από αυτόματους μηχανισμούς που έχουν τη δυνατότητα να φιλτράρουν το 'χάος' του διαδικτύου.

Η έλλειψη ποιότητας στης τάξεις του Διαδικτύου έχει κεντρίσει το ενδιαφέρον της επιστημονικής κοινότητας. Πολλοί ορισμοί που βρίσκονται πλέον στο επίκεντρο του ενδιαφέροντος, περιλαμβάνουν τα data mining, text analysis, text categorization, semantic web και πολλά ακόμα, τα οποία αν και ήταν γνωστά ακόμα και πριν την εξάπλωση του διαδικτύου, φαίνονται να είναι αυτά που δίνουν λύσεις στα μειονεκτήματά του.

Στην παρούσα εργασία δε θα αναλωθούμε στην καταγραφή των πολλών, αν μη τι άλλο, προβλημάτων του Διαδικτύου αλλά θα επικεντρωθούμε σε ένα κομμάτι των προβλημάτων που προκύπτουν από την αέναη, καθημερινή και καταιγιστική δημιουργία δεδομένων και πληροφοριών. Ακόμα περισσότερο, θα εστιάσουμε την προσοχή μας στις πληροφορίες που δημιουργούνται σε καθημερινή βάση από την πληθώρα των ενημερωτικών δικτυακών πυλών που κατακλύζουν στην κυριολεξία το Διαδίκτυο. Ο λόγος για τα γνωστά *news portals*. Πρόκειται για Δικτυακούς τόπους που σαν στόχο έχουν την ενημέρωση των χρηστών του Διαδικτύου για τα φλέγοντα - κυρίως - νέα σε παγκόσμιο επίπεδο. Μερικά και πολύ σημαντικά από αυτά είναι το CNN[6], το BBC[3], το Reuters[40], το FoxNews[11], καθώς και οι υπηρεσίες που προσφέρονται από τους πολυπληθείς και από τους πλέον αναγνωρίσιμους δικτυακούς τόπους Google[16] και Yahoo[49].

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

την ενημέρωσή τους.

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

Μία άλλη παράμετρος που προκύπτει από την χρήση του RSS είναι η εξής: με αυτή την αλλαγή στον τρόπο 'επίσκεψης' μίας σελίδας, τα μεγάλα ειδησεογραφικά πρακτορεία παρατηρούν τη χαμηλή επισκεψιμότητα συγκεκριμένων σελίδων του δικτυακού τους τόπου οι οποίες ουσιαστικά δεν προβάλλονται προς το χρήστη ο οποίος αρκείται στο κανάλι επικοινωνίας που έχει και αποφεύγει κάθε επίσκεψη στο δικτυακό τόπο. Παράλληλα, η ανάπτυξη νέων τεχνολογιών και υπηρεσιών για το Διαδίκτυο κάνει τους διαχειριστές δικτυακών τόπων να επιθυμούν ακόμα μεγαλύτερη επισκεψιμότητα στις σελίδες τους προσφέροντας διαδραστικές υπηρεσίες, υπηρεσίες πολυμέσων κ. α.

Το RSS έχει περιορισμένες δυνατότητες ενώ οι υπηρεσίες προσωποποιημένης πρόσβασης έχουν πολλά να προσφέρουν στους χρήστες. Είναι φανερό πως οι δικτυακοί τόποι, σαν μία κανονική επιχείρηση, επιθυμούν οι χρήστες να 'έρχονται' στο δικτυακό τόπο, να επισκέπτονται όλες τις σελίδες, να βλέπουν τις διαφημίσεις, να αξιοποιούν τις νέες υπηρεσίες, να χρησιμοποιούν κάθε δεδομένο που τους προσφέρεται.

Όσο εντυπωσιακά και αν φαίνονται όλα αυτά, οι σχεδιαστές των υπηρεσιών έχουν παραλείψει σημαντικά στοιχεία. Πόσο εξοικειωμένοι είναι οι χρήστες στη χρήση περίπλοκων συστημάτων; Έχουν όλοι οι χρήστες αρκετά μεγάλη ταχύτητα στην πρόσβαση στο διαδίκτυο προκειμένου να μπορούν να χρησιμοποιούν χωρίς πρόβλημα τις προσφερόμενες υπηρεσίες; Οι χρήστες έχουν ερωτηθεί για τις πληροφορίες που θα επιθυμούσαν να τους διατίθενται; Αποτέλεσμα όλων των παραπάνω είναι: προσωποποιημένες σελίδες δικτυακών τόπων, όπου ο χρήστης αδυνατεί να τις σχεδιάσει όπως επιθυμεί καθότι 'χάνεται' στην πληθώρα δεδομένων που τους παρουσιάζονται, υπερπολλαπλασιασμός των καναλιών RSS των δικτυακών τόπων με αποτέλεσμα ο χρήστης να αντιμετωπίζει το ίδιο χάος. Τρανταχτό παράδειγμα αποτελεί το RSS Feed του CNN που αποτελείται από περισσότερα από 20 επικαλυπτόμενα κανάλια. Τέλος κάτι πολύ σημαντικό, κανείς δεν επιχειρεί να συνδυάσει τις δύο υπηρεσίες οι οποίες δε φαίνεται να διαφέρουν μεταξύ τους. Κανένας δικτυακός τόπος δεν προσπαθεί να συνδυάσει προσωποποιημένες πληροφορίες και παράδοση τους στο desktop του χρήστη.

Συνοψίζοντας λοιπόν, καταλήγουμε στα εξής:

- Προσωποποιημένες σελίδες: Δύσχρηστες - πολύπλοκες. Βασίζονται σε λέξεις κλειδιά ή ακόμα χειρότερα σε γενικές κατηγορίες μόνον. Ο χρήστης σε κάθε περίπτωση παραμένει εκτός της διαδικασίας κατηγοριοποίησης ή κατασκευής περίληψης που παρουσιάζεται στην προσωποποιημένη σελίδα.

- RSS feeds: Ο αριθμός τους είναι υπερβολικά μεγάλος. Ο αριθμός των άρθρων που περιέχουν είναι υπερβολικά μεγάλος. Δεν έχουν στοιχεία φιλτραρίσματος ή προσωποποίησης και

επομένως δε χρησιμοποιούνται σωστά.

Όλα τα παραπάνω έχουν ως αποτέλεσμα οι χρήστες να δυσκολεύονται στην αναζήτηση ειδήσεων και πιο συγκεκριμένα, στην παρακολούθηση αποκλειστικά των ειδήσεων που τους ενδιαφέρουν. Ακόμα περισσότερο, οι χρήστες θα πρέπει με κάποιον τρόπο να γίνουν κομμάτι του πυρήνα ενός τέτοιου συστήματος και να διαμορφώνουν τον τρόπο με τον οποίο πραγματοποιείται η κατηγοριοποίηση αλλά και τον τρόπο με τον οποίο παρουσιάζονται τα αποτελέσματα της αναζήτησης σε αυτούς.

Μία υπηρεσία τέτοιας λογικής παρουσιάζουμε στην παρούσα εργασία. Έναν μηχανισμό που φιλοδοξεί να αποτελέσει πρότυπο για υιοθέτηση από κάθε news portal μιας και οι σύγχρονες επιταγές επιβάλλουν μία ριζική αναθεώρηση του τρόπου που η πληροφορία δίνεται στους χρήστες. Στη συνέχεια θα παρακολουθήσουμε κάθε διαδικασία του συστήματος που αναπτύχθηκε και θα αναλυθούν τα προβλήματα που εντοπίζονται σε κάθε μια από αυτές. Το σύστημά μας ακολουθεί μία σειριακή διαδικασία προκειμένου να παράγει το ζητούμενο αποτέλεσμα το οποίο είναι η παρουσίαση προσωποποιημένων, κατηγοριοποιημένων άρθρων στον τελικό χρήστη. Για να γίνει αυτό θα πρέπει το σύστημα να είναι σε θέση να συλλέγει συνεχώς άρθρα από μεγάλα ειδησεογραφικά πρακτορεία. Η συλλογή των άρθρων δεν είναι αρκετή. Αφού τα άρθρα συγκεντρωθούν, θα πρέπει να εφαρμοστούν σε αυτά μία σειρά από αλγορίθμους προκειμένου να 'καθαριστεί' το κείμενό τους από οποιαδήποτε περιττή πληροφορία. Εν συνεχεία θα πρέπει να εφαρμοστούν αλγόριθμοι κατηγοριοποίησης του κειμένου και εξαγωγής περίληψης. Τέλος θα πρέπει να υπάρχει ένας μηχανισμός ο οποίος θα πραγματοποιεί προσωποποίηση των πιο πρόσφατων άρθρων στον εκάστοτε χρήστη και φυσικά ένα σύστημα παρουσίασης της πληροφορίας στον τελικό χρήστη το οποίο θα προσαρμόζει το περιεχόμενο στις προτιμήσεις του χρήστη προτού η πληροφορία παρουσιαστεί στην επιφάνεια εργασίας του.

1.1.1 Συλλογή δεδομένων

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

Η συλλογή δεδομένων βασίζεται σε μηχανισμούς που περιδιαβαίνουν ολόκληρους τους ειδησεογραφικούς κόμβους και εντοπίζουν τα σημεία εκείνα που περιέχουν αρκετό κείμενο συγκριτικά με άλλες σελίδες που αποτελούν κεντρικούς κόμβους πληροφοριών. Οστόσο, οι νέες τεχνολογίες και κυρίως τα κανάλια επικοινωνίας που χρησιμοποιούνται από τους σύγχρονους δικτυακούς τόπους μπορούν να διευκολύνουν το πρόβλημα της συλλογής δεδομένων. Οι μηχανισμοί δεν είναι υποχρεωμένοι να 'ανακαλύπτουν' τις πολλαπλές δυναμικές σελίδες που ανανεώνονται καθημερινά στους δικτυακούς τόπους. Αρκεί η συλλογή πληροφοριών από τα κανάλια επικοινωνίας (XML/RSS) που υπάρχουν για τη συγκέντρωση των πιο σημαντικών αλλαγών που προκύπτουν καθημερινά και εν προκειμένω τα νέα άρθρα που προστίθενται στα ειδησεογραφικά portals.

1.1.2 Φιλτράρισμα δεδομένων

Η συλλογή πληροφοριών έχει σαν αποτέλεσμα σελίδες που περιέχουν κυρίως HTML κώδικα στον οποίο βεβαίως μπορούμε να εντοπίσουμε και το κείμενο το οποίο επιθυμούμε να εξάγουμε από τη σελίδα και το οποίο αποτελεί το κύριο σώμα του άρθρου. Για το φιλτράρισμα τέτοιου είδους

δεδομένων έχουν γίνει πολλές προτάσεις, κυρίως για τον τρόπο με τον οποίο μπορεί να εξαχθεί και βασικά να εντοπιστεί μέσα στη σελίδα. Το πρόβλημα σε αυτή την περίπτωση είναι η απομόνωση του χρήσιμου μόνο κειμένου το οποίο στην περίπτωση που εξετάζουμε είναι το σώμα του άρθρου αλλά και ο τίτλος του.

1.1.3 Προεπεξεργασία πληροφορίας

Η προεπεξεργασία πληροφορίας είναι μία διαδικασία κατά την οποία το χρήσιμο κείμενο υπόκειται σε διαδικασία αφαίρεσης των σημείων στίξης, των αριθμών που τυχόν περιέχει, αφαίρεση λέξεων οι οποίες δεν περικλείουν κάποιο νόημα και τέλος το πολύ σημαντικό κομμάτι του Stemming το οποίο είναι η διαδικασία εύρεσης της ρίζας μίας λέξης. Σαν αποτέλεσμα έχει την εξαγωγή των λέξεων κλειδιών που υπάρχουν στο κείμενο συνοδευμένα από τη συχνότητα την οποία παρουσιάζουν μέσα στο κείμενο αλλά και το σημείο του κειμένου στο οποίο εντοπίζονται. Για την περαιτέρω ενίσχυση των διαδικασιών ανάκτησης πληροφορίας που ακολουθούν, στις τεχνικές προεπεξεργασίας κειμένου θα εντάξουμε και την ανάκτηση των ουσιαστικών του κειμένου, μιας και είναι γενικά αποδεκτό ότι τα ουσιαστικά του κειμένου φέρουν το μεγαλύτερο ποσοστό της χρήσιμης πληροφορίας αυτου.

Για τους μηχανισμούς εξαγωγής κειμένου και απόρριψης οποιασδήποτε πληροφορίας δεν σχετίζεται με το κείμενο η προεπεξεργασία πληροφορίας είναι μία πρόκληση. Παρά το γεγονός ότι βασίζεται σε συγκεκριμένα και σταθερά βήματα, θα πρέπει να γίνει εκτενής ανάλυση του είδους της πληροφορίας που είναι επιθυμητή προκειμένου το βήμα της προεπεξεργασίας να καταλήξει σε σημαντικά αποτελέσματα και πιο συγκεκριμένα στην εξαγωγή των σωστών λέξεων κλειδιών.

1.1.4 Προσωποποίηση στο χρήστη

Η προσωποποίηση στο χρήστη είναι η διαδικασία κατά την οποία τα αποτελέσματα που εμφανίζονται τελικά στο χρήστη προσαρμόζονται προκειμένου να ανταποκρίνονται 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.

1.1.5 Συμμετοχή του χρήστη στις διαδικασίες του συστήματος

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

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

Εφαρμογή παρουσίασης πληροφορίας στη μεριά του χρήστη

Έχοντας την προσωποποιημένη πληροφορία βάσει του προφίλ χρήστη, ένα σημαντικό τμήμα ενός ολοκληρωμένου συστήματος είναι η παρουσίασή της με σωστό και αποδοτικό τρόπο στον χρήστη. Συγκεκριμένα για να έχει χρηστικότητα σε πραγματικές συνθήκες το σύστημα θα πρέπει να υπάρχει και μια client-side εφαρμογή που θα παρουσιάζει τα άρθρα στην επιφάνεια εργασίας του χρήστη. Αυτό θα πρέπει να γίνεται με τρόπο απόλυτα παραμετροποιήσιμο από τον χρήστη αλλά και αποδοτικά ώστε ο χρήστης να μην αντιλαμβάνεται την επικοινωνία της εφαρμογής με τον κεντρικό server. Προς αυτή την κατεύθυνση χρησιμοποιούνται τεχνικές caching στην μεριά του χρήστη αλλά και πλήρως παραμετροποιήσιμο User Interface.

1.2 Δομή της εργασίας

Η υπόλοιπη εργασία δομείται ως εξής: στο κεφάλαιο 2 παρουσιάζονται τα θέματα που θα μας απασχολήσουν καθώς και οι τρέχουσες εξελίξεις στα ερευνητικά πεδία (State of the Art). Στο κεφάλαιο 3 παρουσιάζεται το κομμάτι των σχετικών εργασιών, ενώ στο κεφάλαιο 4 γίνεται μια γενικότερη περιγραφή της αρχιτεκτονικής και των χαρακτηριστικών του συστήματος που αναπτύχθηκε. Ακολουθεί η παρουσίαση της βάσης δεδομένων που χρησιμοποιήθηκε (κεφάλαιο 5). Στο κεφάλαιο 6 γίνεται μια συνοπτική παρουσίαση των διαθέσιμων τεχνολογιών υλοποίησης καθώς και των επιλογών που έγιναν για τα διάφορα υποσυστήματα του μηχανισμού. Ακολουθεί, στο κεφάλαιο 7, η παρουσίαση των αλγοριθμικών θεμάτων του μηχανισμού καθώς και θέματα της υλοποίησής του. Στο κεφάλαιο 8 δίνονται οι προδιαγραφές και η χρήση του συστήματος. Στο κεφάλαιο 9 παρουσιάζονται τα πειραματικά αποτελέσματα καθώς και η αξιολόγησή του συστήματος. Στο κεφάλαιο 10 δίνονται τα συμπεράσματα που προέκυψαν από την εργασία και τέλος στο κεφάλαιο 11 παρουσιάζονται κάποιες προτάσεις για μελλοντική επέκταση του μηχανισμού καθώς και η μελλοντική εργασία που μπορεί να γίνει.

ΚΕΦΑΛΑΙΟ 2

Ερευνητικά Θέματα

> Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin.
>
> *John von Neumann, American Mathematician, 1957*

Στο συγκεκριμένο κεφάλαιο παρουσιάζονται τα ερευνητικά θέματα με τα οποία καταπιάνεται η εργασία. Εντοπίζονται τα τεχνολογικά προβλήματα που υπάρχουν, τρόποι με τους οποίους έχουν αντιμετωπισθεί καθώς και η δική μας προσέγγιση. Οι παράγραφοι που ακολουθούν δίνουν μια αναλυτική παρουσίαση του state of the art στα θέματα της ανάκτησης πληροφορίας και εξόρυξης δεδομένων.

2.1 Σημασιολογικός ιστός και μεταδεδομένα

Μεγάλοι όγκοι δεδομένων αναζητούνται, ανταλλάσσονται και επεξεργάζονται μέσω του Παγκοσμίου Ιστού. Επειδή όμως ο όγκος των δεδομένων του Ιστού έχει πάρει τεράστιες διαστάσεις, χωρίς να υπάρχει ενιαίος τρόπος οργάνωσης, η ανταλλαγή και η επεξεργασία τους είναι πολύ δύσκολη. Ο *Σημασιολογικός Ιστός* έρχεται ακριβώς να εξυπηρετήσει την ανάγκη για ενιαία οργάνωση των δεδομένων, ώστε το Διαδίκτυο να γίνει μια αποδοτική παγκόσμια πλατφόρμα ανταλλαγής και επεξεργασίας πληροφορίας από ετερογενείς πηγές. 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).

Η λέξη 'Σημασιολογία' έχει ρίζα τις Ελληνικές λέξεις 'σημάδι', 'σημαίνω' και 'σημαντικός' και σήμερα αναφέρεται στο νόημα συχνά σε επίπεδο γλώσσας. Μπορούμε να πούμε ότι ο Σημασιολογικός Ιστός αποτελεί το μεγαλύτερο σε παγκόσμιο επίπεδο έργο έξυπνης ενσωμάτωσης συστημάτων ώστε να συνεργάζονται δια-λειτουργικά. 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 [62] ο οποίος τον όρισε ως: 'μια επέκταση του σημερινού ιστού όπου η πληροφορία έχει καλά καθορισμένο νόημα, καθιστώντας τη συνεργασία μεταξύ

ανθρώπων και υπολογιστών πιο αποτελεσματική'. Από τότε καταβάλλεται μεγάλη προσπάθεια από την επιστημονική κοινότητα για την υλοποίησή του πάνω από τον Παγκόσμιο Ιστό. Το κέντρο βάρους του περιεχομένου του Ιστού μετατοπίζεται συνεχώς από τον άνθρωπο προς προς τα δεδομένα. Για να φτάσει ο Ιστός το μέγιστο των δυνατοτήτων του, πρέπει να εξελιχθεί σε ένα Σημασιολογικό Ιστό, ο οποίος παρέχει μια διεθνώς προσβάσιμη πλατφόρμα που επιτρέπει σε αυτοματοποιημένα εργαλεία αλλά και σε ανθρώπους να μοιράζονται και να επεξεργάζονται δεδομένα. Ο Σημασιολογικός Ιστός αποτελεί πρωτοβουλία της Κοινοπραξίας του Παγκοσμίου Ιστού (W3C) και η σχετική Δραστηριότητα (W3C Semantic Web Activity) έχει δημιουργηθεί για να εξυπηρετήσει έναν ηγετικό ρόλο, τόσο στο σχεδιασμό προδιαγραφών, όσο και στην ανοικτή ανάπτυξη της τεχνολογίας μέσω της συνεργασίας.

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. Στο διάστημα των τριών τελευταίων ετών παρήχθησαν 12 exabytes πληροφορίας υπό τη μορφή έντυπου, οπτικού ή και ηχητικού υλικού. Η αυξανόμενη αυτή παραγωγή και η συνεχής βελτίωση των μεθόδων ψηφιοποίησης συμβάλλουν στην παραγωγή ενός ωκεανού ψηφιακών δεδομένων που προφανώς δύναται να δημιουργήσει μεγάλο αριθμό προβλημάτων. Το πιο σημαντικό ίσως από αυτά είναι ο τρόπος με τον οποίο θα μπορεί κανείς να διαχειριστεί όλη αυτή την πληροφορία. Δε θα πρέπει φυσικά να αμελούμε το γεγονός πως η ικανότητα παραγωγής, αποθήκευσης και μετάδοσης της πληροφορίας έχει ξεπεράσει κατά πολύ τις δυνατότητες αναζήτησης, πρόσβασης και παρουσίασης.

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

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

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

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

Τα τελευταία χρόνια, μια νέα ερευνητική προσπάθεια έχει επικεντρωθεί σε αυτό το πεδίο το οποίο ανήκει στην περιοχή που ονομάζεται Προσαρμοσμένη Πρόσβαση στην Πληροφορία *(Adaptive Information Access)*. Η πρόσβαση στην πληροφορία αφορά 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 μηχανική μάθηση όσο και με αλληλεπίδραση χρηστών με τον υπολογιστή.

Το επόμενο κύμα του Σημασιολογικού Ιστού [160] κάνει σημαντική επαναχρησιμοποίηση των υπάρχουσων οντολογιών και δεδομένων. Είναι ένας διασυνδεδεμένος χώρος πληροφοριών στον οποίο τα δεδομένα εμπλουτίζονται και προστίθενται. Επιτρέπει στους χρήστες να εμπλέκονται με τρόπο που βρίσκουν χωρίς να ψάξουν την σχετική πληροφορία κάτι που αποτελεί πανάκεια για το διαδίκτυο της νέας δεκαετίας. Ήδη παρατηρούμε μια αυξανόμενη ανάγκη και υποχρέωση ανθρώπων και οργανισμών να κάνουν τα δεδομένα που κατέχουν διαθέσιμα. Αυτό το γεγονός καθοδηγείται από τις επιταγές της συνεργατικής επιστήμης, των νέων εμπορικών αναγκών - όπως το να είναι οι λεπτομέρειες των προϊόντων άμεσα διαθέσιμες - αλλά και από ρυθμιστικούς κανόνες των χωρών.

2.2 Εξόρυξη πληροφορίας από το διαδίκτυο

Εξόρυξη πληροφορίας από το Διαδίκτυο ονομάζεται κάθε διαδικασία που έχει σαν αποτέλεσμα ανάκτηση πληροφορίας *(Information Retrieval)* από τον παγκόσμιο ιστό. Στο εξής θα αναφερόμαστε στον όρο ανάκτηση πληροφορίας ως IR για συντομία. Η ανακτώμενη πληροφορία δεν περιορίζεται απλώς σε σελίδες HTML, αλλά μπορεί να αφορά 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 [136] εξέφρασε ανοιχτά σε δημοσίευσή του αυτή την ανάγκη. Αργότερα, στη δεκαετία του 60, το IR είχε γίνει πλέον ένα πολύ δημοφιλές θέμα καθώς πολλοί ερευνητές πίστευαν ότι μπορούν να αυτοματοποιήσουν τις μέχρι τότε χειροκίνητες διαδικασίες όπως η δεικτοδότηση και η αναζήτηση.

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 κερδίζουν επίσης έδαφος, όπως οι τεχνικές αναπαράστασης της πληροφορίας, με σκοπό να βελτιωθεί η αποτελεσματικότητα της ανάκτησης.

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

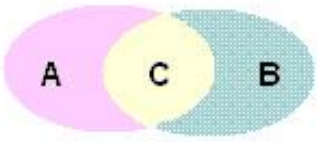
Ένα από τα βασικά στοιχεία του IR είναι η μέτρηση του κατά πόσο τα ανακτημένα κείμενα είναι σχετικά με το ερώτημα που κάνουμε. Έτσι λοιπόν, ένα βασικό στοιχείο στο οποίο εστιάζουμε είναι η εύρεση μετρικών που θα μπορούν να αναπαραστήσουν αριθμητικά τη σχετικότητα των αποτελεσμάτων ενός συστήματος IR.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

Φορμαλιστικά, οι σχέσεις που ισχύουν για τις δύο αυτές μετρικές είναι οι παρακάτω:

$$R = \frac{|\{A\} \cap \{B\}|}{A}$$

$$P = \frac{|\{A\} \cap \{B\}|}{B}$$

όπου R η ανάκληση, P η ακρίβεια, A τα σχετικά κείμενα που βρέθηκαν και B όλα τα άρθρα που ανακτήθηκαν. Οι παραπάνω συσχετίσεις είναι εμφανείς στο σχήμα 2.1.



Σχήμα 2.1: Ακρίβεια - Ανάκληση. Με C είναι τα σχετικά άρθρα που ανακτήθηκαν.

Θα λέγαμε επομένως ότι η ανάκληση μας δίνει ένα μέτρο για το πόσο καλά μια αναζήτηση εντοπίζει αυτό που θέλουμε, ενώ η ακρίβεια μετράει το πόσο καλά απορρίπτουμε αυτό που δεν θέλουμε. Αυτές οι μετρικές, παρότι πολύ χρήσιμες για την αξιολόγηση είναι δύσχρηστες από τη φύση τους. Πρώτα απ΄ όλα η έννοια της ακρίβειας είναι συνήθως αποκλειστικά υποκειμενικό κριτήριο και όχι μια αντικειμενική θετική ή αρνητική απάντηση. Δεύτερον, για κάθε βάση πληροφορίας που είναι αρκετά μεγάλη για να κατασκευαστεί μια μηχανή αναζήτησης πάνω της, θα είναι δύσκολο να υπολογιστούν πραγματικές τιμές ανάκλησης λόγω του μεγέθους της βάσης (για να υπολογιστεί επ΄ ακριβώς η ανάκληση θα πρέπει να γνωρίζουμε ακριβώς πόσα matches έγιναν, και αν γνωρίζαμε κάτι τέτοιο, ποιος ο λόγος να έχουμε μια μηχανή αναζήτησης;). Τρίτον, η ακρίβεια και η ανάκληση δεν είναι στον πραγματικό κόσμο απλά αριθμοί· είναι δύο έννοιες που σχετίζονται στενά. Για παράδειγμα ενώ ψάχνουμε στις σελίδες απάντησης μιας μηχανής αναζήτησης για ένα ερώτημα που δώσαμε, περιμένουμε καθώς περνάμε τις σελίδες η ανάκληση να βελτιώνεται ενώ παράλληλα η ακρίβεια να χειροτερεύει.

2.2.1 Ανάκτηση και φιλτράρισμα πληροφορίας

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

Αυτή η δυσκολία που αντιμετωπίζουν τα συστήματα IR (μικρές τιμές ανάκλησης και ακρίβειας) γεννούν ένα άλλο επιστημονικό πεδίο το οποίο υπάρχει παράλληλα με το IR και είναι το *IF (Information Filtering)*. Σε ένα κλασσικό άρθρο οι Belkin και Croft παρουσίασαν δύο διαφορετικούς ορισμούς για τα δύο παραπάνω θέματα οι οποίοι έχουν κοινές τεχνικές αλλά διαφέρουν σε τρία βασικά στοιχεία [57]. Πρώτον, στο IR όταν ο χρήστης κάνει ένα ερώτημα περιμένει άμεση απόκριση. Στο IF ο χρήστης μπορεί να περιμένει, εν γνώσει του, για μεγάλο χρονικό διάστημα μέχρι να του παρουσιαστεί μία απάντηση. Επιπρόσθετα το IF χειρίζεται και θέματα που από τη φύση τους είναι δυναμικά και εντάσσει στο μηχανισμό του στοιχεία εκμάθησης σύμφωνα με τα κείμενα που προσθέτει στη συλλογή του. Τέλος, το βασικότερο είναι πως το IR αναζητά παραπλήσια κείμενα από μία μεγάλη συλλογή κειμένων σε αντίθεση με το IF το οποίο προσπαθεί να αφαιρέσει από μία συλλογή τα εισερχόμενα κείμενα που δεν είναι σχετικά.

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

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

1. η χρήση μέσων τιμών που προκύπτουν από την εκτέλεση διαφόρων ερωτημάτων μπορεί να αποκρύπτει σημαντικές ανωμαλίες στον αλγόριθμο ανάκτησης,

2. όταν συγκρίνουμε δύο αλγορίθμους, μπορεί να θέλουμε να μελετήσουμε κατά πόσο ο ένας είναι καλύτερος του άλλου για κάθε μία από τις πληροφοριακές ανάγκες που έχουμε και όχι συνολικά.

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

2.2.2 Μοντέλα ανάκτησης πληροφορίας

Τα τρία κλασσικά μοντέλα στην Ανάκτηση Πληροφορίας είναι το *Boolean*, το *Vector Space* και το Πιθανοτικό. Στο μοντέλο Boolean, τόσο τα κείμενα όσο και τα ερωτήματα αντιμετωπίζονται ως ένα σύνολο από όρους δεικτοδότησης. Κατά συνέπεια το μοντέλο μπορεί να θεωρηθεί ως συνολοθεωρητικό. Στο Vector Space, τα κείμενα και τα ερωτήματα αναπαρίστανται ως διανύσματα σε έναν $t-$διάστατο χώρο. Έτσι λέμε ότι το μοντέλο είναι αλγεβρικό. Το Πιθανοτικό μοντέλο εισάγει έναν τρόπο αναπαράστασης, ο οποίος βασίζεται στην πιθανοθεωρία και κατά συνέπεια το μοντέλο είναι πιθανοτικού χαρακτήρα.

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

εκτεταμένο Boolean. Στα αλγεβρικά μοντέλα έχουμε το γενικευμένο vector space, την λανθάνουσα σημασιολογική δεικτοδότηση (LSI) και το μοντέλο των νευρωνικών δικτύων. Στον πιθανοτικό τομέα εμφανίστηκαν τα δίκτυα εξαγωγής συμπεράσματος (inference networks) και τα δίκτυα πεποίθησης (belief networks). Εκτός από την χρήση του περιεχομένου των κειμένων, ορισμένα μοντέλα εκμεταλλεύονται και την εσωτερική δομή που φυσιολογικά υπάρχει στο γραπτό λόγο. Σε αυτή την περίπτωση λέμε ότι έχουμε ένα δομημένο μοντέλο. Για τη δομημένη ανάκτηση κειμένου, συναντούμε δύο μοντέλα, τις μη επικαλυπτόμενες λίστες (non-overlapping lists) και τους κοντινούς κόμβους *(proximal nodes)*.

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 ανάκτησης πληροφορίας.

Ορισμός 2.2.1. Ένα μοντέλο ανάκτησης πληροφορίας είναι η τετράδα $[D, Q, F, R(qi, dj)]$ όπου:

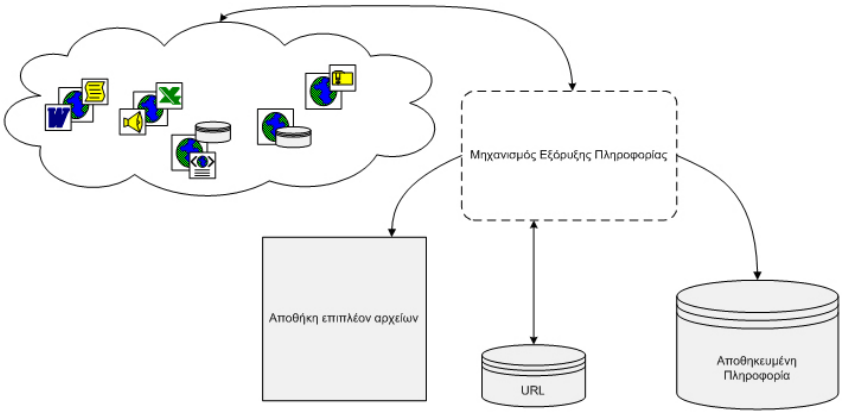
1. D είναι ένα σύνολο από λογικές αναπαραστάσεις για τα κείμενα της συλλογής
2. Q είναι ένα σύνολο από λογικές αναπαραστάσεις για τις πληροφοριακές ανάγκες του χρήστη. Αυτές οι αναπαραστάσεις καλούνται ερωτήματα
3. F είναι ένα υπόβαθρο για την μοντελοποίηση της αναπαράστασης των κειμένων, των ερωτημάτων και των σχέσεων μεταξύ τους
4. $R(q_i, d_j)$ είναι μια συνάρτηση κατάταξης, η οποία συνδέει έναν πραγματικό αριθμό με ένα ερώτημα $q_i \in Q$ και μια αναπαράσταση κειμένου $d_j \in D$. Μια τέτοια κατάταξη ορίζει μια διάταξη πάνω στα κείμενα πάντα με βάση το ερώτημα q_i.

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.

2.2.3 Μηχανισμοί εξόρυξης δεδομένων

Όλες οι μηχανές αναζήτησης πραγματοποιούν ανάκτηση πληροφορίας προκειμένου να μπορούν να εξυπηρετούν τους χρήστες τους. Έτσι, μέχρι σήμερα έχει κατασκευαστεί πληθώρα προγραμμάτων τα οποία είτε λειτουργώντας σαν αυτόνομες μονάδες είτε σε συνεργασία μεταξύ τους πραγματοποιούν εξόρυξη πληροφορίας. Η γενική ιδέα ενός μηχανισμού εξόρυξης πληροφορίας είναι εξαιρετικά απλή και φαίνεται στο παρακάτω σχήμα (Σχήμα 2.2).

Ένας τέτοιος μηχανισμός μπορεί να είναι ένας απλός υπολογιστής ή ακόμα και μερικές χιλιάδες υπολογιστές που λειτουργούν κάτω από την επίβλεψη ενός. Ο μηχανισμός ξεκινά να λειτουργεί περιδιαβαίνοντας σελίδες του διαδικτύου. Οι HTML σελίδες αποθηκεύονται σε μία βάση δεδομένων μαζί με επιπρόσθετες πληροφορίες για αυτές οι οποίες μπορεί να περιλαμβάνουν: το URL, την ώρα που ανακτήθηκε η σελίδα, το μέγεθός της και άλλα. Σε μία ξεχωριστή (συνήθως) βάση δεδομένων αποθηκεύονται όλα τα URL που έχουν ανακτηθεί και τα οποία ανακτώνται ανά τακτά χρονικά



Σχήμα 2.2: Μηχανισμός Εξόρυξης Πληροφορίας.

διαστήματα. Παράλληλα κάθε σελίδα αναλύεται προκειμένου να εξαχθούν από αυτή όλα τα links που περιέχει (σύμβολο $< a >$ στην HTML). Τα links που 'διαβάζει' ο μηχανισμός συγκρίνονται με αυτά που υπάρχουν αποθηκευμένα στη βάση δεδομένων URL και γίνονται οι κατάλληλες προσθήκες. Τέλος, κάποια επιπλέον αρχεία (doc, css, xml, scripts, πολυμέσα) αποθηκεύονται συνήθως σε καταλόγους που ονομάζονται κατάλληλα από τον μηχανισμό, έτσι ώστε να είναι σε θέση να τα προσπελάσει ανά πάσα στιγμή.

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

2.2.4 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 ξεπερνά τις 11 δισεκατομμύρια σελίδες [47].
- Δυναμικός χαρακτήρας
  - Το Internet αλλάζει ώρα με τη ώρα, ενώ στα κλασσικά συστήματα ανάκτησης δεδομένων υπάρχουν σταθερές βάσεις δεδομένων.
- Περιέχει ετερογενές υλικό
  - Υπάρχουν πολλοί διαφορετικοί τύποι αρχείων (κείμενα, εικόνες, βίντεο, ήχος, scripts) με αποτέλεσμα οι αλγόριθμοι ανάκτησης δεδομένων να πρέπει να εφαρμοστούν τόσο σε απλό κείμενο όσο και πολυμεσικά δεδομένα.
- Υπάρχει μεγάλο εύρος γλωσσών

  - Οι γλώσσες που χρησιμοποιούνται στο Διαδίκτυο υπολογίζονται σε πάνω από 100.

- Διπλές εγγραφές
  - Η αντιγραφή είναι ένα βασικό χαρακτηριστικό ου Διαδικτύου. Δεν είναι τυχαίο πως 25-30% των σελίδων του Διαδικτύου αποτελούν αντίγραφα άλλων σελίδων.

- Πολλά links από μία σελίδα σε άλλες
  - Υπολογίζεται πως σε κάθε σελίδα περιέχονται κατά μέσο όρο 10 links προς άλλες σελίδες.

- Πολλοί και διαφορετικών ειδών χρήστες
  - Κάθε χρήστης έχει τα δικές του ανάγκες αλλά και τις δικές του γνώσεις και απαιτήσεις από το Διαδίκτυο.

- Διαφορετική συμπεριφορά από τους χρήστες
  - Έχει υπολογιστεί πως περίπου το 90% των χρηστών του Διαδικτύου παρατηρούν μόνο την πρώτη σελίδα από αυτές που τους 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 σε σύγκριση με τα συνολικά αποτελέσματα 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 να έχουν υψηλή σχετικότητα με το ερώτημα 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 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 της 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. Χαρακτηριστικό παράδειγμα αυτού είναι η μετρική page rank του Google. 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.

2.2.5 Εξόρυξη γνώσης από αποθήκες δεδομένων

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

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

2.2.6 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.

2.2.7 Ανακάλυψη γνώσης από βάσεις δεδομένων σε σχέση με την εξόρυξη γνώσης και δεδομένων

Η ανακάλυψη γνώσης από βάσεις δεδομένων αναφέρεται σε ολόκληρη τη διαδικασία ανακάλυψης χρήσιμης πληροφορίας από μεγάλα σύνολα δεδομένων. Ένας τυπικός ορισμός δόθηκε από τους Frawley, Piatesky-Shaphiro & Matheus [93]:

Ορισμός 2.2.2. 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 E σε μια γλώσσα L η οποία περιγράφει ένα υποσύνολο δεδομένων $F_E \subseteq F$ εκμεταλλευόμενο κοινές ιδιότητες των δεδομένων του.

- Η διαδικασία ανακάλυψη γνώσης από βάσεις δεδομένων είναι μια διαδικασία πολλαπλών βημάτων, η οποία περιλαμβάνει την προ-επεξεργασία των δεδομένων, την αναζήτηση των προτύπων και την αξιολόγηση της εξαγόμενης γνώσης.

- Εγκυρότητα. Το εξαγόμενο πρότυπο (π. χ. περίληψη κειμένου) θα πρέπει να είναι συνεπές σε νέα δεδομένα με κάποιο βαθμό βεβαιότητας. Το ζήτημα της εγκυρότητας αποτελεί ένα από τα βασικά προβλήματα και αντικείμενο έρευνας στην εξόρυξη δεδομένων / πληροφορίας.

- Πιθανά χρήσιμο. Η εξαγωγή των προτύπων θα πρέπει να ακολουθείται από μερικές χρήσιμες διεργασίες όπως η αξιολόγησή τους από κάποιες συναρτήσεις χρησιμότητας. Για παράδειγμα η αυτόματη περίληψη ενός κειμένου θα πρέπει να μπορεί να αξιολογηθεί ως προς την χρησιμότητα / σαφήνειά και την πιστότητά του όσον αφορά το νόημα σε σχέση με το αρχικό κείμενο. Επίσης, θα ήταν χρήσιμο να εμπλουτιστεί η σημασιολογία των προτύπων, διατηρώντας όσο το δυνατόν περισσότερη γνώση από τα αρχικά δεδομένα η οποία μπορεί να φανεί χρήσιμη για τη λήψη αποφάσεων.

- Τελικά κατανοητό. Ο στόχος της εξόρυξης γνώσης είναι να προσδιοριστούν τα πρότυπα και να γίνουν κατανοητά, ώστε να μπορούν να οδηγήσουν ακόμη και τους μη ειδικούς σε χρήσιμα συμπεράσματα και αποφάσεις.

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

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

- Την ολοκλήρωση των δεδομένων. Υπάρχουν διαφορετικά είδη αποθηκών πληροφοριών που μπορούν να χρησιμοποιηθούν στη διαδικασία εξόρυξης γνώσης. Κατά συνέπεια οι πολλαπλές πηγές δεδομένων μπορούν να συνδυαστούν καθορίζοντας το σύνολο στο οποίο τελικά η διαδικασία εξόρυξης πρόκειται να εφαρμοστεί.

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

- Τον καθαρισμό και την προ-επεξεργασία δεδομένων. Αυτό το βήμα περιλαμβάνει βασικές διαδικασίες όπως η αφαίρεση του θορύβου, η συλλογή των απαραίτητων πληροφοριών για τη διαμόρφωση ή τη μέτρηση του θορύβου, η απόφαση σχετικά με τις στρατηγικές διαχείρισης των ελλειπόντων πεδίων δεδομένων.

- Τον μετασχηματισμό των δεδομένων. Τα δεδομένα μετασχηματίζονται ή παγιώνονται σε μορφές κατάλληλες για εξόρυξη. Χρήση των μεθόδων μείωσης διαστάσεων ή μετασχηματισμού για τη μείωση του αριθμού των υπό εξέταση μεταβλητών ή την εύρεση κατάλληλης αντιπροσώπευσης των δεδομένων χωρίς μεταβλητές.

- Την επιλογή των στόχων και των αλγορίθμων εξόρυξης δεδομένων. Σε αυτό το βήμα αποφασίζουμε το στόχο της διαδικασίας εξόρυξης γνώσης, επιλέγοντας τους στόχους εξόρυξης δεδομένων που θέλουμε να επιτύχουμε. Επίσης, επιλέγονται οι μέθοδοι που θα χρησιμοποιηθούν. Αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου και παραμέτρων.

- Την εξόρυξη δεδομένων. Εφαρμόζοντας ευφυείς μεθόδους, ψάχνουμε για ενδιαφέροντα πρότυπα γνώσης. Τα πρότυπα θα μπορούσαν να είναι μιας συγκεκριμένης αντιπροσωπευτικής μορφής ή ενός συνόλου τέτοιων αντιπροσωπεύσεων, όπως κανόνες κατηγοριοποίησης, δέντρα, συσταδοποίηση, κλπ Η απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα προηγούμενα βήματα.

- Την αξιολόγηση των προτύπων. Τα εξαγόμενα πρότυπα αξιολογούνται με κάποια μέτρα, προκειμένου να προσδιοριστούν τα πρότυπα τα οποία αντιπροσωπεύουν τη γνώση, δηλαδή τα αληθινά ενδιαφέροντα πρότυπα.

- Τη σταθεροποίηση και παρουσίαση της γνώσης. Σε αυτό το βήμα, η εξορυγμένη γνώση ενσωματώνεται το σύστημα ή απλά την απεικόνισή μας και κάποιες τεχνικές αντιπροσώπευσης γνώσης χρησιμοποιούνται για να παρουσιάσουν την εξορυγμένη γνώση στο χρήστη. Επίσης, ελέγχουμε για επίλυση τυχών συγκρούσεων με προηγούμενη εξορυγμένη γνώση.

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

2.2.8 Η διαδικασία εξόρυξης δεδομένων

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

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

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

2.2.9 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 μιας 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, κ. λπ. 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.

2.2.10 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 [54], μελέτησαν την ανακάλυψη πολύ απλών προτύπων, που τα ονόμασαν πρότυπα συσχέτισης ζευγών λέξεων k-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. [135]. Προσπαθούν να ανακαλύψουν τις σχέσεις που υπάρχουν ανάμεσα στα διάφορα θέματα που παρουσιάζονται σε εφημερίδες. Επιχειρούν να ανακαλύψουν τον τρόπο που τα θέματα της λεγόμενης πρώτης σελίδας, επηρεάζουν και όλα τα υπόλοιπα θέματα της ειδησιογραφίας.

Οι συσχετίσεις που υπάρχουν ανάμεσα στα διάφορα ειδησιογραφικά θέματα, καλούνται εφήμερες (Ephemeral Associations). Άλλη χαρακτηριστική εφαρμογή, αποτελεί η ανακάλυψη προτύπων σε σύνολα ακολουθιών DNA, που προτείνουν οι Kiem Hoang και Phuc Do [105]. Μελετούν υποακολουθίες που εμφανίζονται πολύ συχνά στο σύνολο των ακολουθιών DNA, για την ανακάλυψη εκείνων των κανόνων συσχέτισης, που βασίζονται στην επανάληψη.

2.2.11 Ανάκτηση γνώσης από βάσεις δεδομένων

Η *ανάκτηση γνώσης από βάσεις δεδομένων* (Knowledge Discovery in Databases - KDD) είναι η μη τετριμμένη διαδικασία της αναγνώρισης έγκυρων, καινότυπων, ενδεχόμενα χρήσιμων και τελικά κατανοητών προτύπων δεδομένων. Τα ακατέργαστα δεδομένα είναι πάντοτε 'ακάθαρτα' με την έννοια ότι πάντα θα υπάρχουν διπλοεγγραφές, ελλειπή πεδία και μη ακριβές τιμές δεδομένων. Είναι επιθυμητό επομένως, τα αποτελέσματα των αναζητήσεων να πρέπει να περάσουν από κάποιο στάδιο εκκαθάρισης πριν παρουσιαστούν στον χρήστη. Η εκκαθάριση δεδομένων στην KDD διαδικασία είναι ένα βασικό βήμα για την αφαίρεση του θορύβου και των outliers[1], την συγκέντρωση των σχετικών πληροφοριών για μοντελοποίηση του θορύβου και την λήψη αποφάσεων για τα ελλειπή 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.

2.3 Προεπεξεργασία δεδομένων και εξαγωγή κωδικολέξεων

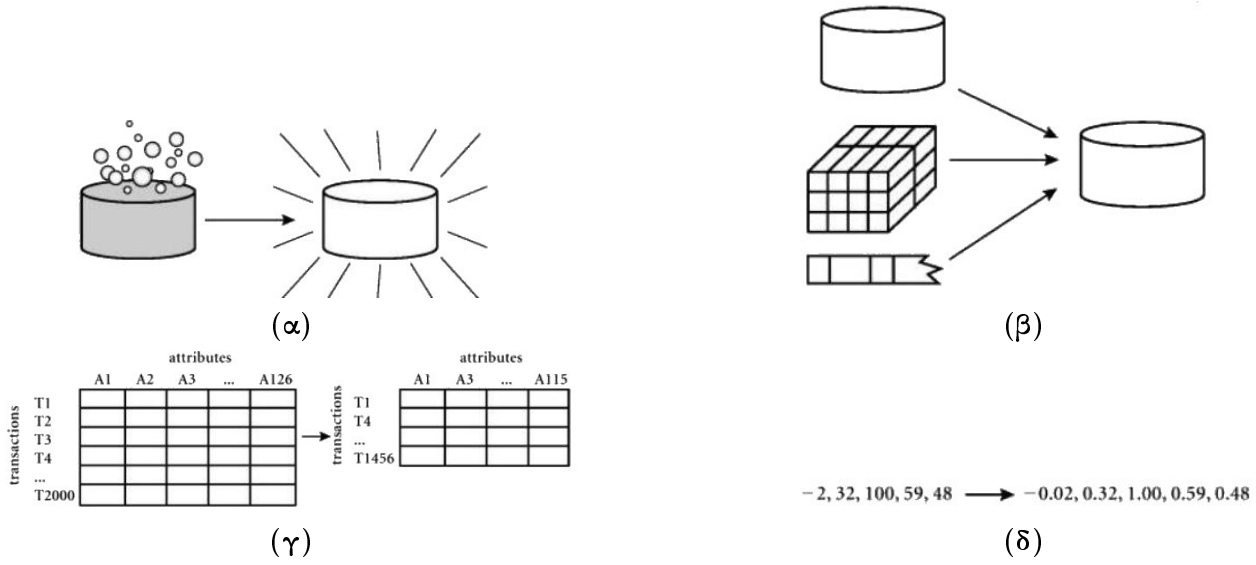
Η εξαγωγή κωδικολέξεων, η διαδικασία δηλαδή κατά την οποία δοθέντος ενός αρχικού κειμένου καταλήγει σε κάποιες λέξεις-κλειδιά που περιγράφουν / αντιπροσωπεύουν το αρχικό κείμενο, αποτελεί τη βάση για κάθε σύστημα ανάκτησης πληροφορίας. Στόχος είναι να εντοπιστούν οι καλύτερες λέξεις του κειμένου και να αντιστοιχιστεί σε αυτές ένα κατάλληλο σκορ που περιγράφει την σημασία τους. Οι 'καλύτερες' λέξεις σε ένα κείμενο είναι ουσιαστικά οι πιο αντιπροσωπευτικές αυτού όσον αφορά στο συνολικό νόημα που εκφράζει το κείμενο. Προκειμένου να καταλήξουμε σε αυτές τις κωδικολέξεις (ή αλλιώς keywords) ένα πλήθος βημάτων πρέπει να ακολουθηθούν και τα οποία κρίνουν ουσιαστικά την αποτελεσματικότητα όλου του μηχανισμού.

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 (Σχήμα 2.3). 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 επίσης να χρησιμοποιηθεί από

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

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.



Σχήμα 2.3: Τεχνικές προεπεξεργασίας δεδομένων (α)Καθαρισμός δεδομένων (β)Ολοκλήρωση δεδομένων (γ)Αφαίρεση δεδομένων (δ)Μετασχηματισμός δεδομένων

2.3.1 Ορθογραφικός έλεγχος

Δεδομένου ότι τα κείμενα που δέχεται ο μηχανισμός προεπεξεργασίας κειμένου πηγάζουν από πηγές του διαδικτύου που μπορεί μην έχουν σαφείς πολιτικές αντιμετώπισης ορθογραφικών λαθών πριν την δημοσίευσή τους, είναι σημαντικό κατά την είσοδο ενός κειμένου στον μηχανισμό αυτό να ανιχνεύονται και να διορθώνονται τυχών ορθογραφικά λάθη. Τούτο προκύπτει από το γεγονός ότι τα ορθογραφικά λάθη είναι πολύ πιθανό να δημιουργήσουν προβλήματα στις διαδικασίες ανάκτησης πληροφορίας που ακολουθούν την προεπεξεργασία δεδομένων.

2.3.2 Αφαίρεση σημείων στίξης

Τα *σημεία στίξης* (*punctuation*) ενός κειμένου δεν προσδίδουν σημασιολογική πληροφορία σε αυτό και άρα δεν δεικτοδοτούνται. Είναι επομένως αναγκαίο, ένα σύστημα ανάκτησης πληροφορίας να αφαιρεί κάθε σημείο στίξης από το αρχικό κείμενο σε πρώιμα στάδια της προεπεξεργασίας. Ιδιαίτερη μέριμνα πρέπει να λαμβάνεται ώστε να συγκρατείται το τέλος της κάθε πρότασης (π. χ. με κάποιο άλλο διαχωριστικό πέραν της τελείας) ώστε να είναι δυνατός ο μετέπειτα διαχωρισμός των προτάσεων. Η διαδικασία θα πρέπει να λαμβάνει όσο το δυνατόν καλύτερα υπ' όψιν τις γλωσ-

σολογικές 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

2.3.3 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 αποθηκευθεί ως meta-δεδομένο για το κείμενο.

2.3.4 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.

2.3.5 Αφαίρεση *Stopwords*

Τα *Stopwords* είναι λέξεις οι οποίες περιέχουν μικρής σημασίας πληροφορία για το κείμενο. Είναι ως επί το πλείστον 'λειτουργικές' λέξεις οι οποίες εμφανίζονται σε ένα μεγάλο μέρος των κειμένων και ως εκ' τούτου, περιέχουν μικρή ικανότητα διάκρισης για δήλωση συσχέτισης. Στην διαδικασία της ανάκτησης πληροφορίας, τα Stopwords συνήθως αγνοούνται και για λόγους αποδοτικότητας, αφού η αποθήκευση των Stopwords σε ένα ευρετήριο λαμβάνει σημαντικό χώρο λόγω της υψηλής συχνότητας εμφάνισής τους. Η αφαίρεση των Stopwords από ένα κείμενο θα μπορούσαμε να πούμε ότι είναι μια τεχνική αφαίρεσης δεδομένων η οποία απαλλάσσει το κείμενο από ένα σημαντικό μέγεθος μη-χρήσιμης πληροφορίας.

2.3.6 *Stemming*

Η διαδικασία του *Stemming* εξάγει τη μορφολογική ρίζα κάθε λέξης. Μερικές φορές όμως ο όρος 'ρίζα' χρησιμοποιείται για να περιγράψει την λέξη χωρίς το επίθεμα αλλά με την λεξικολογική κατάληξή της. Για παράδειγμα η λέξη chatters έχει ως επιθεματική ρίζα τη λέξη chatter αλλά ως λεξικολογική ρίζα την λέξη chat. Οι επιθεματικές ρίζες καλούνται και stems και συνήθως η 'ρίζα' μιας λέξης αντιστοιχεί στην μορφολογική της ρίζα.

Παράλληλα, και ανάλογα με τη λίστα κανόνων που χρησιμοποιούνται, η διαδικασία του Stemming μπορεί να περιλαμβάνει και τη λημματοποίηση του κειμένου· την εύρεση δηλαδή του λήμματος κλιτών λέξεων (π. χ. children→child). Σε καθολικές μηχανές αναζήτησης, το βασικό πρόβλημα της διαδικασίας του Stemming είναι ότι είναι γλώσσο-εξαρτώμενη, και ενώ για την αγγλική γλώσσα υπάρχουν *Stemmers* βασισμένοι σε κανόνες, για άλλες γλώσσες είναι δύσκολη η ανάπτυξη.

2.3.7 Αναγνώριση μερών του λόγου

Με τον όρο 'αναγνώριση μερών του λόγου' (part of speech tagging) εννοούμε τη διαδικασία αντιστοίχησης μοναδικής ετικέτας (tag) σε κάθε λέξη ενός συνόλου κειμένων, ώστε η ετικέτα να παριστάνει το μέρος του λόγου στο οποίο ανήκει η λέξη. Η αναγνώριση μερών του λόγου αποτελεί μέρος του ευρύτερου σταδίου της μορφολογικής ανάλυσης κειμένων και χρησιμοποιείται σε πολλά συστήματα επεξεργασίας φυσικής γλώσσας. Είναι μία ενδιαφέρουσα περιοχή τόσο από πρακτικής όσο και από ερευνητικής πλευράς. Ιδιαίτερο ερευνητικό ενδιαφέρον παρουσιάζει η περίπτωση χρήσης τεχνικών μηχανικής μάθησης, ιδιαίτερα ενεργητικής μάθησης, κατά την οποία το ίδιο το σύστημα συμμετέχει στην επιλογή των παραδειγμάτων εκπαίδευσής του. Αξίζει να σημειωθεί ότι η πλειοψηφία των συστημάτων αναγνώρισης μερών του λόγου χρησιμοποιεί ήδη μηχανική μάθηση αλλά οι τεχνικές ενεργητικής μάθησης δεν έχουν ακόμα αξιοποιηθεί επαρκώς στην περιοχή αυτή.

Η σημαντικότητα της συγκεκριμένης περιοχής έγκειται στο γεγονός ότι η μορφολογική πληροφορία που αποδίδεται σε κάθε λέξη ενός κειμένου αποτελεί τη βάση για την περαιτέρω επεξεργασία του. Συνεπώς ένα τέτοιο σύστημα, μορφολογικής ανάλυσης κειμένων, μπορεί να χρησιμοποιηθεί ως τμήμα διαφόρων άλλων συστημάτων επεξεργασίας φυσικής γλώσσας, όπως συντακτικοί αναλυτές, διορθωτές κειμένων, συστήματα αναγνώρισης φωνής κ. α. Εκτός όμως από πρακτικό ενδιαφέρον, η αναγνώριση μερών του λόγου έχει και ερευνητικό ενδιαφέρον, καθώς ενσωματώνοντας πολλές μορφολογικές πληροφορίες στις ετικέτες δημιουργείται ένα πολύ μεγάλο πλήθος κατηγοριών. Έτσι η μορφολογική ανάλυση κειμένων καθίσταται ένα δύσκολο πρόβλημα κατηγοριοποίησης. Ένα σημαντικό ερώτημα που θα μπορούσε ίσως να προκύψει είναι το γιατί δε χρησιμοποιείται ένα ηλεκτρονικό λεξικό, της εκάστοτε γλώσσας, για τη μορφολογική ανάλυση των λέξεων. Η απάντηση είναι απλή και έγκειται στους περιορισμούς που υπεισέρχονται σε μία τέτοια λύση. Καταρχάς, η κατασκευή ενός ηλεκτρονικού λεξικού είναι μία χρονοβόρα και ιδιαίτερα ακριβή διαδικασία, με συνέπεια τα περισσότερα ηλεκτρονικά λεξικά να μην είναι ελεύθερα διαθέσιμα στο κοινό. Επιπλέον, ένα λεξικό, περιέχοντας πεπερασμένο πλήθος λέξεων, είναι αδύνατο να καλύψει όλες τις πιθανές λέξεις που μπορεί να εμφανιστούν, ιδιαίτερα κύρια ονόματα και νέους τεχνικούς όρους. Τέλος, ένα σύστημα που συμβουλεύεται απλά ένα λεξικό χωρίς να λαμβάνει υπόψη του τα συμφραζόμενα της κάθε λέξης, σε πολλές περιπτώσεις δε μπορεί να αποφασίσει για την ετικέτα που θα αποδώσει σε μία λέξη (π. χ. η λέξη «διατάξεις» εμφανίζεται και ως ρήμα και ως ουσιαστικό).

Σήμερα μπορούμε να πούμε ότι υπάρχουν αρκετά συστήματα που αντιμετωπίζουν το εν' λόγω πρόβλημα με πιθανότητες σωστής πρόβλεψης που αγγίζουν το 100% (φυσικά τα αποτελέσματα εξαρτώνται από την εκάστοτε γλώσσα).

Ανάκτηση ουσιαστικών

Η ανάκτηση ουσιαστικών αποτελεί στην ουσία μία υποκατηγορία της διαδικασίας αναγνώρισης των μερών του λόγου των προτάσεων ενός κειμένου. Είναι ευρέως αποδεκτό ότι όσον αφορά στις λέξεις ενός δοθέντος κειμένου, τα ουσιαστικά καλύπτουν το μεγαλύτερο μέρος της σημασιολογικής πληροφορίας αυτού. Η αναγνώριση των ουσιαστικών ενός κειμένου επομένως, αποτελεί ένα ακόμη σημαντικό βήμα όσον αφορά στην προεπεξεργασία κειμένου ουτοσώστε να γίνει ένα πιο αποτελεσματικό φιλτράρισμα στην πληροφορία που περιέχει το κείμενο.

2.4 Περίληψη πληροφορίας

Η διαδικασία της περίληψης κειμένου *(Text Summarization)*, αποσκοπεί στην παρουσίαση των κύριων σημείων ενός εγγράφου, σε μία περιεκτική μορφή. Μία πραγματική περίληψη, θα πρέπει να εκφράζει την ουσία του εγγράφου, αποκαλύπτοντας το βαθύτερο νόημα του περιεχομένου του.

Σκοπός της είναι, η ανακάλυψη ενδιαφέρουσας και απροσδόκητης πληροφορίας. Σύμφωνα με τον Crangle Colleen [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 *Natural Language Processing - NLP*, συμπεριλαμβανομένων γραμματικών και λεξικών για την ανάλυση του κειμένου, η εξαγωγή μπορεί να θεωρηθεί ως μια διεργασία επιλογής σημαντικών αποσπασμάτων (προτάσεων, παραγράφων, κ.λπ.) από το αρχικό κείμενο και συνένωσής του σε μια νέα πιο σύντομη έκδοση.

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

2.4.1 Χρησιμότητα της περίληψης κειμένου

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

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

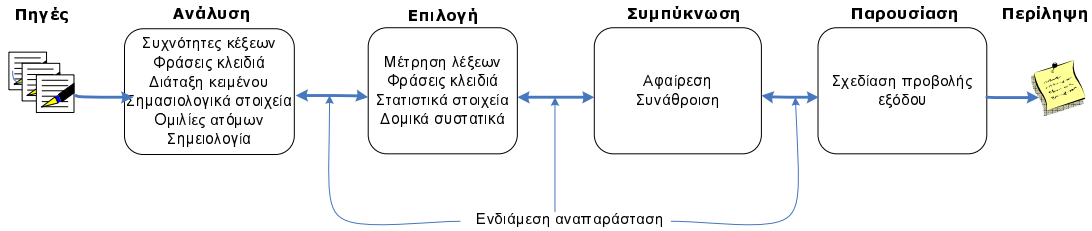
Οι περιλήψεις κειμένων, μπορούν να χρησιμοποιηθούν από αναλυτές πληροφοριών, έτσι ώστε να είναι σε θέση να γνωρίζουν αν θα πρέπει να μελετήσουν κάποια κείμενα στο σύνολο τους, και κάποια άλλα με διαφορετικό και πιο περιεκτικό τρόπο. Οι περιλήψεις μπορούν να αποκαλύψουν ομοιότητες στο περιεχόμενο των κειμένων, οι οποίες μπορούν να χρησιμοποιηθούν για την μετέπειτα ομαδοποίηση ή κατηγοριοποίηση των εγγράφων. Η διαδικασία της κατηγοριοποίησης ή ομαδοποίησης των περιλήψεων περισσοτέρων του ενός εγγράφου, μέσα σε μία συλλογή, μπορεί να

αποκαλύψει αναπάντεχες σχέσεις μεταξύ των εγγράφων. Επιπλέον, η περίληψη μιας συλλογής από σχετιζόμενα έγγραφα, που έχουν επεξεργαστεί μαζί, μπορεί να αποκαλύψει αθροιστική πληροφορία, που υπάρχει μόνο στο επίπεδο της συλλογής των εγγράφων.

2.4.2 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 (Εικόνα 2.4):

1. Ανάλυση του κειμένου
2. Αναγνώριση / Εντοπισμός των σημαντικών τμημάτων του κειμένου
3. Συμπύκνωση πληροφορίας και
4. Παραγωγή της αναπαράστασης της περίληψης που προκύπτει.



Σχήμα 2.4: Γενική διαδικασία παραγωγής περίληψης.

2.4.3 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.

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.

Μέθοδοι βασιζόμενοι σε περιεχόμενο

Αυτές οι μέθοδοι υπολογίζουν την ομοιότητα ανάμεσα σε δύο κείμενα σε ένα πιο λεπτομερές επίπεδο από αυτό των απλών προτάσεων. Η βασική μέθοδος συνίσταται από τον υπολογισμό της

ομοιότητας μεταξύ του αρχικού κειμένου και της περίληψής του με χρήση της μετρικής ομοιότητας συνημιτόνου:

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

όπου τα X και Y βασίζονται στο μοντέλο διανυσματικού χώρου.

Συσχέτιση ομοιότητας

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

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).

2.5 Κατηγοριοποίηση πληροφορίας

Η *κατηγοριοποίηση* αποτελεί μια από τις βασικές εργασίες *εξόρυξης δεδομένων*. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (μη κατηγοριοποιημένο) το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαριστώνται γενικά από τις εγγραφές της βάσης δεδομένων και η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποιες από τις προκαθορισμένες κατηγορίες. Ο στόχος της κατηγοριοποίησης κειμένου είναι η κατάταξη των κειμένων σε μια σταθερή σειρά προκαθορισμένων κατηγοριών. Κάθε κείμενο μπορεί να ανήκει σε καμία, ακριβώς μία, ή περισσότερες κατηγορίες.

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ-κατηγοριοποιημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιεί δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί. Στις περισσότερες περιπτώσεις, υπάρχει ένας περιορισμένος αριθμός (προκαθορισμένων) κατηγοριών και ο αλγόριθμος αναθέτει κάθε εγγραφή στην κατάλληλη κατηγορία. Για το σκοπό αυτό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο κατηγορίες. Η πρώτη χρησιμοποιεί *δέντρα απόφασης* (Decision Trees) και η δεύτερη *Νευρωνικά δίκτυα* (*Neural Networks*). Και οι δύο στηρίζονται στην ιδέα της εκπαίδευσης (training) με τη βοήθεια ενός υποσυνόλου δεδομένων που ονομάζεται *σύνολο εκπαίδευσης* (*training set*). Το υποσύνολο αυτό επιλέγεται σαν αντιπροσωπευτικό δείγμα του συνολικού όγκου δεδομένων. Με την εφαρμογή της διαδικασίας εκπαίδευσης καθορίζονται κάποια πρότυπα για τις κατηγορίες δεδομένων. Έτσι, όταν προκύψει ένα νέο στοιχείο μπορεί εύκολα να κατηγοριοποιηθεί.

2.5.1 Αλγόριθμοι για κατηγοριοποίηση πληροφορίας

Η κατηγοριοποίηση χαρακτηρίζεται από ένα καλά καθορισμένο σύνολο κατηγοριών καθώς και ένα σύνολο από κατηγοριοποιημένα (pre-classified) παραδείγματα. Αντίθετα, η διαδικασία συσταδοποίησης δεν στηρίζεται σε προκαθορισμένες κατηγορίες ή παραδείγματα. Γενικά, ο στόχος της

διαδικασίας κατηγοριοποίησης είναι η δημιουργία ενός μοντέλου που θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων των οποίων η κατηγοριοποίηση είναι άγνωστη. Πιο συγκεκριμένα, η κατηγοριοποίηση μπορεί να περιγραφεί ως μια διαδικασία δύο βημάτων:

1. Εκμάθηση (Learning). Σε αυτό το βήμα χτίζεται ένα μοντέλο περιγράφοντας ένα προκαθορισμένο σύνολο από κατηγορίες δεδομένων. Τα δεδομένα εκπαίδευσης (training data) αναλύονται από έναν αλγόριθμο κατηγοριοποίησης για να κατασκευάσουν στη συνέχεια το μοντέλο. Τα στοιχεία που αποτελούν το σύνολο κατάρτισης επιλέγονται τυχαία από έναν πληθυσμό δεδομένων και ανήκουν σε μία από τις προκαθορισμένες κατηγορίες. Δεδομένου ότι η κατηγορία των δειγμάτων εκπαίδευσης είναι γνωστή, αυτό το βήμα είναι επίσης γνωστό σας '*εποπτευόμενη μάθηση*' (*supervised learning*).

2. Κατηγοριοποίηση (Classification). Σε αυτό το βήμα χρησιμοποιούνται τα δοκιμαστικά δεδομένα (data set) για να υπολογίσουν την ακρίβεια του μοντέλου. Υπάρχουν διάφορες μέθοδοι για να εκτιμηθεί η ακρίβεια του κατηγοριοποιητή. Τα δεδομένα εκπαίδευσης επιλέγονται τυχαία και είναι ανεξάρτητα. Το μοντέλο κατηγοριοποιεί κάθε ένα από τα δοκιμαστικά παραδείγματα (training samples). Στη συνέχεια η κατηγορία που ανήκουν τα δεδομένα με βάση το σύνολο δοκιμαστικών δεδομένων συγκρίνεται με την πρόβλεψη που έκανε το μοντέλο για την κατηγορία. Η ακρίβεια του μοντέλου σε ένα καθορισμένο σύνολο δεδομένων δοκιμής είναι το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά από το υπό εκπαίδευση μοντέλο. Εάν η ακρίβεια θεωρείται ως αποδεκτή, το μοντέλο μπορεί πλέον να χρησιμοποιηθεί για να κατηγοριοποιήσει και τα μελλοντικά δείγματα δεδομένων, των οποίων η κατηγοριοποίηση είναι άγνωστη.

Bayesian κατηγοριοποίηση

Η *Bayesian* κατηγοριοποίηση βασίζεται στη στατιστική θεωρία κατηγοριοποίησης του Bayes. Ο στόχος είναι να κατηγοριοποιηθεί ένα δείγμα X σε μια από τις δεδομένες κατηγορίες $C_1, C_2, ..., C_n$ χρησιμοποιώντας ένα μοντέλο πιθανότητας που ορίζεται σύμφωνα με τη θεωρία Bayes. Κάθε κατηγορία χαρακτηρίζεται από μια εκ των προτέρων πιθανότητα (a priori probability) παρατήρησης της κλάσης C_i. Επίσης, υποθέτουμε ότι το δεδομένο δείγμα Q ανήκει σε μια κλάση C_i, με την υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας: $p(X/C_i) \in [0, 1]$. Κατόπιν, χρησιμοποιώντας τους ανωτέρω ορισμούς και βασιζόμενοι στη θεωρία Bayes, καθορίζουμε την εκ των υστέρων (posterior) πιθανότητα $p(C_i/x)$ ως:

$$p(C_i|X) = \frac{p(X|C_i)p(C_i)}{p(X)}$$

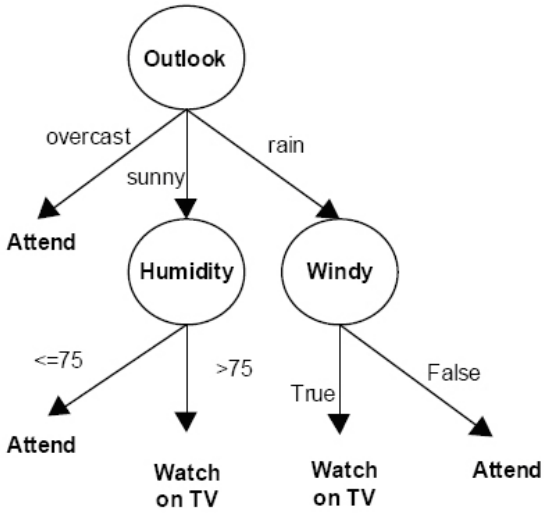
Ο απλούστερος Bayesian κατηγοριοποιητής είναι ο *Naive Bayesian*. Αυτός υποθέτει ότι η επίδραση ενός γνωρίσματος σε μια δεδομένη κατηγορία είναι ανεξάρτητη από τις τιμές των άλλων γνωρισμάτων. Αυτή η υπόθεση γίνεται για να απλοποιήσει τους υπολογισμούς που εμπλέκονται και καλείται *υπό συνθήκη ανεξαρτησία κατηγορίας*. Παρότι αυτή η υπόθεση δεν ισχύει συνήθως σε πραγματικά δεδομένα, ο αλγόριθμος είναι αρκετά αποτελεσματικός. Πρόσφατη ανάλυση του προβλήματος της Bayesian κατηγοριοποίησης έδειξε ότι υπάρχουν κάποιοι θεωρητικοί λόγοι για την εμφανώς περίεργη καλή απόδοση του έχει ο naive τρόπος [179]. Το πλεονέκτημα δε του naive κατηγοριοποιητή είναι ότι απαιτεί μικρό μέγεθος δεδομένων εκπαίδευσης για τον υπολογισμό των παραμέτρων κατηγοριοποίησης. Δεδομένου ότι οι μεταβλητές θεωρούνται ανεξάρτητες, μόνο οι τυπικές απόκλισεις των μεταβλητών για κάθε κλάση κατηγοριοποίησης χρειάζονται να καθοριστούν.

Ένας άλλος Bayesian κατηγοριοποιητής είναι τα Bayesian Belief Networks, τα οποία προσδιορίζουν τις συνδεδεμένες υπό συνθήκη κατανομές πιθανότητας στοχεύοντας στο να λάβουν υπόψη τις εξαρτήσεις που μπορούν να υπάρξουν μεταξύ των μεταβλητών.

Δέντρα απόφασης

Τα *δέντρα απόφασης* είναι μια από τις ευρέως χρησιμοποιούμενες τεχνικές για την κατηγοριοποίηση και την πρόβλεψη. Ένα δέντρο απόφασης κατασκευάζεται με βάση ένα σύνολο εκπαίδευσης προ-κατηγοριοποιημένων δεδομένων. Κάθε ένας από τους εσωτερικούς κόμβους του δέντρου απόφασης προσδιορίζει τον έλεγχο ενός γνωρίσματος και κάθε κλειδί που 'κατεβαίνει' από εκείνον τον κόμβο αντιστοιχεί σε μία από τις πιθανές τιμές για το συγκεκριμένο γνώρισμα. Επίσης, κάθε φύλλο αντιστοιχεί σε μια από τις κατηγορίες που έχουν ορισθεί.

Η διαδικασία για την κατηγοριοποίηση ενός νέου δείγματος με βάση ένα δέντρο απόφασης είναι η ακόλουθη: ξεκινώντας από τη ρίζα του δέντρου και εξετάζοντας τα γνωρίσματα που καθορίζονται από τον κόμβο αυτό προσδιορίζονται διαδοχικά οι εσωτερικοί κόμβοι που θα επισκεφθούμε έως ότου καταλήξουμε σε ένα φύλλο. Σε κάθε εσωτερικό κόμβο εξετάζεται εάν το δείγμα ικανοποιεί το συγκεκριμένο κόμβο. Η έκβαση αυτής της δοκιμής σ' έναν εσωτερικό κόμβο καθορίζει το κλαδί που θα διασχίσουμε στη συνέχεια καθώς και τον επόμενο κόμβο που θα επισκεφθούμε. Η κατηγορία του υπό μελέτη δείγματος είναι η κατηγορία του τελικού κόμβου ο οποίος αντιστοιχεί σε φύλλο του δέντρου.



Σχήμα 2.5: Δέντρο Απόφασης.

Διάφοροι αλγόριθμοι κατασκευής των δέντρων απόφασης έχουν αναπτυχθεί κατά τη διάρκεια των τελευταίων ετών. Μερικοί από τους πιο γνωστούς είναι οι:

- ID3
- C4.5
- SPRINT
- SLIQ

- CART
- RainForest

Γενικά, οι περισσότεροι από τους αλγορίθμους έχουν δύο διακριτές φάσεις: τη φάση οικοδόμησης και τη φάση περικοπής. Στη φάση οικοδόμησης, το σύνολο των δεδομένων εκπαίδευσης χωρίζεται κατ΄ επανάληψη μέχρις ότου όλα τα δείγματα σ΄ ένα τμήμα να ανήκουν στην ίδια κατηγορία. Το αποτέλεσμα είναι ένα δέντρο που κατηγοριοποιεί κάθε στοιχείο του συνόλου εκπαίδευσης. Ωστόσο, το δέντρο που έχει κατασκευαστεί μπορεί να είναι ευαίσθητο στις στατιστικές παρατυπίες του συνόλου κατάρτισης. Κατά συνέπεια, οι περισσότεροι από τους αλγορίθμους εκτελούν μια φάση περικοπής μετά από τη φάση κατασκευής του δέντρου, στην οποία οι κόμβοι περικόπτονται για να αποτραπούν οι επικαλύψεις και για να δημιουργηθεί ένα δέντρο με υψηλότερη ακρίβεια. Οι διάφοροι αλγόριθμοι κατασκευής δέντρων απόφασης χρησιμοποιούν διαφορετικούς αλγορίθμους για την επιλογή του κριτηρίου ελέγχου για την κατηγοριοποίηση ενός συνόλου δεδομένων. Ένας από τους πιο πρόσφατους αλγορίθμους, ο *CLS*, εξετάζει όλα τα δυνατά δέντρα αποφάσεων σ΄ ένα συγκεκριμένο βάθος και στη συνέχεια επιλέγει τον έλεγχο που ελαχιστοποιεί το υπολογιστικό κόστος κατηγοριοποίησης ενός στοιχείου. Ο ορισμός αυτού του κόστους αποτελείται από το κόστος καθορισμού των τιμών των χαρακτηριστικών για έλεγχο καθώς και το κόστος λανθασμένης κατηγοριοποίησης.

Οι αλγόριθμοι *ID3* και *C4.5* βασίζονται σε μια στατιστική ιδιότητα, καλούμενη κέρδος πληροφορίας (information gain), προκειμένου να επιλέξουμε το γνώρισμα που θα ελέγξουμε σε κάθε κόμβο του δέντρου. Ο ορισμός του μέτρου βασίζεται στην εντροπία, η οποία χαρακτηρίζει την καθαρότητα μιας αφηρημένης επιλογής των δειγμάτων. Εναλλακτικά οι αλγόριθμοι όπως ο SLIQ, SPRINT επιλέγουν το γνώρισμα που θα ελεγχθεί με βάση το δείκτη *GINI* και όχι το μέτρο εντροπίας. Το καλύτερο γνώρισμα για τον έλεγχο δίνει και τη χαμηλότερη τιμή για τον δείκτη GINI.

Νευρωνικά δίκτυα

Μια άλλη προσέγγιση της κατηγοριοποίησης που χρησιμοποιείται σε πολλές εφαρμογές εξόρυξης γνώσης για πρόβλεψη και κατηγοριοποίηση βασίζεται στα *νευρωνικά δίκτυα*. Οι μέθοδοι αυτής της προσέγγισης χρησιμοποιούν τα νευρωνικά δίκτυα για να κατασκευάσουν ένα μοντέλο κατηγοριοποίησης ή πρόβλεψης. Τα κύρια βήματα της διαδικασίας είναι:

- Αναγνώριση των χαρακτηριστικών εισόδου και εξόδου.
- Κατασκευή ενός δικτύου με την κατάλληλη τοπολογία.
- Επιλογή του σωστού συνόλου εκπαίδευσης.
- Εκπαίδευση του δικτύου με βάση ένα αντιπροσωπευτικό σύνολο δεδομένων. Τα δεδομένα πρέπει να απεικονίζονται με τέτοιον τρόπο ώστε να μεγιστοποιηθεί η δυνατότητα του δικτύου να αναγνωρίζει πρότυπα.
- Έλεγχος του δικτύου χρησιμοποιώντας ένα σύνολο ελέγχου το οποίο είναι ανεξάρτητο από το σύνολο εκπαίδευσης.

Το μοντέλο που παράγεται από το δίκτυο εφαρμόζεται για να προβλέψει τις κατηγορίες των μη κατηγοριοποιημένων δειγμάτων.

Τα νευρωνικά δίκτυα αποτελούνται από 'νευρώνες' με βάση τη νευρωνική δομή του εγκεφάλου. Επεξεργάζονται τα στοιχεία ένα κάθε φορά και 'μαθαίνουν' συγκρίνοντας την κατηγοριοποίησή

τους για μια εγγραφή (που, στην έναρξη, είναι κατά ένα μεγάλο μέρος αυθαίρετη) με τη γνωστή πραγματική κατηγοριοποίηση της εγγραφής. Τα λάθη από την αρχική κατηγοριοποίηση της πρώτης εγγραφής επανατροφοδοτούνται στο δίκτυο, και χρησιμοποιούνται για να τροποποιήσουν τον αλγόριθμο δικτύων τη δεύτερη φορά. Η διαδικασία αυτή συνεχίζεται επαναληπτικά.

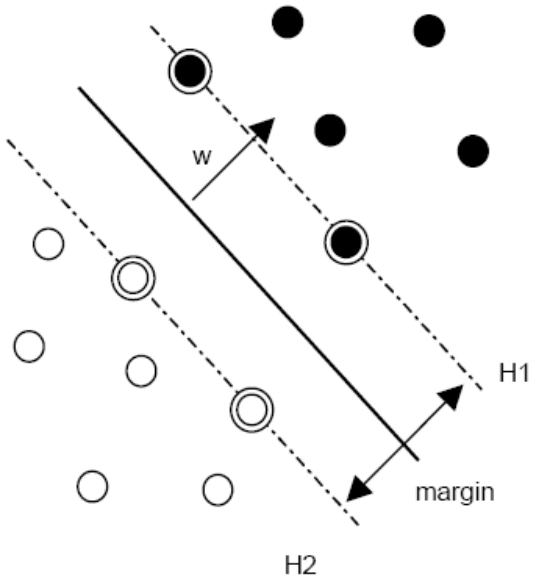
Κοντινότεροι γείτονες (*NearestNeighbors* - *NN*)

Η τεχνική των *κοντινότερων γειτόνων* είναι μια απλή προσέγγιση του προβλήματος της κατηγοριοποίησης. Σύμφωνα με αυτή, ένα νέο στοιχείο κατηγοριοποιείται χρησιμοποιώντας την πλειοψηφία μεταξύ των κατηγοριών από τα K παραδείγματα που είναι τα πιο κοντινά σ' αυτό που δίνεται να κατηγοριοποιηθεί. Μια τέτοια μέθοδος παράγει συνεχείς και επικαλυπτόμενες, παρά σταθερές γειτονιές. Επιπρόσθετα, έχει αποδειχθεί ότι ένας NN κανόνας έχει ασυμπτωτικό ποσοστό σφάλματος που είναι δύο φορές το ποσοστό σφάλματος Bayes, ανεξάρτητα από το μέτρο απόστασης που χρησιμοποιείται.

Η τεχνική NN έχει μειονεκτήματα σε χώρους υψηλών διαστάσεων. Η αυστηρή πόλωση μπορεί να εισαχθεί στην τεχνική NN όταν υπάρχει ένας πεπερασμένος αριθμός από τα παραδείγματα σε χώρο υψηλών διαστάσεων.

Support Vector Machines

Τα *SVMs* είναι μια καινούρια μέθοδος κατηγοριοποίησης η οποία προτάθηκε από τον Vapnik [167], και έχει ήδη αποκτήσει μεγάλη δημοσιότητα. Στην πιο απλή του μορφή, ένα SVM ορίζεται σαν έναν υπερεπίπεδο που δύναται να διαχωρίσει ένα σύνολο θετικών από ένα σύνολο αρνητικών στοιχεία που αφορούν μια συγκεκριμένη κατηγορία. Αυτό φαίνεται και στο παρακάτω σχήμα όπου υποθέτοντας ότι οι μαύρες κουκίδες αφορούν τα θετικά στοιχεία και οι άσπρες τα αρνητικά στοιχεία, ορίζεται με τη βοήθεια του SVM ένα μέγιστο υπερεπίπεδο που αποτελεί το διαχωριστικό ανάμεσα στα στοιχεία.



Σχήμα 2.6: Γραμμικά χωρισμένα υπερεπίπεδα.

Στη γραμμική μορφή του αλγορίθμου, το περιθώριο μεταξύ των στοιχείων μπορεί να οριστεί σαν η απόσταση του υπερεπιπέδου από τα κοντινότερα θετικά και αρνητικά στοιχεία. Η μεγιστοποίηση αυτού του περιθωρίου μπορεί να αποτελέσει ένα πρόβλημα βελτιστοποίησης. Φυσικά τα περισσότερα παραδείγματα δε μπορούν να διαχωριστούν με τη χρήση της γραμμικής μορφής του αλγορίθμου γι' αυτό χρησιμοποιούνται πίνακες προκειμένου να υπολογιστούν τα περιθώρια και οι αποστάσεις. Οι αλγόριθμοι για SVM έχουν αποδειχθεί ότι έχουν καλή γενικά απόδοση ακόμα και σε δύσκολα προβλήματα κατηγοριοποίησης μερικά από τα οποία είναι η αναγνώριση γραφικού χαρακτήρα, η αναγνώριση προσώπου, η κατηγοριοποίηση κειμένων. Η απλή γραμμική μορφή έχει πολύ καλή απόδοση, υφίσταται γρήγορη εκμάθηση και παράλληλα μπορεί να κατηγοριοποιεί εξαιρετικά γρήγορα. Περισσότερα στοιχεία για το SVM μπορούν να βρεθούν στο [58].

Ασαφής κατηγοριοποίηση (*Fuzzy Classification*)

Οι προηγούμενες μέθοδοι κατηγοριοποίησης παράγουν μια αυστηρή κατηγοριοποίηση με την έννοια ότι ένα αντικείμενο είτε ανήκει σε μια κατηγορία είτε όχι. Αυτό σημαίνει ότι όλα τα αντικείμενα θεωρούνται ότι ανήκουν σε μια κατηγορία με τον ίδιο βαθμό πίστης. Επιπλέον, οι κατηγορίες θεωρούνται ως μη επικαλυπτόμενες. Είναι προφανές ότι δεν λαμβάνεται υπόψη η έννοια της αβεβαιότητας σε αυτές τις μεθόδους.

Μια εναλλακτική προσέγγιση στο πρόβλημα αφορά την ασαφή κατηγοριοποίηση η οποία βασίζεται στην ασαφή λογική. Η βασική ιδέα είναι η εξαγωγή των ασαφών κανόνων προκειμένου να αναγνωριστεί κάθε κατηγορία δεδομένων. Οι μέθοδοι εξαγωγής κανόνων βασίζονται στον υπολογισμό των ομάδων (κατηγοριών) στα δεδομένα και κάθε κατηγορία που ορίζεται αντιστοιχεί σε έναν ασαφή κανόνα που συσχετίζει μια περιοχή στο χώρο εισόδου με μια κατηγορία εξόδου. Κατά συνέπεια, για κάθε κατηγορία C_i καθορίζεται το κέντρο των ομάδων έτσι ώστε να προκύψει ένας κανόνας της μορφής:
if {input is near x_i} then class is C_i
Κατόπιν για ένα δεδομένο διάνυσμα εισόδου x, το σύστημα ορίζει το βαθμό ικανοποίησης κάθε κανόνα και τα συμπεράσματα του κανόνα με τον υψηλότερο βαθμό ικανοποίησης επιλέγεται ως η έξοδος του ασαφούς συστήματος. Έτσι η προσέγγιση χρησιμοποιεί ασαφή λογική για να καθορίσει την καλύτερη κατηγορία μέσα στην οποία ένα δεδομένο μπορεί να κατηγοριοποιηθεί, αλλά το τελικό αποτέλεσμα είναι η κατηγοριοποίηση κάθε στοιχείου σε μια από τις κατηγορίες.

Παραγωγή κανόνων κατηγοριοποίησης

Η γνώση που παράγεται κατά τη διάρκεια της διαδικασίας της κατηγοριοποίησης μπορεί να εξαχθεί και να αναπαρασταθεί υπό τη μορφή κανόνων. Μια κοινή προσέγγιση της κατηγοριοποίησης είναι τα δέντρα απόφασης. Σε αυτή την περίπτωση τα πρότυπα γνώσης που εξάγονται περιγράφονται υπό τη μορφή ενός δέντρου. Ωστόσο οι κανόνες είναι ευκολότερα αντιληπτοί από τους ανθρώπους, ιδιαίτερα εάν το δέντρο είναι πολύ μεγάλο.

2.6 Αξιοποίηση πληροφορίας

Είναι γεγονός ότι η πληροφορία που ανακτάται τόσο από έναν μηχανισμό εξόρυξης όσο και ένα μηχανισμό κατηγοριοποίησης είναι συνήθως υπέρογκη. Για το λόγο αυτό θα πρέπει να υπάρχει ένας ισχυρός μηχανισμός που να είναι σε θέση να αξιοποιήσει τη συγκεκριμένη πληροφορία και να μπορεί να βελτιώσει τους τρόπους που γίνονται ερωτήματα στη βάση και προσθήκες νέων εγγραφών. Οι αλγόριθμοι που θα πρέπει να χρησιμοποιούνται θα πρέπει να είναι αρκετά έξυπνοι ώστε να λαμβάνουν υπ' όψιν τους την συνεχή ανανεώση της πληροφορίας. Παρόμοια θα πρέπει ανελλιπώς

να διαγράφονται ή να τροποποιούνται τα στοιχεία τα οποία δε συγκεντρώνουν το ενδιαφέρον των χρηστών του συστήματος και γενικότερα επιβαρύνουν το σύστημα.

Προκειμένου να αξιοποιηθεί η πληροφορία θα πρέπει να δημιουργηθούν περιβάλλοντα διαχείρισης και μηχανισμοί ανάλυσης των ερωτημάτων και εύρεσης απάντησης. Παράλληλα θα πρέπει να υπάρχει τρόπος με τον οποίο να είναι εφικτή η ανάλυση πληροφορίας από τις κινήσεις του χρήστη. Είναι ουσιαστικά μία προσπάθεια να προσεγγίσουμε περισσότερα πραγματικά δεδομένα ξεφεύγοντας από την πληροφορία εκπαίδευσης. Επίσης, με αυτό τον τρόπο θα κάνουμε το μηχανισμό μας πιο διάφανο προς το χρήστη καθώς και πιο φιλικό με την έννοια ότι το σύστημα θα μπορεί να αναγνωρίζει και να προσαρμόζεται στην ανάγκες του χρήστη χωρίς να χρειάζεται ο ίδιος να δηλώσει ρητά τις προτιμήσεις του. Τα εργαλεία διαχείρισης δε θα πρέπει να περιέχουν πολύπλοκες συναρτήσεις, μα ούτε και πολύπλοκο περιβάλλον. Ο όγκος της πληροφορίας κάνει απαγορευτική την άμεση προσέγγισή της, συνεπώς ο διαχειριστής του συστήματος θα πρέπει να είναι σε θέση να έχει μια γενική εποπτεία του συστήματος διατηρώντας παράλληλα ανεκτά τα επίπεδα πρόσβασης σε εξειδικευμένα στοιχεία του συστήματος.

2.7 Προσωποποίηση στο χρήστη

Η προσωποποίηση στο χρήστη είναι διαδικασία κατά την οποία τα αποτελέσματα που εμφανίζονται τελικά στο χρήστη προσαρμόζονται προκειμένου να ανταποκρίνονται 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. Παράλληλα, η προσωποποίηση περνάει σε ένα ακόμη επίπεδο αφού λαμβάνει υπ' όψιν και τις δυνατότητες απεικόνισης της τελικής συσκευής του χρήστη ώστε να στέλνεται κάθε φορά το κατάλληλο μέγεθος περιλήψεων για την όσο το δυνατόν πιο εύληπτη απεικόνιση στη συσκευή. Το πρόβλημα που τίθεται είναι η εύρεση ενός 'έξυπνου' αλγορίθμου ο οποίος θα μπορεί να αξιοποιεί όλες τις πληροφορίες που μπορούν προέρχονται από τον χρήστη προκειμένου να του επιστραφούν όσο το δυνατόν καλύτερα και ποιοτικότερα αποτελέσματα.

2.7.1 Συμμετοχή του χρήστη στις διαδικασίες του συστήματος

Ο χρήστης είναι αυτός που δέχεται την τελική πληροφορία και αυτός που ουσιαστικά διαμορφώνει την πληροφορία για τον εαυτό του. Αυτό σημαίνει πως ο χρήστης θα πρέπει να είναι αναπόσπαστο κομμάτι του συστήματος. Θα πρέπει να είναι σε θέση να διαμορφώσει διαδικασίες του πυρήνα του συστήματος όπως είναι η κατηγοριοποίηση και η εξαγωγή περίληψης. Στα περισσότερα συστήματα τα οποία αντιμετωπίστηκαν κατά τη διάρκεια της μελέτης για τη συγκεκριμένη εργασία, παρατηρήθηκε πως ο χρήστης συμμετέχει μόνο στα επιτελικά στάδια των συστημάτων ενώ έχουν ήδη εκτελεστεί τα βασικά βήματα του πυρήνα των μηχανισμών. Η συμμετοχή του χρήστη στις διαδικασίες πυρήνα ενός large scale συστήματος είναι επίπονη διαδικασία η οποία απαιτεί αλγορίθμους που θα μπορούν να εκτελούνται αποδοτικά σε πραγματικό χρόνο προκειμένου ο χρήστης να διαμορφώνει όχι μόνον τα τελικά αποτελέσματα που εμφανίζονται σε αυτόν αλλά και συγκεκριμένες διαδικασίες ολόκληρου του συστήματος.

2.7.2 Προσωποποίηση περιεχομένου

Προκειμένου ένα σύστημα αυτόματης εξαγωγής περιεχομένου να γνωρίζει τις προτιμήσεις του χρήστη, είναι απαραίτητη η δημιουργία κάποιου είδους μοντέλου γι' αυτόν. Μέσω αυτού, ο χρήστης μοντελοποιείται από το σύστημα καθορίζοντας έτσι τις μεταβαλλόμενες απαιτήσεις που έχει απ' αυτό. Για παράδειγμα, κάποιος χρήστης επιθυμεί πληροφορία που έχει να κάνει με πολιτική (γενικά)

ή με κάποια άλλη κατηγορία του συστήματος. Κάποιος άλλος χρήστης επιθυμεί μεν πολιτική, αλλά φαίνεται να ενδιαφέρεται περισσότερο για κάποιο συγκεκριμένο πολιτικό ρεύμα. Το πως το σύστημα θα κατορθώσει να μοντελοποιήσει τον χρήστη είναι ένα σημαντικό ζήτημα που πρέπει να αντιμετωπισθεί με ιδιαίτερη προσοχή από ένα σύστημα που φιλοδοξεί να ικανοποιεί κάθε χρήστη του προβλέποντας παράλληλα τις μεταβαλλόμενες ανάγκες του.

2.7.3 Προφίλ χρήστη για δυναμικά περιβάλλοντα

Ένα πολύ σημαντικό στοιχείο της εργασίας είναι το προφίλ χρήστη σε δυναμικό περιβάλλον. Είναι το στοιχείο που πρέπει να χαρακτηρίζει κάθε εφαρμογή ποιοτικού περιεχομένου του διαδικτύου. Το δυναμικό περιβάλλον της πρέπει να 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 (π. χ. 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.

2.7.4 Προσωποποίηση εμφάνισης περιεχομένου

Ένα βήμα παραπέρα όσον αφορά στην προσωποποίηση στον χρήστη έχει να κάνει ο τρόπος που η πληροφορία θα παρουσιάζεται σε αυτόν. Είναι προφανές ότι ο χρήστης δεν επιθυμεί να κατακλύζεται από κάθε άρθρο ή πληροφορία που ταιριάζει στο προφίλ του. Είναι επίσης εύλογο να μπορεί να έχει με μερικές μόνο ενέργειες στη διάθεσή του αυτή την πληροφορία. Σκοπός επομένως είναι να βρεθεί μια χρυσή τομή μεταξύ στον όγκο της πληροφορίας που παρουσιάζεται στον χρήστη και στην χρηστικότητα αλλά και την αναγνωσιμότητα που αυτή έχει. Είναι επίσης σημαντικό να τονίσουμε ότι το παραπάνω γεγονός θα πρέπει να γίνεται με τρόπο εύκολα προσαρμοζόμενο από τον χρήστη. Ο χρήστης θα πρέπει να επιλέγει εύκολα τα κανάλια επικοινωνίας που επιθυμεί να εμφανίζονται στην οθόνη του ουτοσώστε να ρυθμίζει την ποσότητα της πληροφορίας.

ΚΕΦΑΛΑΙΟ 3

Σχετικές εργασίες

> I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.
>
> *Alan Turing, English Mathematician, 1954*

Στο παρών κεφάλαιο παρουσιάζεται η ερευνητική δραστηριότητα στο χώρο για τα θέματα με τα οποία καταπιάνεται η εργασία. Ακολουθεί ουσιαστικά μια συνοπτική παρουσίαση των σχετικών εργασιών για θέματα συλλογής, φιλτραρίσματος, προεπεξεργασίας δεδομένων, κατηγοριοποίησης και αυτόματης εξαγωγής περίληψης κειμένου. Τέλος γίνεται μια παρουσίαση των τρεχόντων ερευνητικών θεμάτων που έχουν να κάνουν με προσωποποίηση περιεχομένου στο χρήστη.

3.1 Συλλογή δεδομένων

Ο Web crawler, γνωστός και ως Web spider, Web robot είναι ένα αυτοματοποιημένο πρόγραμμα που διαπερνάει τον παγκόσμιο ιστό με κάποια συγκεκριμένη τακτική. Η διαδικασία που επιτελεί ένας Web crawler λέγεται Web crawling και είναι μία διαδικασία που χρησιμοποιείται κατά κόρων από τις υπηρεσίες δεικτοδότησης ώστε να κατεβάσουν τις σελίδες του διαδικτύου. Γενικά, ένας crawler ξεκινά από μία λίστα URLs που πρόκειται να επισκεφτεί. Συνεχίζει αναδρομικά βρίσκοντας τα links στις σελίδες που επισκέπτεται και τερματίζει όταν κάποιες παράμετροι, π. χ. χρόνος εκτέλεση, ποσότητα δεδομένων, κλπ καλυφθούν.

Το πλήθος των Web crawlers είναι αρκετά μεγάλο και αν εξαιρέσουμε τους εξειδικευμένους crawlers (focused crawlers) παρατηρούμε πως οι περισσότεροι έχουν σαν σκοπό να συλλέξουν όλες τις HTML σελίδες από τις οποίες απαρτίζεται ένας δικτυακός τόπος μαζί με τα βοηθητικά αρχεία (pdf, εικόνες, video, css, javascript) και ουσιαστικά να δημιουργήσουν ένα offline-instance του δικτυακού τόπου τον οποίο προσπελαύνουν.

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

3.1.1 Γνωστοί *Crawlers*

WebCrawler

Πρόκειται για έναν από τους πρώτους crawlers που κατασκευάστηκαν από τον Pinkerton το 1994 [146]. 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. Ο WebCrawler από το 2001 είναι τμήμα της Infospace η οποία τον χρησιμοποιεί ως την βάση για την ομώνυμη μετα-μηχανή αναζήτησης[45].

Google Crawler

Ο Google Crawler ή αλλιώς Googlebot είναι ένας από τους πιο σημαντικούς crawlers που κατασκευάστηκαν και διατηρούνται ακόμα και σήμερα, με σημαντικές βέβαια βελτιώσεις. Ο *Google Crawler* των Brin και Page [72], 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. Συγκεκριμένες πληροφορίες υλοποίησης για τον crawler που χρησιμοποιεί η μηχανή αναζήτησης google δεν είναι διαθέσιμες στο ευρύ κοινό μιας και ο αλγόριθμος του crawling που χρησιμοποιείται μεταβάλλεται διαρκώς προκειμένου α) να ανταποκρίνεται στις μεταβαλλόμενες απαιτήσεις του Web και β) να αντιμετωπίζει τις επιθέσεις των spammers

Mercator

Ο *Mercator* [104] είναι ένας κατανεμημένος και 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++. Περιλαμβάνει έναν κεντρικό μηχανισμό και μία σειρά από "ant" (μερμήγκι) μηχανισμούς [87]. 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 οι σελίδες που διαθέτει ανταποκρίνονται στις πραγματικές σελίδες που βρίσκονται on-line στους δικτυακούς τόπος και όχι απλά σε 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 [178]. 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* [63] είναι ένας κατανεμημένος crawler γραμμένος σε Java και δε διαθέτει κεντρικοποιημένη διαδικασία. 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 επισκεφθεί 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 έχει 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 [15]
- Heritrix [20]
- ht://Dig [8]
- HTTrack [21]
- Larbin [25]
- Methabot [28]
- Nutch [33]
- WebSPHINX [46]
- WIRE - Web Information Retrieval Environment [44]

3.1.2 Εστιασμένο *crawling*

Ένας γενικού σκοπού Web crawler συγκεντρώνει όσο περισσότερες σελίδες μπορεί από ένα δεδομένο σύνολο από URL's. Αντίθετα, ένας εστιασμένος ή αλλιώς focused crawler είναι σχεδιασμένος για να συγκεντρώνει μόνο έγγραφα ενός συγκεκριμένου θέματος ή ενδιαφέροντος και επομένως ελαττώνει την κίνηση του δικτύου και κατά συνέπεια και τον απαιτούμενο χρόνο περάτωσης. Ο σκοπός του focused crawler είναι να αναζητεί επιλεκτικά σελίδες που είναι σχετικές με ένα προκαθορισμένο σύνολο από θεματικές έννοιες. Οι έννοιες καθορίζονται όχι με βάσεις κάποιες λέξεις-κλειδιά, αλλά χρησιμοποιώντας έγγραφα-παραδείγματα. Αντί λοιπόν να συλλέγονται και να δεικτοδοτούνται όλα τα διαθέσιμα έγγραφα του ιστού ώστε να υπάρχει η δυνατότητα να μπορούμε να απαντήσουμε σε κάθε πιθανή τυχαία ερώτηση, ένας focused crawler αναλύει τα όρια του crawling εντοπίζοντας δεσμούς που είναι πολύ πιθανό να συσχετίζονται με το crawling που γίνεται, αποφεύγοντας παράλληλα άσχετες θεματικές περιοχές του ιστού. Αυτό έχει ως συνέπεια την σημαντική ελλάτωση των απαιτούμενων πόρων σε υλικό και εύρος ζώνης και βοηθάει την διαδικασία του crawling ώστε να έχει πιο επικαιροποιημένο περιεχόμενο.

Ο focused crawler περιέχει τρία βασικά συστατικά:

1. έναν ταξινομητή (clasifier), ο οποίος λαμβάνει αποφάσεις συσχέτισης όσων αφορά στις σελίδες που διαπερνούνται ουτοσώστε να αποφασιστεί αν θα γίνει το λεγόμενο link expansion (επέκταση και ακολούθηση των δεσμών της σελίδας)
2. έναν distiller, ο οποίος καθορίζει το μέτρο του κατά πόσον οι σελίδες που διαπερνιούνται παραμένουν εντός θεματικών εννοιών
3. έναν crawler με δυναμικά αναπροσαρμοζόμενη προτεραιότητα που ελέγχεται από τον ταξινομητή και τον distiller

Η πιο σημαντική εκτίμηση της ικανότητας του focused crawler δίνεται από το harvest ratio που αυτός έχει. Η μετρική αυτή μας δίνει το ρυθμό κατά τον οποίο οι σχετικές σελίδες συλλέγονται και οι μη σχετικές απορρίπτονται από τη διαδικασία του crawling. Το harvest ratio θα πρέπει να είναι μεγάλο αλλιώς ο focused crawler περνάει πολύ χρόνο απλά απορρίπτοντας μη σχετικές σελίδες και πιθανά να είναι καλύτερα σε αυτή την περίπτωση να χρησιμοποιηθεί ένας κλασικός crawler.

3.1.3 Αλγόριθμοι για εστιασμένο *crawling*

Οι Focused crawlers βασίζονται σε δύο ειδών αλγορίθμους για την διατήρηση του περιεχομένου τους εντός των προκαθορισμένων εννοιών [165]. Οι αλγόριθμοι που αναλύουν τον ιστό χρησιμοποιούνται για να κρίνουν την συσχέτιση και την ποιότητα των ιστοσελίδων, ενώ οι αλγόριθμοι αναζήτησης καθορίζουν την βέλτιστη σειρά με την οποία τα URLs πρέπει να επισκεφθούν.

Αλγόριθμοι ανάλυσης του ιστού

Γενικά, οι αλγόριθμοι αυτού του τύπου μπορούν να κατηγοριοποιηθούν σε δύο κατηγορίες: σε αυτούς που βασίζονται στην ανάλυση του περιεχομένου και σε αυτούς που βασίζονται στην ανάλυση των δεσμών των σελίδων.

Οι αλγόριθμοι που κάνουν ανάλυση περιεχομένου, εφαρμόζουν τεχνικές δεικτοδότησης για την ανάλυση αλλά και την εξαγωγή λέξεων - κλειδιών ουτοσώστε να καθορίσουν αν το περιεχόμενο μιας σελίδας είναι σχετικό με το πεδίο του crawler. Η κατηγορία αυτή ενσωματώνει την γνώση για το πεδίο στην ανάλυση των σελίδων βελτιώνοντας τα αποτελέσματα. Για παράδειγμα, ελέγχονται οι λέξεις μιας ιστοσελίδες σε σύγκριση με μια προκαθορισμένη λίστα λέξεων του πεδίου. Συχνά ανατίθεται επίσης μεγαλύτερο βάρος σε λέξεις και φράσεις που ανήκουν στον τίτλο ή σε κεφαλίδες της σελίδας (πληροφορία που εντοπίζεται με βάση τα HTML tags). Η URL διεύθυνση επίσης περιέχει σημαντική πληροφορία για την σελίδα που έχει να κάνει με τον προορισμό της ή για το πεδίο γνώσης της. Παράλληλα η τοποθέτηση των δεσμών έχει ιδιαίτερη σημασία όσον αφορά το πεδίο γνώσης μιας ιστοσελίδας [98]. Πιο συγκεκριμένα, ο συγγραφέας της σελίδας A, που τοποθετεί ένα link προς τη σελίδα B, θεωρεί ότι η σελίδα B σχετίζεται με την A.

Ο όρος 'δεσμοί εισόδου' αναφέρεται στα links που 'δείχνουν' προς τη σελίδα. Συνήθως, όσο μεγαλύτερος είναι αυτός ο αριθμός, τόσο πιο υψηλά βαθμολογείται μία σελίδα. Η λογική πίσω απ' αυτό είναι παρόμοια με αυτή της ανάλυσης των αναφορών των συγγραφέων: ένα κείμενο που γίνεται αναφορά συχνά από άλλους συγγραφείς, θεωρείται καλύτερο από κάποιο που δεν έχει καμία αναφορά από άλλους. Η υπόθεση είναι ότι αν δύο σελίδες έχουν ένα link μεταξύ τους είναι πολύ πιθανό να έχουν το ίδιο θεματικό πεδίο. Συγκεκριμένα, στο [83], υπολογίζεται ότι η πιθανότητα οι σελίδες που έχουν link μεταξύ τους να έχουν παρόμοιο περιεχόμενο κειμένου είναι υψηλή αν επιλέγονται τυχαία σελίδες από τον ιστό.

Το anchor text είναι η λέξη ή η φράση που έχει ένας δεσμός ως το κείμενο που εμφανίζεται στον browser. Το κείμενο αυτό μπορεί να δώσει μία καλή πηγή πληροφορίας σχετικά με την σελίδα που δείχνει ο δεσμός, ένα θέμα που έχει θιγεί από πολλές μελέτες, π. χ. [51]. Τέλος είναι σχετικά λογικό να δίνεται ένα επιπλέον βάρος σε κάποιον δεσμό που 'έρχεται' από κάποια διάσημη πηγή (π. χ. yahoo.com). Οι γνωστότεροι αλγόριθμοι που βασίζονται στην ανάλυση των δεσμών είναι ο PageRank [72] και ο HITS [113].

Αλγόριθμοι αναζήτησης του ιστού

Οι αλγόριθμοι αυτού του τύπου χρησιμοποιούνται για να καθοριστεί η βέλτιστη σειρά με την οποία τα URLs πρέπει να επισκεφθούν. Παρότι πολλοί διαφορετικοί αλγόριθμοι αναζήτησης έχουν προταθεί, οι δύο πιο συνηθισμένοι είναι ο κατά πλάτος και ο κατά βάθος. Ο αλγόριθμος αναζήτησης

κατά πλάτος είναι η απλούστερη στρατηγική για το crawling μιας και δεν χρησιμοποιεί ευρετικά για την απόφαση σχετικά με το επόμενο link που πρόκειται να ακολουθηθεί. Όλα τα URLs στο τρέχον επίπεδο θα επισκεφθούν με τη σειρά που ανακαλύπτονται. Παρότι αυτή η στρατηγική δεν διαφοροποιεί ιστοσελίδες διαφορετικής ποιότητας ή διαφορετικού πεδίου, εντούτοις είναι αρκετά καλή για το χτίσιμο συλλογής μιας γενικού σκοπού μηχανής αναζήτησης. Πρόσφατα όμως [138], δείχθηκε ότι παρότι απλή, η στρατηγική αυτή μπορεί να χρησιμοποιηθεί και για συλλογές συγκεκριμένου πεδίου. Η λογική είναι ότι αν τα URLs εκκίνησης είναι σχετικά με το πεδίο που αναζητούμε, είναι πιθανό οι σελίδες του επόμενου επιπέδου να είναι επίσης σχετικές με το πεδίο, κ.ο.κ. Η στρατηγική αναζήτησης κατά πλάτος έχει επίσης χρησιμοποιηθεί και σε συνδυασμό με το εστιασμένο crawling [91] όπου οι σελίδες διαπερνιούνται αρχικά κατά πλάτος και στη συνέχεια οι μη σχετικές σελίδες φιλτράρονται από τη συλλογή με χρήση αλγορίθμου που αναλύει το περιεχόμενο. Σε σχέση με την στρατηγική αναζήτησης κατά πλάτος, η τεχνική αυτή μπορεί να χρίσει πολύ μεγαλύτερες συλλογές συγκεκριμένου πεδίου με πολύ λιγότερο θόρυβο. Παρόλα αυτά, επειδή πολλές μη σχετικές σελίδες ανακτώνται και επεξεργάζονται με τον αλγόριθμο της ανάλυσης του περιεχομένου, η μέθοδος αυτή έχει χαμηλή αποδοτικότητα.

Η στρατηγική αναζήτησης 'το καλύτερο πρώτα' είναι η πιο γνωστή την τρέχουσα περίοδο στον τομέα των εστιασμένων crawlers [91][60][114][130]. Σε αυτή τη στρατηγική τα URLs επισκέπτονται απλά με τη σειρά που εντοπίζονται. Αντίθετα, μερικά ευρετικά (συχνά αποτελέσματα που προκύπτουν από αλγόριθμους ανάλυσης του ιστού) χρησιμοποιούνται για την κατάταξη των URLs στην σειρά με την οποία λαμβάνει χώρα το crawling. Τα αποτελέσματα που μοιάζουν πιο 'ελπιδοφόρα', γίνονται crawl πρώτα και επομένως η τεχνική αυτή πλεονεκτεί σε σχέση με την κατά πλάτος. Παρόλα αυτά, έχει και ορισμένα μειονεκτήματα. Στο [61] δείχνεται ότι οι crawlers με τη στρατηγική αυτή ενδέχεται να χάνουν σημαντικές σελίδες και να έχουν ως αποτέλεσμα χαμηλή ανάκληση όσον αφορά στην τελική συλλογή και αυτό διότι το ευρετικό που χρησιμοποιείται είναι γενικά τοπικό (είναι εφικτό να ελεγχθούν μόνο οι κοντινοί γείτονες μιας σελίδας πριν παρθεί η απόφαση για το πόσο σχετική είναι).

3.1.4 Κατανεμημένο *crawling*

Η δεικτοδότηση του ιστού είναι μία μεγάλη πρόκληση λόγω της αύξησής του και λόγω της δυναμικής φύσης του. Καθώς το μέγεθος του παγκόσμιου ιστού αυξάνει, έχει γίνει επιβεβλημένη η παραλληλοποίηση της διαδικασίας του crawling προκειμένου να ολοκληρώνεται το κατέβασμα των σελίδων μέσα σε λογικά χρονικά πλαίσια. Μία μοναδική διεργασία crawling, ακόμα και αν είναι πολυνυματική, παραμένει μη επαρκής για μεγάλες μηχανές αναζήτησης που χρειάζεται να κατεβάζουν μεγάλες ποσότητες δεδομένων περιοδικά. Επίσης, όταν χρησιμοποιείται μία διεργασία, όλα τα δεδομένα που ανακτώνται περνάνε από το ίδιο μοναδικό physical link. Κατανέμοντας την διαδικασία του crawling σε πολλές διεργασίες και πολλά συστήματα μπορεί να βοηθήσει στην κατασκευή ενός κλιμακώσιμου, εύκολα παραμετροποιήσιμου συστήματος που στην ουσία είναι ανεκτικό στις βλάβες. Παράλληλα, ο διαχωρισμός του φόρτου ελαττώνει τις απαιτήσεις σε υλικό και ταυτόχρονα αυξάνει την συνολική ταχύτητα και αξιοπιστία. Κάθε εργασία λαμβάνει χώρα με πλήρως κατανεμημένο τρόπο και επομένως δεν χρειάζεται κεντρικοποιημένος έλεγχος του crawling.

3.1.5 Ο κρυμμένος ιστός

Οι δομημένες βάσεις δεδομένων του Web έχουν εξελιχθεί στον βασικό αποθηκευτικό χώρο που χρησιμοποιείται στο διαδίκτυο για την αποθήκευση σχεσιακών δεδομένων. Δεδομένου όμως το ότι αυτά τα δεδομένα δεν είναι άμεσα προσπελάσιμα στους παραδοσιακούς Web Crawlers βάσει των διεπαφών που αυτοί έχουν, το κομμάτι αυτό του ιστού αναφέρεται συχνά ως κρυμμένος ιστός

(deep Web ή hidden Web) [29].

Για να μπορέσουμε να έχουμε πρόσβαση τελικά σε αυτόν τον τεράστιο όγκο δεδομένων υπάρχουν δύο βασικές σχεδιαστικές επιλογές. Η πρώτη είναι η data warehouse-like προσέγγιση [18],[30], στην οποία τα δεδομένα συγκεντρώνονται από ένα μεγάλο πλήθος πηγών του Web και αξιοποιούνται (για εξόρυξη πληροφορίας) με κεντρικοποιημένο τρόπο. Η δεύτερη προσέγγιση είναι η MetaQuerier [77] που προσφέρει ένα πεδίο 'αφαίρεσης' των βάσεων δεδομένων του ιστού δίνοντας έτσι ένα ενδιάμεσο σχήμα στους χρήστες. Μαζί με πολλές ακόμη διαφορές μεταξύ αυτών των προσεγγίσεων, έχουν διαφορετικούς στόχους όσον αφορά στην off-line απόκτηση δεδομένων. Για την περίπτωση της data-warehouse, η ενσωμάτωση δεδομένων από δομημένες πηγές του ιστού και η συγκέντρωσή τους σε ένα κεντρικό σημείο είναι ένα σημαντικό αρχικό βήμα εφόσον τα ερωτήματα των χρηστών απαντώνται αποκλειστικά από τα δεδομένα που αποθηκεύονται στο κεντρικοποιημένο warehouse. Σε αντίθεση, η τεχνική του MetaQuerier θέτει λιγότερες απαιτήσεις στην απόκτησε των δεδομένων διότι αρκετά ερωτήματα τυπικά επαρκούν για αντιστοίχηση ερωτημάτων μέσω κάποιου σχήματος [169].

Ενώ δεν είναι ακόμη σαφές αν η προσέγγιση τύπου MetaQuerier ή αυτή που βασίζεται στο warehousing είναι πιο κατάλληλη για την πρόσβαση στον κρυμμένο ιστό, η απόκτηση δεδομένων από δομημένες πηγές του ιστού είναι ένα ενδιαφέρον πεδίο από μόνο του. Πολλές εφαρμογές μπορούν να αναπτυχθούν (και έχουν ήδη αναπτυχθεί) για να αξιοποιούν τα δομημένα δεδομένα από τις κρυμμένες πηγές του ιστού. Για παράδειγμα στα [123] και [162] οι εφαρμογές αξιοποιούν τα δεδομένα για τεχνικές εκπαίδευσης συγκεκριμένων πεδίων γνώσης. Επίσης εφαρμογές που εφαρμόζουν 'σύγκριση αγορών' ενσωματώνοντας δεδομένα από διαφορετικούς (πιθανά ανταγωνιστικούς) προμηθευτές. Τα δεδομένα του κρυμμένου ιστού μπορούν επίσης να βοηθήσουν στην δόμηση άλλων λιγότερο δομημένων εγγράφων. Για την ακρίβεια πολλές μηχανές αναζήτησης έχουν αρχίσει να παρέχουν υπηρεσίες αναζήτησης προϊόντων βασισμένες σε δομημένα δεδομένα που μαζεύουν από πλήθος πηγών του ιστού [18],[30].

Στην βιβλιογραφία, υπάρχουν δύο τρόποι για να ανακτηθούν τα δεδομένα από την διάφορες πηγές του ιστού. Η πιο αποτελεσματική μέθοδος είναι αν αφήσουμε τις ίδιες τις πηγές να εξαγάγουν τις βάσεις τους με βάσει κάποιου είδους αδειοδότηση· έτσι τα δεδομένα τους μπορούν να δεικτοδοτηθούν άμεσα. Δυστυχώς, στο αυτόνομο, μη-συνεργατικό και ανταγωνιστικό περιβάλλον του παγκόσμιου ιστού, αυτή η προσέγγιση δύσκολα κλιμακώνεται λόγω του πλήθους των ιστότοπων και λόγω του ότι απαιτεί ένα σημαντικό χρονικό διάστημα χώρια τον ανθρώπινο παράγοντα. Η εναλλακτική προσέγγιση βασίζεται σε έναν κρυφό Web Crawler που θέτει ερωτήματα διαρκώς στην κρυμμένη βάση δεδομένων 'ξετυλίγοντας' το περιεχόμενό της. Τα ερωτήματα μπορούν να τίθενται είτε μέσω των φορμών που έχουν ήδη αυτοί οι ιστότοποι ή μέσω των δημοσιευμένων Web Services Interfaces που έχουν πολλοί ιστότοποι (π. χ. Amazon Web Service).

Σε αντίθεση με τις παραδοσιακές τεχνικές Crawling, το Crawling που βασίζεται σε ερωτήματα χαρακτηρίζεται από βρόχο που κάνει την αλλαγή των ερωτημάτων (Query-harvest Decomposite loop) και που επαναληπτικά αποκαλύπτει την πληροφορία. Ένας τέτοιος Crawler αρχίζει με μερικά αρχικά ερωτήματα με μορφή attribute-value, π. χ. Actors, Hanks, Tom, Brand, IBM ανακτώντας και αποθηκεύοντας την πληροφορία που επιστρέφεται είτε σε HTML είτε σε XML μορφή τοπικά. Τα δεδομένα που επιστρέφονται με κάθε ερώτημα θεωρούνται πιθανά για να σχηματίσουν ένα μελλοντικό ερώτημα. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να γίνουν όλα τα πιθανά ερωτήματα ή μέχρι να λάβει χώρα κάποιο άλλο κριτήριο τερματισμού.

Σημαντική προσπάθεια έχει γίνει για να αυτοματοποιηθεί η προηγούμενη διαδικασία. Προς αυτή την κατεύθυνση, οι τεχνικές προκλήσεις έγκειται στην αυτόματη συμπλήρωση φορμών [148] και στην εξαγωγή δομημένων δεδομένων [53]. Όμως ένα σημαντικό και γενικά δύσκολο πρόβλημα είναι το εξής: πως μπορούμε να επιλέξουμε καλά ερωτήματα ώστε να έχουμε ικανοποιητική κάλυψη της γνώσης που περιέχεται στην βάση δεδομένων μέσα πάντα σε ανεκτό κόστος επικοινωνίας με

τον server;

Διαισθητικά, ενώ η τελική κάλυψη της γνώσης που είναι εφικτή είναι ουσιαστικά προκαθορισμένη από τα αρχικά ερωτήματα, τα κόστη επικοινωνίας είναι ευθέως εξαρτώμενα από την μέθοδο επιλογής των ερωτημάτων που τελικά χρησιμοποιείται. Στην πράξη, είναι συχνά αδύνατο για έναν Crawler να εξαντλήσει κάθε πιθανό ερώτημα στη βάση δεδομένων. Επομένως, μία μέθοδος αποτελεσματικής επιλογής ερωτημάτων είναι απαραίτητη για να πετύχουμε 'καλή' κάλυψη με λογικό κόστος επικοινωνίας. Παρά την εύκολη διατύπωση, το παραπάνω πρόβλημα δεν είναι απλό. Για παράδειγμα στο [173], οι συγγραφείς δείχνουν ότι μία επαρκής λύση για το πρόβλημα είναι τεχνικά non-trivial. Επίσης για να επιτευχθεί ισοδύναμη κάλυψη της βάσης δεδομένων, μία καλή μέθοδος επιλογής των queries μπορεί να έχει σημαντικά λιγότερο overhead σε σχέση με την naive μέθοδο. Στην ίδια εργασία επίσης αποδεικνύεται ότι το σημαντικότερο ζήτημα για το deep Web crawling είναι στην κατάλληλη επιλογή των queries και το πρόβλημα μοντελοποιείται ως ένα πρόβλημα διάτρεξης γράφου. Υπό αυτή την έννοια, ο στόχος είναι η εύρεση ενός Weighted Minimum Dominating Set στον αντίστοιχο γράφο που αντιστοιχίζει πεδία με τιμές.

3.2 Εξαγωγή χρήσιμου κειμένου

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 δέντρο 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.

3.3 Προεπεξεργασία δεδομένων

Στη θεωρία, τα βασισμένα σε κείμενο χαρακτηριστικά ενός εγγράφου μπορούν να περιλαμβάνουν κάθε λέξη / φράση η οποία μπορεί να εμφανίζεται σε ένα δεδομένο σύνολο κειμένων. Όμως, επειδή κάτι τέτοιο είναι υπολογιστικά μη-ρεαλιστικό, χρειαζόμαστε κάποια μέθοδο προεπεξεργασίας κειμένων για την αναγνώριση των λέξεων - κλειδιών (κωδικολέξεων ή αλλιώς keywords) και φράσεων οι οποίες μπορεί να μας είναι χρήσιμες. Διάφορες τεχνικές έχουν προταθεί για την αναγνώριση των keywords ενός κειμένου όπως τα Hidden Markov Models [79], η Naive Bayes [140] και τα Support Vector Machines [112]· όμως όλες αυτές οι μέθοδοι τείνουν να κάνουν χρήση συγκεκριμένης γνώσης μετα-πληροφορίας για τη γλώσσα του κειμένου. Άλλες μέθοδοι χρησιμοποιούν στατιστικές πληροφορίες, όπως η συχνότητα μιας λέξης. Μια ευρέως γνωστή τεχνική είναι η TF-IDF (Term Frequency - Inverse Document Frequency), όπου TF είναι το πλήθος των εμφανίσεων ενός όρου σε ένα δεδομένο σύνολο κείμενων συγκρινόμενο με το πλήθος των κειμένων που περιέχουν το συγκεκριμένο όρο, και IDF είναι ένα μέτρο των συνολικών κειμένων σε μια συλλογή κειμένων, συγκρινόμενο με το συνολικό αριθμό κειμένων που περιέχουν μια δεδομένη λέξη [110]. Σχετικές τεχνικές, οι οποίες περιλαμβάνουν άλλες στατιστικές που πηγάζουν από το σύνολο των κειμένων, έχουν επίσης προταθεί τα πρόσφατα χρόνια· π. χ. κέρδος πληροφορίας [176], odds ratio [132], CORI [94], κλπ. Οι τεχνικές αυτές προσφέρουν μια βελτιωμένη προσέγγιση.

3.3.1 Αναγνώριση μερών του λόγου

Πολλές μεθοδολογίες έχουν προταθεί για την αυτόματη διαδικασία αναγνώρισης των μερών του λόγου που αντιστοιχούν στις λέξεις που απαρτίζουν τις προτάσεις ενός κειμένου. Ο σκοπός είναι αφενός μεν για συντακτικούς λόγους, αφετέρου δε για καλύτερη εκτίμηση των σημαντικότερων λέξεων του κειμένου. Πιο συγκεκριμένα και για την περίπτωση των συστημάτων που μας αφορούν, η διαδικασία αναγνώρισης μερών του λόγου μπορεί να μας δώσει τα ουσιαστικά του κειμένου προς επεξεργασία, τις λέξεις επομένως, που κατά γενική ομολογία, εκφράζουν το μεγαλύτερο ποσοστό της νοηματική πληροφορίας που περιέχει το κείμενο.

Η εύρεση των ουσιαστικών ενός κειμένου, μία υποκατηγορία της αναγνώρισης μερών του λόγου, έχει επίσης διάφορες προτεινόμενες μεθοδολογίες ανάμεσα στις οποίες βρίσκονται γλωσσολογικές [111], συμβολικές [82], ή βασιζόμενες σε support vector machines [99] και μπορούν να κατηγοριοποιηθούν σε αυτές που κάνουν μορφολογική ανάλυση του κειμένου, και σε αυτές που βασίζονται στην αναγνώριση των μερών του λόγου. Οι πρώτες προσπαθούν να παράγουν όλες τις πιθανές ερμηνείες μιας φράσης υλοποιώντας έναν μορφολογικό αναλυτή ή πιο απλά χρησιμοποιώντας λεξικολογικά λεξικά. Είναι πιθανό να υπερ-παράγουν ή να παράγουν μη έγκυρα ουσιαστικά λόγω της αμφισημίας των λέξεων και επίσης χαρακτηρίζονται από μικρό επίπεδο ακρίβειας. Από την άλλη μεριά, οι μέθοδοι που βασίζονται στην αναγνώριση μερών του λόγου επιλέγουν την πιο πιθανή ανάλυση μεταξύ των αποτελεσμάτων που παράγονται από τον μορφολογικό αναλυτή. Η μέθοδος αυτή, λόγω της επίλυσης των αμφισημιών μπορεί να δώσει καλύτερα αποτελέσματα. Παρόλα αυτά, και αυτή έχει ορισμένα προβλήματα, όπως τα λάθη που δημιουργούνται από το module που αναγνωρίζει τα μέρη του λόγου ή από τον μορφολογικό αναλυτή που προηγείται.

3.3.2 *Stemming* ή *rooting*;

Η ρίζα μίας λέξη είναι η πρωτεύουσα λεξικολογική μονάδα αυτής, η οποία περιλαμβάνει το πιο σημαντικό σημασιολογικό περιεχόμενο της και παράλληλα δεν μπορεί να μειωθεί ακόμα περισσότερο. Μερικές φορές όμως ο όρος 'ρίζα' χρησιμοποιείται για να περιγράψει την λέξη χωρίς το επίθεμα αλλά με την λεξικολογική κατάληξή της. Για παράδειγμα η λέξη chatters έχει ως επιθεματική ρίζα

τη λέξη chatter αλλά ως λεξικολογική ρίζα την λέξη chat. Οι επιθεματικές ρίζες καλούνται και stems και συνήθως η 'ρίζα' μιας λέξης αντιστοιχεί στην μορφολογική της ρίζα.

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

Το πρόβλημα του stemming έχει προσεγγιστεί από μια μεγάλη ποικιλία μεθόδων που περιγράφονται στο [120] και περιλαμβάνουν αφαίρεση της κατάληξης, τμηματοποίηση λέξης και λεξιλογική μορφοποίηση. Δύο από τους διασημότερους αλγορίθμους, ο Lovins[122] και ο Porter[147], βασίζονται στην αφαίρεση της κατάληξης. Ο αλγόριθμος Lovins βρίσκει το μακρύτερο ταίριασμα από μια μεγάλη λίστα καταλήξεων, ενώ ο Porter [38] χρησιμοποιεί έναν επαναληπτικό αλγόριθμο με μικρότερο αριθμό καταλήξεων και μερικούς κανόνες. Ένας ακόμη αλγόριθμος, ο Paice/Husk [143], χρησιμοποιεί αποκλειστικά ένα σύνολο κανόνων ενώ ακολουθεί επαναληπτική προσέγγιση.

Στο [119] 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 [92] 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.

3.4 Κατηγοριοποίηση πληροφορίας

Η αυτόματη *κατηγοριοποίηση* κειμένων είναι η διαδικασία ανάθεσης ετικετών κατηγορίας (προκαθορισμένων) σε νέα κείμενα που καταφθάνουν, στηριζόμενη στην πιθανότητα η οποία προτείνεται από τη βάση γνώσης που προϋπάρχει. Οι στατιστικοί αλγόριθμοι για το πεδίο έχουν μία ιστορία περίπου 40 ετών. Την τελευταία δεκαετία περίπου, η στατιστική προσέγγιση έχει κυριαρχήσει στη

βιβλιογραφία. Οι στατιστικές προσεγγίσεις για την κατηγοριοποίηση κειμένων, ή αλλιώς οι προσεγγίσεις επιβλεπόμενης εκμάθησης, εμπεριέχουν την έννοια 'εκμάθησης' του κατηγοριοποιητή (του κανόνα δηλαδή που αποκρίνεται για το αν ένα κείμενο πρέπει να ανατεθεί σε μία κατηγορία ή όχι) από ένα σύνολο κειμένων που έχουν ήδη κατηγοριοποιηθεί.

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 [86][175], κατηγοριοποίηση κοντινότερων γειτόνων [127][174], πιθανοτικές προσεγγίσεις με μεθόδους Bayes [166][121], επαγωγική εκμάθηση κανόνων [52][78], νευρωνικά δίκτυα [139], on-line εκμάθηση [78] και Support Vector Machines [108]. Παρά την 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 [150]. Άλλες εναλλακτικές επιλογές, όπως συντακτικές και στατιστικές εκφράσεις έχουν επίσης χρησιμοποιηθεί [74][96][151]. Το βασικό θέμα της αναγνώρισης του θέματος με χρήση NLP έχει αναλυθεί διεξοδικά στο [107].

Άλλες επαναστατικές τεχνικές, όπως η χρήση κωδικών ελέγχου [59], η χρήση αιτιολογικών δικτύων έχουν προταθεί και αποτελούν ουσιαστικά μια τροποποιημένη έκδοση του Bayes αλγορίθμου του [156] που αποδίδουν καλά σε εργασίες κατηγοριοποίησης κειμένων. Καμία από τις προηγούμενες τεχνικές δεν αντιμετωπίζει τα σημασιολογικά θέματα.

3.4.1 Ταξινόμηση κειμένων

Δεδομένου ενός συνόλου πινάκων κειμένων $\{d_1, d_2, ..., d_n\}$ και των συσχετιζόμενων με αυτά ετικετών $c(d_i) \in \{c_1, c_2, ..., c_l\}$, η διαδικασία της ταξινόμησης αφορά στον καθορισμό της σωστής ετικέτας του νέου κειμένου d. Η ταξινόμηση κειμένων (text classification) έχει μελετηθεί σε μεγάλο βαθμό, ιδιαίτερα ύστερα από την εμφάνιση του διαδικτύου. Οι περισσότεροι αλγόριθμοι βασίζονται στο μοντέλο 'συνόλου λέξεων' του κειμένου [155]. Ένας απλός και συνάμα αποτελεσματικός αλγόριθμος είναι αυτός του Naive Bayes [131]. Για το πρόβλημα της ταξινόμησης κειμένων, διάφορες παραλλαγές του Naive Bayes έχουν χρησιμοποιηθεί αλλά έχει βρεθεί [129] ότι η παραλλαγή που βασίζεται στο πολυωνυμικό μοντέλο οδηγεί σε καλύτερα αποτελέσματα.

Η μέθοδος των Support Vector Machines (SVMs) έχει επίσης χρησιμοποιηθεί επίσης με καλά αποτελέσματα [108][73]. Για ιεραρχικά δεδομένα κειμένων, όπως οι ιεραρχίες θεμάτων του Yahoo! [49] και το Open Directory Project [34], έχει μελετηθεί στα [118][75][85].

Για να αποφευχθούν οι πολλές διαστάσεις στην αναπαράσταση των κειμένων, πολλές μέθοδοι επιλογής χαρακτηριστικών έχουν προταθεί [176][118][75]. Επίσης συχνά επιζητείται η ιδιότητα της 'ισχυρής' ταξινόμησης όπου η κάθε λέξη του κειμένου μπορεί να αντιπροσωπευθεί από τη μοναδική ομάδα που ανήκει. Τέτοια ιδιότητα αξιοποιείται στα [129][163]. Η επιλογή του μεγίστου πλήθους των λέξεων που θα απαρτίζουν ένα cluster είναι επίσης κάτι σημαντικό [168][152].

3.5 Αυτόματη εξαγωγή περίληψης

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

Η έρευνα στον τομέα της αυτόματης περίληψης, θεωρούμενη ως εξαγωγή, αφαίρεση ή περίληψη χρήσιμου κειμένου, έχει μεγάλη ιστορία με αρχικό 'ξέσπασμα' τις προσπάθειες στη δεκαετία του 60 της πρωτοποριακής εργασίας του Luhn, ακολουθείται από τις δύο επόμενες δεκαετίες με σχετικά μικρή έρευνα στο θέμα, και κορυφώνεται τη δεκαετία του 90 και ως της μέρες μας με πολλές ερευνητικές προσπάθειες [144],[88],[124]. 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 [116],[56] 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 [154] 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 [158] προσεγγίζουν το πρόβλημα με Poisson και αρνητικές διωνυμικές κατανομές ή με χρήση του k-mixture μοντέλου [157] 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 [177]. 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 [89], 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. Παράδειγμα της προσπάθειας αυτής αποτελεί η Rhetorical Structure Theory[126]. 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.

3.5.1 Συστήματα περίληψης βασισμένα στη γνώση

Από την γέννηση τους, η ανάπτυξη των συστημάτων αντίληψης κειμένων ήταν άρρηκτα συνδεδεμένη με το πεδίο της αναπαράστασης γνώσης και των μεθόδων λογικής [159]. 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 [172]. Αυτές οι αρχές χρησιμοποιήθηκαν σε συστήματα περίληψης όπως τα: SUSY [95], SCISOR [106] και TOPIC [102]. Αλλά ακόμη και αυτού του είδους τα συστήματα αδυνατούσαν να εξάγουν αποτελεσματικά αξιόλογες μεταφράσεις.

3.5.2 Αναγνώριση θεμάτων

Το θέμα της *αναγνώρισης θεμάτων (Topic Identification)*, αναφέρεται στην διαδικασία της έρευνας σε έγγραφα κειμένου, για την ανακάλυψη συγκεκριμένων δομών. Σύμφωνα με τους Mather A. Laura και Note Jarrod [128], 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.

3.5.3 Περίληψη κειμένου βασισμένη στο χρόνο

Παρότι είναι λίγη σχετικά η έρευνα στο συγκεκριμένο τομέα, ορισμένοι ερευνητές έχουν ασχοληθεί με το πως είναι δυνατή η εξαγωγή προσωρινών εκφράσεων από ένα κείμενο, αναζητώντας και κανονικοποιώντας αναφορές σε ημερομηνίες, χρόνο και παρερχόμενο χρόνο [125]. 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 [50]. 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 [164] 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.

3.5.4 Αξιολόγηση της περίληψης κειμένου

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

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

- Χρήση αρκετών πρωτοτύπων παραδειγμάτων από τεχνικές περίληψης κειμένου για τις οποίες γνωρίζουμε την απόδοσή τους
- Συμμετοχή ανθρώπων [71][142] με την ανάγνωση των περιλήψεων και την βαθμολόγησή τους με κριτήριο το πόσο αντιπροσωπευτική θεωρείται σε σχέση με το αρχικό κείμενο
- Θεωρούμε ότι η περίληψη του κειμένου είναι ένα υποσύνολο του κειμένου και ελέγχουμε εάν μπορεί να αντιπροσωπεύσει επαρκώς το αρχικό κείμενο σε θέματα όπως: είναι δυνατό να κατηγοριοποιηθεί το κείμενο με βάση την περίληψή του ή να εντοπιστεί εάν ανταποκρίνεται στις προτιμήσεις του χρήστη χωρίς να εξεταστεί το αρχικό κείμενο [90][141]; Μπορεί ένας χρήστης να εμπεδώσει σωστά το κείμενο έχοντας διαβάσει μόνο την περίληψή του και απαντώντας σε tests [137]; Μπορεί ο χρήστης να αντιστοιχίσει σωστές λέξεις - κλειδιά σε μια περίληψη [153];
- Συγκρίνουμε την ομοιότητα μεταξύ προτάσεων επιλεγμένων από ανθρώπους, ως αντιπροσωπευτικές για το κείμενο, και των προτάσεων που προέκυψαν από την αυτοματοποιημένη περίληψη [101][149], ή συγκρίνουμε το βαθμό αντιπροσωπευτικότητας που δίνουν οι χρήστες σε μια πρόταση σε σχέση με αυτόν που δίνει ο μηχανισμός [84]. Οι τεχνικές αυτού του είδους αναφέρονται συνήθως και ως corpus-based.

3.5.5 Παραδείγματα συστημάτων

Ακολουθούν ορισμένα σημαντικά συστήματα αυτόματης εξαγωγής περίληψης που χρίζουν αναφοράς.

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 γλωσσολογική 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

Η εφαρμογή MS Word στις πιο πρόσφατες εκδόσεις της 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

Ο MEAD περιλήπτης είναι μια ελεύθερα διαθέσιμη σειρά εργαλειών για πολυ-γλωσσική περίληψη και αξιολόγηση. Χρησιμοποιεί πολλούς αλγορίθμους περίληψης π. χ. keyword-based, TF*IDF. Είναι γραμμένος σε γλώσσα Perl και πρόκειται ίσως για τον πιο ολοκληρωμένο μηχανισμό αυτόματης εξαγωγής περίληψης.

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:

$$\underline{\text{Κατηγοριοποίηση} = \text{Εντοπισμός τίτλου} + \text{μετάφραση} + \text{παραγωγή}}$$

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 τρία 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.

3.6 Προσωποποίηση στο χρήστη

Σύμφωνα με τον Mobasher [133], η προσωποποίηση στο διαδίκτυο μπορεί να περιγραφεί 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 αναφερόμαστε 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 [35, 7], προσωποποίηση στη μεριά του χρήστη [109] 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) [103] 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) [117] 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) [115] 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) [81] 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 προτού αυτοί επισκεφθούν το 'κατάστημά' τους. Έτσι, σε πολλούς δικτυακούς τόπους, όπως για παράδειγμα το Amazon.com [1], η προσωποποίηση και οι συστάσεις που παρουσιάζονται,

εφαρμόζονται 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 [35] και το Liberty Alliance [134]. Αυτά παρέχουν μία ενιαία βάση δεδομένων που περιέχει τα προσωπικά στοιχεία και τις επιλογές του. Οι χρήστες προσθέτουν από μόνοι τους στοιχεία στη βάση δεδομένων στα οποία έχουν ελεύθερη πρόσβαση εταιρίες που είναι συμβεβλημένες με τα εκάστοτε SSO συστήματα.

Βασικό πρόβλημα αυτής της προσέγγισης είναι η εξασφάλιση της ασφάλειας του συστήματος καθότι ο χρήστης μπορεί να αποθηκεύει ευαίσθητα δεδομένα. Το συγκεκριμένο θέμα τονίζεται ακόμα και στα προϊόντα των εταιριών (για παράδειγμα η Sun το τονίζει ιδιαίτερα στο πρόγραμμα Liberty. Πως θα εμπιστευτεί ένας χρήστης το πρόγραμμα το οποίο του τονίζει ιδιαίτερα πως δεν είναι ασφαλές· Τα νεότερα SSO συστήματα όπως το Liberty Alliance και το SXIP έχουν δώσει ιδιαίτερη προσοχή στο συγκεκριμένα θέμα προκειμένου να βελτιωθούν. Μάλιστα το SIXP επιτρέπει σε ένα χρήστη να διαθέτει πολλαπλά προφίλ ανάλογα με το μέγεθος των δεδομένων που επιθυμεί να είναι ορατά σε διάφορους δικτυακούς τόπους ορίζοντας με αυτό τον τρόπο αυτόνομα το επίπεδο ασφάλειας. Παράλληλα είναι ένα σύστημα ανοιχτού κώδικα προκειμένου οι χρήστες να μπορούν να δουν επακριβώς τι στοιχεία τους διαμοιράζονται και με ποιον τρόπο. Αυτό βέβαια δεν ξεπερνά τα προβλήματα που παρουσιάζονται. Οι χρήστες πρέπει να αποφασίσουν αν οι εταιρίες στις οποίες θα εμπιστευτούν τα προσωπικά τους δεδομένα είναι έμπιστες ή όχι. Αυτό συνεπάγεται και την αποτυχία τέτοιων συστημάτων με χαρακτηριστικό παράδειγμα το σύστημα Passport σαν τεχνολογία καθότι οι χρήστες δεν έχουν κάποια ιδιαίτερη προτίμηση στα SSO συστήματα. Παράλληλα, όπως αναφέρει και ο Gartner [12], 'όσο οι χρήστες δε δείχνουν να αποδέχονται τέτοια συστήματα οι εταιρίες δεν πρόκειται να κάνουν απολύτως καμία επένδυση'.

Υπάρχουν βέβαια και συστήματα τα οποία δεν απαιτούν την εισαγωγή στοιχείων από το χρήστη αλλά χρησιμοποιούν μεταδεδομένα που υπάρχουν από τα ίχνη που αφήνει ένας χρήστης καθώς πραγματοποιεί περιαγωγή σε σελίδες του διαδικτύου. Το WAWA (Wisconsin Adaptive Web Assistant) [161] 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 [145] 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 [76] 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 [170] 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 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 [100] 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.

3.7 Παραδείγματα συστημάτων αποδελτίωσης

Δεδομένης της έκτασης που έχει πάρει η διακίνηση της πληροφορίας στο διαδίκτυο, ιδιαίτερα στα ειδησεογραφικά sites, έχει ήδη γίνει εμφανής η ανάγκη για την ύπαρξη συστημάτων τα οποία συγκεντρώνουν και συνοψίζουν τα άρθρα προς όφελος των χρηστών. Τα συστήματα αυτά που είναι γνωστά και ως συστήματα αποδελτίωσης - indexing, έχουν ως στόχο την ολοκληρωμένη κάλυψη του χρήστη από άποψη ενημέρωσης σε σχέση με τα ενδιαφέροντά του με την ελάχιστη δυνατή επιπλέον πληροφορία η οποία, ως γνωστών, αποθαρρύνει τους χρήστες. Ορισμένα πειραματικά συστήματα τα οποία υπάρχουν στο διαδίκτυο είναι τα παρακάτω:

1. Newsjunkie [97]. Πρόκειται για ένα σύστημα που βασίζεται στον εντοπισμό ενδιαφερόντων άρθρων, κάτι που καλύπτεται με ευρετικές τεχνικές μέσω του όρου information novelty. Αυτό που γίνεται είναι να εντοπίζονται ανανεώσεις (updates) σε κάποιο προυπάρχον θέμα αποφεύγοντας παράλληλα τις άσκοπες επαναλήψεις πληροφορίας. Για να συγκρίνει δύο κείμενα, το newsjunkie βρίσκει λέξεις-κλειδιά εξορύσσοντας named-entities, δηλαδή ονόματα ανθρώπων, οργανισμών και γεωγραφικών τοποθεσιών. Η έννοια αυτή είναι ουσιαστικά το κλειδί για να βρεθεί το update ενός θέματος βασιζόμενοι στην υπόθεση ότι το novelty συνήθως καλύπτεται εισάγοντας νέα named entities. Παρότι το σύστημα υπόσχεται πολλά, δεν υπάρχει διαθέσιμη έκδοση στο internet προς δοκιμή.

2. NewsMe [171]. Ουσιαστικά πρόκειται για ένα σύστημα όπου εξασκούνται τεχνικές προσωποποίησης χωρίς την ανάγκη για άμεσο feedback από τον χρήστη. Το σύστημα αναγνωρίζει τις κινήσεις του χρήστη ανανεώνοντας έμμεσα το προφίλ του κάτι που φαίνεται να είναι και πιο λογικό δεδομένου ότι ο χρήστης συνήθως δεν επιθυμεί να μπαίνει σε μεγάλη λεπτομέρεια σε σχέση με την περιγραφή των προτιμήσεών του. Το NewsMe χρησιμοποιεί δύο κανάλια (πρόσφατα νέα και προτεινόμενα νέα) και ο χρήστης χειροκίνητα μπορεί να προσθέσει ένα νέο άρθρο σε αυτά που παρακολουθεί αν θεωρεί ότι τον ενδιαφέρει ή να το κάνει blacklist αν ισχύει το αντίθετο. Επίσης μπορεί να ανανεώσει την βαθμολογία των άρθρων που παρακολουθεί ή που έχει απορρίψει. Η προσωποποίηση γίνεται μόνο στα πρόσφατα άρθρα και βασίζεται σε TF-IDF τεχνική πινάκων. Για την κατηγοριοποίηση το σύστημα χρησιμοποιεί το μοντέλο των κοντινότερων γειτόνων.

3. PersoNews [55]. Στο συγκεκριμένο σύστημα, ο χρήστης επιλέγει ένα πεδίο ενδιαφέροντος από μία θεματική ιεραρχία. Η επιλογή αυτή είναι ζωτική για το σύστημα μιας και οι πηγές που μπορούν να αφορούν ένα πεδίο μπορούν να είναι πάρα πολλές. Οι πηγές στη συνέχεια που αντιστοιχούν στις επιλογές του χρήστη παρακολουθούνται, κατηγοριοποιούνται και προσωποποιούνται από το σύστημα στον κάθε χρήστη. Παρότι το σύστημα διαφημίζει τα παραπάνω χαρακτηριστικά, δεν φαίνεται η online εκδοχή του [36] να προσφέρει κάτι παραπάνω από έναν news feeder.

4. GoogleNews [17]. Το διασημότερο σύστημα αποδελτίωσης που αυτή τη στιγμή υπάρχει στο διαδίκτυο είναι αυτό της Google. Πρόκειται για έναν αυτοματοποιημένο 'συσσωρευτή' νέων που συγκεντρώνει και παρουσιάζει άρθρα από γνωστά ειδησεογραφικά πρακτορεία. Ανάμεσα στα χαρακτηριστικά της υπηρεσίας είναι η αναζήτηση σε άρθρα, οι ειδοποιήσεις μέσω ηλεκτρονικού ταχυδρομείου για θέματα ενδιαφέροντος (Google News Alerts), RSS feeds. Το σημαντικότερο χαρακτηριστικό είναι η δυνατότητα προσωποποίησης που προσφέρει το Google

News όσον αφορά στα θέματα που εμφανίζονται στη σελίδα της υπηρεσίας και που έχουν τη μορφή widgets. Ο χρήστης μπορεί να μετακινεί τα widget, να προσθέτει και να αφαιρεί άρθρα και να έχει μια γενικότερη εποπτεία του τρόπου με τον οποίο παρουσιάζεται η πληροφορία. Επίσης νέα από διαφορετικές πηγές του Google News μπορούν να συνδυάζονται σε μία προσωποποιημένη σελίδα. Παρότι δεν εμπλέκεται ανθρώπινος παράγοντας για την λειτουργία του Google News, δεν είναι σαφές τι τεχνικές ανάκτησης πληροφορίας χρησιμοποιεί για την κατηγοριοποίηση και την προσωποποίηση. Είναι σαφές όμως ότι δυνατότητα περίληψης των άρθρων δεν υπάρχει ούτε υπάρχει και δυνατότητα προσωποποίησης όσον αφορά στο περιεχόμενο των άρθρων που εμφανίζονται. Για την κατηγοριοποίηση το σύστημα χρησιμοποιεί το μοντέλο των κοντινότερων γειτόνων.

ΚΕΦΑΛΑΙΟ 4

Το σύστημα *PeRSSonal*, αρχιτεκτονική και χαρακτηριστικά

> Part of the inhumanity of the computer is that, once it is competently programmed and working smoothly, it is completely honest.
>
> *Isaac Asimov, American Scientist, 1992*

Στο τρέχον κεφάλαιο, περιγράφεται η αρχιτεκτονική του συστήματος που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας, με το όνομα PeRSSonal, καθώς και οι στόχοι πάνω στους οποίους βασίζεται. Γίνεται παρουσίαση όλων των στοιχείων από τα οποία αποτελείται το σύστημα (υποσυστήματα), ενώ παράλληλα παρουσιάζεται ο τρόπος με τον οποίο γίνεται η εσωτερική διασύνδεση όλων των υποσυστημάτων καθώς και ο τρόπος με τον οποίο το σύστημα μπορεί να αξιοποιηθεί για χρήση από την client side εφαρμογή.

4.1 Χαρακτηριστικά του συστήματος

Το σύστημα που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας είναι αρκετά πολύπλοκο και περιλαμβάνει αρκετά υποσυστήματα που επιτελούν τις επιμέρους λειτουργίες. Αποτελεί επομένως έναν τμηματοποιημένο μηχανισμό, κάθε κομμάτι του οποίου σχεδιάστηκε με σκοπό να μπορεί να λειτουργήσει και αυτόνομα. Η επιθυμητή αυτή ιδιότητα επιτυγχάνεται με τη χρήση γενικά αποδεκτών προτύπων για τη διασύνδεση των διαφόρων υποσυστημάτων. Κάθε υποσύστημα δέχεται είσοδο σε μορφή XML και παράγει έξοδο στην ίδια μορφή κάνοντας έτσι εύκολη τη διαχείριση τους. Είναι επομένως εύκολο να αντικατασταθεί ένα τμήμα (module) του συστήματος από ένα νεότερο ή καλύτερο, μπορούμε π. χ. σε μελλοντική προσπάθεια να αντικαταστήσουμε το μηχανισμό της περίληψης με έναν νεότερο και αποτελεσματικότερο μηχανισμό χωρίς να χρειαστεί να πειράξουμε τα υπόλοιπα υποσυστήματα. Η παραπάνω λογική σχεδίασης αναφέρεται συχνά ως modular.

4.1.1 Στόχοι του συστήματος

Ο κεντρικός στόχος του συστήματος που αναπτύχθηκε είναι να παρέχει ως έξοδο, στο χρήστη ή σε άλλα συστήματα, ποιοτική πληροφορία. Όπως έχει ήδη αναφερθεί στα προηγούμενα κεφάλαια, η πληροφορία του παγκοσμίου ιστού είναι σχεδόν χαοτική με αποτέλεσμα οι χρήστες να μην είναι εφικτό να προσεγγίσουν πληροφορία που τους είναι χρήσιμη και επιθυμητή. Σκοπός του συστήματός μας είναι να δημιουργήσουμε την κατάλληλη υποδομή ούτως ώστε να πραγματοποιείται φιλτράρισμα στην πληροφορία που κινείται στο διαδίκτυο και να αξιοποιείται κατά το μέγιστο δυνατό βαθμό προτού φτάσει στο χρήστη. Αυτό που θέλουμε να επιτύχουμε λοιπόν μέσα από αυτή τη διαδικασία είναι να αξιολογήσουμε κατά κάποιο τρόπο την πληροφορία που μας γίνεται διαθέσιμη από τα διάφορα ειδησεογραφικά πρακτορεία και να δημιουργήσουμε τα κανάλια επικοινωνίας εκείνα που θα παρέχουν ποιοτική πληροφορία, ικανή να παρέχει σαφή και σφαιρική ενημέρωση στον χρήστη.

Συχνά στις μέρες μας έχει παρατηρηθεί να μιλούμε για την ποιότητα στην ενημέρωση που παρέχει το διαδίκτυο. Το σύστημά μας θα αντλεί και θα επεξεργάζεται περιεχόμενο που εντοπίζεται σε ειδησεογραφικούς δικτυακούς τόπους. Το περιεχόμενό τους θα παραλαμβάνεται καθημερινά, θα φιλτράρεται, θα αναλύεται, θα κατηγοριοποιείται και θα περιλήπτεται. Έτσι στο τέλος θα έχουμε στα χέρια μας μια κατηγοριοποιημένη περίληψη της πληροφορίας, έτοιμη να παρουσιαστεί στο χρήστη ανάλογα με τις προτιμήσεις του, ή να μεταδοθεί σε άλλα ενδιάμεσα συστήματα, όπως η εφαρμογή client side για την επιφάνεια εργασίας του χρήστη που επίσης αναπτύσσεται στα πλαίσια της παρούσας εργασίας. Παράλληλα, το σύστημα δε θα αφήνει τον τελικό χρήστη έξω από τη διαδικασία μιας και το περιεχόμενο θα είναι πλήρως προσωποποιημένο και παραμετροποιήσιμο σε αυτόν/ην.

Η κατηγοριοποίηση που θα πραγματοποιείται θα είναι σε δύο επίπεδα, τόσο ένα οριζόντιο που θα αφορά το γλωσσολογικό και εννοιολογικό κομμάτι, όσο και το κομμάτι της σχετικότητας και της ανάλυσης. Για κάθε κείμενο που εμφανίζεται στη συλλογή μας, μπορούμε να το κατατάξουμε σε μία κατηγορία βάση της έννοιας την οποία εμπεριέχει ή σε κάποια κατηγορία ανάλυσης ανάλογα με τη διαδικασία που ακολουθήσαμε για να επεξεργαστούμε την πληροφορία του κειμένου. Αυτό θα γίνει πιο σαφές μέσα από ένα παράδειγμα. Μπορούμε σε ένα κείμενο να κάνουμε μια εκτενέστατη ανάλυση και να εντοπίσουμε πως αναφέρεται στην κατηγορία ποδόσφαιρο από τη γενικότερη κατηγορία αθλητικά. Το βάρος που θα έχει για το κείμενο η έννοια ποδόσφαιρο θα είναι μεγάλη σε αυτή την ανάλυση, όμως θα περιοριστεί με αυτό τον τρόπο το κοινό στο οποίο απευθύνεται το κείμενο καθώς έχουμε μικρό βαθμό αφαίρεσης. Από την άλλη μεριά μπορούμε να κάνουμε μια πιο γενική ανάλυση του κειμένου, εντοπίζοντας απλά την κατηγορία στην οποία ανήκουν ενδεικτικές προτάσεις μέσα από το κείμενο. Με αυτό τον τρόπο οι λέξεις κλειδιά μέσα στα κείμενα ταξινομούνται βάσει του αφαιρετικού βάρους, δηλαδή βάση της ικανότητάς τους να περιγράψουν το κείμενο σαν λέξεις που επελέγησαν από μέρος του κειμένου και όχι από το σύνολό του. Τα κομμάτια που επελέγησαν από το κείμενο μπορούν να αποτελέσουν και μια σύντομη περιγραφή του κειμένου, η αλλιώς μια περίληψή του. Ως εκ' τούτου, φαίνεται να δημιουργούνται δύο είδη γενικών κατηγοριών. Πρόκειται για τις κατηγορίες που δημιουργούνται για τους χρήστες που επιθυμούν να έχουν πρόσβαση σε πολύ εξειδικευμένη πληροφορία και σε αυτούς που επιθυμούν να έχουν πρόσβαση σε πληροφορία γενικά.

Όσον αφορά την έννοια της ανάλυσης είναι ένα κομμάτι το οποίο θα γίνεται αλγοριθμικά βασισμένο στη μέθοδο Support Vector Machine. Το θέμα που αφορά στο αφαιρετικό βάρος δε θα μπορούσε να ανήκει αποκλειστικά σε μία μηχανή αλλά περισσότερο σε έναν άνθρωπο, σε ένα χρήστη του συστήματος. Οι χρήστες θα είναι αυτοί που θα έχουν τον κύριο λόγο στη δημιουργία των αφαιρετικών βαρών ανάλογα με τον τρόπο με τον οποίο αντιμετωπίζουν τα αποτελέσματα. Στόχος είναι να ενταχθεί ο χρήστης του συστήματος στη διαδικασία με την οποίο λαμβάνει ο ίδιος αποτελέ-

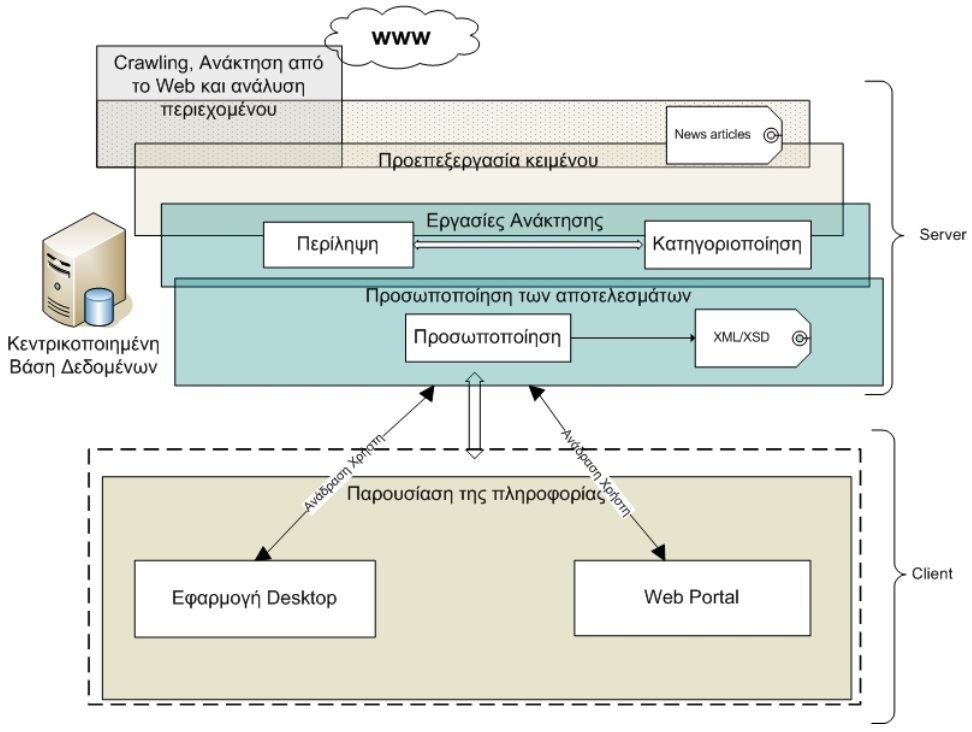
σματα, είτε αυτά αφορούν τη διαδικασία της περίληψης, είτε τη διαδικασία της κατηγοριοποίησης. Απώτερος στόχος είναι να γίνει ξεχωριστή περίληψη και κατηγοριοποίηση της πληροφορίας για κάθε χρήστη, προκειμένου ο καθένας ο οποίος χρησιμοποιεί το σύστημα να είναι σε θέση να έρθει πιο κοντά στα αποτελέσματα που επιθυμεί να βρει.

4.2 Γενική αρχιτεκτονική του συστήματος

Το PeRSSonal βασίστηκε σε κατανεμημένη αρχιτεκτονική και αυτόνομα υποσυστήματα, αλλά η διαδικασία που ακολουθείται για να παραχθεί το επιθυμητό αποτέλεσμα έχει έννοιες ακολουθιακές. Αυτό σημαίνει ότι η ροή πληροφορίας είναι αντιπροσωπευτική των υποσυστημάτων από τα οποία αποτελείται το σύστημα. Ένα άλλο σημαντικό αρχιτεκτονικό ζήτημα είναι η τμηματικότητα του μηχανισμού.

Εστιάζουμε στο τμήμα της περίληψης κειμένου παρότι θα παρουσιαστεί αναλυτικά το τμήμα κατηγοριοποίησης καθώς και το τμήμα προσωποποίησης του μηχανισμού, με στόχο να παρουσιαστεί η αλληλοσύνδεση των τμημάτων αυτών του συνολικού συστήματος. Όπως έχει ήδη αναφερθεί, η διαδικασία περίληψης δέχεται πληροφορία από την διαδικασία προ-επεξεργασίας και ανταλλάσσει γνώση με τους μηχανισμούς κατηγοριοποίησης και προσωποποίησης, με στόχο την δημιουργία της περίληψης κειμένου σύμφωνα με τις ανάγκες του κάθε χρήστη.

Ο μηχανισμός αποτελείται από μια σειρά από συστήματα για την παραγωγή του επιθυμητού αποτελέσματος. Η συνεργασία μεταξύ των κατανεμημένων υποσυστημάτων βασίζεται σε ανοιχτά πρότυπα για είσοδο και έξοδο τα οποία υποστηρίζονται από κάθε τμήμα του συστήματος αλλά και από την επικοινωνία με την κεντρικοποιημένη βάση δεδομένων. Το σχήμα 4.1 παρουσιάζει την αρχιτεκτονική του συνολικού μηχανισμού.



Σχήμα 4.1: Βασική Αρχιτεκτονική του Συστήματος.

Ο μηχανισμός, όπως παρουσιάζεται στο σχήμα 4.1 ακολουθεί τις παρακάτω διαδικασίες:

I. Συλλογή ιστοσελίδων που περιέχουν ειδήσεις από τον Παγκόσμιο Ιστό και εξαγωγή του χρήσιμου κείμενου από αυτές

II. Ανάλυση του εξαγόμενου κείμενου

III. Εφαρμογή αλγορίθμων κατηγοριοποίησης και περίληψης στο κείμενο

IV. Εφαρμογή αλγορίθμων προσωποποίησης για τον εκάστοτε χρήστη στο κείμενο

V. Παρουσίαση των αποτελεσμάτων ως έξοδο στον χρήστη είτε μέσω της Desktop εφαρμογής είτε μέσω του Web interface που προσφέρει το PeRSSonal

Για να συλλεχθούν οι ιστοσελίδες από τον παγκόσμιο ιστό, ένας απλός εστιασμένος Crawler[1] χρησιμοποιείται. Οι διευθύνσεις οι οποίες χρησιμοποιούνται ως είσοδος για τον crawler εξάγονται από ροές νέων (RSS Feeds). Οι ροές νέων 'δείχνουν' απευθείας σε σελίδες όπου υπάρχουν τα άρθρα με νέα. Ο crawler αποθηκεύει τις ιστοσελίδες σε μορφή html χωρίς άλλα στοιχεία της ιστοσελίδας (εικόνες, css, javascript, κ.λπ. παραλείπονται). Αποθηκεύοντας μόνο την σελίδα σε μορφή html, η βάση δεδομένων γεμίζει με σελίδες που είναι έτοιμες για είσοδο στο πρώτο επίπεδο ανάλυσης. Τα RSS feeds-πηγές μας παρέχουν επίσης μία αρχική πληροφορία κατηγοριοποίησης για τα άρθρα που προέρχονται από την κάθε πηγή, μιας και συνήθως, τα μεγάλα ειδησεογραφικά sites κάνουν από μόνα τους μια καταγοριοποίηση των άρθρων που εισάγονται. Η διαδικασία του crawling είναι κατανεμημένη σε πολλά υποσυστήματα τα οποία συγχρονίζονται βάσει της κεντρικοποιημένης βάσης δεδομένων.

Ένας crawler έχει ως έξοδο συχνά κώδικα HTML χωρίς καμία επεξεργασία. Φυσικά η χρήση του κώδικα στη διαδικασία κατηγοριοποίησης ή περίληψης είναι κάτι το απαγορευτικό. Προκειμένου λοιπόν να μπορέσουμε να προχωρήσουμε στο επόμενο βήμα θα πρέπει να έχουμε στα χέρια μας καθαρό κείμενο. Ένας μηχανισμός ανάλυσης και εξαγωγής του χρήσιμου μόνο κειμένου από σελίδες του διαδικτύου κατασκευάστηκε, προκειμένου να μπορέσουμε να παρέχουμε στις διαδικασίες του συστήματος 'καθαρό' κείμενο (χωρίς στοιχεία κώδικα). Το εξαγόμενο 'καθαρό' κείμενο αποθηκεύεται επίσης στη βάση απ' όπου χρησιμοποιείται ασύγχρονα από τα επόμενα βήματα.

Κατά τη διάρκεια του πρώτου επιπέδου ανάλυσης, το σύστημά μας απομονώνει το 'χρήσιμο κείμενο' από την html σελίδα. Ως χρήσιμο κείμενο κρατάμε τον τίτλο και το κυρίως σώμα του άρθρου (article body). Το δεύτερο επίπεδο ανάλυσης δέχεται ως είσοδο αρχεία σε μορφή XML τα οποία περιλαμβάνουν τον τίτλο και το σώμα των άρθρων. Ο κύριος σκοπός του είναι να εφαρμόσει αλγορίθμους προ-επεξεργασίας πάνω στο κείμενο και να παράγει ως έξοδο λέξεις-κλειδιά, την θέση τους στο κείμενο καθώς και την συχνότητα εμφάνισης τους μέσα σε αυτό. Αυτά τα αποτελέσματα είναι απαραίτητα για να προχωρήσουμε στο τρίτο επίπεδο ανάλυσης.

Η καρδία του μηχανισμού μας βρίσκεται στο τρίτο επίπεδο ανάλυσης, όπου τα υποσυστήματα της περίληψης και κατηγοριοποίησης εντοπίζονται. Ο κύριος στόχος τους είναι να χαρακτηρίζουν ένα άρθρο με μία ετικέτα (κατηγορία) και να παράγουν μια περίληψή του. Τα αποτελέσματα μπορούν στη συνέχεια είτε να παρουσιαστούν στους τελικούς χρήστες μέσω ενός προσωποποιημένου portal, είτε να οδηγηθούν προς χρήση σε άλλα υποσυστήματα που ακολουθούν. Η έξοδος αυτή ακολουθεί επίσης τα ανοιχτά πρότυπα και δίνεται σε μορφή XML και επομένως εύκολα αναγνωρίσιμη από οποιοδήποτε υποσύστημα μπορεί να ακολουθεί.

Όσον αφορά την διαδικασία της κατηγοριοποίησης, αρχικό και βασικό χαρακτηριστικό του συστήματος είναι το σύνολο των κειμένων εκπαίδευσης. Προκειμένου να είναι εφικτή η δυνατότητα

[1] Focused Crawler: Ένας εστιασμένος μηχανισμός αυτόματης αναζήτησης και καταγραφής περιεχομένου του διαδικτύου που σε αντίθεση με έναν απλό Crawler δεν διαπερνά ότι βρεθεί στο δρόμο του, παρά σελίδες που είναι σχετικές με ένα προκαθορισμένο σετ θεμάτων

αυτόματης κατηγοριοποίησης θα πρέπει να αρχικοποιηθούν κάποιες βασικές κατηγορίες με κείμενα αντιπροσωπευτικά αυτών. Έτσι το κομμάτι εκείνο το οποίο θα είναι υπεύθυνο για την κατηγοριοποίηση των κειμένων, θα πρέπει αρχικά να αναλάβει να δημιουργήσει τις διαφορετικές κατηγορίες. Εν συνεχεία θα είναι σε θέση να παραλάβει μη κατηγοριοποιημένα κείμενα και να προσπαθήσει να τα εντάξει σε κάποια από τις ήδη υπάρχουσες κατηγορίες.

Η κατηγοριοποίηση της πληροφορίας περνά από συγκεκριμένα στάδια τα οποία αποτελούν και διαφορετικά υποσυστήματα που λειτουργούν σειριακά για κάθε ξεχωριστό κείμενο αλλά συνολικά παράλληλα. Έτσι υπάρχει ξεχωριστός μηχανισμός που πραγματοποιεί την προεπεξεργασία και ξεχωριστός μηχανισμός που αναλαμβάνει να 'τρέξει' τον αλγόριθμο κατηγοριοποίησης για κάθε επεξεργασμένο κείμενο.

Οι παραπάνω ξεχωριστοί μηχανισμοί όπως προαναφέρθηκε πρέπει να λειτουργήσουν σειριακά πάνω σε κάθε ξεχωριστό κείμενο προκειμένου να είναι επιτυχημένη τόσο η κατηγοριοποίηση όσο και η δημιουργία του δυναμικού προφίλ. Ωστόσο κάθε μηχανισμός εσωτερικά είναι δημιουργημένος ώστε να μπορεί να δουλεύει σαν ένα παράλληλο σύστημα αφού καθένας είναι ανεξάρτητος από τους υπολοίπους. Στην ουσία δηλαδή εκτελούμε το crawling και το κατέβασμα των σελίδων σε διαφορετική χρονική στιγμή από το βήμα της προεπεξεργασίας κειμένου για εξαγωγή κωδικολέξεων γλιτώνοντας έτσι πολύτιμο χρόνο που αφορά στο κατέβασμα της σελίδας. Τέλος, όλοι οι μηχανισμοί που έχουν αναπτυχθεί, χρησιμοποιούν τοπικά μία μικρή μνήμη του συστήματος (κατά την εκτέλεσή τους) προκειμένου να αποθηκεύουν (τοπικά) συγκεκριμένα αποτελέσματα από τις διαδικασίες τους, η μνήμη αυτή περιλαμβάνει είτε τη φυσική μνήμη συστήματος που καταλαμβάνουν τα προγράμματα κατά την εκτέλεσή τους, είτε κάποια προσωρινά αρχεία. Ωστόσο, όλα τα αποτελέσματα συγκεντρώνονται σε μία κεντρικοποιημένη βάση δεδομένων, προκειμένου να εξασφαλιστεί η ακεραιότητα τους αλλά και η διαθεσιμότητα τους όσο το δυνατόν νωρίτερα στους υπόλοιπους μηχανισμούς. Η κεντρική βάση δεδομένων μπορεί να προκαλεί αρκετή καθυστέρηση σε συγκεκριμένα κομμάτια του συστήματος, ωστόσο μιλούμε για ένα σύστημα το οποίο απαιτεί απόλυτη ακρίβεια στα δεδομένα και αποφυγή διπλοεγγραφών η σφαλμάτων. Η επιλογή της κεντρικοποιημένης βάσης δεδομένων έγινε για λόγους απλότητας και εφαρμοσιμότητας μιας και δεν υπάρχει κάποιο open source framework διαθέσιμο για αξιόπιστη κατανεμημένη αρχιτεκτονική για την βάση δεδομένων. Με αυτό εννοούμε επίσης ότι το σύστημα μπορεί να μεταβεί σε μία τέτοια αρχιτεκτονική μελλοντικά και αν αυτό κριθεί αναγκαίο για λόγους απόδοσης και διαθεσιμότητας.

Λίγο πριν την παρουσίαση των αποτελεσμάτων στο χρήστη, υπάρχει ο μηχανισμός ο οποίος αναλαμβάνει να διαχειριστεί το προφίλ του κάθε χρήστη ούτως ώστε να παράγει το προσωποποιημένο περιεχόμενο που θα του αποσταλεί. Πρόκειται για ένα μηχανισμό ο οποίος λαμβάνει υπόψη του τις προτιμήσεις του χρήστη όσον αφορά τις κατηγορίες νέων ή κάποια ξεχωριστής σημασίας λέξεων-κλειδιών που τον ενδιαφέρουν, αλλά και τις δυνατότητες απεικόνισης που διαθέτει η συσκευή του, π. χ. πρόκειται για pda, κινητό τηλέφωνο, φορητό υπολογιστή ή για κάποια άλλη συσκευή. Οι πληροφορίες αυτές οργανώνονται και αποθηκεύονται επίσης στην κεντρικοποιημένη βάση δεδομένων, ώστε να κατασκευαστεί το δυναμικό προφίλ του χρήστη.

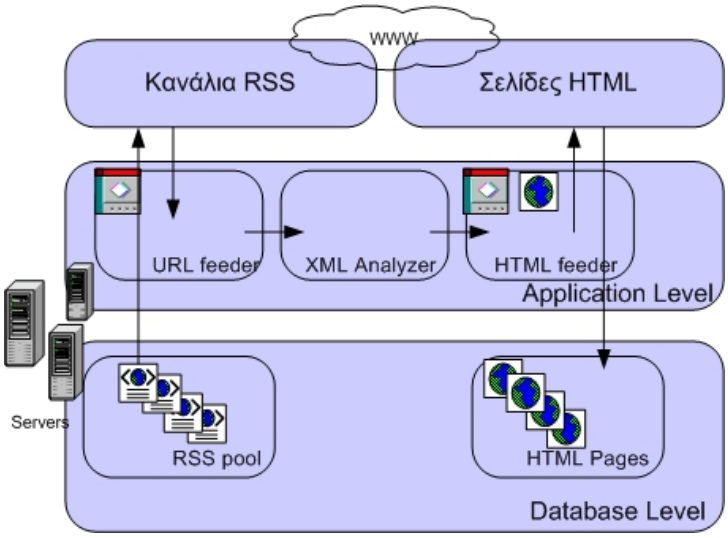
4.3 Υποσυστήματα

Στη συνεχεία ακολουθεί μία περισσότερο αναλυτική περιγραφή των υποσυστημάτων του μηχανισμού προκειμένου να γίνει κατανοητή η λειτουργία του σε κάθε διαφορετικό επίπεδο υλοποίησης.

4.3.1 Συλλογή πληροφορίας

Η συλλογή πληροφορίας από το σύστημά μας αποτελεί ουσιαστικά την αρχική διεργασία που επιτελείται μιας και είναι αυτή που εξασφαλίζει την συνεχή και αδιάκοπη ροή άρθρων από το

Διαδίκτυο. Για την διεργασία αυτή εκμεταλλευόμαστε την τάση που επικρατεί σε όλους τους δικτυακούς τόπους να προσφέρουν κανάλια άμεσης επικοινωνίας με τους χρήστες, και δε μιλούμε για κάτι διαφορετικό από τα RSS Feeds.



Σχήμα 4.2: Μηχανισμός Συλλογής Πληροφορίας.

Η αρχιτεκτονική του μηχανισμού είναι απλή και εκμεταλλεύεται πλήρως το γεγονός ότι οι μεγαλύτερες δικτυακές πύλες ενημέρωσης προσφέρουν στους χρήστες RSS feeds. Όπως φαίνεται από το σχήμα 4.2, ένας απλοϊκός mixed selective crawler χρησιμοποιείται προκειμένου να λαμβάνει το σύστημά μας HTML σελίδες. Πρόκειται για έναν mixed crawler διότι συνδυάζει τη χρήση wrapper και crawler. Ο wrapper είναι ένας μηχανισμός αναγνώρισης προτύπων που συνήθως ακολουθείται από επεξεργασία αυτών. Στην περίπτωσή μας ο wrapper στο μηχανισμό συλλογής πληροφορίας εντοπίζει μέσα στα XML αρχεία εκείνα τα σημεία τα οποία περιέχουν πληροφορίες για τα άρθρα που θέλουμε να εξάγουμε. Μέσα από αυτά τα αρχεία προκύπτουν τα URL seeds τα οποία επανατροφοδοτούν το ίδιο μηχανισμό για να προχωρήσει στο 'κατέβασμα' των σελίδων HTML, που περιέχουν άρθρα, από τη φυσική τους θέση χωρίς να χρειαστεί καμία απολύτως αναζήτηση. Ο wrapper συνεπώς χρησιμοποιείται για να μπορέσουμε να εξάγουμε τον τίτλο του άρθρου και τη διεύθυνση στην οποία βρίσκεται με τη βοήθεια των RSS feeds και εν συνεχεία το πρόγραμμα αλλάζει μορφή και μετατρέπεται σε crawler ο οποίος 'επισκέπτεται' τα URLs που έχει εξάγει ο wrapper και από αυτά λαμβάνει τον HTML κώδικα. Η βάση δεδομένων δε χρειάζεται τις ενδιάμεσες πληροφορίες και έτσι οι πληροφορίες που έχει είναι η λίστα με τα RSS. Οι πληροφορίες που αποθηκεύονται για κάθε άρθρο και για κάθε RSS Feed που διαπερνάται φαίνονται στους πίνακες 4.3.1 και 4.3.1 αντίστοιχα.

Άρθρο	
τίτλος	ο τίτλος του άρθρου που εξάγεται από το RSS Feed
HTML	ο κώδικας της σελίδας του άρθου
γλώσσα	Η γλώσσα στην οποία είναι γραμμένο το άρθρο
ημερομηνία	η ημερομηνία κατά την οποία ανακτήθηκε από τον crawler το άρθρο
κατηγορία	η κατηγορία στην οποία ανήκει το άρθρο βάσει του RSS Feed από το οποίο προέρχεται

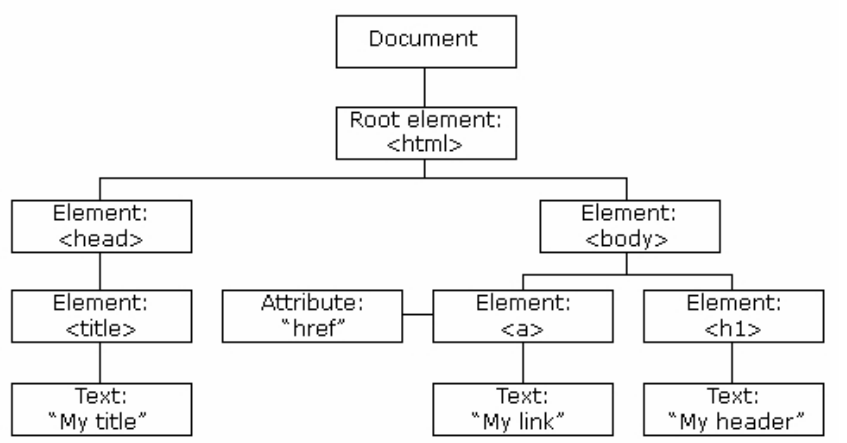
Πίνακας 4.1: Πληροφορίες που ανακτώνται για κάθε άρθρο

RSS Feed	
περιγραφή	μία σύντομη περιγραφή για το RSS Feed. Η πληροφορία βρίσκεται στα μεταδεδομένα του feed
url	το url του RSS Feed, ουσιαστικά ένα XML αρχείο
γλώσσα	Η γλώσσα στην οποία τα άρθρα του RSS Feed είναι γραμμένα
κατηγορία	η κατηγορία στην οποία ανήκει το άρθρο RSS Feed

Πίνακας 4.2: Πληροφορίες που ανακτώνται για κάθε *RSS Feed*

4.3.2 Εξαγωγή χρήσιμου κειμένου

Η εξαγωγή χρήσιμου κειμένου, μια διαδικασία που συχνά αναφέρεται και ως φιλτράρισμα, βασίζεται στην ιδιότητα της HTML να μπορεί να αναπαρασταθεί σε δενδρική μορφή σύμφωνα με το DOM (Document Object Model) [9] μοντέλο, όπως φαίνεται και στο σχήμα 4.3.

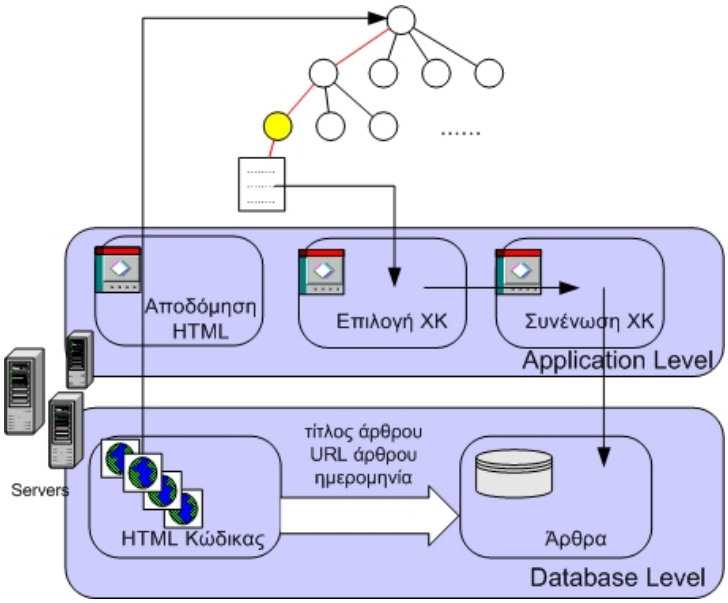


Σχήμα 4.3: *HTML Document Object Model (DOM).*

Το σχήμα 4.3 είναι η DOM αναπαράσταση του παρακάτω HTML κώδικα.

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

Βασιζόμενοι λοιπόν στο γεγονός ότι κάθε HTML κώδικας μπορεί να αποδομηθεί στα βασικά του στοιχεία σε δενδρική μορφή, χρησιμοποιούμε ένα μηχανισμό όπως αυτός που φαίνεται στο σχήμα 4.4, προκειμένου να εξάγουμε το χρήσιμο κείμενο από τις HTML σελίδες.

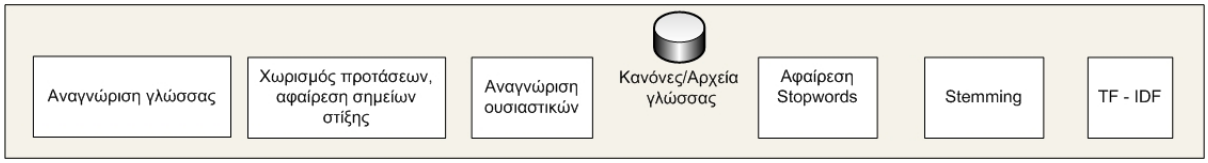


Σχήμα 4.4: Εξαγωγή Χρήσιμου Κειμένου.

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

4.3.3 Προεπεξεργασία κειμένου

Ο μηχανισμός προεπεξεργασίας κειμένου είναι ένα σημαντικό τμήμα του συνολικού μηχανισμού ο οποίος αναλαμβάνει το καθάρισμα του σώματος του κειμένου και την εξαγωγή κωδικολέξεων (keywords). Η διαδικασία για την προεπεξεργασία κειμένου και την εξαγωγή των λέξεων κλειδιών φαίνεται στο Σχήμα 4.5. Η είσοδος στο υποσύστημα αυτό είναι μορφής XML που περιέχει τα απαραίτητα μόνο στοιχεία: τίτλος και σώμα κειμένου.



Σχήμα 4.5: Προεπεξεργασία κειμένου και εξαγωγή κωδικολέξεων.

Εκτός από το αρχείο (ή τη δομή) XML, σαν είσοδος στον μηχανισμό δίνεται ένας αριθμός από παραμέτρους:

- το ελάχιστο μήκος λέξης (οι λέξεις που είναι μικρότερες από αυτό το μήκος θα αφαιρεθούν)

- καθορισμός εάν τα αριθμητικά δεδομένα θα κρατηθούν ή θα αφαιρεθούν
- καθορισμός μιας λίστας από λέξεις τετριμμένες και συνηθισμένες οι οποίες δεν εκφράζουν κάποιο συγκεκριμένο νόημα και μπορούν να θεωρηθούν ως 'σκουπίδια' (stopwords)
- καθορισμός του αλγορίθμου stemming που θα χρησιμοποιηθεί
- καθορισμός της βαρύτητας που δίνεται στα ουσιαστικά του κειμένου

Η διαδικασία που ακολουθείται από τον μηχανισμό προεπεξεργασίας κειμένου έχει ως εξής. Αρχικά, η γλώσσα του κειμένου αναγνωρίζεται κάτι που γίνεται είτε με ειδικό λογισμικό αναγνώρισης είτε έμμεσα χρησιμοποιώντας την προκαθορισμένη γλώσσα του RSS feed από το οποίο προέρχεται το άρθρο. Ακολουθεί η διαδικασία χωρισμού των προτάσεων, ο ορθογραφικός έλεγχος, και έπειτα η αφαίρεση των σημείων στίξης που υπάρχουν. Στη συνέχεια λαμβάνει χώρα η διεργασία αναγνώρισης των ουσιαστικών του κειμένου χρησιμοποιώντας τον POS SVM-based tagger [99], και που μπορεί να καθορίσει με μεγάλη ακρίβεια τα ουσιαστικά που περιέχει η κάθε πρόταση. Μερικές κοινότυπες τεχνικές εξαγωγής κωδικολέξεων ακολουθούν που σκοπό έχουν να περιορίσουν τον θόρυβο των αποτελεσμάτων: η αφαίρεση των stopwords και το stemming. Είναι σημαντικό να τονιστεί ότι η διαδικασία εύρεσης των ουσιαστικών του κειμένου πρέπει να προηγείται αυτών των διεργασιών αν επιθυμούμε να επιτύχει με μεγάλη πιθανότητα (μιας και οι λέξεις μπορούν εύκολα να τακτοποιηθούν ως μέρη του λόγου μέσα στην πρόταση στην οποία ανήκουν). Ένα εξίσου σημαντικό στοιχείο είναι ότι οι διαδικασίες της αναγνώρισης των ουσιαστικών, της αφαίρεσης των stopwords και του stemming είναι ισχυρά εξαρτώμενες από την γλώσσα του κειμένου. Γνωρίζοντας επομένως την γλώσσα του κειμένου, μπορούμε να λάβουμε τις σωστές αποφάσεις προεπεξεργασίας του: να αποφασίσουμε ποια θα πρέπει να είναι η λίστα με τα stopwords που θα πρέπει να αφαιρεθούν, ποιοι θα πρέπει να είναι οι κανόνες για το POS tagging που θα εφαρμόσει ο SVM tagger, ποιοι θα είναι οι κανόνες για την διαδικασία stemming που θα εφαρμοστεί και τελικά ποιο θα είναι το μέγεθος των αρχικών λέξεων που θα πρέπει να κρατηθούν, μιας και ορισμένες γλώσσες περιέχουν κατά κόρων μεγαλύτερες λέξεις από κάποιες άλλες.

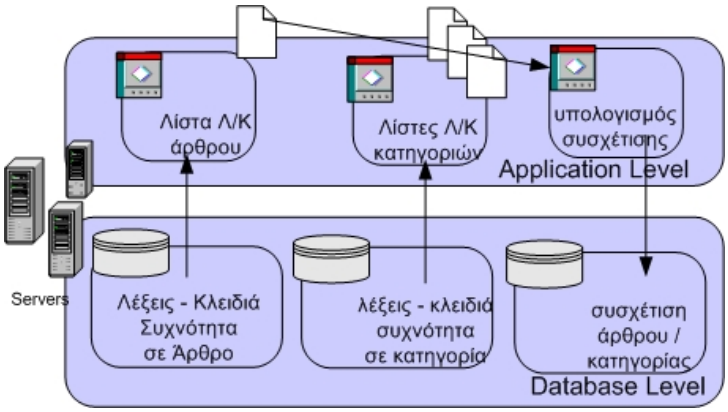
Η έξοδος του μηχανισμού προεπεξεργασίας κειμένου μπορεί είτε να αποθηκεύεται στη βάση δεδομένων του συστήματος, είτε να δρομολογείται σε άλλα υποσυστήματα που ακολουθούν. Στην δεύτερη περίπτωση, η έξοδος έχει τη μορφή XML αρχείου έτσι ώστε να είναι εύκολη η διασύνδεση με άλλα υποσυστήματα. Η έξοδος περιλαμβάνει:

- τις κωδικολέξεις που προέκυψαν από την διαδικασία του keyword extraction
- τις θέσεις των keywords στο αρχικό κείμενο, σε ποιες προτάσεις δηλαδή εμφανίζονται
- το πλήθος με το οποίο εμφανίζονται τα keywords κάτι που εκφράζεται είτε ως απόλυτη συχνότητα εμφάνισης (π. χ. ένα keyword εμφανίζεται 5 φορές στο κείμενο), είτε ως σχετική συχνότητα εμφάνισης (π. χ. ένα keyword εμφανίζεται 5 φορές σε ένα κείμενο 50 λέξεων, άρα με σχετική συχνότητα 0,1).
- την πληροφορία για το αν το keyword είναι ουσιαστικό ή όχι

Τα παραπάνω αναπαριστώνται μέσω πινάκων term frequency – inverse document frequency (TF-IDF) οι οποίοι αποθηκεύονται στην βάση δεδομένων και αξιοποιούνται από τις διαδικασίες του επόμενου επιπέδου.

4.3.4 Κατηγοριοποίηση κειμένου

Η κατηγοριοποίηση των κειμένων από το σύστημα PeRSSonal αποτελεί ένα κεντρικό συστατικό του μηχανισμού που αναπτύχθηκε και σε συνδυασμό με εκείνο της εξαγωγής περίληψης, βρίσκονται στο δεύτερο επίπεδο ανάλυσης του συστήματος αποτελώντας τον πυρήνα του μηχανισμού. Το υποσύστημα περιγράφεται από το Σχήμα 4.6.



Σχήμα 4.6: Μηχανισμός κατηγοριοποίησης κειμένου.

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

Το υποσύστημα κατηγοριοποίησης βασίζεται στην μετρική ομοιότητας συνημιτόνου, σε εσωτερικά γινόμενα καθώς και σε υπολογισμούς ζυγίσματος όρων. Η χρήση αυτών των μετρικών γίνεται ύστερα από την αρχικοποίηση του training set της βάσης γνώσης και μέσω μιας διαδικασίας η οποία on the fly ελέγχει τη συσχέτιση του κάθε keyword του προς κατηγοριοποίηση κειμένου με τις υπάρχουσες κατηγορίες. Η ομοιότητα συνημιτόνου υπολογίζει ουσιαστικά ένα όρο της μορφής:

$$similarity = \frac{AB}{||A||||B||}$$

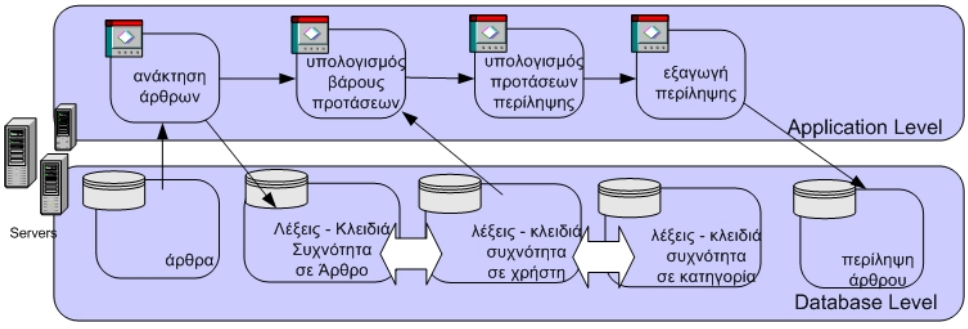
όπου ο όρος A είναι ο πίνακας που περιέχει τα keywords που κρατούνται από το κείμενο και ο πίνακας B περιέχει τα κοινά keywords που έχει η ενίοτε κατηγορία προς σύγκριση με το συγκεκριμένο κείμενο. Οι συσχετίσεις που θα βρεθούν αθροίζονται και κανονικοποιούνται με αποτέλεσμα να προκύπτει για κάθε κείμενο ένα ποσοστό ομοιότητας (relativity) με κάθε μια από τις υπάρχουσες κατηγορίες. Εάν το training set είναι αποτελεσματικό και αξιόπιστο, τα άρθρα που περιέχουν πολλά keywords σχετικά με κάποια από τις κατηγορίες, θα πρέπει να έχουν κατηγοριοποιηθεί με συσχέτιση μεγαλύτερη ως προς αυτή. Στην πράξη βέβαια, ακόμη και αν το κείμενο είναι εντελώς αντιπροσωπευτικό κάποιας κατηγορίας, δεν αποκτά συσχέτιση 100% με μία και μόνο κατηγορία,

αφού είναι φυσικό να περιέχει ορισμένα keywords τα οποία συσχετίζονται και με τις υπόλοιπες κατηγορίες (το άθροισμα των συσχετίσεων ενός κειμένου με όλες τις κατηγορίες, προφανώς σε κάθε περίπτωση είναι 1). Η έξοδος του υποσυστήματος κατηγοριοποίησης, οι συσχετίσεις δηλαδή του κειμένου με κάθε κατηγορία, αποθηκεύονται στη βάση δεδομένων του συστήματος ολοκληρώνοντας τη διαδικασία της κατηγοριοποίησης.

4.3.5 Εξαγωγή περίληψης κειμένου

Το υποσύστημα εξαγωγής περίληψης (Σχήμα 4.7) κειμένου του μηχανισμού αποτελεί ένα ανεξάρτητο υποσύστημα το οποίο δέχεται ως είσοδο τα αποτελέσματα του keyword extraction (αποθηκευμένα στη βάση ή σε μορφή XML) που περιέχουν: τα keywords που κρατήθηκαν, τη συχνότητα εμφάνισής τους στο κείμενο, τις θέσεις τους (σε ποιες προτάσεις εμφανίζονται, π. χ. 1η, 3η, κ.ο.κ.), την πληροφορία για το αν πρόκειται για keywords-ουσιαστικά ή όχι και το πόσες προτάσεις πρέπει να κρατηθούν για την τελική περίληψη. Τα στοιχεία αυτά, μαζί με την πληροφορία για τον τίτλο του κειμένου, είναι αρκετά ώστε να μπορεί το υποσύστημα αυτό να επιχειρεί μια βαθμολόγηση των προτάσεων του κειμένου.

Θα πρέπει να πούμε σε αυτό το σημείο ότι, ο μηχανισμός αυτόματης εξαγωγής περίληψης δεν χρειάζεται απαραίτητα αυτό καθ' αυτό το κείμενο αν και για να παραχθεί η τελική περίληψη ενός κειμένου αυτό είναι αναγκαίο. Με το προηγούμενο εννοούμε ότι, το υποσύστημα αυτό μπορεί να παράγει μια τελική κατάταξη των προτάσεων του κειμένου απλά και μόνο με τις εισόδους που περιγράφηκαν νωρίτερα και ενώ το αρχικό κείμενο βρίσκεται αποθηκευμένο μία φορά μόνο στην βάση δεδομένων. Το τελευταίο δεδομένο εισόδου του υποσυστήματος περιγράφει πόσες προτάσεις επιθυμούμε να έχουμε ως έξοδο για περίληψη του αρχικού κειμένου. Το πλήθος των προτάσεων μπορεί να καθοριστεί είτε ως ποσοστό % των προτάσεων του αρχικού κειμένου είτε ως συνολικό πλήθος χαρακτήρων. Για παράδειγμα, αν το αρχικό κείμενο είχε 20 προτάσεις και κρατάμε ένα ποσοστό 30% επί των προτάσεων, στην περίληψη θα κρατηθούν οι 6 σημαντικότερες προτάσεις του κειμένου, αντίθετα, εάν επιθυμούμε η περίληψη του κειμένου να περιέχει περίπου ένα συγκεκριμένο πλήθος χαρακτήρων, θα επιλεχθούν τόσες προτάσεις από τις σημαντικότερες ώστε και να καλύπτεται το πλήθος χαρακτήρων που τέθηκε και να μην ξεπερνιέται κατά πολύ αυτό. Στην ουσία επιλέγεται η βέλτιστη επιλογή μήκους χαρακτήρων στο όριο να επιλεχθεί μια παραπάνω πρόταση ή μια λιγότερη.



Σχήμα 4.7: Μηχανισμός περίληψης κειμένου.

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

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

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

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

4.3.6 Παρουσίαση πληροφορίας και προσωποποίηση στο χρήστη

Σκοπός του υποσυστήματος προσωποποίησης που αναπτύχθηκε είναι ο χρήστης να μην αντιλαμβάνεται τις διεργασίες που λαμβάνουν χώρα (transparency) και να απολαμβάνει ποιοτικά και γρήγορα αποτελέσματα βάση των προσωπικών του επιλογών. Για την προσωποποίηση στο χρήστη μπορούν να χρησιμοποιηθούν δύο μέθοδοι:

1. Ο χρήστης να δώσει κάποια πληροφορία στο σύστημα και το σύστημα να ξεκινήσει παρουσιάζοντας εξ΄ αρχής προσωποποιημένα αποτελέσματα και να συγκλίνει γρήγορα στις ανάγκες του χρήστη.
2. Ο χρήστης να μη δώσει καθόλου πληροφορία στο σύστημα και το σύστημα να ξεκινήσει παρουσιάζοντας γενικές πληροφορίες και να συγκλίνει σταδιακά στις στο προφίλ του χρήστη βάσει των επιλογών που κάνει.

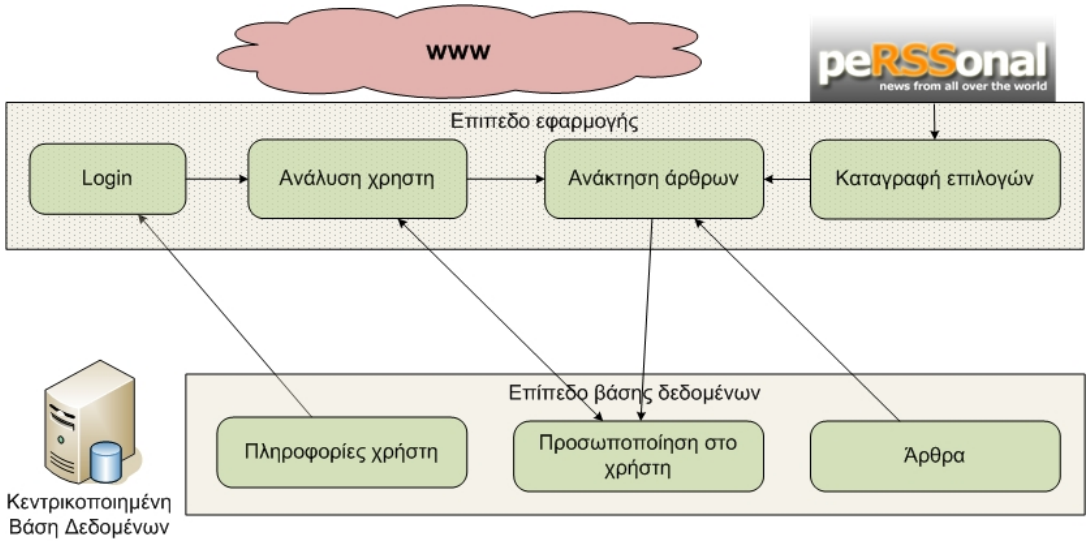
Το υποσύστημα προσωποποίησης στον χρήστη είναι εύκολα προσαρμόσιμο, μπορεί δηλαδή να καταλαβαίνει τις αλλαγές που παρουσιάζονται ανά περιόδους στις θεματικές επιλογές που κάνει ο χρήστης. Για να επιτευχθεί αυτό, απαιτείται μία αλγοριθμική διαδικασία η οποία λαμβάνει υπ΄ όψιν της πολλές παραμέτρους ώστε να φιλτράρει την πληροφορία για τον χρήστη. Η αλγοριθμική αυτή διαδικασία θα περιγραφεί σε επόμενο κεφάλαιο.

Σε κάθε περίπτωση το επιθυμητό επιτυγχάνεται και πρόκειται για τη σύγκλιση των πληροφοριών που παρουσιάζονται στις ανάγκες του χρήστη. Το σχήμα 4.8 απεικονίζει την αρχιτεκτονική του συγκεκριμένου μηχανισμού.

Ο χρήστης, εφόσον έχει εγγραφεί στο σύστημα, συνδέεται με το σύστημα χρησιμοποιώντας το αντίστοιχο module της εφαρμογής ή του web interface. Το σύστημα δέχεται την αίτησή του και εξάγει τις προσωπικές του προτιμήσεις από τη βάση δεδομένων. Στη συνέχεια, και βάσει των προτιμήσεων του χρήστη, ανακτώνται τα άρθρα που τον ενδιαφέρουν και επιτελείται την εξαγωγή προσωποποιημένης περίληψης πάνω σε αυτά. Παράλληλα, ο χρήστης λαμβάνει επιπλέον πληροφορίες για το άρθρο που έχουν να κάνουν με την κατηγοριοποίηση που έχει κάνει το PeRSSonal γι΄ αυτό καθώς και τα σχετικά με αυτό άρθρα. Στη συνέχεια, και λόγω της ανάγκης για συνεχή ανανέωση των επιλογών και προτιμήσεων του χρήστη, το σύστημα καταγράφει τις επισκέψεις των χρηστών

στα άρθρα που του δόθηκαν. Αυτό γίνεται ως εξής: στα URLs που δίνονται ως απάντηση στον χρήστη, τα links που περιέχουν την πηγή του περιληπτημένου άρθρου ανακατευθύνονται μέσω του συστήματος μέσω απλών σελίδων PHP. Η ανακατεύθυνση αυτή (redirect) των χρηστών μπορεί να μας δώσει στοιχεία για το ποια άρθρα από το RSS feed αποφάσισε να επισκεφθεί ο χρήστης και πότε, δίνοντάς μας έτσι την δυνατότητα διαρκούς ανανέωσης του δυναμικού προφίλ του. Αποτέλεσμα αυτού είναι το σύστημα, ακόμη και αν έχει ξεκινήσει από ένα προφίλ εντελώς άσχετο για κάποιον χρήστη (λόγω π. χ. εσκεμμένα λανθασμένης βαθμολόγησης κατηγοριών), να μπορεί να συγκλίνει πολύ γρήγορα στο πραγματικό προφίλ που εκφράζει τον χρήστη.

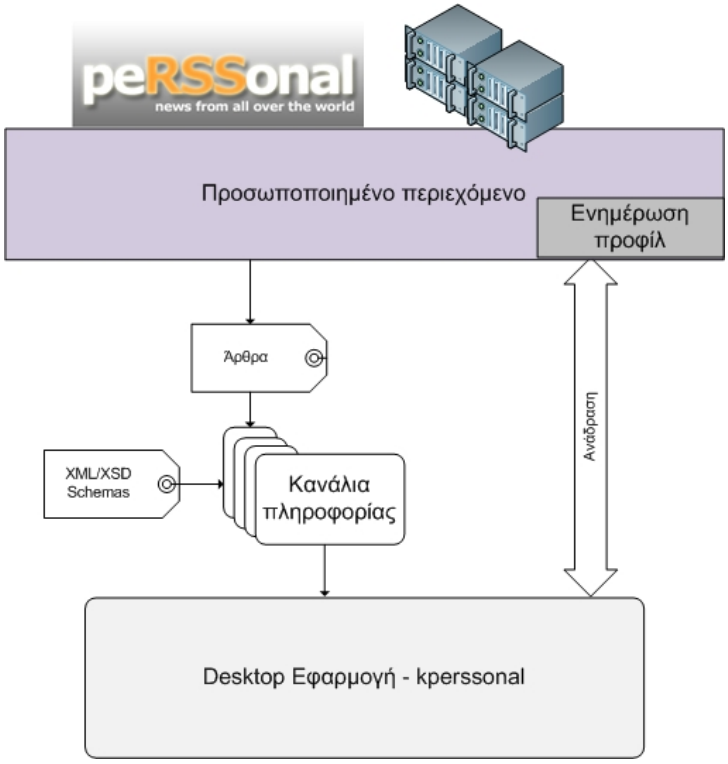
Τα ζητήματα της προσωποποίησης στον χρήστη, αλλά και της διαρκούς ανανέωσης του προφίλ του περιγράφονται σχηματικά από το υποσύστημα PeRSSonal του Σχήματος 4.8.



Σχήμα 4.8: Αρχιτεκτονική της προσωποποίησης στον χρήστη.

4.3.7 *Client side* εφαρμογή

Η desktop εφαρμογή που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας έχει στόχο συμπληρωματικό ως προς το web interface του PeRSSonal που ήδη υπάρχει. Ο σκοπός είναι να καταστεί εφικτό για τον χρήστη οποιουδήποτε λειτουργικού συστήματος να μπορεί να έχει στην επιφάνεια εργασίας του τα προσωποποιημένα άρθρα που επιθυμεί βάσει του προφίλ του που διατηρείται στον server. Με την αρχιτεκτονική που προτείνεται, ο χρήστης μπορεί να έχει μια πληθώρα προσωποποιημένων πληροφοριών για τις οποίες έχει εκφράσει ή εκτιμάται ότι θα εκφράσει ενδιαφέρον, απευθείας στην επιφάνεια εργασίας του. Ανάμεσα στις πληροφορίες που μεταφέρονται είναι: προσωποποιημένες περιλήψεις άρθρων, προτάσεις για σχετικά άρθρα, κατηγοριοποιημένα άρθρα, καθώς και πολλά ακόμα κανάλια τα οποία μπορούν να προστίθεται δυναμικά. Πιο αναλυτικά, η αρχιτεκτονική της εφαρμογής, που φαίνεται στο σχήμα 4.9 δείχνει την επικοινωνία που έχει με το σύστημα PeRSSonal. Η μεταφορά των άρθρων γίνεται σε μορφή XML και μορφοποιείται σύμφωνα με XSD schemas ή CSS τα οποία έρχονται από τον server ή τροποποιεί ο χρήστης με βάσει τις προτιμήσεις του. Το παραπάνω είναι ένα ιδιαίτερο χαρακτηριστικό της εφαρμογής που αναπτύχθηκε καθώς δίνει τη δυνατότητα στο χρήστη να έχει πλήρη έλεγχο στον τρόπο εμφάνισης των καναλιών πληροφορίας που μεταδίδονται. Ένα επίσης σημαντικό στοιχείο που παρουσιάζεται στο σχήμα 4.9 είναι η ανάδραση (feedback) που υπάρχει από την εφαρμογή προς τον server και που δηλώνει τις



Σχήμα 4.9: Αρχιτεκτονική της εφαρμογής *kperssonal*.

επιλογές που κάνει ο χρήστης. Οι επιλογές μεταδίδονται έτσι στον server ο οποίος με τη σειρά του ανανεώνει το προφίλ του χρήστη.

Η εν΄ λόγω εφαρμογή επιφάνειας εργασίας ονομάστηκε kperssonal λόγω του ότι προορίζεται για το παραθυρικό περιβάλλον ανοιχτού κώδικα KDE.

ΚΕΦΑΛΑΙΟ 5

Βάση δεδομένων

> Research is what I'm doing when I don't know what I'm doing.
>
> *Wernher von Braun, German Scientist, 1977*

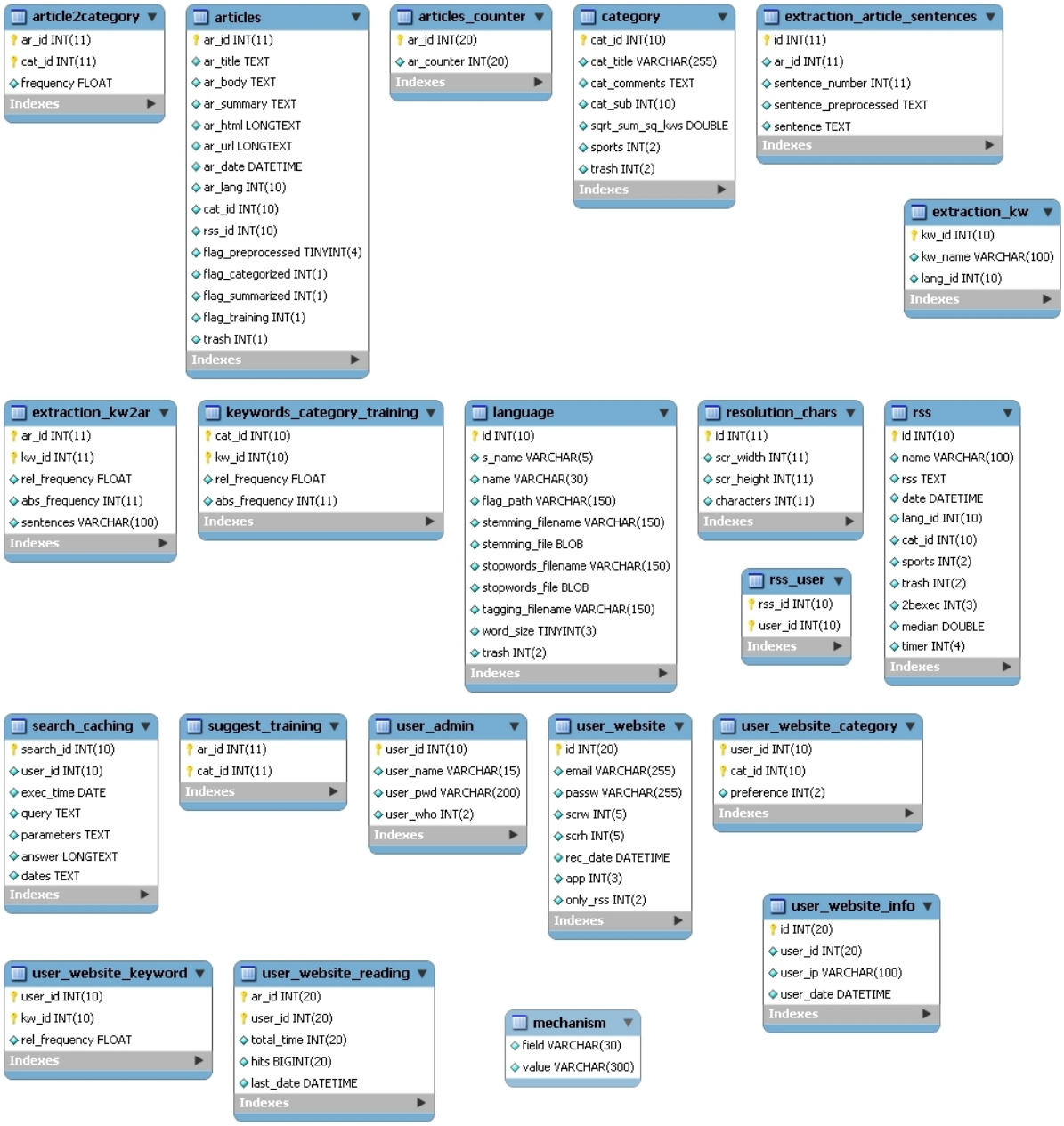
Στο παρόν κεφάλαιο δίνεται η παρουσίαση της βάσης δεδομένων που χρησιμοποιεί το σύστημα. Περιγράφονται αναλυτικά τα πεδία κάθε πίνακα της βάσης και αναφέρεται η χρησιμότητά τους.

5.1 Η βάση δεδομένων του *PeRSSonal*

Το σύστημα PeRSSonal στην παρούσα έκδοσή του χρησιμοποιεί την έκδοση 5.0.60 της MySQL 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 5.1. Η βάση δεδομένων του συστήματος έχει δεχθεί πολλές σχεδιαστικές αλλαγές σε σχέση με αυτή που χρησιμοποιήθηκε στα πλαίσια της διπλωματικής εργασίας. Αυτή είναι ένα στοιχείο θετικό για το PeRSSonal καθώς με αυτό τον τρόπο έχουμε μία πιο συνοπτική αναπαράσταση των δεδομένων που αποθηκεύει το σύστημά μας και επομένως καλύτερη απόδοση όσον αφορά στην εκτέλεση των ερωτημάτων.

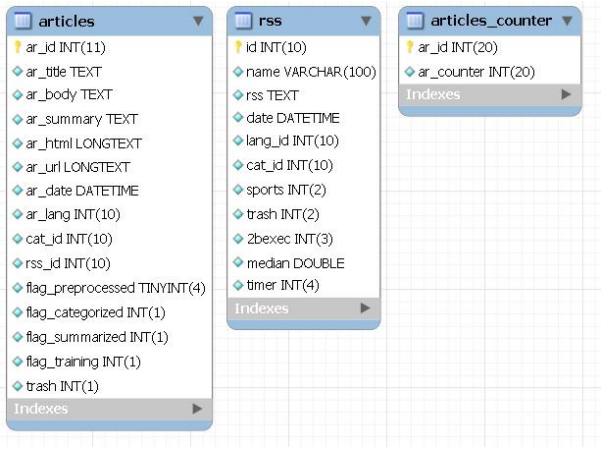
Η παραπάνω εικόνα της βάσης δεδομένων είναι αρκετά γενική και μη κατανοητή με μία πρώτη ανάγνωση. Παρόλα αυτά, οι πίνακές της μπορούν να ομαδοποιηθούν προκειμένου να παρουσιαστεί ο ακριβής τρόπος με τον οποίο γίνεται η αλληλεπίδραση μεταξύ τους. Η βάση δεδομένων λοιπόν μπορεί να ομαδοποιηθεί στις εξής κατηγορίες πινάκων:

- που περιγράφουν τα άρθρα και γενικότερα την είσοδο που δέχεται ο μηχανισμός από το διαδίκτυο 5.2
- που αφορούν στο τμήμα εξαγωγής κωδικολέξεων 5.3
- που αφορούν στο τμήμα κατηγοριοποίησης και εκπαίδευσης του συστήματος 5.4
- που περιγράφουν τους χρήστες 5.5

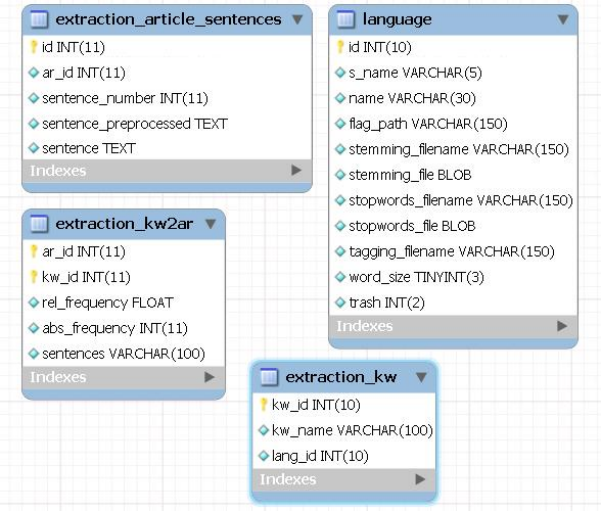


Σχήμα 5.1: Οι πίνακες της βάσης δεδομένων.

- 'γενικοί' πίνακες 5.6



Σχήμα 5.2: Πίνακες που αφορούν τα άρθρα και την είσοδο που δέχεται το σύστημα από το διαδίκτυο.



Σχήμα 5.3: Πίνακες που αφορούν στο υποσύστημα εξαγωγής κωδικολέξεων.

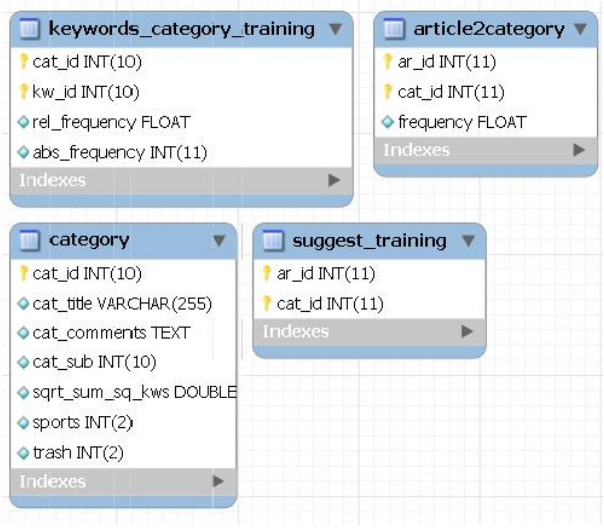
Στο παρόν κεφάλαιο θα επιχειρήσουμε μια πιο αναλυτική περιγραφή των πινάκων απο τους οποίους αποτελείται η βάση δεδομένων του συστήματος PeRSSonal. Η ανάλυση γίνεται με βάσει την κατηγοριοποίησή τους που έγινε πιο πάνω.

5.2 Πίνακες άρθρων και διεπαφής με το διαδίκτυο

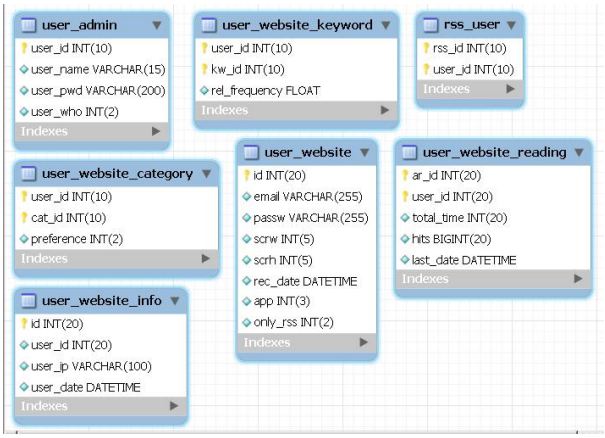
5.2.1 *articles*

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

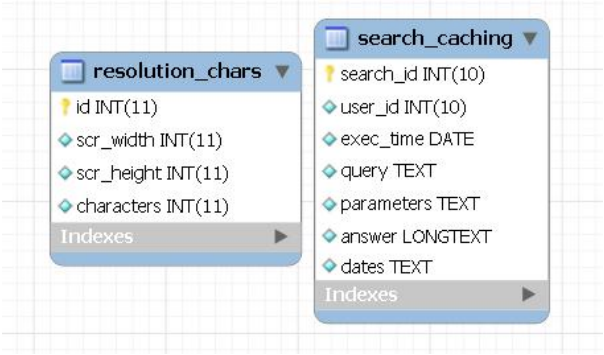
ar_id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα.



Σχήμα 5.4: Πίνακες που αφορούν στην κατηγοριοποίηση και την εκπαίδευση του συστήματος.



Σχήμα 5.5: Πίνακες που αφορούν τους χρήστες του συστήματος.



Σχήμα 5.6: Γενικοί πίνακες της βάσης δεδομένων του συστήματος.

ar_title Ο τίτλος του άρθρου όπως αυτός αναγνωρίστηκε μέσα από τις σελίδες των RSS Feeds και όχι από την ανάλυση των σελίδων.

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

ar_summary Η γενική περίληψη του άρθρου όπως προκύπτει από το μηχανισμό αυτόματης εξαγωγής περίληψης. Στην περίπτωση της δυναμικής δημιουργίας περίληψης το σύστημα τη συνθέτει σε πραγματικό χρόνο και δεν την ανακτά από τη βάση δεδομένων. Επομένως η περίληψη που αποθηκεύεται στο παρόν πεδίο αφορά την λεγόμενη 'γενική' περίληψη

ar_url Το URL το οποίο οδηγεί στη σελίδα του άρθρου όπως αυτό ανακτήθηκε μέσα από το RSS Feed.

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

ar_lang Ακέραιος που αντιστοιχεί στην γλώσσα του συγκεκριμένου άρθρου

cat_id Ακέραιος που αντιστοιχεί στην κατηγορία στην οποία ανήκει το συγκεκριμένο άρθρο βάσει της κατηγοριοποίησης που γίνεται λόγω του RSS Feed στο οποίο ανήκει. Η κατηγορία αυτή, παρότι δεν έχει να κάνει σε τίποτα με την κατηγοριοποίηση που κάνει το PeRSSonal μας δίνει μία σημαντική εκτίμηση για την απόδοση του υποσυστήματος κατηγοριοποίησης.

rss_id Ακέραιος που αντιστοιχεί στο RSS από το οποίο προέρχεται το παρόν άρθρο

flag_preprocessed Πρόκειται για μία μεταβλητή αναγνώρισης προκειμένου να γνωρίζουμε σε ποιό σημείο βρίσκεται η διαδικασία προεπεξεργασίας κειμένου (και εξαγωγής κωδικολέξεων) για το παρόν άρθρο

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

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

flag_training Επίσης μια μεταβλητή αναγνώρισης που έχει να κάνει με το αν το συγκεκριμένο άρθρο προτείνεται για προσθήκη στο training set για την κατηγορία στην οποία ανήκει

trash Μεταβλητή αναγνώρισης που αντιπροσωπεύει αν η συγκεκριμένη εγγραφή είναι διεγραμμένη ή ενεργή

5.2.2 *rss*

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

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

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

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

date Η ημερομηνία κατά την οποία προστέθηκε το rss feed

lang_id Η γλώσσα στην οποία γράφονται τα άρθρα που προέρχονται συνήθως από το συγκεκριμένο RSS Feed (ξένο κλειδί από τον πίνακα language. Θα πρέπει να τονιστεί ότι παρότι το πεδίο αυτό υπάρχει και τον πίνακα των άρθρων, χρησιμοποιείται και εδώ λόγω του γεγονότος ότι υπάρχει η (ακραία μεν, εφικτή δε) πιθανότητα κάποιο ή κάποια άρθρα από δεδομένο RSS Feed να μην έχουν την ίδια γλώσσα με το RSS τους.

cat_id Η κατηγορία στην οποία ανήκει το RSS Feed (ξένο κλειδί από τον πίνακα category. Η πληροφορία αυτή συνήθως υπάρχει στα news portals από τα οποία πηγάζει το feed και αξιοποιείται από το σύστημά μας

sports Αναγνωριστικό που δείχνει αν το συγκεκριμένο RSS Feed ανήκει στην γενικότερη κατηγορία των Sports. Η γενικότερη αυτή διάκριση γίνεται από το PeRSSonal προκειμένου να υπάρχει υποστήριξη για δημιουργία εντελώς ξεχωριστού συστήματος για αποδελτίωση αθλητικών άρθρων και μόνο. Διατηρόντας όμως όλα τα άρθρα στους ίδιους πίνακες της βάσης μας, χρειαζόμαστε αυτό το flag προκειμένου να είναι δυνατή η διάκριση

trash Μεταβλητή αναγνώρισης που αντιπροσωπεύει αν η συγκεκριμένη εγγραφή είναι διεγραμμένη ή ενεργή. Τα RSS Feeds που είναι διεργαμμένα, δεν λαμβάνονται υπ' όψιν κατά την λειτουργία του crawler του PeRSSonal

2*bexec* Η ακέραια μεταβήτή αυτή εκφράζει το πλήθος των εκτελέσεων που πρέπει να γίνουν από τον crawler (ανάμεσα από τα sleeps) προκειμένου να χρησιμοποιηθεί ως είσοδος το συγκεκριμένο RSS Feed

median Τιμή που προκύπτει από τον αλγόριθμο προσαρμογής στη συμπεριφορά του RSS. Ουσιαστικά κρατάει τον ρυθμό ανανέωσης του RSS Feed δεδομένων των αλλαγών που έχουν παρατηρηθεί στο παρελθόν

timer Μετρητής χρόνου προκειμένου να γίνει trigger η χρήση του συγκεκριμένου RSS Feed από τον crawler (όταν ο timer μηδενίζεται

5.2.3 *articles_counter*

Ο πίνακας articles_counter αποθηκεύει τα λεγόμενα hits που γίνονται για τα άρθρα από όλους τους χρήστες - εγγεγραμμένους και μη - του συστήματος.

ar_id Το αναγνωριστικό του άρθρου που αφορά η συγκεκριμένη εγγραφή, ξένο κλειδί από τον πίνακα articles

ar_counter Το συνολικό πλήθος αιτημάτων που έχουν γίνει για ανάγνωση του άρθρου

5.3 Πίνακες υποσυστήματος εξαγωγής κωδικολέξεων

5.3.1 *extraction_kw*

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

kw_id Μοναδικό αναγνωριστικό κλειδί για τις εγγραφές του συγκεκριμένου πίνακα

kw_name Η λέξη κλειδί (stemmed)

lang_id Η γλώσσα στην οποία ανήκει το συγκεκριμένο keyword

5.3.2 *extraction_kw2ar*

Ο πίνακας αυτός κρατάει της πληροφορίες συσχέτισης των άρθρων με τα keywords

ar_id Το αναγνωριστικό του άρθρου που αφορά η τρέχουσα εγγραφή. Ξένο κλειδί από τον πίνακα articles

kw_id Το αναγνωριστικό με το οποίο το άρθρο βρέθηκε να έχει συσχέτιση, ξένο κλειδί από τον πίνακα extraction_kw

rel_frequency Η σχετική συχνότητα εμφάνισης του keyword στο άρθρο

abs_frequency Ακέραιος που δίνει την απόλυτη συχνότητα εμφάνισης του keyword στο άρθρο

sentences Μία λίστα από αριθμούς που αντιπροσωπεύουν τις προτάσεις του άρθρου στις οποίες βρέθηκε το keyword

5.3.3 *extraction_article_sentences*

Ο πίνακας αυτός κρατάει της πληροφορίες εξαγωγής των προτάσεων που έχουν γίνει από το υποσύστημα εξαγωγής κωδικολέξεων.

id Το αναγνωριστικό της τρέχουσας εγγραφής

ar_id Το αναγνωριστικό του άρθρου, ξένο κλειδί από τον πίνακα articles

sentence_number Ο αριθμός που αντιστοιχεί στην συγκεκριμένη πρόταση. Ουσιαστικά δίνει τη σειρά εμφάνισης της πρότασης στο άρθρο

sentence_preprocessed Η πρόταση έχοντας περάσει από το μηχανισμό προεπεξεργασίας κειμένου

sentence Η αρχική πρόταση του κειμένου. Προφανώς, αν για κάθε κείμενο ενώσουμε με τη σειρά τα πεδία sentence αυτού του πίνακα, θα πάρουμε το αρχικό κείμενο του άρθρου

5.3.4 *language*

Ο πίνακας language περιέχει τις απαραίτητες πληροφορίες που σχετίζονται με τη γλώσσα επεξεργασίας του κειμένου και χρησιμοποιείται από το υποσύστημα προεπεξεργασίας κειμένου και εξαγωγής κωδικολέξεων. Ουσιαστικά, ο συγκεκριμένος πίνακας παρέχει τις δυνατότητες πολύγλωσσης υποστήριξης που παρέχει το PeRSSonal για συλλογή άρθρων από πολλούς ειδησεογραφικούς δικτυακούς τόπους ανά τον κόσμο.

id Το αναγνωριστικό κάθε εγγραφής στον πίνακα

s_name Short name, το 'μικρό' όνομα που έχει η γλώσσα, π.χ. για τα ελληνικά: GR

name Το πλήρες όνομα της γλώσσας

flag_path Το path μιας (προαιρετικής) εικόνας για την γλώσσα

stemming_filename Το όνομα του αρχείου που περιέχει τους κανόνες stemming

stemming_file Το ίδιο το αρχείο που περιέχει τους κανόνες για την διαδικασία του stemming. Το αρχείο αυτό αποθηκεύεται σε μορφή blob στον πίνακα

tagging_filename Το όνομα του αρχείου που περιέχει τους κανόνες αναγνώρισης των μερών του λόγου ενός κειμένου για την συγκεκριμένη γλώσσα. Το αρχείο αυτό ανακτάται και χρησιμοποιείται από την βιβλιοθήκη του SVM POS tagger

word_size Το ελάχιστο πλήθος γραμμάτων για τις λέξεις κάτω από το οποίο οι λέξεις του κειμένου της δεδομένης γλώσσας απορρίπτονται

trash Μεταβλητή αναγνώρισης που αντιπροσωπεύει αν η συγκεκριμένη εγγραφή είναι διεγραμμένη ή ενεργή. Οι διεγραμμένες γλώσσες δεν χρησιμοποιούνται από το σύστημα PeRSSonal

5.4 Πίνακες κατηγοριοποίησης και εκπαίδευσης του συστήματος

5.4.1 *category*

Ο πίνακας αυτός περιέχει τα στοιχεία των κατηγοριών που υπάρχουν στο σύστημα. Οι κατηγορίες είναι προκαθορισμένες στο σύστημά μας και η εκπαίδευσή τους γίνεται με βάση το training set.

cat_id Το μοναδικό αναγνωριστικό κλειδί για τις εγγραφές του συγκεκριμένου πίνακα. Ξένο κλειδί από τον πίνακα category

cat_title Το όνομα της συγκεκριμένης κατηγορίας.

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

cat_sub Μας δίνει το αναγνωριστικό της κατηγορίας της οποίας η τρέχουσα είναι υποκατηγορία. Αν η κατηγορία δεν έχει 'γονέα', το πεδίο αυτό παίρνει την τιμή 0

sqrt_sum_sq_kws Το πεδίο αυτό κρατάει την τετραγωνική ρίζα του αθροίσματος των τετραγώνων των keywordss της κατηγορίας. Η τιμή αυτή ανανεώνεται με κάθε αλλαγή στη βάση γνώσης του συστήματος και γίνεται για να μην υπολογίζεται διαρκών κάθε φορά που απαιτείται συσχέτιση άρθρου με κατηγορία

sports Εκφράζει το αν η κατηγορία είναι αθλητικά ή κάποια άλλη υποκατηγορία των sports

trash Μεταβλητή αναγνώρισης που αντιπροσωπεύει αν η συγκεκριμένη εγγραφή είναι διεγραμμένη ή ενεργή. Οι κατηγορίες που είναι διεγραμμένες, δεν λαμβάνονται υπ' όψιν από το PeRSSonal

5.4.2 *keywords_category_training*

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

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

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

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

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

5.4.3 *article2category*

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

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

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

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

5.4.4 *suggest_training*

Ο πίνακας sugest_training αποθηκεύει τα άρθρα που βρέθηκαν από το υποσύστημα κατηγοριοποίησης του PeRSSonal να είναι πολύ αντιπροσωπευτικά με κάποια από τις κατηγορίες. Τα άρθρα κρατούνται εδώ προκειμένου να δοθούν στο administrative web interface απ' όπου ο διαχειριστής του συστήματος μπορεί να αποδεχθεί την ένταξή τους στη βάση γνώσης ή όχι. Σκοπός της παραπάνω διαδικασίας είναι η βάση γνώσης του PeRSSonal να μπορεί να εξελίσσεται (αλλά να μην γιγαντώνεται) περιέχοντας μια αντιπροσωπευτική γνώση για τα άρθρα που κυκλοφορούν την εκάστοτε περίοδο στο διαδίκτυο.

ar_id Ξένο κλειδί στον πίνακα articles που αντιστοιχεί σε ποίο άρθρο αναφέρεται η εγγραφή

cat_id Το αναγνωριστικό της κατηγορίας (ξένο κλειδί στον πίνακα category) στην οποία προτείνεται για εισαγωγή το άρθρο

5.5 Πίνακες προσωποποίησης στους χρήστες

Οι πίνακες που ακολουθούν αφορούν στο κομμάτι προσωποποίησης του συστήματος PeRSSonal. Στην ουσία, περιέχουν χρήσιμες πληροφορίες που αφορούν στις προτιμήσεις των χρηστών και μπορούν να αξιοποιηθούν για την μεταφορά ποιοτικότερης πληροφορίας στο υποσύστημα παρουσίασης του μηχανισμού.

5.5.1 *user_admin*

Ο πίνακας user_admin καταγράφει τους χρήστες του συστήματος που έχουν administrative προνόμια

user_id Το αναγνωριστικό του χρήστη

user_name Το συνθηματικό του χρήστη

user_pwd Ο κωδικός πρόσβασης που έχει ο χρήστης

user_who Ο ρόλος που έχει ο χρήστης

5.5.2 *user_website_keyword*

Στον πίνακα user_website_keywords καταγράφονται οι προτιμήσεις που έχει ο χρήστης για keywords που έχει δείξει αρνητική ή θετική προδιάθεση. Οι εγγραφές αυτές ανανεώνονται διαρκώς με βάσει τις επιλογές που κάνει ο χρήστης όταν είναι συνδεδεμένος στο σύστημα και ουσιαστικά αποτελεί το διάνυσμα προτιμήσεων του χρήστη.

user_id Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τους χρήστες.

kw_id Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τις λέξεις κλειδιά.

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

5.5.3 *user_rss*

Ο πίνακας user_rss αποθηκεύει τα RSS Feeds που επιθυμούν οι χρήστες να εμπεριέχονται στην διαδικασία δεικτοδότησης που επιτελεί το PeRSSonal.

rss_id Αναγνωριστικό γι' αυτό το RSS Feed χρήστη

rss_id Το αναγνωριστικό του χρήστη. Ξένο κλειδί από τον πίνακα user_website_info

5.5.4 *user_website_category*

Ο πίνακας αυτός αποθηκεύει τις πρωταρχικές επιλογές του χρήστη που αφορούν τις κατηγορίες προτίμησης των χρηστών. Παράλληλα, ενημερώνεται ανά τακτά χρονικά διαστήματα λαμβάνοντας υπ' όψιν τα στοιχεία που αποθηκεύονται στον πίνακα user_website_keyword ώστε να αντιστοιχίζει το προφίλ του χρήστη με τις υπάρχουσες κατηγορίες του συστήματος

user_id Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τους χρήστες.

cat_id Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τις κατηγορίες.

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

5.5.5 *user_website*

Πρόκειται για τον πίνακα που αποθηκεύει τις προσωπικές πληροφορίες κάθε χρήστη.

id Το μοναδικό αναγνωριστικό πρωτεύον κλειδί για τις εγγραφές του συγκεκριμένου πίνακα.

emai Η διεύθυνση ηλεκτρονικού ταχυδρομείου του χρήστη που ουσιαστικά αποτελεί και το αναγνωριστικό του

passw Ο κωδικός του χρήστη. Για τη βελτιστοποίηση της ασφάλειας του συστήματος ο κωδικός είναι κωδικοποιημένος με md5 κωδικοποίηση. $SHA1$ κωδικοποίηση είναι επίσης δυνατή.

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

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

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

app Το πλήθος των άρθρων που επιστρέφονται στον χρήστη

only_rss Το πεδίο αυτό εκφράζει αν ο χρήστης θα παρακολουθεί μόνο τα δικά του RSS Feeds είτε και τα δικά του και του συστήματος

5.5.6 *user_website_reading*

Ο πίνακας χρησιμεύει για να καταμετρήσουμε το χρόνο που κάθε χρήστης 'σπαταλά' για να διαβάσει ένα άρθρο καθώς και το ποια άρθρα επιλέγει να δει. Τα παραπάνω χρησιμοποιούνται από τον αλγόριθμο προσωποποίησης προκειμένου να διατηρούν το προφίλ του χρήστη ενημερωμένο

ar_id Το ξένο κλειδί που αντιπροσωπεύει το άρθρο

user_id Το ξένο κλειδί που αντιπροσωπεύει το χρήστη

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

hits Πόσες φορές διάβασε ο χρήστης το άρθρο

last_date Η τελευταία ημερομηνία (timestamp) που ο χρήστης διάβασε το άρθρο

5.5.7 *user_website_info*

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

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

user_id Το μοναδικό ξένο κλειδί που αντιπροσωπεύει τους χρήστες.

user_ip Η IP από την οποία έχει συνδεθεί ο χρήστης.

user_date Η ημερομηνία που συνδέθηκε ο χρήστης.

5.6 Γενικοί πίνακες

Στην παρούσα ενότητα παρουσιάζονται ορισμένοι πίνακες της βάσης δεδομένων του συστήματος που δεν καλύφθηκαν από τις προηγούμενες ενότητες.

5.6.1 *resolution_chars*

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

id Το μοναδικό αναγνωριστικό κλειδί που αφορά το συγκεκριμένο πίνακα.

scr_width Το μήκος της οθόνης σε pixels (screen width).

scr_height Το ύψος της οθόνης σε pixels (screen height).

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

5.6.2 *search_caching*

Ο πίνακας search_caching αποτελεί μία server-side cache που χρησιμοποιείται από το σύστημα PeRSSonal προκειμένου να αυξάνεται η ταχύτητα εκτέλεσης ορισμένων συχνών ερωτημάτων. Η χρήση της είναι πειραματική ακόμα, η ανάλυσή της ξεφεύγει από τα πλαίσια της παρούσας εργασίας και η ερεύνα στο συγκεκριμένο τμήμα του μηχανισμού συνεχίζεται.

search_id Αναγνωριστικό για το συγκεκριμένο ερώτημα

user_id Το αναγνωριστικό του χρήστη που έκανε το ερώτημα. Ξένο κλειδί από τον πίνακα user_website

exec_time Ο χρόνος που εκτελέσθηκε το ερώτημα

query Το ερώτημα με μορφή SQL Query

parameters Οι παράμετροι που χρησιμοποιεί το ερώτημα

answer Η απάντηση που επιστράφηκε από τη βάση δεδομένων για το συγκεκριμένο ερώτημα

dates Οι ημερομηνίες που αφορούν το συγκεκριμένο ερώτημα, ουσιαστικά το χρονικό παράθυρο που καλύπτει

5.6.3 *mechanism*

Ο πίνακας mechanism περιέχει πληροφορίες που χαρακτηρίζουν τον server που τρέχει ο μηχανισμός PeRSSonal. Τα πεδία αφορούν σε μεταβλητές που αξιοποιεί η desktop εφαρμογή προκειμένου να γνωρίζει ποιές φόρμες πρέπει να καλέσει για να αντλήσει την πληροφορία από τον server. Ουσιαστικά πρόκειτε για το description του server.

field Το όνομα της μεταβλητής

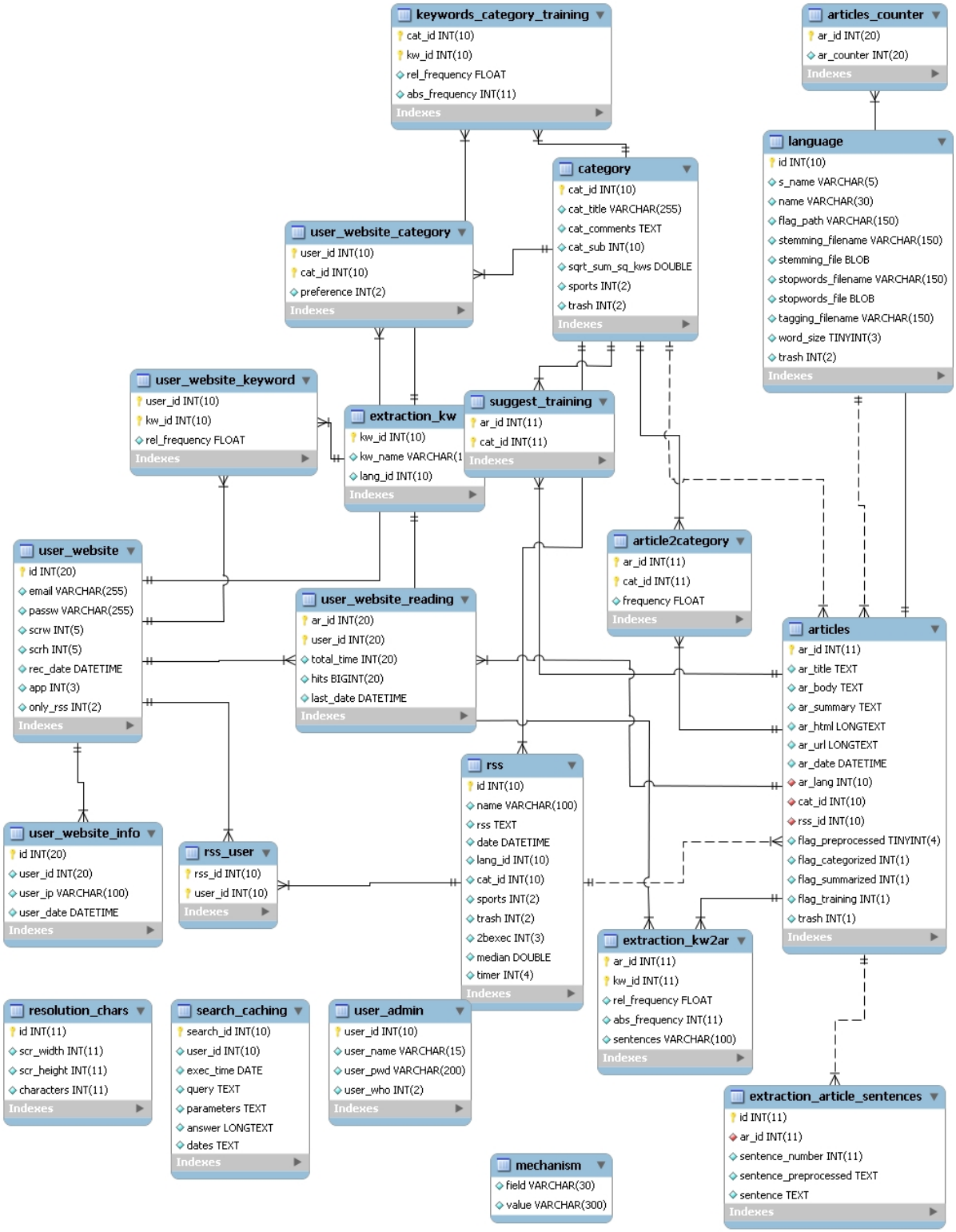
value Η τιμή της μεταβλητής

Ορισμένα παραδείγματα global μεταβλητών που αποθηκεύονται στον πίνακα είναι:

- baseUrl Το βασικό url του server
- recentUrl Το url που καλείται για να ανακτηθούν άρθρα από τον server
- relatedUrl Το url που καλείται προκειμένου να ανακτηθούν πληροφορίες σχετικών άρθρων
- loginUrl Το url που καλείται για να κάνει login ο χρήστης στο σύστημα
- registerUrl Το url που καλείται για να εγγραφεί ο χρήστης στο σύστημα
- profileUpdateUrl Η φόρμα που ανανεώνει το προφίλ του χρήστη με βάσει τις επιλογές που κάνει

5.7 Σχεδιάγραμμα $E-R$

Το σχεδιάγραμμα entities-relationships της βάσης δεδομένων φαίνεται στο σχήμα 5.7



Σχήμα 5.7: Διάγραμμα $E-R$.

ΚΕΦΑΛΑΙΟ 6

Τεχνολογίες υλοποίησης

> Microsoft isn't evil, they just make really crappy operating systems.
>
> *Linus Torvalds, Programmer and kernel hacker*

Οι τεχνολογίες που θα πρέπει να χρησιμοποιηθούν κατά την κατασκευή ενός σύνθετου συστήματος είναι εξαιρετικά σημαντικές προκειμένου να δημιουργηθεί ένα καθολικό σύστημα το οποίο να είναι ευέλικτο, να υποστηρίζει εύκολα αλλαγές και αναβαθμίσεις, να αποτελείται από υποσυστήματα και τέλος να βασίζεται σε ανοιχτά πρότυπα.

Το σύστημα που υλοποιήθηκε είναι εξαιρετικά σύνθετο καθότι έχει βάση το διαδίκτυο αλλά ένα σημαντικό κομμάτι του, ίσως ο πυρήνας, κρύβεται στο μηχανισμό που πραγματοποιεί την κατηγοριοποίηση και την περίληψη κειμένου και γενικότερα τη διαχείριση πληροφορίας. Ο τελευταίος μηχανισμός ουσιαστικά δεν έχει καμία επαφή με το διαδίκτυο και φυσικά δεν είναι και απαραίτητο να έχει. Βέβαια, τα δεδομένα που δέχεται προέρχονται από εξόρυξη πληροφορίας στο διαδίκτυο (HTML σελίδες) ενώ τα δεδομένα που εξάγει χρησιμοποιούνται προκειμένου να τροφοδοτήσουν τους χρήστες είτε του Web interface είτε της desktop εφαρμογής.

6.1 Τεχνολογίες υλοποίησης μηχανισμού

Για την κατασκευή του συστήματος PeRSSonal καθώς και της client side εφαρμογής, χρησιμοποιήθηκε πληθώρα τεχνολογιών σε κάθε επίπεδό του προκειμένου να επιτευχθεί η μέγιστη απόδοση του συστήματος συνολικά με τη χρήση κάθε μίας από αυτές.

6.1.1 Βάση Δεδομένων

MySQL

Η MySQL είναι η δημοφιλέστερη Βάση Δεδομένων ανοιχτού κώδικα που προσφέρεται από το Δίκτυο MySQL. Η αρχιτεκτονική της την κάνει να είναι εξαιρετικά γρήγορη και πολύ εύκολη σε αλλαγές και αναβαθμίσεις. Επιτρέπει επαναχρησιμοποίηση κώδικα όπου αυτό είναι αναγκαίο και

παρέχει ένα μινιμαλιστικό τρόπο δημιουργίας στοιχείων διαχείρισης βάσης δεδομένων τέτοιο ώστε να κάνει τη MySQL ασύγκριτη σε ταχύτητα, σε κατάληψη χώρου, σταθερότητα και ευκολία. Ο μοναδικός στο είδος του διαχωρισμός του κεντρικού πυρήνα του server από το μηχανισμό αποθήκευσης κάνει δυνατή την ύπαρξη αυστηρού ελέγχου σε συναλλαγές και μείωση ταχύτητας ή ύπαρξη θεαματικά μεγάλης ταχύτητας με απευθείας προσπέλαση των δεδομένων, στοιχεία που μπορούν να χρησιμοποιηθούν ανάλογα με τις ανάγκες των χρηστών. Η MySQL περιλαμβάνει αποθήκευση σε μηχανή InnoDB, η οποία υποστηρίζει ασφάλεια στις συναλλαγές και ACID-συμβατή μηχανή αποθήκευσης με commit, rollback, crash recovery και low-level locking δυνατότητες. Η έκδοση της MySQL που βρίσκεται αυτή τη στιγμή σε σταθερή κατάσταση είναι η 5.0.60 και υποστηρίζει πολλά στοιχεία που αφορούν την απόδοση, τη διεθνοποίηση και τη δυνατότητα ένταξης του MySQL server σε άλλα στοιχεία υλικού και λογισμικού. Τα πιο βασικά στοιχεία που χαρακτηρίζουν τη MySQL είναι:

- Υποερωτήματα, που επιτρέπουν στους χρήστες να κάνουν σύνθετα ερωτήματα με μεγάλη ευκολία και αποδοτικά.
- Γρήγορη επικοινωνία μεταξύ server και client μέσα από ένα καινούριο πρωτόκολλο.
- Μικρότερη κατανάλωση πόρων από το server μέσα από βελτιστοποίηση στις βιβλιοθήκες.
- Υποστήριξη Unicode, διεθνείς χαρακτήρες και υποστήριξη αποθήκευσης στην πλειοψηφία των συνόλων χαρακτήρων.
- Υποστήριξη τύπων GIS για ερωτήματα που αφορούν χάρτες και γεωγραφικά δεδομένα.

Τα παραπάνω στοιχεία κάνουν τη MySQL ένα υπερ-πολύτιμο εργαλείο στα χέρια κάποιου χρήστη και τη θέτουν στην 1η θέση για επιλογή ως βάση δεδομένων του συστήματός μας [32]

PostgreSQL

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

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

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

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

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

Τέλος, αξίζει να τονιστεί η άδεια χρήσης κάτω από την οποία βρίσκεται η PostgreSQL σύμφωνα με την οποία μπορεί να χρησιμοποιηθεί, αλλαχθεί και διακινηθεί από τον καθένα χωρίς κανένα κόστος [39]. Το σημαντικό όμως στοιχείο της ταχύτητας εκτέλεσης των ερωτημάτων υστερεί στην PostgreSQL σε σχέση με ίδιες βάσεις για την περίπτωση της MySQLl ή την Oracle γεγονός που την καθιστά μη αποδοτική επιλογή για την περίπτωση μας όπου τα ερωτήματα είναι διαρκή και περίπλοκα.

Oracle

Η βάση δεδομένων Oracle είναι επίσης μια σχεσιακή βάση δεδομένων όπως και οι προηγούμενες επιλογές. Η τρέχουσα έκδοση της Oracle είναι η 10g. Οι βασικές διαφορές με τις προηγούμενα συστήματα διαχείρισης βάσεων δεδομένων έγκειται στα εξής χαρακτηριστικά:

- Είναι propriatery, χρειάζεται άδεια λειτουργίας προκειμένου να χρησιμοποιηθεί
- Είναι πολύ ακριβή και χρειάζεται αρκετές γνώσεις προκειμένου να χρησιμοποιηθεί
- Χρειάζεται πολλούς πόρους του συστήματος (είναι σχετικά 'βαριά')
- Μπορεί να διαχειρίζεται πολύ περισσότερα transactions σε σχέση με τα άλλα ΣΔΒΔ.
- Η διαχείριση των αντιγράφων ασφαλείας και γενικά εργασιών συντήρησης της βάσης δεδομένων είναι γενικά απλή διαδικασία.

Από τα παραπάνω στοιχεία το γεγονός ότι η Oracle παρέχει εξαιρετικές δυνατότητες για διαχείριση transactions κάνει την επιλογή αυτή ιδιαίτερα ελκυστική. Παρόλα αυτά, το κόστος είναι απαγορευτικό για μη επαγγελματικούς σκοπούς.

Επιλέγοντας τη Βάση Δεδομένων

Σύμφωνα με τα παραπάνω αλλά και λαμβάνοντας υπόψη μας τους σκοπούς που έχει το σύστημά μας καταλήξαμε στην επιλογή της MySQL σαν τη βάση δεδομένων που θα χρησιμοποιηθεί στο σύστημα. Συγκρίνοντας τα τρία DBMS μπορούμε να καταλήξουμε στο ότι διαθέτουν πολλά κοινά στοιχεία, ωστόσο η MySQL φαίνεται αρκετά γρήγορη στην εξυπηρέτηση των queries και transactions καθώς και πιο διαδεδομένη, λόγοι οι οποίοι την κάνουν πιο ισχυρή. Επιπρόσθετα τα στοιχεία διεθνοποίησης που διαθέτει φαίνονται πολύ χρήσιμα για το σύστημα μας το οποίο διαθέτει πολύγλωσση υποστήριξη. Τέλος θα πρέπει να λάβουμε υπόψη μας το γεγονός πως δημιουργούμε ένα σύστημα πολυεπίπεδο με τη βάση δεδομένων να είναι ο ουσιαστικός σύνδεσμος μεταξύ των περισσοτέρων κομματιών και συνεπώς μία βάση δεδομένων με μεγάλη σταθερότητα και αξιοπιστία θα προσέδιδε κύρος και ασφάλεια στο συνολικό σύστημα.

Καταλήγουμε λοιπόν στη χρήση Mysql Server, έκδοση 5.0.60 [32].

6.1.2 Ορθογραφικός έλεγχος

Ο ορθογραφικός έλεγχος του κειμένου είναι ένα σημαντικό χαρακτηριστικό του υποσυστήματος προεπεξεργασίας κειμένου που αντιλαμβάνεται τυχόν λανθασμένες λέξεις και τις διορθώνει. Με αυτό τον τρόπο αποφεύγονται λανθασμένα keywords που μπορούν να προκύψουν κατά τη διαδικασία.

GNU Aspell

Το GNU Aspell [19] είναι ένας ανοιχτού κώδικα ορθογράφος σχεδιασμένος για την αντικατάσταση του παλαιού Ispell. Μπορεί είτε να χρησιμοποιηθεί ως βιβλιοθήκη είτε ως ανεξάρτητος ορθογράφος (πρόγραμμα). Το βασικό του χαρακτηριστικό είναι ότι είναι εξαιρετικά αποτελεσματικός στο να προτείνει πιθανές αντικαταστάσεις για μία λανθασμένη λέξη τουλάχιστον για την Αγγλική γλώσσα. Μπορεί επίσης να διατρέξει κείμενα με UTF-8 κωδικοποίηση και λαμβάνοντας υπ' όψιν του τις τρέχουσες ρυθμίσεις εντοπιότητας του χρήστη (locale). Ένα ακόμη σημαντικό χαρακτηριστικό του Aspell είναι ότι μπορεί να χρησιμοποιεί πολλαπλά λεξικά ταυτόχρονα αλλά και να διαχειρίζεται τα ατομικά λεξικά που έχει ο κάθε χρήστης. Συνοπτικά θα λέγαμε ότι ο Aspell είναι ένας γρήγορος ανοιχτού κώδικα και αρκετά αποτελεσματικός ορθογράφος, στοιχεία που τον καθιστούν ιδανική επιλογή για την χρήση που επιθυμούμε.

6.1.3 Μηχανισμός περίληψης και κατηγοριοποίησης

Ο μηχανισμός περίληψης είναι ένα υποσύστημα το οποίο αναλαμβάνει μια πολύ μεγάλη και επίπονη διαδικασία. Προκειμένου να καταλάβουμε τι τεχνολογία πρέπει να χρησιμοποιηθεί θα συνοψίσουμε της εργασίες του μηχανισμού στην παράγραφο που ακολουθεί.

Ο μηχανισμός περίληψης δέχεται ως είσοδο αρχεία, ή καλύτερα, δομημένη μορφή XML με στοιχεία για το κείμενο και προχωράει σε μια διαδικασία εξαγωγής λέξεων - κλειδιά για αυτό. Ακολουθεί η διαδικασία αντιστοίχησης λέξεων σε προτάσεις και ακολούθως η βαθμολόγηση των προτάσεων για την εξαγωγή των σημαντικότερων αυτών ώστε να προκύψει η περίληψη του κειμένου. Μια αντίστοιχη διαδικασία ακολουθείται και στην περίπτωση κατηγοριοποίησης ενός κειμένου. Ο μηχανισμός συνεχίζει με ένα επίπεδο προσωποποίησης όπου παράγεται μια προσωποποιημένη περίληψη του κειμένου και αποστέλλεται στο χρήστη σε κατάλληλη μορφή (π. χ. XML channels). Η επικοινωνία με τη βάση δεδομένων είναι διαρκής σε κάθε φάση του μηχανισμού (αποθήκευση/ανάκτηση keywords και συχνοτήτων, κειμένων, κατηγοριών, στοιχείων προσωποποίησης κ.λπ.).

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

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. Επιπλέον η διάθεση του GNU C/C++ Compiler με την GPL άδεια χρήσης κάνει την ανάπτυξη ενός συστήματος με χρήση της γλώσσας C ιδιαίτερα ελκυστική, αφού σε συνδυασμό με το λειτουργικό σύστημα Linux και κάποιο από τα πολλά υπάρχοντα ολοκληρωμένα περιβάλλοντα (IDEs) προγραμματισμού για αυτό, είναι μιας πρώτης τάξεως λύση για το ζήτημα.

C++

Πρόκειται για την αντικειμενοστραφή εξέλιξη της γλώσσας C. 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 και να χρησιμοποιηθούν 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.

Η C++ με τις πάμπολλες βιβλιοθήκες που διαθέτει, είτε ανήκουν στην STL, είναι είναι ξεχωριστές (π. χ. boost, mysql++, cgicc, κ. α.), και με τα πλεονεκτήματα ως γλώσσα προγραμματισμού που κληρονομεί από την C, αποτελεί την ιδανική τεχνολογία υλοποίησης για ένα αποτελεσματικό, γρήγορο, real time μηχανισμό, σας αυτό που αναπτύχθηκε στα πλαίσια αυτής της διπλωματικής.

Java

Η Java 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 θεωρείται 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. Μία αναλυτική αναφορά στο χρονικό της εξέλιξης της γλώσσας μπορείτε να βρεθεί στο [23].

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.

Η χρήση της perl ταιριάζει με τα προβλήματα εύρεσης προτύπου και κανονικών εκφράσεων που αντιμετωπίζονται από τον μηχανισμό προεπεξεργασίας. Τα εργαλεία που χρησιμοποιεί ή perl κάνουν χρήση βέλτιστων αλγορίθμων και η γλώσσα μπορεί να χρησιμοποιηθεί κατά κόρων για εργασίες που έχουν να κάνουν με διαχείριση συμβολοσειρών (string manipulation). Από την άλλη όμως, η χρήση της perl για ανάπτυξη προγραμμάτων οδηγεί σε κώδικα δύσκολα κατανοητό και συντηρήσιμο, ενώ ο κώδικας απαιτεί όχι και τόσο συνηθισμένη σύνταξη.

6.2 Μηχανισμός συλλογής ειδήσεων

Για το μηχανισμό συλλογής ειδήσεων χρησιμοποιήθηκε επίσης η γλώσσα προγραμματισμού C++. Η επιλογή αυτής της γλώσσας για το συγκεκριμένο μηχανισμό έγινε γιατί προσφέρει ταχύτατη επικοινωνία τόσο με την διεπαφή του διαδικτύου όσο και με την βάση δεδομένων MySQL που χρησιμοποιείται. Παράλληλα, προσφέρει μεγάλη ευελιξία στη διαχείριση πόρων του διαδικτύου αλλά και γιατί διαθέτει APIs ανάλυσης βάση του DOM μοντέλου των σελίδων HTML χρησιμοποιώντας την πληθώρα βιβλιοθηκών που υπάρχουν ευρέως διαθέσιμες.

6.3 Μηχανισμός εξαγωγής χρήσιμου κειμένου

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

6.4 Μηχανισμός παρουσίασης πληροφορίας και προσωποποίησης

Ο μηχανισμός παρουσίασης πληροφορίας και προσωποποίησης έγινε με χρήση α) PHP σελίδων για την καταγραφή των επισκέψεων σε άρθρα από τους χρήστες του συστήματος (τα links στα οποία δρομολογούνται μέσω του συστήματος) και β) κώδικα σε C++.

Σε αυτό το κομμάτι έχει προκύψει αρκετές φορές το ζήτημα επιλογής τεχνολογίας καθότι μία πιο ολοκληρωμένη πρόταση θα ήταν υλοποίηση όλων των μηχανισμών σε Java και επιλογή JSP με Enterprise Java beans για το δικτυακό κομμάτι. Ωστόσο, η απόκριση της γλώσσας προγραμματισμού Java στις διαδικασίες πυρήνα του συστήματος μας είναι πολύ πιο αργή από τη C++. Αυτό συμβαίνει κυρίως, όπως έχει ήδη αναφερθεί, στην καλύτερη αντιμετώπιση που έχει η C++ όταν εκτελεί διαδικασίες χαμηλού επιπέδου.

6.4.1 *XML*

Η XML είναι μια γλώσσα μορφοποίησης (markup language) για κείμενα τα οποία περιέχουν δομημένη πληροφορία. Η δομημένη πληροφορία περιέχει τόσο περιεχόμενο (π. χ. λέξεις, εικόνες, κ.λπ.), όσο και το τι ρόλο παίζει αυτό το περιεχόμενο. Μια γλώσσα μορφοποίησης είναι ένας μηχανισμός για αναγνώριση δομής σε ένα κείμενο. Η XML προδιαγραφή ορίζει έναν στάνταρ τρόπο για προσθήκη μορφοποίησης σε έγγραφα.

Η XML δεν είναι HTML. Στην HTML, τόσο οι ετικέτες tags όσο και τα στοιχεία τους είναι προκαθορισμένα, κάτι τέτοιο δεν ισχύει για την XML όπου οι ετικέτες αλλά και τα στοιχεία τους ορίζονται από τον χρήστη. Η XML δεν καθορίζει ούτε εννοιολογικά δεδομένα ούτε και ένα σύνολο ετικετών. Για την ακρίβεια, είναι μια meta-γλώσσα που χρησιμοποιείται για την περιγραφή γλωσσών μορφοποίησης. Με άλλα λόγια, η XML παρέχει τη δυνατότητα να καθορίζονται tags και πληροφορίες δομής μεταξύ αυτών. Εφόσον δεν υπάρχει κάποιο προκαθορισμένο σύνολο από ετικέτες, δεν υπάρχει και προκαθορισμένη σημασιολογία ετικετών. Όλες οι εννοιολογικές πληροφορίες ενός XML εγγράφου, θα παρέχονται είτε από την εφαρμογή που το επεξεργάζεται, είτε από ξεχωριστά αρχεία που καθορίζουν το στυλ (stylesheets).

Η γλώσσα μορφοποίησης XML περιγράφει μια κατηγορία πληροφοριών (data objects) που καλούνται XML έγγραφα (documents) καθώς επίσης περιγράφει τμηματικά τη συμπεριφορά των προγραμμάτων που τα επεξεργάζονται. Τα XML έγγραφα αποτελούνται από μονάδες αποθήκευσης που καλούνται entities (οντότητες), οι οποίες περιέχουν πληροφορίες αναλυμένες ή μη. Οι αναλυμένες πληροφορίες αποτελούνται από χαρακτήρες (characters) οι οποίοι συνθέτουν character data και άλλοι οι οποίοι συνθέτουν markup. Η μορφή markup κωδικοποιεί την περιγραφή της τελικής αποθήκευσης του εγγράφου καθώς και τη λογική δομή.

Ένα λογισμικό μοντέλο που καλείται επεξεργαστής XML χρησιμοποιείται να διαβάζει XML έγγραφα και παρέχει πρόσβαση στο περιεχόμενο και τη δομή τους. Υποτίθεται ότι ο επεξεργαστής XML λειτουργεί εκ μέρους ενός άλλου μοντέλου που καλείται application (εφαρμογή). Αυτή η προδιαγραφή περιγράφει την απαιτούμενη συμπεριφορά του επεξεργαστή και συγκεκριμένα πως θα πρέπει να διαβάζει τα XML δεδομένα και ποιες πληροφορίες πρέπει να παρέχει στην εφαρμογή.

Οι προσχεδιασμένοι στόχοι της XML σύμφωνα με το W3C [43] είναι:

1. Η XML πρέπει να είναι εύχρηστη στο Internet.
2. Η XML πρέπει να υποστηρίζει μεγάλη ποικιλία από εφαρμογές.
3. Η XML πρέπει να είναι συμβατή με την SGML.
4. Θα είναι εύκολο να γράφονται προγράμματα που επεξεργάζονται XML έγγραφα.
5. Ο αριθμός των προαιρετικών χαρακτηριστικών στην XML θα είναι όσο το δυνατόν πιο μικρός, ιδανικό επίπεδο το μηδέν.
6. Τα XML έγγραφα θα πρέπει να είναι ευανάγνωστα.
7. Ο σχεδιασμός XML θα πρέπει να προετοιμάζεται γρήγορα.
8. Ο σχεδιασμός XML θα πρέπει να είναι τυπικός και περιεκτικός.
9. Τα XML έγγραφα θα πρέπει να δημιουργούνται εύκολα.
10. Η περιεκτικότητα στον XML συμβολισμό είναι μικρής σημασίας.

6.4.2 *RSS*

Το πρότυπο RSS [41], όπως έχει ήδη αναφερθεί, είναι μια οικογένεια από σχήματα τροφοδότησης περιεχομένου στους χρήστες που χρησιμοποιούνται για να δημοσιεύουν συχνά ενημερωμένο περιεχόμενο όπως οι καταχωρήσεις blog, οι τίτλοι ειδήσεων ή τα podcasts. Ένα έγγραφο RSS, που καλείται επίσης και 'feed', περιέχει είτε μια περίληψη του περιεχομένου του σχετικού ιστοχώρου, είτε το πλήρες κείμενο. Το RSS καθιστά δυνατό για τους χρήστες να παρακολουθούν τους αγαπημένους ιστοχώρους τους με έναν αυτοματοποιημένο τρόπο που δεν απαιτεί την πλοήγηση σε

αυτούς. Το περιεχόμενο RSS μπορεί να διαβαστεί χρησιμοποιώντας λογισμικό γνωστό και ως 'feed reader' ή 'aggregator'. Από τη στιγμή που ο χρήστης γίνεται συνδρομητής σε ένα feed, ο aggregator αναλαμβάνει να λαμβάνει τα νέα που προέρχονται από το feed ανά τακτά χρονικά διαστήματα. Υπάρχουν διάφορα πρότυπα RSS, RSS 2.0, RSS 1.0, RSS 0.91, όλα όμως χρησιμοποιούν μια XML δομή για τη σύνταξη των δεδομένων.

Το βασικότερο πλεονεκτήματα του προτύπου RSS (όποια έκδοση και αν επιλεγεί), είναι ότι εξοικονομεί χρόνο και εύρος ζώνης στους τελικούς χρήστες, ιδιαίτερα σε αυτούς που χρησιμοποιούν συσκευές μικρού μεγέθους. Η χρήση μάλιστα του προτύπου XML από την τεχνολογία RSS προσδίδει περαιτέρω πλεονεκτήματα στο πρότυπο RSS. Σαν μειονέκτημα του προτύπου RSS θα μπορούσαμε να πούμε ότι είναι η ύπαρξη πολλών 'εκδόσεων' που δημιουργήθηκαν ανά τον χρόνο και ανάλογα με τις απαιτήσεις των εφαρμογών. Η κατάσταση αυτή δημιουργεί ορισμένες ασυμβατότητες στις διάφορες εφαρμογές aggregator, ή τους web browsers που έχουν δυνατότητα απεικόνισης των RSS feeds. Το μειονέκτημα όμως μπορεί να αντιμετωπιστεί χρησιμοποιώντας μόνο τα πρότυπα όπως περιγράφονται από το W3C και εφαρμογές που τα ακολουθούνε πιστά.

6.4.3 *CGI*

Τα CGI (Common Gateway Interface) scripts επιτρέπουν να τρέξει ένα εκτελέσιμο πρόγραμμα στον HTTP server. Οι περιπτώσεις στις οποίες χρησιμοποιούνται είναι όταν θέλουμε να επεξεργαστούμε δεδομένα που έρχονται ως αποτελέσματα συμπλήρωσης μιας φόρμας, για τη δημιουργία δυναμικών HTML εγγραφών, για μετρητές προσπελάσεων (counters). Τα CGI scripts μπορούν να γραφούν σε οποιαδήποτε γλώσσα μπορεί να παράγει εκτελέσιμο αρχείο στη μηχανή που τρέχει ο server. Ανάλογα με την πλατφόρμα υλοποίησης , επιλέγεται και η γλώσσα. Έτσι, σε Unix χρησιμοποιούνται PERL, C-shell, C/C++ ενώ σε Windows 95/NT, PERL, Visual Basic, Visual C++.

Ένα CGI script είναι ένα πρόγραμμα το οποίο στο standard output παράγει (συνήθως) HTML ή μορφής XML κώδικα. Σε κάποιες περιπτώσεις παράγει στο standard output κώδικα GIF αρχείου (χρησιμοποιείται στην περίπτωση γραφικών counters σελίδων). Γιάυτό το λόγο αρχικά πρέπει πάντα να τίθεται μια γραμμή προσδιορισμού του περιεχομένου που θα ακολουθήσει. Στο υπόλοιπο μέρος του προγράμματος υπάρχουν εντολές εκτύπωσης του περιεχομένου στο standard output το οποίο και διαβάζει ο Browser ή οποιαδήποτε εφαρμογή επικοινωνεί με τον Web Server (π. χ. RSS Reader). Τα CGI προγράμματα συνήθως αποθηκεύονται σε ένα συγκεκριμένο χώρο. Το directory το οποίο τα περιέχει συνήθως ονομάζεται 'cgi-bin'. Τα αρχεία που αποθηκεύονται εκεί είναι εκτελέσιμα αρχεία που μπορεί να τα τρέξει ένα σύστημα.

Τα CGI scripts αποτελούν μια εύκολη λύση για να προσαρμόσουμε τον υπάρχοντα κώδικα παραδοσιακών γλωσσών προγραμματισμού ώστε να δέχεται και να στέλνει περιεχόμενο μέσω ενός Web Server.

6.5 *Client Side* εφαρμογή

Η εφαρμογή χρήστη του PeRSSonal, όπως και κάθε εφαρμογή που προορίζεται για χρήση στην επιφάνεια εργασίας του χρήστη, απαιτεί μία προσεγμένη αλλά και αποδοτική σχεδιαστική προσέγγιση ώστε να είναι φιλική προς τον χρήστη, εύκολα και πλήρως παραμετροποιήσιμη, αποδοτική στην επικοινωνία με τον server καθώς και αυτή η επικοινωνία να είναι όσο πιο διαφανής γίνεται (transparency). Προς αυτή την κατεύθυνση χρειαζόμαστε ένα toolkit σχεδίασης της εφαρμογής που θα καλύπτει τα παραπάνω χαρακτηριστικά, θα είναι σύγχρονο και επιπλέον θα είναι cross-platform.

6.5.1 *Qt Toolkit*

Το περιβάλλον προγραμματισμού (toolkit) της Trolltech [42], Qt είναι ένα διαλειτουργικό περιβάλλον ανάπτυξης που επικεντρώνεται στα Graphical User Interface προγράμματα. Παρότι χρησιμοποιείται για τον προγραμματισμό και μη-GUI προγραμμάτων, το Qt toolkit είναι ευρύτερα γνωστό για την χρήση του στο παραθυρικό σύστημα KDE, στον Web browser Opera και σε πολλές ακόμη εφαρμογές.

Η Qt επεκτείνει την γλώσσα προγραμματισμού C++ με αρκετές προσθήκες που υλοποιούνται από έναν επιπλέον προ-επεξεργαστή που παράγει κλασικό C++ κώδικα πριν το compilation. Η Qt έχει επίσης και πολλά επιπλέον bindings για διάφορες γλώσσες: Ada (QtAda) C# (Qyoto/-Kimono) Java (Qt Jambi), Pascal, Perl, PHP (PHP-Qt), Ruby (RubyQt), Python (PyQt). Στα βασικό πλεονεκτήματα της Qt είναι η ταχύτητα εκτέλεσης (ως C++ κώδικας) καθώς και ότι τρέχει σε όλες τις βασικές υπολογιστικές πλατφόρμες. Η τρέχουσα έκδοση της Qt στην οποία στηρίχθηκε και η ανάπτυξη της client side εφαρμογής είναι η 4.4.2.

6.6 Διασύνδεση μηχανισμών

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

ΚΕΦΑΛΑΙΟ 7

Ανάπτυξη του συστήματος

> Experience is simply the name we give our mistakes.
>
> *Oscar Wilde, Irish Dramatist, 1900*

Στο παρόν κεφάλαιο δίνεται μια αναλυτική περιγραφή των υποσυστημάτων που αναπτύχθηκαν για τη λειτουργία του συστήματος PeRSSonal καθώς και της εφαρμογής για την επιφάνεια εργασίας του χρήστη. Ιδιαίτερη έμφαση δίνεται στη διαδικασία που ακολουθείται από την είσοδο του κειμένου, οι παράμετροι που χρησιμοποιούνται, και όλα τα ενδιάμεσα στάδια μέχρι να φτάσουμε σε μία αποδεδειγμένα καλή περίληψη του κειμένου. Παρουσιάζονται αλγοριθμικά θέματα καθώς και οι διαδικασίες (ροή) που ακολουθείται από το σύστημα. Δεδομένου ότι οι γραμμές κώδικα που γράφηκαν συνολικά για την κατασκευή του συστήματος είναι υπερβολικά πολλές για να παρουσιαστούν, θα εστιάσουμε την προσοχή μας στα πιο σημαντικά σημεία καθώς και στις τεχνικές με τις οποίες υλοποιήθηκε κάθε αλγόριθμος σε κάθε σημείο.

7.1 Αλγοριθμικά θέματα

Ο αλγόριθμος 7.1.1 που ακολουθεί, δίνει μια γενική εποπτεία των διαδικασιών που εκτελούνται κατά την λειτουργία του μηχανισμού. Παρότι αρκετά γενικός, τονίζει τα σημαντικά στοιχεία τα οποία και θα αναλυθούν στη συνέχεια.

Η αλγοριθμική διαδικασία 7.1.1 παρουσιάζει τις βασικές διεργασίες που φέρνει σε πέρας ο μηχανισμός. Μέσω αυτής επιτυγχάνουμε τους τρεις βασικούς στόχους: κατηγοριοποίηση, περίληψη και αλληλεπίδραση μεταξύ των μηχανισμών. Ξεκινάμε κάνοντας την προεπεξεργασία των άρθρων και την εξαγωγή των keywords που περιέχει. Το επόμενο βήμα αφορά στην κατηγοριοποίηση του άρθρου βάσει του συνόλου εκμάθησης που προϋπάρχει στη βάση δεδομένων. Για την διαδικασία αυτή χρησιμοποιούνται οι αντιπροσωπευτικές κωδικολέξεις (οι οποίες είναι stemmed από την διαδικασία προεπεξεργασίας) μαζί με την συχνότητα εμφάνισής τους από το συγκεκριμένο άρθρο καθώς και οι αντίστοιχες λίστες για όλες τις κατηγορίες που υπάρχουν στη βάση δεδομένων. Οι αυτές λίστες αποτελούνται από τις ίδιες κωδικολέξεις ακολουθούμενες από την συχνότητά τους στην εκάστοτε κατηγορία. Εξετάζουμε την ομοιότητα συνημιτόνου αυτών των λιστών με σκοπό να καθορίσουμε

Αλγόριθμος 7.1.1 perssonal()

```
Crawler = fork();
while 1 do
  String Text = fetch_next_article();
  String PreprocessedText = Preptocess(Text);
  List kwfr(text) = create_keyword_frequency_list(text);
  List kwfr_cat(category) = create_keyword_frequency_list(text, category);
  Categorize (kwfr(text), kwfr_cat(category));
  if !Categorize then
    String Generic_summary = Summarize(text);
    List kwfr(Generic_summary) = create_keyword_frequency_list(Generic_summary);
    List kwfr_cat(category) = create_keyword_frequency_list(Generic_summary, category);
    Categorize (kwfr(Generic_summary), kwfr_cat(category));
    if !Categorize then
      StoreInitialCategorizationResults();
    end if
  else
    String Categorization_Enhanced_Summary = Summarize(text, categorization_results);
  end if
  StoreResults();
end while
```

την κατηγορία του κειμένου. Παράδειγμα του αποτελέσματος που προκύπτει φαίνεται στον Πίνακα 7.1.

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

Πίνακας 7.1: Ομοιότητα μεταξύ κειμένου και κατηγορίας

Εάν το κείμενο δεν μπορεί να ταξινομηθεί σε κάποια από τις υπάρχουσες κατηγορίες, χρησιμοποιείται για την κατηγοριοποίηση η γενική περίληψη του άρθρου, την οποία παράγει ο μηχανισμός βασιζόμενος στις πληροφορίες για την κατανομή των keywords στο άρθρο. Η περίληψη αυτή είναι 'γενικής' φύσεως μιας και δεν αντιπροσωπεύει συγκεκριμένες προτιμήσεις χρήστη. Το στοιχείο αυτό αποτελεί μία καινοτομία του μηχανισμού κατηγοριοποίησης του συστήματος PeRSSonal: κατά γενική ομολογία, η περίληψη ενός άρθρου αποτελεί μια σύντομη εκδοχή του, τόσο από άποψη μεγέθους, όσο και από άποψη νοήματος, συνεπώς απαλλαγμένη από θόρυβο. Η χρήση επομένως της περίληψης ενός κειμένου για μία διαδικασία ανάκτησης πληροφορίας, όπως εν' προκειμένω η κατηγοριοποίηση έχει πολλαπλά οφέλη μεταξύ των οποίων ταχύτερη έξοδο και με μεγαλύτερη νοηματική συνοχή. Το τελευταίο γεγονός φαίνεται (με βάσει και τα πειράματα αξιολόγησης που έγιναν

για τη συγκεκριμένη διαδικασία) ότι μας δίνει πιο σαφή αποτελέσματα για να αποφασίσουμε πως θα πρέπει να κατηγοριοποιηθεί ένα άρθρο.

Στο σημείο αυτό πρέπει να αναφέρουμε τις συνθήκες που πρέπει να ισχύσουν προκειμένου ένα άρθρο να μπορεί να κατηγοριοποιηθεί επιτυχώς από το υποσύστημα κατηγοριοποίησης. Θα πρέπει:

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

Άν λοιπόν καλύπτονται οι παραπάνω συνθήκες (το κείμενο μπορεί να ταξινομηθεί), και έχουμε να κάνουμε με πλήρες κείμενο (δεν έχει γίνει ακόμη περίληψή του) προχωρούμε στη διαδικασία της περίληψης αξιοποιώντας τις πληροφορίες που προέκυψαν από την κατηγοριοποίηση. Αυτή αποτελεί επίσης μία καινοτομία του συστήματος PeRSSonal: η περίληψη κειμένου αξιοποιεί τις πληροφορίες επιτυχούς κατηγοριοποίησης για την βελτίωση των αποτελεσμάτων της. Σε επόμενη ενότητα θα περιγράψουμε αναλυτικότερα αυτή την περίπτωση.

Στην τελευταία περίπτωση, όπου η ομοιότητα συνημιτόνου μεταξύ του κειμένου και της αντιπροσωπευτικής του κατηγορίας είναι πολύ μεγάλη, και παρόμοια η διαφορά των ομοιοτήτων συνημιτόνου μεταξύ της ισχυρότερης και των υπολοίπων κατηγοριών είναι επίσης πολύ μεγάλη, το κείμενο προτείνεται για προσθήκη στο δυναμικό σύνολο κειμένων εκπαίδευσης που χρησιμοποιεί ο μηχανισμός. Οι προηγούμενες διαδικασίες αποτυπώνονται και στο διάγραμμα ροής του Σχήματος 7.1.

Στη συνέχεια της παρούσας ενότητας θα παρουσιάσουμε μία πιο λεπτομερή προσέγγιση των διαδικασιών του μηχανισμού. Η παρουσίαση αφορά τους αλγορίθμους που υλοποιούνται σε κάθε επίπεδο προκειμένου να καταλήξουμε σε μία συνοπτική και περιεκτική, προσωποποιημένη παρουσίαση της πληροφορίας στον χρήστη.

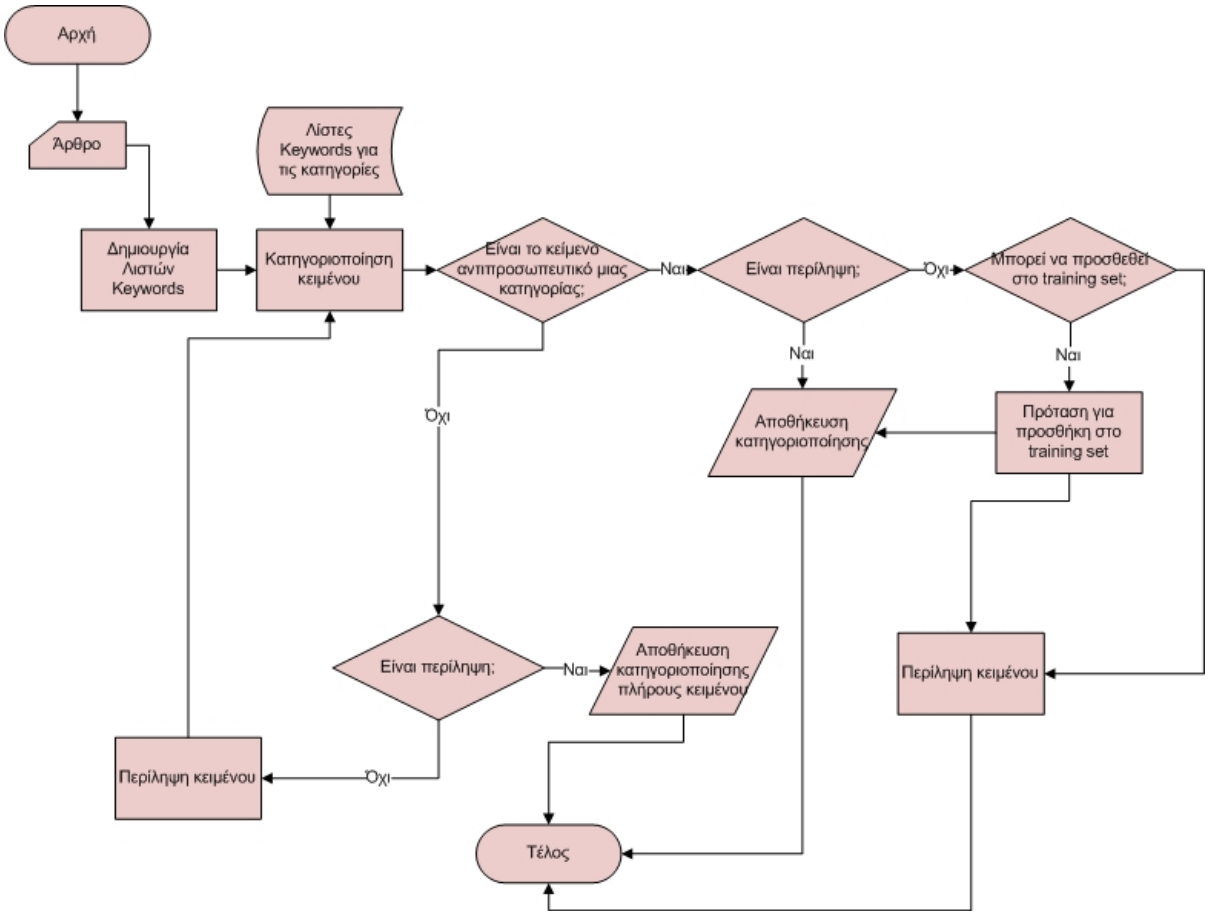
7.1.1 Εξόρυξη άρθρων - *crawling*

Ο μηχανισμός εξόρυξης πληροφορίας χρησιμοποιείται ως τροφοδότης άρθρων από το διαδίκτυο και αποτελεί τον μηχανισμό εισόδου για το PeRSSonal. Ακολουθεί μία αλγοριθμική περιγραφή του υποσυστήματος (Αλγόριθμος 7.1.2)

Αλγόριθμος 7.1.2 crawl()

```
feed_urls = Database_Query();
for all feed_urls as feed do
  if Content(feed)==changed then
    XML_Code = Fetch_Data(feed);
    Extracted_articles = Analyze_Data(XML_Code);
    for all Extracted_articles as article do
      if Not_Exists_in_DB(article) then
        Add_to_DB(article);
      end if
    end for
  end if
end for
```

Η παραπάνω διαδικασία εκτελείται από το μηχανισμό κάθε 6 με 10 λεπτά για κάθε RSS Feed



Σχήμα 7.1: Το διάγραμμα ροής των διεργασιών του βασικού αλγορίθμου.

προκειμένου να εντοπίσει νέα άρθρα. Ο τρόπος που επιλέγεται το χρονικό διάστημα μεταξύ διαδοχικών διαβασμάτων στο ίδιο RSS Feed θα αναλυθεί στη συνέχεια.

Κάθε RSS Feed αποθηκεύεται στην κεντρική βάση δεδομένων μαζί με τρεις τιμές που αλλάζουν δυναμικά από τον αλγόριθμο του crawler με βάση το πόσο συχνά αλλάζει το feed. Οι μεταβλητές αυτές είναι:

- ένας timer (T)
- μία τιμή μέσου όρου (median - M)
- ο χρόνος εκτέλεσης (ToE)

Η τιμή median μεταβάλλεται ανάλογα με την προσωρινή συμπεριφορά που έχει το RSS Feed. Ο timer είναι ένας μετρητής που αυξάνεται + median για κάθε φορά που το RSS που ελέγχεται και δεν βρίσκεται αλλαγμένο. Η μεταβλητή του χρόνου εκτέλεσης τέλος είναι μία βοηθητική μεταβλητή που μειώνεται κατά ένα κάθε φορά που ο crawler εκτελείται (μεταξύ διαδοχικών περιόδων αδράνειας). Όταν η μεταβλητή αυτή είναι μηδέν, το RSS Feed πρέπει να διαπεραστεί από τον crawler στην επόμενη εκτέλεσή του και μετά τη διαδικασία αυτή ο χρόνος εκτέλεσης τίθεται ίσος με την μεταβλητή median.

Προκειμένου ο μηχανισμός ανάκτησης άρθρων να γνωρίζει αν ύστερα από την ανάκτηση των RSS Feeds έχουμε κάποια αλλαγή στα άρθρα και να γίνει το crawling και σε αυτά, χρησιμοποιείται ένα είδος τοπικής cache στην οποία διατηρείται το hash του τελευταίου XML αρχείου που ανακτήθηκε. Εξετάζοντας αν το hash της προηγούμενης και της τελευταίας εκτέλεσης έχει μεταβληθεί, ο μηχανισμός αποφαίνεται για την διαπέραση των άρθρων που περιγράφονται από το δεδομένο RSS Feed και επιπλέον η μεταβλητή του χρόνου εκτέλεσης τίθεται ίση με 1. Αν αντίθετα εντοπιστεί ότι το hash του αρχείου δεν έχει αλλάξει, η μεταβλητή χρόνου εκτέλεσης αυξάνεται κατά median. Η παραπάνω διαδικασία περιγράφεται στον αλγόριθμο 7.1.3.

Αλγόριθμος 7.1.3 fetch_rss()

```
loop
  feeds_to_parse = Select RSS feeds from database with ToE equal to zero;
  for all feeds_to_parse as url do
    xml_code = fetch_rss(url);
    if hash_file(url) == hash(xml_code) then
      continue;
    else
      articles_in_rss = extract_info(xml_code);
      for all articles_in_rss as article do
        if title_not_found_in_last_articles(article) then
          add_to_db(article);
        end if
      end for
    end if
  end for
  sleep(X);
end loop
```

Για κάθε άρθρο που διαπερνά ο crawler, εξετάζεται αν το ζεύγος τίτλος-url βρίσκεται ήδη στον πίνακα των άρθρων. Με αυτό τον τρόπο αποφεύγονται άσκοπες ανακτήσεις άρθρων που ήδη

υπάρχουν και που επιπλέον θα δυσχέραιναν το έργο του υποσυστήματος παρουσίασης.

Ο αλγόριθμος που ανανεώνει τις μεταβλητές που περιγράφηκαν προηγουμένως για κάθε RSS Feed δίνεται στη συνέχεια (Αλγόριθμος 7.1.4).

Αλγόριθμος 7.1.4 update()

```
feeds=Fetch_rss_having_zero_ToE();
for all feeds as url do
  if hash_file(url) == hash(xml_code) then
    T = T + M;
    M = 70%M + 30%T;
  else
    M = 20%M + 80%T;
    T = 1;
  end if
  ToE = M;
end for
```

7.1.2 Προεπεξεργασία κειμένου

Ένα από τα υποσυστήματα 'πυρήνα' του συστήματος PeRSSonal είναι αυτό της προεπεξεργασίας κειμένου και εξαγωγής κωδικολέξεων. Το υποσύστημα αυτό αποτελεί ένα ανεξάρτητο μηχανισμό που φέρνει εις πέρας μια σημαντική διεργασία του όλου συστήματος, καθώς τροφοδοτεί τους μηχανισμούς που ακολουθούν με την απαραίτητη είσοδο: λέξεις κλειδιά και πληροφορίες για το κείμενο. Πρόκειται για μια αλγοριθμική, ακολουθιακή διαδικασία η οποία περιγράφεται από τον Αλγόριθμο 7.1.5.

Αλγόριθμος 7.1.5 create_keyword_frequency_list(XML,options)

```
String Text = fetch_next_text(XML);
String Title = fetch_next_title(XML);
parseTitle(Title);
Nouns = retrieveNouns(Text);
Text = removePunctuation(Text);
Text = removeStopwords(Text);
list Keywords = keepKeywordsPercentage(Text);
Keywords = stemming(Keywords);
list keyword_frequency_list = measure_keyword_frequencies(Text,Keywords);
list keyword_positions = get_keywords_positions(Text,Keywords);
return keyword_frequency_list, keyword_positions, Nouns;
```

Η βασική συνάρτηση measure_keyword_frequencies(Text,Keywords) περιγράφεται από τον Αλγόριθμο 7.1.6

και η διαδικασία get_keywords_positions(Text,Keywords) παρουσιάζεται στον Αλγόριθμο 7.1.7

Η διαδικασία της αναγνώρισης των ουσιαστικών ενός κειμένου εμπεριέχει ένα off-line βήμα εκμάθησης για τον POS tagger (SVM) χρησιμοποιώντας κανόνες που έχουν να κάνουν με τη συγκεκριμένη γλώσσα του κειμένο. Πριν από τη διαδικασία του tagging επομένως τα SVM μοντέλα

Αλγόριθμος 7.1.6 measure_keyword_frequencies(Text,Keywords)

```
for all kw in Keywords do
  if kw is found in Text then
    keyword_frequency_list[kw][appearances]++;
  end if
end for
return keyword_frequency_list;
```

Αλγόριθμος 7.1.7 get_keywords_positions(Text,Keywords)

```
for all kw in Keywords do
  for all sentence in Text do
    if kw is found in sentence then
      keyword_positions_list[kw][positions].push_back(position);
    end if
  end for
end for
return keyword_positions_list;
```

κατασκευάζονται από ένα σύνολο κειμένων εκμάθησης χρησιμοποιώντας το λογισμικό εκμάθησης (training) που παρέχεται με τη βιβλιοθήκη του SVM tagger [99]. Όταν η διαδικασία του training έχει ολοκληρωθεί, το σώμα του κειμένου προωθείται στον tagger όπου σημειώνονται τα ουσιαστικά του.

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

7.1.3 Μηχανισμός περίληψης

Περιγραφή

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. Κατά τη διαδικασία αυτή χρησιμοποιούνται ένα πλήθος παραμέτρων κάτι που κάνει το ζύγισμα αυτών και την εύρεση της βέλτιστης περίληψης μια περίπλοκη διαδικασία.

Οι παράμετροι που αξιοποιούνται είναι οι εξής:

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

(β′) η συχνότητα εμφάνισης του keyword στον τίτλο του κειμένου

(γ′) το ποσοστό των keywords μέσα στην πρόταση

(δ′) το ποσοστό των keywords στο κείμενο

(ε′) η πληροφορία για το αν το keyword είναι ουσιαστικό ή όχι

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

(ζ′) 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 [(α′) και (β′)], παράγουμε την πρώτη και αρχική εξίσωση για μια γενική βαθμολόγηση των προτάσεων:

$$S_i = \sum w_{k,i}(k_1 + k_2) \tag{7.1.1}$$

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

Ανάλυση

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

$$k_1 = 1 + 0.1x \tag{7.1.2}$$

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

$$k_2 = 1 + 1.2y \tag{7.1.3}$$

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

$$y = \frac{n}{t}\frac{m}{t} = \frac{nm}{t^2} \tag{7.1.4}$$

Για να κανονικοποιήσουμε τις τιμές που προκύπτουν από την εξίσωση (7.1.1), προτείνουμε την χρήση των παραγόντων (γ′) και (δ′). Η κανονικοποίηση χρειάζεται διότι, οι μεγάλες σε μήκος προτάσεις του κειμένου, τείνουν να βαθμολογούνται υψηλότερα σε σχέση με τις μικρές σε μήκος. Ο παράγοντας (γ′) 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 (γ′) ισούται με τρία πέμπτα (3/5) και ο παράγοντας (δ′) 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 (7.1.1), αφού το μήκος της την 'βοηθά' να έχει περισσότερα keywords. Οι δύο παράγοντες που προτείνονται, κανονικοποιούν αυτή την πιθανή 'αδικία'. Η μεγάλη πρόταση θα έχει 5/20 και 5/30 αντίστοιχα, ενώ η μικρή πρόταση θα έχει 4/4 και 4/30 για τους παράγοντες (γ΄) και (δ΄) 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 (7.1.1) και το $S_i' = S_i * Norm$, όπου το $Norm$ είναι ο παράγοντας κανονικοποίησης που ισούται με το γινόμενο των (γ΄) και (δ΄) υπολογίζεται απο τη διαδικασία μία φορά για κάθε πρόταση και για κάθε άρθρο.

Επεκτείνοντας την ανάλυση στον παράγοντα (ε΄), ο οποίος αφορά στο αν το keyword είναι ουσιαστικό ή όχι, θεωρούμε, όπως έχει αναφερθεί και σε προηγούμενο κεφάλαιο, ότι τα ουσιαστικά μίας πρότασης έχουνε εξέχουσα θέση σε ότι έχει να κάνει με την καλύτερη περιγραφή του νοήματος της πρότασης και κατ΄ επέκταση του κειμένου. Σκοπός επομένως του συγκεκριμένου ευρετικού είναι να βρεθεί ένας τρόπος συνυπολογισμού της πληροφορίας αυτής στη διαδικασία της εξαγωγής περίληψης. Καταλήγουμε λοιπόν στον παράγοντα N, όπου:

$$N = L * s \tag{7.1.5}$$

με $s = 0$ αν το keyword είναι ουσιαστικό και $s = 1$ αλλιώς. Ο παράγοντας L εκφράζει το επιθυμητό επιπλέον βάρος που θα πρέπει ένα ουσιαστικό να έχει σε μία πρόταση. Όπως θα δούμε και στο επόμενο κεφάλαιο, η πειραματική διαδικασία μας έδειξε ότι η τιμή του L δεν θα πρέπει να ξεπερνάει το 1.5 καθώς θα οδηγεί στο φαινόμενο οι προτάσεις που έχουν λίγα ουσιαστικά να αποκλείονται ουσιαστικά από τη διαδικασία εξαγωγής περίληψης. Τυπικές τιμές για το L είναι μεταξύ 0 και 1 με το 0 να εκφράζει την περίπτωση που επιθυμούμε η διαδικασία να μην λαμβάνει υπ΄ όψιν της τα αποτελέσματα της εξαγωγής ουσιαστικών.

Βασισμένοι στα προαναφερθέντα ευρετικά, παράγουμε μία περίληψη που αποτελείται από τις πιο αντιπροσωπευτικές προτάσεις του κειμένου. Για να τις καθορίσουμε, χρησιμοποιούμε την ακόλουθη εξίσωση:

$$S_i = \sum w_{k,i}(k_1 + k_2 + N) \tag{7.1.6}$$

Οι παράγοντες (στ΄), η ικανότητα του keyword να αντιπροσωπεύει την κατηγορία, και (ζ΄), 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.

7.1.4 Μηχανισμός κατηγοριοποίησης

Περιγραφή

Το υποσύστημα κατηγοριοποίησης βασίζεται στην μετρική ομοιότητας συνημιτόνου, σε εσωτερικά γινόμενα πινάκων και σε υπολογισμούς ζυγίσματος βαρών. Πιο συγκεκριμένα, το σύστημα αρχικοποιείται με ένα σύνολο κειμένων (άρθρα ειδήσεων) εκμάθησης τα οποία συλλέγονται από σημαντικές ειδησεογραφικές ιστοσελίδες (major news portals). Τα κείμενα αυτά είναι προκατηγοριοποιημένα 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) τα stemmed keywords, η απόλυτη και σχετική συχνότητα εμφάνισής τους αλλά και η θέση τους στο κείμενο καθώς και την πληροφορία αν είναι ουσιαστικά ή όχι, 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.

Ανάλυση

Αφού η διαδικασία περίληψης κειμένου του συστήματος βασίζεται στην επιλογή των πιο αντιπροσωπευτικών προτάσεων οι οποίες επιλέγονται ζυγίζοντας 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 βάρος 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 εξίσωση (7.1.7) χρησιμοποιείται για τον υπολογισμό της επίδρασης της διαδικασίας της κατηγοριοποίησης σε αυτήν της περίληψης.

$$k_3 = \begin{cases} A \cdot cw_i & \text{όπου } A > 1 \text{ και } cw \text{ το βάρος κατηγορίας} \\ -A \cdot cw_i & \text{όπου } A > 1 \text{ και } cw \text{ το βάρος κατηγορίας} \\ 1 & \text{για ουδέτερα ή μη βαθμολογημένα από το σύστημα keywords ή εάν } A = 0 \end{cases} \tag{7.1.7}$$

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

Με την χρήση της εξίσωσης (7.1.7), η εξίσωση (7.1.6) γίνεται:

$$S_i = \sum w_{k,i}(k_1 + k_2 + N)k_3 \tag{7.1.8}$$

7.1.5 Μηχανισμός προσωποποίησης

Προσωποποίηση περίληψης

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

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

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

$$k_4 = \begin{cases} B \cdot uw & \text{όπου } B > 1 \text{ και } uw \text{ το βάρος χρήστη} \\ -B \cdot uw & \text{όπου } B > 1 \text{ και } uw \text{ το βάρος χρήστη} \\ 1 & \text{για ουδέτερα ή μη βαθμολογημένα από τον χρήστη keywords ή εάν } B = 0 \end{cases} \tag{7.1.9}$$

Η παράμετρος uw αποτυπώνει τη σχετική συχνότητα ενός keyword για τον χρήστη. Αυτή μπορεί να μας παρέχει πληροφορία για το πόσο σημαντικό (ισχυρό) είναι ένα keyword για τον χρήστη. Αυτή η παράμετρος προστίθενται στην εξίσωση (7.1.8) η οποία γίνεται:

$$S'_i = \sum w_{k,i}(k_1 + k_2 + N)k_3k_4 \tag{7.1.10}$$

Πίνακας 7.2: Επίδραση των παραμέτρων A και B στο ζύγισμα των προτάσεων

A	**B**	**Αποτέλεσμα**
0	0	Οι παράγοντες προσωποποίησης και κατηγοριοποίησης δε υπολογίζονται στο αποτέλεσμα
0	1	Μόνο ο παράγοντας προσωποποίησης έχει επίδραση στο ζύγισμα των προτάσεων
1	0	Μόνο ο παράγοντας κατηγοριοποίησης έχει επίδραση στο ζύγισμα των προτάσεων
1	2	Ο παράγοντας προσωποποίησης έχει διπλάσια επίδραση σε σχέση με τον παράγοντα κατηγοριοποίησης στο αποτέλεσμα
1	10	Ο παράγοντας προσωποποίησης είναι τόσο μεγαλύτερος από τον παράγοντα κατηγοριοποίησης που η επίδραση του δευτέρου είναι ασήμαντη
1	1	Η ίδια επίδραση και για τον παράγοντα προσωποποίησης και για τον παράγοντα κατηγοριοποίησης
1.2	1.8	Οι τιμές που χρησιμοποιούνται από το μηχανισμό

Οι παράμετροι A και B στις εξισώσεις (7.1.7) και (7.1.9) αντίστοιχα, χρησιμοποιούνται σε συνδυασμό μεταξύ τους. Εάν δεν σκοπεύουμε να χρησιμοποιήσουμε κάποιον από τον παράγοντα κατηγοριοποίησης ή προσωποποίησης, μπορούμε να θέσουμε την τιμή 0 για την αντίστοιχη παράμετρο. Εάν θέλουμε να εστιάσουμε την προσοχή μας κυρίως στον παράγοντα προσωποποίησης και λιγότερο στην κατηγοριοποίησης, τότε μπορούμε να θέσουμε $B = 2$ και $A = 1$. Αυτό σημαίνει ότι ο παράγοντας k_4 θα έχει διπλάσια επίδραση από τον k_3. Ο πίνακας 7.2 δείχνει την επίδραση των παραμέτρων (ε′) και (στ′) σύμφωνα με τις τιμές των A και B.

Όπως παρατηρείται από την εξίσωση 7.1.10, μερικές 'ειδικές' περιπτώσεις μπορούν να λάβουν χώρα από τους μηδενισμούς που εισάγουν οι παράμετροι k_3 και k_4. Ο πίνακας 7.3 δείχνει την αντίδραση του αλγορίθμου στις τέσσερις διαφορετικές καταστάσεις.

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

Πίνακας 7.3: Αντίδραση του αλγορίθμου περίληψης στις μεταβλητές k_3 και k_4

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

Προσωποποίηση παρουσίασης στο χρήστη

Το υποσύστημα προσωποποίησης στον χρήστη αφορά τόσο την προσωποποιημένη περίληψη που θα στέλνεται σε αυτόν/ήν (και που αναλύθηκε στην προηγούμενη ενότητα) όσο και στο προσωποποιημένο περιεχόμενο γενικότερα αλλά και στην δυνατότητα πλήρους παραμετροποίησης από τη μεριά του χρήστη. Η εφαρμογή της αλγοριθμικής διαδικασίας που ακολουθεί, αφορά στην παρουσίαση της πληροφορίας στην εφαρμογή desktop που αναπτύχθηκε.

Ο μηχανισμός που αναπτύχθηκε βασίζεται στη διαρκή ανατροφοδότηση από τη μεριά του χρήστη. Η αλγοριθμική διαδικασία που αξιοποιείται προς αυτή την κατεύθυνση φαίνεται στον Αλγόριθμο 7.1.8.

Όταν ένας νέος χρήστης εγγράφεται στο PeRSSonal, δηλώνει τα keywords/κατηγορίες που προτιμάει καθώς και τη βαθμολογία που δίνει σε αυτά, στοιχεία που δίνουν μία αρχικοποίηση για το προφίλ. Η διαδικασία αυτή είναι τετριμμένη και μπορεί να αποφευχθεί εντελώς μιας και ο μηχανισμός προσωποποίησης καταγράφει τις προτιμήσεις του χρήστη, το ιστορικό περιήγησης ανανεώνοντας το προφίλ. Το προφίλ κάθε χρήστη αποτελείται από δύο λίστες keywords: μία θετική όπου οι λέξεις-κλειδιά που προτιμούνται από τον χρήστη κρατούνται, και μία αρνητική όπου οι μη-ενδιαφέρουσες λέξεις-κλειδιά για τον χρήστη αποθηκεύονται. Χρησιμοποιώντας αυτές τις λίστες, όπως είδαμε και στην προηγούμενη ενότητα, ο μηχανισμός προσωποποίησης της περίληψης μπορεί να παράγει βελτιωμένα αποτελέσματα.

Η διαδικασία ανανέωσης του προφίλ ενός χρήστη που περιγράφεται από τους αλγορίθμους 7.1.8, 7.1.9 και 7.1.10 τρέχει διαρκώς για κάθε χρήστη λαμβάνοντας υπ' όψιν τις εξής παραμέτρους:

- τα άρθρα που κάνει browse ο χρήστης
- το συνολικό χρόνο που ξοδεύει ο χρήστης βλέποντας την περίληψη ή το πλήρες κείμενο του άρθρου
- τα άρθρα (από τα προτεινόμενα) που ο χρήστης αποφεύγει να κάνει browse

Τα παραπάνω πηγάζουν από τις εξής απλές διαπιστώσεις. Ένας χρήστης συνήθως ξοδεύει χρόνο από ένα όριο και πάνω, έστω R_{ar_thr1} και R_{ar_sum1} διαβάζοντας το πλήρες κείμενο ή την περίληψη

Αλγόριθμος 7.1.8 Update_profile(a, b, c)

```
Get_articles(a,b)
for all articles do
  if article is full text then
    if time_viewed > Rar_thr1&&time_viewed < Rar_thr2 then
      Keywords_positive = top 5 frequent keywords
      Update_list(Positive, Keywords_positive)
    end if
  else
    if time_viewed > Rsum_thr1&&time_viewd < Rsum_thr2) then
      Keywords_positive = select top 5 frequent keywords
      Update_list(Positive, Keywords_positive)
    end if
  end if
  Get_articles(c)
end for
for all articles do
  Keywords_negative = select top 5 frequent keywords
  Update_list(Negative, Keywords_negative)
end for
```

Αλγόριθμος 7.1.9 Get_article(lists)

```
Βρες τα άρθρα που έχει δει το χρήστης
και τον χρόνο που πέρασε σε αυτά
είτε στο πλήρες άρθρο είτε στην περίληψή του
Βρες τα αρνητικά άρθρα (c)
```

Αλγόριθμος 7.1.10 Update_list(list, keywords)

```
Get_articles(a,b)
for all keyword in keywords do
  if (keyword not in list[]) then
    list.add(keywords[keyword])
  else
    list.update_freq(keywords[keyword])
  end if
end for
```

ενός άρθρου αντίστοιχα που βρίσκει ενδιαφέρον (παράγοντας α). Όμως ένα άνω όριο για τους προηγούμενους χρόνους είναι αναγκαίο, έστω R_{ar_thr2} και R_{ar_sum2}, μιας και δεν θέλουμε ο μηχανισμός να μπερδεύει τα 'ξεχασμένα' ανοιχτά άρθρα με τα πραγματικά ενδιαφέροντα. Τα όρια που χρησιμοποιούνται για τα R_{ar_thr1} και R_{ar_thr2} είναι 30 δευτερόλεπτα και 3 λεπτά αντίστοιχα καθορίζοντας έτσι ποια keywords από τα άρθρα πρέπει να κρατηθούν και να προστεθούν (τα ίδια αν δεν υπάρχουν, η συχνότητά τους στις υπάρχουσες συχνότητες αν υπάρχουν) στη λίστα με τα θετικά keywords για τον χρήστη.

Παρόμοια υπολογίζονται και τα όρια για το browsing των περιλήψεων των άρθρων:

$$R_{sum\_thr1} = R_{ar\_thr1} * S_{ratio}$$

και

$$R_{sum\_thr2} = R_{ar\_thr2} * S_{ratio}$$

όπου το S_{ratio} εκφράζει το ποσοστό 'συμπίεσης' που έχει επιτύχει επί του αρχικού κειμένου η περίληψή του:

$$S_{ratio} = \frac{total\_summary\_words}{total\_text\_words}$$

Επιπλέον, τις περισσότερες φορές παρατηρείται ότι ένας χρήστης επιλέγει να δει άρθρα μιας κατηγορίας που βρίσκει ενδιαφέρουσα (παράγοντας β) όπως αυτά διαφημίζονται από τον τίτλο και την περίληψη που έχουν. Επίσης, ένας χρήστης συνήθως αποφεύγει άρθρα που βρίσκει μηενδιαφέροντα και επομένως τα keywords που αναπαριστούν αυτά τα άρθρα θα πρέπει να λαμβάνουν αρνητικό βάρος (παράγοντας γ)

Από τους παράγοντες που περιγράφηκαν προηγουμένως, ο αλγόριθμος προσωποποίησης παρακολουθεί τα keywords για τα οποία ο χρήστης έχει εκφράσει ενδιαφέρον και επομένως τα άρθρα (που περιέχουν αυτά τα keywords) τα οποία ο χρήστης θα ήταν πρόθυμος να διαβάσει στο μέλλον. Η παράμετρος που αποτυπώνει την προτίμηση του χρήστη για ένα keyword βάσει των παραμέτρων α-γ είναι ο uw_i, που αναφέρθηκε και προηγουμένως. Αυτός βασίζεται στη σχετική συχνότητα που έχει το keyword στη λίστα, μία συχνότητα που μεταβάλλεται διαρκώς βάσει των επιλογών του χρήστη. Η παράμετρος uw_i πηγάζει από την εξίσωση που ακολουθεί

$$uw_i = relative\_frequency(kw_i) * (1 + T_{kw_i})$$

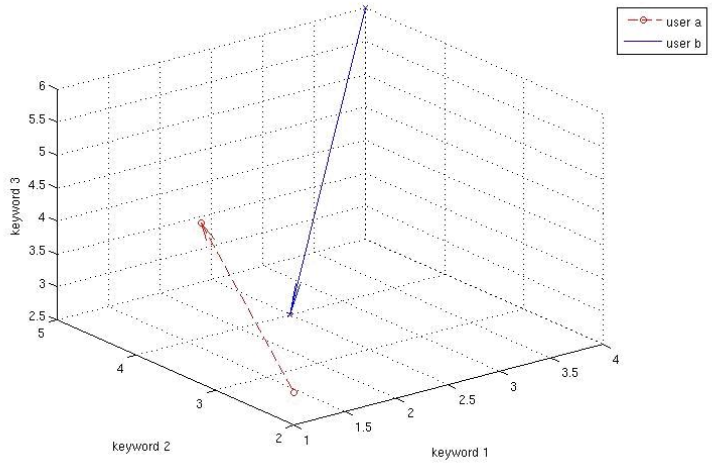
όπου T_{kw_i} είναι ο κανονικοποιημένος συνολικός χρόνος που ο χρήστης έχει ξοδέψει για το keyword i αν ανήκει στην λίστα με τα θετικά keywords. Αν το keyword είναι στη λίστα με τα αρνητικά keywords, το T_{kw_i} τίθεται ίσο με 0 αφού στην πραγματικότητα ο χρήστης δεν έχει ξοδέψει καθόλου χρόνο σε αυτό. Παράλληλα, περιμένουμε ότι όταν το προφίλ χρήστη φτάνει σε μία σταθερή κατάσταση, οι μέσοι χρόνοι για τις προτιμήσεις που έχει ο χρήστης για τα keywords θα απαλείψουν τυχών στατιστικές ανωμαλίες εκφράζοντας έτσι τις συνολικές προτιμήσεις που έχει ο χρήστης.

Ο παράγοντας προσωποποίησης συνολικά για κάθε keyword i, καλούμενος $k4$ βρίσκεται ως εξής:

$$k_{4i} = B * u_{wi}$$

όπου η παράμετρος $B > 1$ αν το keyword ανήκει στη λίστα με τα θετικά keywords και $B < 1$ αν το keyword ανήκει στη λίστα με τα αρνητικά keywords. Η νόρμα της παραμέτρου B μπορεί να πάρει κάθε δυνατή τιμή αυξάνοντας ή μειώνοντας έτσι την επίδραση που έχει η προσωποποίηση και το δυναμικό προφίλ στην διαδικασία ζύγισης των προτάσεων. Από τα προηγούμενα είναι σαφές ότι το $k4_i$ μπορεί να είναι θετικό, αρνητικό ή και μηδέν αν δεν έχουμε πληροφορία για το ενδιαφέρον του χρήστη για το keyword i.

Θα μπορούσαμε να αναπαραστήσουμε το συνολικό προφίλ χρήστη σαν ένα πολυδιάστατο μητρώο με πλήθος διαστάσεων τόσες όσα και τα keywords του προφίλ του, που αποτελείται από όλες τις $k4$ παραμέτρους (για κάθε keyword). Αυτό το μητρώο μεταβάλλεται διαρκώς καθώς αλλάζουν και οι προτιμήσεις του χρήστη. Μία γραφική απεικόνιση φαίνεται και στην εικόνα 7.2



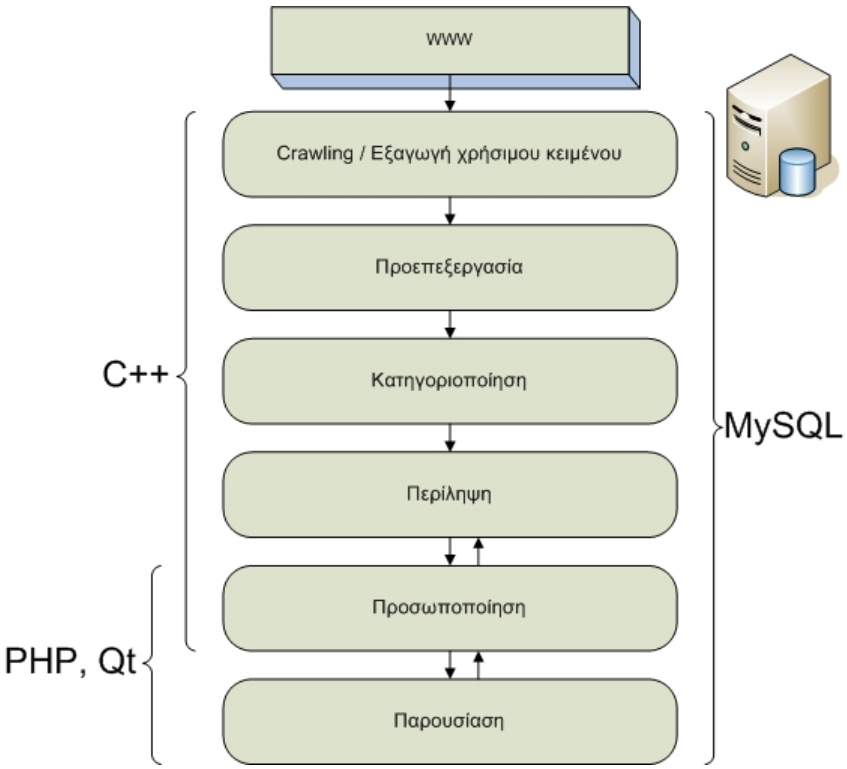
Σχήμα 7.2: Διανυσματική αναπαράσταση των προτιμήσεων του χρήστη.

7.2 Υλοποίηση του συστήματος

Προκειμένου το σύστημα PeRSSonal να διαθέτει τα χαρακτηριστικά της ταχύτατης απόκρισης στα ερωτήματα που γίνονται, αλλά και της διαλειτουργικότητας μεταξύ των διαφόρων υποσυστημάτων τα οποία όπως έχουμε αναφέρει και προηγουμένως μπορούν να αντικατασταθούν αυτόνομα, η υλοποίηση των υποσυστημάτων έγινε με χρήση τεχνολογιών C++ και PHP. Η διάκριση μεταξύ των δύο αφορά στο σημείο ενδιαφέροντος του μηχανισμού: προκειμένου να υπάρχει αλληλεπίδραση με τον χρήστη, έγινε χρήση PHP ενώ για τις υπόλοιπες εσωτερικές διαδικασίες του PeRSSonal χρησιμοποιήθηκε C++. Παράλληλα για την εφαρμογή desktop χρησιμοποιήθηκε Qt. Σημαντικό είναι επίσης ότι το τμήμα της προσωποποίησης του σύστηματος PeRSSonal είναι επιφορτισμένο με δύο λειτουργίες: τόσο για με την προσωποποιημένη περίληψη όσο και με την την καταγραφή των προτιμήσεων του χρήστη. Για την προσωποποίηση της περίληψης και του περιεχομένου γενικότερα χρησιμοποιείται η γλώσσα C++, ενώ για την καταγραφή των προτιμήσεων του χρήστη χρησιμοποιούνται σελίδες PHP κώδικα. Για τη διασύνδεση όλων των συστημάτων χρησιμοποιούμε τη βάση δεδομένων η οποία αποτελεί το κανάλι επικοινωνίας όλων των συστημάτων που κατασκευάζουμε, καθότι το σύστημα αντλεί πληροφορίες από αυτή και σε αυτή τις αποθηκεύει. Τα παραπάνω φαίνονται και στο σχήμα υλοποίησης 7.3.

7.2.1 Συλλογή άρθρων από το διαδίκτυο

Για τη συλλογή των πιο πρόσφατων άρθρων από το διαδίκτυο, υλοποιήθηκε ένας μηχανισμός που 'τρέχει' ανά τακτά χρονικά διαστήματα συγκεντρώνει άρθρα από ειδησεογραφικούς τόπους. Η ιδέα για την υλοποίηση του συγκεκριμένου μηχανισμού στηρίζεται στο γεγονός πως οι μεγαλύτεροι ειδησεογραφικοί δικτυακοί τόποι διαθέτουν RSS feeds τα οποία ανανεώνονται συνέχεια με τις νέες ειδήσεις που προκύπτουν. Έτσι, ελέγχονται με συγκεκριμένο αλγοριθμικό τρόπο (που περιγράφηκε



Σχήμα 7.3: Τεχνολογίες υλοποίησης ανά υποσύστημα.

στο κεφάλαιο 7.1.1) τα RSS feeds των ειδησεογραφικών πρακτορείων και αν υπάρχουν νέες καταχωρήσεις, τότε ο μηχανισμός διαβάζει όλα τα νέα άρθρα και τα προσθέτει στη βάση δεδομένων. Δεδομένης της μορφής που έχουν τα RSS feeds ο μηχανισμός αυτός μπορεί να συλλέξει την εξής πληροφορία:

- Τίτλος άρθρου
- URL άρθρου
- Περιγραφή άρθρου
- Ημερομηνία δημοσίευσης άρθρου

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

        <description>Read full story for latest details.
                <a href=http://rss.cnn.com/~a/rss/edition?a=CfB4PD>
                        <img src="http://rss.cnn.com/~a/rss/edition?i=CfB4PD" border="0">
                        </img>
                </a>                                                                    10
        </description>
        <pubDate>Sun, 22 Jul 2007 11:27:25 EDT</pubDate>
</item>
```

Προκειμένου να εξαχθούν τα συγκεκριμένα στοιχεία χρησιμοποιείται ένας XML Parser ο οποίος προσπαθεί να εντοπίσει όλα τα στοιχεία $< title >$ και όλα τα στοιχεία $< link >$. Οι υπόλοιπες πληροφορίες ($< description >$ και $< pubDate >$) δεν είναι σημαντικές για το μηχανισμό αλλά αποθηκεύονται ως μεταδεδομένα. Προκειμένου να αποφασιστεί εάν κάθε άρθρο θα πρέπει ή όχι να προστεθεί στη βάση δεδομένων, ελέγχεται αν στον πίνακα των άρθρων της βάσης δεδομένων υπάρχει ήδη το ζεύγος 'τίτλος-url'. Με αυτό τον τρόπο επιτυγχάνονται τα εξής ζητούμενα:

- Άρθρα που προέρχονται από διαφορετικά RSS Feeds και που έχουν τον ίδιο τίτλο δεν απορρίπτονται (αφού έχουν διαφορετικό url)
- Ανανεώσεις άρθρων με ίδιο τίτλο αντιλαμβάνονται ως διαφορετικά άρθρα και δεικτοδοτούνται από το σύστημα μιας και έχουν συνήθως διαφορετικό url
- Δεν εισάγεται δύο φορές το ίδιο άρθρο

Έχοντας εντοπίσει τα άρθρα που πρόκειται να προστεθούν στη βάση δεδομένων (μιας και είναι καινούργια), απομένει η αναδρομική επίσκεψη του url τους από τον crawler προκειμένου να ανακτηθεί και να αναλυθεί ο HTML κώδικας τους, διαδικασία που αναφέρεται ως εξαγωγή χρήσιμου κειμένου.

7.2.2 Εξαγωγή χρήσιμου κειμένου

Η διαδικασία εξαγωγής χρήσιμου κειμένου περιλαμβάνει την απομόνωση των χρήσιμων κομματιών μίας ιστοσελίδας τα οποία στη συγκεκριμένη περίπτωση είναι τα άρθρα - ειδήσεις. Η ανάλυση και εξαγωγή του κειμένου βασίζεται στον τρόπο με τον οποίο είναι δομημένες οι σελίδες που περιέχουν άρθρα - ειδήσεις αλλά και στο DOM μοντέλο στο οποίο μπορεί να αποδομηθεί μία HTML σελίδα. Ο μηχανισμός εξαγωγής χρήσιμου κειμένου ακολουθεί μετά τη διαδικασία συλλογής άρθρων από το διαδίκτυο (crawling) ενώ για μεγαλύτερη ταχύτητα μπορεί να εκτελείται παράλληλα από τη στιγμή που έστω και μία νέα σελίδα συλλέγεται από τους ειδησεογραφικούς δικτυακούς τόπους.

Όπως έχουμε ήδη αναφέρει, 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. Οστόσο 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 ($< b >$, $< i >$, $< h1 >$, $< h2 >$, κ. λπ.) 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.

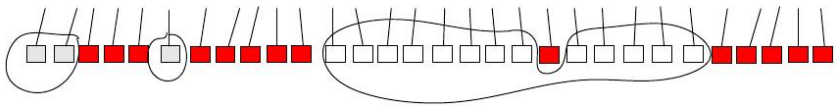
$TP = SLx/SH$. Πρόκειται για το Text Percentage το οποίο είναι μία ποσότητα που μας δείχνει πόσο κείμενο περιέχει ένα φύλλο συγκριτικά με το κείμενο ολόκληρης της σελίδας. Αν αυτή η ποσότητα είναι μεγάλη τότε συνεπάγεται πως το κείμενο αυτού του φύλλου ενδέχεται να είναι 'χρήσιμο κείμενο'. Αφού απορρίψουμε 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. $DX,Y = IY - IX$.

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 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 7.4.



Σχήμα 7.4: Ομάδες γειτονικών φύλλων.

Όπως μπορούμε να δούμε υπάρχουν αρχικά δύο φύλλα τα οποία περιέχουν αρκετό κείμενο ώστε να χαρακτηριστεί χρήσιμο κείμενο αλλά είναι πολύ μακριά από άλλα τέτοια φύλλα. Στη συνέχεια παρουσιάζεται ένα μεμονωμένο και έπειτα μία συστάδα από φύλλα τα οποία έχουν χαρακτηριστεί σαν φύλλα με χρήσιμο κείμενο και τα αποδέχεται ο μηχανισμός. Το συγκεκριμένο παράδειγμα θα μπορούσε να είναι της σελίδας 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.

7.2.3 Προεπεξεργασία κειμένου

Μια βασική υπο-διαδικασία του συστήματος PeRSSonal (όπως και κάθε συστήματος ανάκτησης πληροφορίας) είναι η προεπεξεργασία των κειμένων που δέχεται ο μηχανισμός ως είσοδο. Πρόκειται για την διαδικασία 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. Η διαδικασία αποτελείται από την υπο-διαδικασία της εξαγωγής κωδικολέξεων (keyword extraction) και πρόκειται ουσιαστικά για μια ακολουθιακή διαδικασία, η οποία μπορεί να θεωρηθεί ως ένα module του όλου συστήματος (και επομένως να αντιμετωπιστεί ξεχωριστά από αυτό).

Το υποσύστημα προεπεξεργασίας δέχεται ως είσοδο ένα πλήθος παραμέτρων:

- Το όνομα του XML αρχείου που περιέχει τα απαραίτητα στοιχεία του κειμένου (τίτλος, σώμα, ID και ενδεχόμενα την κατηγορία του)
- Το ελάχιστο μήκος λέξεων που πρέπει να κρατηθούν
- Η γλώσσα του κειμένου
- 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 βαρύτητα που πρέπει να δοθεί στα ουσιαστικά 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
- Ορθογραφικός έλεγχος του κειμένου και διόρθωση λαθών
- Αφαίρεση των σημείων στίξης του κειμένου
- Εύρεση και μαρκάρισμα των ουσιαστικών (noun retrieval)
- Αφαίρεση των μεγάλων κενών που υπάρχουν στις προτάσεις του κειμένου. Πλέον κάθε λέξη έχει απόσταση ενός κενού από την επόμενη
- Διαγραφή των stopwords με σύγκριση των λέξεων των προτάσεων με αυτές που έχουν δοθεί ως είσοδος
- Εξαγωγή μεμονωμένων λέξεων από τις προτάσεις (keywords)
- Πέρασμα των keywords του κειμένου από τη διαδικασία του stemming
- Αντιστοίχηση των keywords με τις αρχικές προτάσεις του κειμένου και εύρεση απόλυτης συχνότητας εμφάνισης του κάθε keyword μέσα στο κείμενο
- Κράτημα του ποσοστού των keywords που μας ενδιαφέρει (εξαρτάται από τις διαδικασίες που ακολουθούν το k/w extraction και είναι συνήθως 30-50% των συνολικών keywords)

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

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

Οι παραπάνω έξοδοι του μηχανισμού keyword extraction, αποθηκεύονται στους κατάλληλους πίνακες της βάσης δεδομένων απ΄ όπου αξιοποιούνται από τα συστήματα που ακολουθούν (περίληψη/κατηγοριοποίηση) με ασύγχρονο τρόπο.

7.2.4 Κατηγοριοποίηση κειμένου

Το υποσύστημα κατηγοριοποίησης κειμένου που υλοποιήθηκε για τον μηχανισμό PeRSSonal επιτελεί μια ουσιαστική διαδικασία καθώς μπορεί, δεδομένης μιας βάσης γνώσης κειμένων και ενός συνόλου κατηγοριών, να χαρακτηρίσει ένα νέο κείμενο ταξινομώντας το κατάλληλα με κάποια συσχέτιση στις κατηγορίες. Συνήθως, και εφόσον το κείμενο ανήκει σε κάποια από τις κατηγορίες, η συσχέτιση με αυτή την κατηγορία θα είναι σχετικά μεγάλη, ενώ με τις υπόλοιπες θα είναι πολύ μικρότερη. Είναι σημαντικό να τονίσουμε ότι η διαδικασία κατηγοριοποίησης προσπαθεί να ταξινομήσει τα νεοεισερχόμενα άρθρα σε κάποια από τις ήδη υπάρχουσες κατηγορίες, και όχι να ανιχνεύσει τυχών ομοιότητες και κοινά πρότυπα μεταξύ άρθρων ούτως ώστε να παράγει νέες κατηγορίες (θεμελιώδης διαφορά σε σχέση με την διαδικασία συσταδοποίησης που χρησιμοποιείται επίσης συχνά).

Το υποσύστημα κατηγοριοποίησης μπορεί να λειτουργήσει με δύο τρόπους: είτε κατηγοριοποιώντας το κείμενο που δίνεται στην είσοδο, είτε προσθέτοντας το κείμενο της εισόδου στην δυναμική βάση γνώσης (training set) του συστήματος. Οι είσοδοι του υποσυστήματος κατηγοριοποίησης κειμένου περιλαμβάνουν τα εξής:

- Τις πληροφορίες από τη διαδικασία του keyword extraction που περιέχει τα keywords του κειμένου, τις συχνότητες εμφάνισής τους και τις θέσεις τους στο κείμενο (το τελευταίο στοιχείο δεν χρειάζεται για την κατηγοριοποίηση).

- Ένα training set flag που αντιπροσωπεύει τον τρόπο λειτουργίας από τους δύο που περιγράφηκαν προηγουμένως.

- Το ποσοστό των keywords που θα πρέπει να κρατηθούν από την λίστα με τα keywords του κειμένου προκειμένου να προχωρίσει αποδοτικά η διαδικασία. Αυτό δίνεται ξεχωριστά και ανεξάρτητα από τη διαδικασία του keyword extraction γιατί κρατάμε διαφορετικά μεγέθη του συνόλου των keywords σε κάθε περίπτωση. Οι λόγοι και οι επιλογές οι οποίες γίνονται περιγράφονται αναλυτικά στη συνέχεια.

Εκπαίδευση συστήματος κατηγοριοποίησης

Η διαδικασία εκπαίδευσης του συστήματος κατηγοριοποίησης έχει να κάνει με την αρχική δημιουργία και την εν συνεχεία συντήρηση των υπαρχόντων εγγεγραμμένων κατηγοριών. Προκειμένου το υποσύστημα κατηγοριοποίησης να είναι σε θέση να αποφανθεί για την κατηγορία στην οποία ανήκουν τα νεοεισερχόμενα άρθρα, το σωστό αρχικό σύνολο εκπαίδευσης έχει μείζονα σημασία. Ακολουθεί η περιγραφή της διαδικασίας εκπαίδευσης του συστήματος κατηγοριοποίησης.

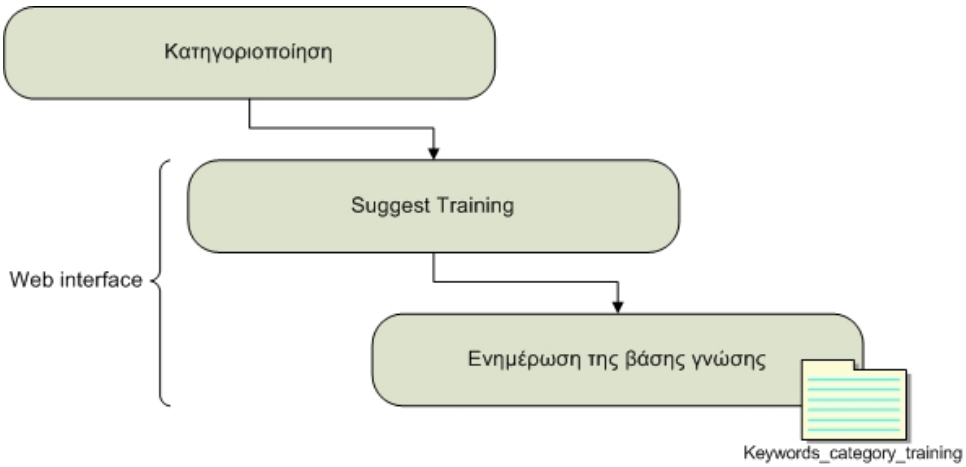
Διαδικασία προσθήκης στο *training set*

Όπως ήδη αναφέρθηκε, το τμήμα της κατηγοριοποίησης είναι εκείνο που διαχειρίζεται τη βάση γνώσης του συστήματος και επομένως έχει τη δυνατότητα προσθήκης νέων άρθρων, αντιπροσωπευτικών των κατηγοριών, σε αυτή. Η διαδικασία που ακολουθείται είναι απλή: έχοντας ως δεδομένα τα keywords του κειμένου (50% των συνολικών) και την κατηγορία την οποία αντιπροσωπεύει το κείμενο, εισάγονται (εάν δεν υπάρχουν ήδη) τα keywords του κειμένου και ανανεώνονται (ή εισάγονται) οι συσχετίσεις των keywords με την κατηγορία, στους κατάλληλους πίνακες εκπαίδευσης. Τέλος εισάγεται και η συσχέτιση των keywords που εισήχθησαν (η τροποποιήθηκαν) με το κείμενο που μόλις μπήκε στο training set. Φυσικά σε κάθε περίπτωση εξασφαλίζεται η μη ύπαρξη διπλοεγγραφών στη βάση και η γενικότερη συνέπεια των δεδομένων. Η διαδικασία εκκίνησης και

ανανέωσης του training set είναι ευκολότερη αφού δεν εμπεριέχει υπολογισμούς εσωτερικών γινομένων ή ομοιοτήτων συνημιτόνου όπως η διαδικασία της κατηγοριοποίησης, παρά μόνο συναλλαγές με τη βάση δεδομένων. Για την εκκίνηση του συνόλου εκπαίδευσης χρησιμοποιείται ειδικό module του υποσυστήματος κατηγοριοποίησης το οποίο δεδομένου ενός συνόλου κειμένων (corpus) πραγματοποιεί την αρχική δημιουργία των κατηγοριών. Παρόμοια για την ανανέωση του training set, έστω ότι έχουμε μία νέα κατηγορία για προσθήκη στην βάση δεδομένων, αυτό που είναι απαραίτητο είναι η ύπαρξη ενός συνόλου κειμένων για τα οποία γνωρίζουμε ότι είναι αντιπροσωπευτικά της δεδομένης κατηγορίας.

Πρόταση για προσθήκη στο *training set*

Προκειμένου το σύστημα κατηγοριοποίησης να είναι εύκολα προσαρμοζόμενο στη διαρκώς μεταβαλλόμενη πληροφορία άρθρων του διαδικτύου που δεικτοδοτούνται, δημιουργήθηκε ένας μηχανισμός αυτόματης πρότασης νέων άρθρων για προσθήκη στο σύνολο εκπαίδευσης. Κατά την κατηγοριοποίηση των νέων άρθρων, μπορούμε να εντοπίσουμε άρθρα που ανήκουν με σχετικά μεγάλη συσχέτιση σε μία κατηγορία (και σχετικά μικρή στις υπόλοιπες). Τα άρθρα αυτά προτείνονται εν΄ συνεχεία για προσθήκη στο training set, μια διεργασία που πραγματοποιείται μέσω των σελίδων διαχείρισης του PeRSSonal (administrative web interface). Η συνολική διαδικασία ενημέρωσης της βάσης γνώσης του συστήματος κατηγοριοποίησης φαίνεται στο σχήμα 7.5.



Σχήμα 7.5: Διαδικασία ενημέρωσης βάσης γνώσης.

Χρήση της ανάκτησης ουσιαστικών

Μετά την αρχικοποίηση του συνόλου εκπαίδευσης του κατηγοριοποιητή, δημιουργούνται λίστες με keywords - ουσιαστικά που είναι αντιπροσωπευτικά της κάθε κατηγορίας. Οι λίστες αυτές περιέχουν τα ουσιαστικά των κειμένων που δίνονται προς εκπαίδευση και που ανήκουν με υψηλή συσχέτιση σε μία κατηγορία και αντίστοιχα μικρή ή μηδενική σε άλλες. Η χρήση επομένως της ανάκτησης των ουσιαστικών του κειμένου σε ότι έχει να κάνει με τη διαδικασία της κατηγοριοποίησης εμπλέκεται στη διαδικασία εκμάθησης των κατηγοριών του συστήματος μιας και χρησιμοποιούνται μόνο λέξεις - ουσιαστικά για αυτή.

Ποσοστό των *keywords* για *training set*

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

Ποσοστό των *keywords* για κατηγοριοποίηση

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

Διαδικασία κατηγοριοποίησης

Για την κατηγοριοποίησης ενός νέου άρθρου ακολουθούνται τα παρακάτω βήματα:

1. Ανακτούνται οι συχνότητες των keyword του κειμένου με κάθε κατηγορία από τη βάση δεδομένων (τα keyword που εμφανίζονται στο κείμενο και υπάρχουν και στην κατηγορία). Έχουμε επομένως, εκτός από τον αρχικό πίνακα συχνοτήτων των keywords του κειμένου, και ένα πίνακα για κάθε κατηγορία που περιέχει συχνότητα εμφάνισης για την κατηγορία του κάθε keyword του κειμένου που εμφανίζεται και στην κατηγορία. Προφανώς, αν κάποιο από τα keywords του κειμένου δεν εμφανίζεται στην εκάστοτε κατηγορία, η αντίστοιχη θέση στον πίνακα συχνοτήτων της κατηγορίας θα έχει την τιμή 0 (μη εμφάνιση).

2. Υπολογίζονται τα μέτρα των προηγούμενων πινάκων, το εσωτερικό τους γινόμενο και από αυτά, η ομοιότητα συνημιτόνου μεταξύ τους. Έχουμε επομένως για κάθε κατηγορία, μια συσχέτιση του κειμένου, κάτι που είναι και το ζητούμενο.

3. Οι συσχετίσεις κειμένου-κατηγορίας ταξινομούνται κατά φθίνουσα σειρά.

4. Ακολουθούνται οι συνθήκες του αλγόριθμου 7.1.1 προκειμένου να αποφασιστεί αν η κατηγοριοποίηση έχει πετύχει ή όχι ακολουθώντας τα αντίστοιχα βήματα.

7.2.5 Αυτόματη εξαγωγή περίληψης

Ο μηχανισμός εξαγωγής περίληψης κειμένου του PeRSSonal, δέχεται ως είσοδο την έξοδο του keyword extraction μηχανισμού, δηλαδή τις εξαγόμενες κωδικολέξεις μαζί με τις συχνότητες εμφάνισής τους στο κείμενο, την πληροφορία για το αν αποτελούν ή όχι ουσιαστικό, καθώς και τις θέσεις τους στις προτάσεις. Επίσης ως είσοδος δίνεται το μέγεθος της απάντησης που επιθυμούμε και αν υπάρχει, πληροφορία για την κατηγορία του κειμένου.

Έχει ήδη αναφερθεί, ότι η διαδικασία αυτόματης εξαγωγής περίληψης δίνει ένα βαθμό, ή αλλιώς ένα σκορ, σε κάθε πρόταση του κειμένου ανάλογα με τη σχετικότητα που εκτιμά πως έχει. Το

σκορ της κάθε πρότασης σχηματίζεται με τη βοήθεια ενός πλήθους παραμέτρων που αφορούν στη συχνότητα εμφάνισης του keyword στο κείμενο, στην πιθανότητα εμφάνισης του keyword στον τίτλο του κειμένου, στο αν πρόκειται για ουσιαστικό, στην κατηγορία στην οποία ανήκει το κείμενο και τέλος στις ιδιαίτερες προτιμήσεις του χρήστη για τις κατηγορίες και επομένως για ορισμένα keywords. Το σύνολο των παραμέτρων συνοψίζεται στη σχέση 7.1.10. Για την υλοποίησή μας, και ύστερα από πειράματα, καταλήξαμε στο να θέσουμε τον μεν παράγοντα $k_1 = 1,4$, ενώ τον παράγοντα $k_2 = 1,2$. Ο πρώτος αφορά στην περίπτωση εμφάνισης του keyword στον τίτλο, ενώ ο δεύτερος στη συχνότητα εμφάνισης του keyword στο κείμενο. Η διαδικασία έχοντας αυτές τις δύο παραμέτρους μπορεί να δώσει μια βασική βαθμολόγηση για τις προτάσεις του κειμένου και επομένως μια περίληψη.

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

1. Γνωρίζοντας τα ουσιαστικά τα οποία απαρτίζουν τις προτάσεις του κειμένου, μπορούμε να βαθμολογίσουμε περισσότερο keywords που είναι ουσιαστικά.

2. Έχοντας την πληροφορία για την κατηγορία του κειμένου, η διαδικασία αναζητεί για κάθε keyword κάθε πρότασης την σχετικότητα που έχει το keyword με την κατηγορία. Σημειώνοντας το πόσο σχετικό είναι το κάθε keyword ή όχι με την κατηγορία, οι προτάσεις βαθμολογούνται με θετικό βάρος για κάθε keyword σχετικό που περιέχουν και με αρνητικό βάρος για κάθε keyword μη σχετικό που έχουν. Το τελικό σκορ των προτάσεων προκύπτει ύστερα από την προσθαφαίρεση όλων των βαρών. Με αυτό τον τρόπο, οι προτάσεις που περιέχουν keywords αντιπροσωπευτικά της κατηγορίας επιτυγχάνουν υψηλότερο σκορ σε σχέση με άλλες που δεν περιέχουν πολλά αντιπροσωπευτικά keywords και επιτυγχάνουν ουδέτερο σκορ (κοντά στο 0), ή με άλλες που έχουν πολλά αντιπροσωπευτικά άλλων κατηγοριών keywords και επιτυγχάνουν αρνητικό σκορ.

3. Θέλοντας να παράγουμε μια προσωποποιημένη περίληψη για κάποιον χρήστη με δεδομένο προφίλ στο σύστημα, βαθμολογούμε υψηλότερα προτάσεις που περιέχουν keywords αντιπροσωπευτικά των προτιμήσεων του χρήστη (keywords που ανήκουν με υψηλή θετική βαρύτητα στο προφίλ του χρήστη), ή βαθμολογούμε χαμηλότερα ή αρνητικά προτάσεις που περιέχουν keywords μη αντιπροσωπευτικά των προτιμήσεων του χρήστη. Η βαθμολόγηση γίνεται όπως και στην περίπτωση της κατηγοριοποιημένης περίληψης αναζητώντας για κάθε keyword, κάθε πρότασης, τη σημαντικότητα που έχει για τον χρήστη.

Τα προηγούμενα ευρετικά καθορίζουν τα N, k_3, k_4 της σχέσης 7.1.10 για την κάθε πρόταση. Το k_4 αποφασίστηκε να έχει μεγαλύτερο βάρος από το k_3 κατά έναν λόγο 2 (διπλάσια επίδραση)· δηλαδή το A της σχέσης 7.1.7 τέθηκε ίσο με 1 και το B της σχέσης 7.1.9 τέθηκε ίσο με 2. Το τελικό σκορ κάθε πρότασης προκύπτει συνολικά από τις παραμέτρους k_1, k_2, N, k_3, k_4, οι βαθμολογίες ταξινομούνται σε φθίνουσα σειρά και υπολογίζεται η απάντηση που πρέπει να επιστρέψει η διαδικασία. Το μέγεθος της εξόδου μπορεί είτε να καθορίζεται ως ποσοστό % επί των προτάσεων του κειμένου εισόδου, είτε ως επιθυμητό πλήθος χαρακτήρων. Οι προτάσεις που τελικά θα αποσταλούν ως περίληψη, ταξινομούνται σύμφωνα με τη σειρά εμφάνισής τους στο κείμενο, διατηρώντας έτσι την νοηματική συνοχή της απάντησης, και τελικά επιστρέφονται ως έξοδος της διαδικασίας.

7.2.6 Προσωποποίηση στο χρήστη

Μια εξαιρετικής σημασίας δυνατότητα για το σύστημα PeRSSonal είναι η προσωποποίηση που παρέχει στις προτιμήσεις του χρήστη. Σε αυτό το στάδιο διαμορφώνεται το δυναμικό προφίλ και όλα τα αποτελέσματα των προηγούμενων μηχανισμών προβάλλονται πίσω στο χρήστη.

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

- Οι επιλογές του χρήστη που αφορούν τις κατηγορίες που έχει το σύστημα (μόλις κάνει εγγραφή)
  - Βαθμολόγηση των κατηγοριών ανάλογα με το πόσο ενδιαφέρουν το χρήστη
- Οι επιλογές του χρήστη μόλις του εμφανίζονται άρθρα
  - Επιλογή του χρήστη να διαβάσει ένα άρθρο
  - Επιλογή του χρήστη να μη διαβάσει ένα άρθρο
  - Ο χρόνος που καταναλώνει ο χρήστης διαβάζοντας την περίληψη ενός άρθρου
  - Ο χρόνος που καταναλώνει ο χρήστης διαβάζοντας το πλήρες κείμενο ενός άρθρου
  - Την πιθανή επιλογή που κάνει ο χρήστης να διαβάσει σχετικά άρθρα (με το παρόν)

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

Ως δεδομένα έχουμε στο σύστημά μας έχουμε 7 κατηγορίες τις οποίες τις χαρακτηρίζουν λέξεις κλειδιά με συγκεκριμένα βάρη. Ο πίνακας 7.4 δείχνει ένα τέτοιο παράδειγμα για μία από τις κατηγορίες του συστήματός μας.

Cat id	Kw id	Rel frequency	Abs frequency
1	42	0.00105974	298
1	43	0.000927275	201
1	44	0.00172208	201
1	41	0.0103325	188
1	37	0.00516625	150
1	228	0.0149689	148
1	45	0.00251689	141

Πίνακας 7.4: Συσχέτιση λέξεων κλειδιών με κατηγορία

Το σκεπτικό είναι πως κάθε χρήστης αντιπροσωπεύεται με ένα διάνυσμα στον χώρο των κατηγοριών (keywords) που φανερώνει την 'τάση' που έχει στο να διαβάζει άρθρα που περιέχουν τα αντίστοιχα keywords ή που ανήκουν στην αντίστοιχες κατηγορίες. Αυτό σημαίνει πως εφόσον οι λίστες με τις λέξεις κλειδιά δύνανται να χαρακτηρίσουν μία κατηγορία αυτό συνεπάγεται και πως λίστες με λέξεις κλειδιά δύνανται να χαρακτηρίσουν τις επιλογές και τις προτιμήσεις ενός χρήστη. Αυτό που μας ενδιαφέρει συνεπώς είναι να μπορέσουμε από τις διαδικασίες που περιγράψαμε παραπάνω να καταλήγουμε σε λέξεις κλειδιά και συγκεκριμένα βάρη σε κάθε μία προκειμένου να χαρακτηρίσουμε το χρήστη. Σε πρώτη φάση αυτό που κάνουμε είναι να διαμορφώσουμε κάποιο

αρχικό προφίλ για το χρήστη κατά τη διάρκεια που πραγματοποιεί εγγραφή στο σύστημα. Δεδομένου ότι θέλουμε να κρατήσουμε τις διαδικασίες όσο το δυνατόν πιο διαφανείς προς τους χρήστες είναι ίσως το μόνο σημείο που μπορούμε ανώδυνα να βάλουμε το χρήστη στη διαδικασία του να συμπληρώσει κάποια στοιχεία για το προφίλ του.

Η διαδικασία εγγραφής και γενικά το περιβάλλον διεπαφής αποτελούν την βασική μονάδα επικοινωνίας του χρήστη με το σύστημα. Ένας χρήστης εγγράφεται στο σύστημα δίνοντας πληροφορίες για τις κατηγορίες που θέλει να παρακολουθεί. Ο χρήστης είναι δυνατόν να αλλάξει τα στοιχεία του μελλοντικά, κάτι που βέβαια δεν επηρεάζει άμεσα τα στοιχεία που έχουν ήδη συλλεχθεί για το προφίλ του, εκτός κι αν ο ίδιος επιθυμεί δημιουργία από την αρχή του προφίλ που ήδη έχει. Οι πληροφορίες αποθηκεύονται στην κεντρικοποιημένη βάση δεδομένων και ανανεώνονται συνεχώς με το δυναμικό προφίλ του όπως θα δούμε στη συνέχεια. Όταν ο χρήστης βρίσκεται στη διαδικασία εγγραφής στο σύστημα του παρουσιάζονται όλες οι κατηγορίες του συστήματος και του ζητείται να δηλώσει την προτίμησή του για κάθε κατηγορία. Ο χρήστης καλείται να επιλέξει μία βαθμολογία για κάθε κατηγορία από -5 έως 5. Το -5 μεταφράζεται σαν η κατηγορία δε με αντιπροσωπεύει καθόλου ενώ το +5 σημαίνει πως η κατηγορία αντιπροσωπεύει απόλυτα το χρήστη. Η επιλογή του 0 σαν προτίμηση κατηγορίας μεταφράζεται σαν ουδέτερη στάση απέναντι στην κατηγορία.

Εκμεταλλευόμενοι τις απαντήσεις των χρηστών μπορούμε να διαμορφώσουμε ένα αρχικό προφίλ για το χρήστη. Αυτό γίνεται ως εξής. Αρχικά δημιουργούμε εγγραφές για τις κατηγορίες που αρέσουν στο χρήστη και γι αυτές που ο χρήστης δεν προτιμά. Αυτό θα μας βοηθήσει να κάνουμε ένα πρώτο ξεκαθάρισμα των άρθρων ανάμεσα σε αυτά που ο χρήστης θέλει να δει και σε αυτά που δεν τον ενδιαφέρουν, ανάλογα με τις γενικές κατηγορίες που έχει επιλέξει. Ο χρήστης όμως πέρα από τις επιλογές για το τι θέλει να βλέπει και τι δε θέλει, δίνει και κάποια βαθμολογία για κάθε κατηγορία. Χρησιμοποιώντας αυτά τα δεδομένα μπορούμε να δημιουργήσουμε μία πιο αναλυτική περιγραφή του προφίλ. Το αναλυτικό προφίλ όπως έχει ήδη αναφερθεί περιλαμβάνει λίστες με λέξεις κλειδιά όπως αυτές που υπάρχουν για τις κατηγορίες που δείχνουν ποιες λέξεις κλειδιά ενδιαφέρουν το χρήστη και ποιες δεν τον αφορούν. Σε αυτή την περίπτωση επιτρέπονται τόσο θετικά βάρη όσο και αρνητικά. Ο υπολογισμός των βαρών για τις λέξεις κλειδιά του χρήστη υπολογίζονται από τον αλγόριθμο 7.2.1.

Αλγόριθμος 7.2.1 extract_user_keywords

```
for all (selection s) do
  if (s!=0) then
    Keyword_name_usr = select 20*s keywords from category keywords;
    Keyword_weight_usr = select (2*s*relative frequency) from category keywords;
  else
    Keyword_name_usr = select 10 keywords from category keywords;
    Keyword_weight_usr = select relative_frequency from category.keywords;
  end if
  Insert_into_user_profile_keyword_name_usr, keyword_weight_usr;
  if exists then
    Update_user_profile_set_keyword_weight += keyword_weight_usr where ke-
    yword_name = keyword_name_usr;
  end if
end for
```

Υποθέτουμε ότι ο χρήστης κάνει κάποιες επιλογές για τις κατηγορίες και επιλέγει από -5 έως 5. Από αυτές τις επιλογές επιλέγουμε $20*s$ λέξεις κλειδιά, όπου s είναι η επιλογή του χρήστη ($s \in$

[−5...5]) από τη λίστα με τις λέξεις κλειδιά που αφορούν την κατηγορία, όπως ο πίνακας που είδαμε παραπάνω. Εν συνεχεία, επιλέγουμε τη σχετική συχνότητα κάθε λέξης και την πολλαπλασιάζουμε με $2*s$. Αν για παράδειγμα ο χρήστης έχει επιλέξει για μία κατηγορία την επιλογή -3 και μία συγκεκριμένη λέξη κλειδί για την κατηγορία έχει σχετική συχνότητα 0,12 τότε στον πίνακα του χρήστη η συγκεκριμένη λέξη θα πάρει σχετική συχνότητα -0,72. αυτός ο αριθμός μας δείχνει και το πόσο ο χρήστης ενδιαφέρεται για τη συγκεκριμένη λέξη κλειδί. Στο παράδειγμα που δείξαμε ο χρήστης δεν ενδιαφέρεται για τη συγκεκριμένη λέξη. Πραγματοποιώντας αυτή τη διαδικασία καταλήγουμε σε μία αρχική λίστα με λέξεις κλειδιά και σχετικές συχνότητες για το χρήστη οι οποίες μας δίνουν τα παρακάτω στοιχεία:

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

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

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

Η προαναφερθείσα διαδικασία, συμπεριλαμβανομένης και της κατασκευής της λίστας με τις λέξεις κλειδιά πραγματοποιήθηκε προκειμένου να έχουμε κάποια πρώτα στοιχεία για το αρχικό προφίλ του χρήστη. Στη συνέχεια θα περάσουμε στην κατασκευή του δυναμικού προφίλ χρήστη, το οποίο μεταβάλλεται διαρκώς βάσει των επιλογών που κάνει ο χρήστης είτε χρησιμοποιεί το Web interface του PeRSSonal είτε την desktop εφαρμογή. Είναι σημαντικό να τονιστεί το γεγονός πως όσο περισσότερο χρησιμοποιεί ο χρήστης το σύστημα, τόσο καλύτερα διαμορφώνεται το προφίλ του και επομένως τόσο καλύτερα φιλτραρισμένη είναι η πληροφορία που του προσφέρεται.

Δυναμική διαμόρφωση προφίλ χρήστη

Όσο ο χρήστης χρησιμοποιεί τις υπηρεσίες του συστήματος PeRSSonal, τόσο καλύτερα διαμορφώνεται το προφίλ του από τα στοιχεία που συλλέγονται από τις επιλογές του. Κατά την σύνδεση ενός εγγεγραμμένου χρήστη περιμένουμε όταν του εμφανιστούν τα τελευταία (προσωποποιημένα) άρθρα, κάποια από αυτά να τα διαβάσει και άλλα να μην τα δει καθόλου. Το ίδιο μπορεί να κάνει ο χρήστης για όλα τα άρθρα που δεικτοδοτεί το PeRSSonal μιας και μπορεί εύκολα να εντοπίσει μέσω των προσφερόμενων καναλιών και των λειτουργιών αναζήτησης κάθε άρθρο. Και οι δύο αντιδράσεις (ανάγνωσης άρθρου ή μη) καταγράφονται και αποτελούν αντικείμενο μελέτης για το μηχανισμό μας. Αυτό που μπορούμε να καταλάβουμε δημιουργώντας εικονικά προφίλ στο

μηχανισμό μας είναι πως ο χρήστης θα επιλέξει να διαβάσει τα άρθρα που τον ενδιαφέρουν ενώ στα υπόλοιπα δε θα δώσει σημασία. Αυτή τη συμπεριφορά χρήστη την καταγράφουμε και την εκμεταλλευόμαστε προκειμένου να διαμορφώσουμε το προφίλ του.

Για κάθε άρθρο που επιλέγει ο χρήστης να διαβάσει προσθέτουμε τις συγκεκριμένες λέξεις κλειδιά στο προφίλ του βάση της σχετικής συχνότητας που παρουσιάζουν στο συγκεκριμένο άρθρο. Πρόκειται για μία μεγάλη σχετική συχνότητα κάτι που είναι επιθυμητό καθότι πρόκειται για λέξεις κλειδιά σε ένα άρθρο που ενδιαφέρει το χρήστη. Όσον αφορά τα άρθρα που δεν επέλεξε ο χρήστης, συγκεντρώνουμε τις λέξεις κλειδιά από αυτά τα άρθρα και ανανεώνουμε τις λέξεις κλειδιά του προφίλ χρήστη με αρνητική σχετική συχνότητα. Σε αυτή την περίπτωση και προκειμένου να διατηρηθεί η ακεραιότητα του μηχανισμού δεν αφαιρούμε με την πολύ μεγάλη σχετική συχνότητα που έχουν οι λέξεις κλειδιά αλλά με το $\frac{1}{4}$ αυτής. Έτσι σε περίπτωση που ένας χρήστης δεν διάβασε ένα άρθρο που τον ενδιέφερε επειδή του διέφυγε δεν υπάρχει μεγάλη διαφορά στο προφίλ του. Αντίθετα, για τα άρθρα που επιλέγει ο χρήστης παρατηρείται σημαντική αλλαγή στο προφίλ του.

Την ώρα που ο χρήστης επιλέγει να διαβάσει ένα άρθρο, όπως ήδη είπαμε, οι λέξεις κλειδιά αυτού του άρθρου προστίθενται στο προφίλ του. Η διαδικασία αυτή όμως μπορεί να οδηγήσει σε λανθασμένη ανανέωση του προφίλ αν ο χρήστης εσφαλμένα επιλέξει να διαβάσει ένα άρθρο που δεν τον ενδιαφέρει και απομακρυνθεί στα πρώτα δευτερόλεπτα από το άρθρο. Για το λόγο αυτό υπάρχει μία δικλείδα ασφαλείας, όπου αν ο χρήστης ανοίξει ένα άρθρο για να το διαβάσει και το κλείσει μέσα σε κάποιο δεδομένο χρονικό διάστημα, τότε αυτό δεν προσμετράται σε αυτά που έχει διαβάσει. Επιπλέον υπάρχει μία δικλείδα ασφαλείας για την περίπτωση που ο χρήστης 'ξεχάσει' για οποιονδήποτε λόγο ανοιχτό το άρθρο (πλήρες κείμενο ή περίληψη). Αν ο χρήστης ξεπεράσει αυτό το όριο θεωρείται πως έχει 'ξεχάσει' ανοιχτό το άρθρο και έτσι αυτός ο χρόνος δεν είναι αντιπροσωπευτικός για το χρόνο που δαπάνησε ο χρήστης στο συγκεκριμένο άρθρο. Τα όρια αυτά αναλύθηκαν στην ενότητα 7.1.5.

Ο χρόνος που καταναλώνει ο χρήστης σε ένα άρθρο ή σε μία περίληψη είναι φυσικά ευθέως ανάλογος με το μέγεθος του άρθρου ή της περίληψης. Ας υπενθυμίσουμε σε αυτό το σημείο πως είτε έχουμε να κάνουμε με το Web interface του PeRSSonal είτε με την desktop εφαρμογή, στο χρήστη προβάλλεται η εξής πληροφορία μόλις διαβάζει ένα άρθρο:

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

Παράλληλα, στον χρήστη παρουσιάζονται ταυτόσημα, σχετικά και παρόμοια άρθρα που έχουν δεικτοδοτηθεί από το μηχανισμό. Τα ταυτόσημα άρθρα αφορούν εκείνα που έχουν ποσοστό συσχέτισης άνω του 85% και είναι εντός των τελευταίων 8 ωρών. Τα παρόμοια άρθρα είναι εκείνα που παρουσιάζουν ποσοστό συσχέτισης άνω του 70% και είναι εντός των 24 ωρών. Τέλος τα σχετικά άρθρα είναι εκείνα με ποσοστό συσχέτισης άνω του 50% και έχουν εισαχθεί στο σύστημα εντός των τελευταίων 3 ημερών. Η επιλογή ενός χρήστη για μετάβαση σε κάποιο ταυτόσημο, σχετικό ή παρόμοιο άρθρο με δεδομένο το τρέχον έχει αντίστοιχη επίπτωση στο προφίλ του χρήστη καθώς φανερώνει το ενδιαφέρον για κάποιο θέμα που είναι κοινό στα άρθρα αυτά.

Οι παραπάνω ενέργειες του χρήστη έχουν άμεση επίδραση στον τρόπο με τον οποίο ανανεώνονται οι λέξεις κλειδιά στο προφίλ του. Κάποιες ενέργειες θεωρούνται πιο σημαντικές από άλλες

και έτσι δεν αυξάνονται ομοιόμορφα οι συχνότητες των λέξεων κλειδιών του προφίλ του χρήστη. Όπως έχουμε ήδη δει η διαμόρφωση του προφίλ συνίσταται στην καταγραφή λέξεων κλειδιών με κάποιο βάρος. Η αρχική τιμή αυτού του βάρους συλλέγεται από τις λέξεις κλειδιών των κατηγοριών ενώ στην πορεία από τον πίνακα που περιέχει τις λέξεις κλειδιά του συγκεκριμένου άρθρου μαζί με τα βάρη τους.

Σύμφωνα με τα παραπάνω και βάση των ενεργειών του χρήστη, οι λέξεις κλειδιά στο προφίλ του χρήστη διαμορφώνονται βάσει του πίνακα 7.5.

Ενέργεια	**Επίδραση**	**Πολλαπλασιαστής**
Μη επιλογή	Αρνητική	-0,3
Επιλογή	Θετική	+1
Ανάγνωση σώματος	Θετική	+0,5
Ανάγνωση περίληψης	Θετική	+0,3
Μετάβαση στο δικτυακό τόπο που φιλεξενεί	Θετική	+0,25
Ταυτόσημα	Θετική	+0,25
Παρόμοια	Θετική	+0,17
Σχετικά	Θετική	+0,14

Πίνακας 7.5: Ανανέωση των βαρών των *keywords* του προφίλ χρήστη

7.2.7 Εφαρμογή παρουσίασης πληροφορίας στην επιφάνεια εργασίας

Για την υλοποίηση της εφαρμογής χρήστη που αναπτύχθηκε ως ένα τμήμα παρουσίασης της πληροφορίας για το σύστημα PeRSSonal, έγινε χρήση της γλώσσας προγραμματισμού Qt. Τα πλεονεκτήματα τις συγκεκριμένης επιλογής είναι η διαλειτουργικότητα που έχει σε διάφορες πλατφόρμες καθώς και η υψηλή απόδοση. Η διαθεσιμότητα είναι ένα στοιχείο - κλειδί ώστε κάθε δικτυακή εφαρμογή να πετύχει λαμβάνοντας υπ' όψιν ότι η 'αντίπερα όχθη', οι εφαρμογές δηλαδή που τρέχουν σε επίπεδο web browser είναι εξαιρετικά διαλειτουργικές από την φύση τους. Η συγκεκριμένη επιλογή υλοποίησης δίνει στην εφαρμογή χρήστη ένα σημαντικό πλεονέκτημα απόδοσης σε σχέση με το web interface μιας και είναι multithreaded και υλοποιημένη σε γλώσσα C++. Η πλευρά του χρήστη με τη συγκεκριμένη υλοποίηση γίνεται επίσης πιο άμεση σε απόκριση και πιο φιλική προς τον χρήστη σε σχέση με το web interface του PeRSSonal. Η εφαρμογή παρουσίασης στην επιφάνεια εργασίας του χρήστη παρέχει πλήρεις δυνατότητες απο λειτουργικής άποψης στον χρήστη όσον αφορά τόσο στην προσπέλαση την πληροφορίας, όσο και στο κανάλι ανατροφοδότησης για την ενημέρωση του προφίλ του χρήστη με βάση τις επιλογές που κάνει και που αναλύθηκαν στην προηγούμενη ενότητα. Για τις επιλογές που κάνει ο χρήστης γίνεται κλήση στα ίδια αρχεία PHP που χρησιμοποιεί και το Web inteface έχοντας έτσι κοινό υπόβαθρο. Ένα ακόμη σημαντικό χαρακτηριστικό της εφαρμογής είναι η ευκολία προσαρμοστικότητας της εμφάνισης που δίνει στον χρήστη. Τα διάφορα κανάλια πληροφορίας έχουν την μορφή widgets και μπορούν να μετακινούνται εύκολα. Αντίστοιχα απλή ειναι και η παραμετροποίηση της εφαρμογής με εύκολα μενού ρυθμίσεων.

Όπως έχει ήδη αναφερθεί, για την μετάδοση των δεδομένων μεταξύ της εφαρμογής χρήστη και τον server χρησιμοποιείται XML μορφοποίηση. Τα αρχεία XML δημιουργούνται δυναμικά από τον server του PeRSSonal όταν ζητούνται από τον χρήστη και τα αντίστοιχα XSD σχήματα είναι επίσης διαθέσιμα για την μορφοποίηση και παρουσίαση στην εφαρμογή. Με αυτό τον τρόπο, οι δυνατότητες εξυπηρέτησης του μηχανισμού αυξάνονται σημαντικά: κάθε συνδεδεμένος χρήστης επικοινωνεί μέσω των XSD σχημάτων και των δυναμικών XML δεδομένων. Η χρήση επίσης XML αρχείων προς μετάδοση δίνει την δυνατότητα μελλοντικής προσθήκης δυνατοτήτων caching στην εφαρμογή.

ΚΕΦΑΛΑΙΟ 8

Προδιαγραφές και χρήση του συστήματος

> Program testing can be used to show the presence of bugs, but never to show their absence!
>
> *Edsger Dijkstra, Dutch Scientist, 2002*

Στο παρόν κεφάλαιο δίνονται οι προδιαγραφές του συστήματος PeRSSonal ώστε αυτό να είναι σε θέση να λειτουργεί σωστά και να παράγει αποτελέσματα που έχουν αξία. Επίσης δίνονται και ορισμένα στοιχεία που έχουν να κάνουν με τις απαιτήσεις του μηχανισμού σε υλικό και λογισμικό ώστε να μπορεί να λειτουργεί αποτελεσματικά.

8.1 Προδιαγραφές

8.1.1 Συλλογή άρθρων και εξαγωγή χρήσιμου κειμένου

Το σύστημα PeRSSonal ξεκινά την διαδικασία δεικτοδότησης άρθρων με τον μηχανισμό συλλογής άρθρων από το διαδίκτυο ο οποίος τρέχει ανεξάρτητα από τα υπόλοιπα υποσυστήματα που έχουν αλληλεπίδραση με τον χρήστη. Σε αυτόν περιλαμβάνονται η συλλογή άρθρων από τον ιστό και η εξαγωγή του χρήσιμου κειμένου από αυτά. Η λειτουργία είναι αυτοματοποιημένη ώστε να αλληλεπιδρά με τη βάση δεδομένων και η ανθρώπινη επίδραση μπορεί να είναι μόνο έμμεση. Το συγκεκριμένο υποσύστημα δέχεται σαν είσοδο τα RSS Feeds που καταγράφονται στη βάση δεδομένων και για την ακρίβεια τα urls των RSS feeds των news portals τα οποία πρέπει να διαπεράσει ο crawler. Είναι εύλογο πως υπόκειται στον διαχειριστή του συστήματος ο καθορισμός έγκυρων RSS Feeds για την τροφοδότηση του μηχανισμού με άρθρα, κάτι που είναι εφικτό μέσω της διεπαφής διαχείρισης του PeRSSonal. Ο μηχανισμός εξαγωγής χρήσιμου κειμένου είναι σχεδιασμένος ώστε να εξάγει κείμενα άρθρων από τη σελίδα· δεν έχει επομένως νόημα, και για την ακρίβεια γεμίζει τη βάση δεδομένων με 'σκουπίδια', η εισαγωγή urls από RSS feeds που δεν περιέχουν σώμα. Παρόμοια, πρέπει να αποφεύγεται η χρήση urls που δεν υπάρχουν (dead links) καθώς οδηγούν τον crawler και όλο συνολικά το σύστημα σε χάσιμο χρόνου.

8.1.2 Προεπεξεργασία κειμένου

Ως γνωστών, ο μηχανισμός προεπεξεργασίας κειμένου είναι αυτοματοποιημένος ώστε να αλληλεπιδρά με τα κείμενα της βάσης δεδομένων. Η ορθή λειτουργία του επομένως εναπόκειται στην ορθή κατάσταση της βάσης δεδομένων και τις συναλλαγές που γίνονται με αυτή. Δεδομένου ότι όλα τα απαραίτητα πεδία των πινάκων της βάσης δεδομένων περιέχουν ορθές πληροφορίες, η εξαγωγή κωδικολέξεων προχωράει βάσει αυτών. Πρέπει να σημειωθεί επίσης ότι η διαδικασία της προεπεξεργασίας κειμένου (αφαίρεση στίξης και αριθμών, ανάκτηση ουσιαστικών, αφαίρεση stopwords, stemming) εκτελείται σειριακά και πριν τις διαδικασίες κατηγοριοποίησης και περίληψης για το κάθε άρθρο. Τα αποτελέσματα του υποσυστήματος εξαγωγής κωδικολέξεων, όπως έχουμε πει, αποθηκεύονται στους κατάλληλους πίνακες της βάσης δεδομένων του συστήματος για να είναι διαθέσιμα στα υποσυστήματα που ακολουθούν.

8.1.3 Κατηγοριοποίηση και εξαγωγή περίληψης

Τα υποσυστήματα κατηγοριοποίησης και εξαγωγής περίληψης, που αποτελούν και τον πυρήνα του συστήματος μαζί με αυτό της προσωποποίησης, είναι σχεδιασμένα ώστε να δέχονται ως είσοδο τα δεδομένα της προεπεξεργασίας κειμένου. Όπως έχει ήδη αναφερθεί, η διαδικασία που ακολουθείται μετά την προεπεξεργασία κειμένου είναι: προσπάθεια για κατηγοριοποίηση του κειμένου βάσει κάποιων κριτηρίων και της βάσης γνώσης που έχουμε, αν η κατηγοριοποίηση είναι επιτυχής (το κείμενο είναι πολύ σχετικό με μία κατηγορία), προχωρούμε σε εξαγωγή γενικής περίληψης υποβοηθούμενη από την κατηγορία του κειμένου. Αν η κατηγοριοποίηση δεν είναι εφικτή, προχωρούμε σε εξαγωγή γενικής περίληψης και επιχειρούμε την κατηγοριοποίηση αυτής. Αν η δεύτερη απόπειρα κατηγοριοποίσης δώσει καλύτερα αποτελέσματα, αποθηκεύουμε αυτά στη βάση δεδομένων, αλλιώς τα πρώτα. Φυσικά τα υποσυστήματα μπορούν να κληθούν και αυτόνομα, π. χ. να ζητήσουμε περίληψη ή κατηγοριοποίηση ενός άρθρου που έχουμε στην κατοχή μας.

Η προσπάθεια κατηγοριοποίησης ενός άρθρου μοιάζει με την Linear Least Squares Fit - LLSF τεχνική και προχωράει ως εξής: η κατηγοριοποίηση των άρθρων γίνεται χρησιμοποιώντας την λίστα με τα πιο αντιπροσωπευτικά (stemmed) keywords του κειμένου μαζί με τις συχνότητες εμφάνισής τους. Έχοντας ήδη στη διάθεσή μας παρόμοιες λίστες που αφορούν στα πιο αντιπροσωπευτικά keywords της κάθε κατηγορίας, συγκρίνουμε τις λίστες χρησιμοποιόντας την ομοιότητα συνημιτόνου. Ένα επιπλέον σημαντικό χαρακτηριστικό είναι ότι η ανάλυση είναι διαφορετική για τους διαφορετικούς χρήστες. Όσο μεγαλύτερη είναι η αφαίρεση πληροφορίας σε τόσο λιγότερες προτάσεις ενός κειμένου πραγματοποιείται κατηγοριοποίηση του κειμένου και συνεπώς η κατηγορία στην οποία εντάσσεται ένα κείμενο είναι πιο γενική. Η παραπάνω διαδικασία έχει σαν αποτέλεσμα να δημιουργηθεί πολλαπλού είδους κατηγοριοποίηση στα κείμενα τα οποία θα διαθέτει το σύστημα με αποτέλεσμα να είναι διαφορετικά τα αποτελέσματα για κάθε χρήστη ανάλογα με τη λεπτομέρεια της αναζήτησης που πραγματοποιούν. Το ένα είδος κατηγοριοποίησης θα είναι καθαρά αλγοριθμικό ενώ το δεύτερο κομμάτι θα βασίζεται κυρίως στις προσωπικές επιλογές του χρήστη, οι οποίες δημιουργούν κατηγορίες αφαίρεσης πληροφορίας.

Έχει ήδη αναφερθεί αρκετές φορές ποια είναι η λειτουργία του μηχανισμού κατηγοριοποίησης. Αξίζει όμως να τονίσουμε κάποια βασικά στοιχεία της λειτουργίας αυτού του μηχανισμού. Ο μηχανισμός αυτός από τη στιγμή που θα αρχικοποιηθεί με ένα σύνολο πρότυπων κειμένων για τη δημιουργία μίας κατηγορίας μπορεί να λειτουργεί ανεξάρτητα από το υπόλοιπο σύστημα κατηγοριοποιώντας συνεχώς κείμενα. Είναι πολύ βασικό για την καλή λειτουργία του συστήματος να ανανεώνεται συχνά η βάση γνώσης με επικαιροποιημένα κείμενα χρησιμοποιώντας το τμήμα της ανανέωσης της βάσης γνώσης του μηχανισμού (suggest training).

8.1.4 Προσωποποίηση

Η λειτουργία του υποσυστήματος προσωποποίησης είναι πολυδιάστατη: αφορά τόσο στο προσωποποιημένο περιεχόμενο που παρουσιάζεται στο χρήστη όσο και στον τρόπο απεικόνισης αυτού στην εφαρμογή desktop. Προκειμένου η πληροφορία που μεταφέρεται μέσω των καναλιών XML να καλύπτει κατά το καλύτερο δυνατό τις προτιμήσεις του χρήστη, είναι σημαντικό το σύστημα να αντιλαμβάνεται εγκαίρως αλλαγές στο προφίλ του. Οι χρήστες σπάνια ξοδεύουν χρόνο για να δηλώσουν ρητά τι επιθυμούν, πολλές φορές λόγω του ότι δεν εμπιστεύονται τις προτιμήσεις που έχουν σε ένα απρόσωπο σύστημα που ζητάει υπερβολικά πολλά στοιχεία γι' αυτούς. Ο μόνος δρόμος επομένως είναι οι πληροφορίες αυτές να συλλέγονται έμμεσα καταγράφοντας τις επιλογές που κάνει ο χρήστης κατά την διάρκεια παραμονής του στο σύστημα. Η ερμηνεία όμως αυτών των συμπεριφορών που φαίνεται να παρουσιάζουν οι χρήστες πρέπει να ερμηνεύονται και κατάλληλα από το σύστημα βάσει σωστών παραμέτρων και μετρικών. Ήδη αναφέραμε τις παραμέτρους που αξιοποιεί το PeRSSonal προκειμένου να εξάγει το μητρώο με τα keywords και τις προτιμήσεις για καθένα που έχει ο χρήστης. Η διαδικασία όμως αυτή είναι επιρρεπής σε λάθη μεσοπρόθεσμα: ο χρήστης αρχικά πιθανών να μην βλέπει όλα τα νέα άρθρα που επιθυμεί ή πιθανών να βλέπει και κάποια που θεωρεί ότι αντιτίθενται στο προφίλ του. Μακροπρόθεσμα όμως, έχοντας αρκετά στοιχεία για την συμπεριφορά του χρήστη το σύστημα φαίνεται να προσαρμόζεται αρκετά καλά στις προτιμήσεις, κάτι που θα γίνει ορατό και στο επόμενο κεφάλαιο μέσα από την πειραματική διαδικασία.

8.2 Απαιτήσεις του συστήματος

Στην ενότητα αυτή παρουσιάζονται οι απαιτήσεις του συστήματος από άποψη λογισμικού και υλικού.

Λογισμικό και βιβλιοθήκες

Για την ανάπτυξη του συστήματος χρησιμοποιήθηκαν τα παρακάτω πακέτα λογισμικού και βιβλιοθήκες 8.1:

Kdevelop-3.5.1 [24]
GCC-4.1.2 [14]
Qt-4.4.2 [42]
MySQL-5.0.60 [32]
Apache-2.2.9 [2]
PHP-5.2.6 [37]
Boost-filesystem-1.34.1 [4]
Boost-regex-1.34.1 [4]
cgicc-3.2.5 [5]
mysql++-2.3.2 [31]
libcurl-7.18.2 [26]
expat-2.0.1 [10]
xerces-2.7.0 [48]
libstemmer [27]

Πίνακας 8.1: Σύνθεση υλικού για ανάπτυξη του συστήματος

Η ανάπτυξη του συστήματος έγινε εξ' ολοκλήρου σε open source λογισμικό και λειτουργικό σύστημα Gentoo Linux [13].

Υλικό

Το σύστημα που αναπτύχθηκε δεν έχει υψηλές απαιτήσεις υλικού. Μπορεί να στηθεί σε κάποιον υπολογιστή γενιάς Pentium II και νεότερο. Φυσικά εάν οι απαιτήσεις μας έχουν να κάνουν με ένα σύστημα που θα πραγματοποιεί real time κατηγοριοποίηση και εξαγωγή προσωποποιημένης περίληψης κειμένων είναι εύλογο να χρησιμοποιηθεί ένα πιο σύγχρονο σύστημα στο οποίο η βάση δεδομένων (η οποία και αποτελεί το bottleneck του συστήματος λόγω των πολλών συναλλαγών) θα έχει καλύτερους χρόνους εξυπηρέτησης. Για την ανάπτυξη των μηχανισμών χρησιμοποιήθηκε η παρακάτω σύνθεση υλικού (Πίνακας 8.2):

CPU	Intel Centrino 1.6 GHz
RAM	1256MB 333MHz
Cache	2048KB
Hard Disk	80GB, 7200rpm

Πίνακας 8.2: Σύνθεση υλικού για ανάπτυξη του συστήματος

ενώ για την καθημερινή λειτουργία του εξυπηρετητή PeRSSonal χρησιμοποιείται η παρακάτω σύνθεση Virtual Server (Πίνακας 8.3):

Dual Core CPU	Intel(R) Xeon(R) E5430 @ 2.66GHz
RAM	2056MB
Hard Disk	100GB, 7200rpm

Πίνακας 8.3: Σύνθεση υλικού του εξυπηρετή *PeRSSonal*

ΚΕΦΑΛΑΙΟ 9

Το σύστημα σε πλήρη λειτουργία

> No amount of experimentation can ever prove me right; a single experiment can prove me wrong.
>
> *Albert Einstein, German Physicist, 1955*

Το σύστημα PeRSSonal αναπτύχθηκε τμηματικά και κάθε τμήμα αυτού αξιολογήθηκε με πειραματική διαδικασία που έγινε είτε στα πλαίσια της προπτυχιακής διπλωματικής εργασίας, είτε στα πλαίσια της παρούσας μεταπτυχιακής εργασίας. Η διαδικασία αξιολόγησης που έγινε για κάθε τμήμα του μηχανισμού σκόπευε στην καταγραφή της αποτελεσματικότητας του ως ξεχωριστή οντότητα και στον προσδιορισμό των απαραίτητων παραμέτρων που πρέπει να χρησιμοποιηθούν σε κάθε βήμα ώστε ο μηχανισμός, ως σύνολο, να παράγει το βέλτιστο αποτέλεσμα. Ακολουθεί μια αναλυτική παρουσίαση των πειραματικών διαδικασιών και αξιολογήσεων που έλαβαν μέρος και που αφορούν στα βασικά υποσυστήματα του μηχανισμού: τη διαδικασία του keyword extraction, τους μηχανισμούς κατηγοριοποίησης και περίληψης μαζί με τις μεταξύ τους αλληλεπιδράσεις, στο υποσύστημα παρουσίασης πληροφορίας καθώς και στην εφαρμογή desktop που αναπτύχθηκε.

9.1 Μηχανισμός εξαγωγής κωδικολέξεων

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

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

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

Για να 'μετρηθεί' η διαφορά του νοήματος μεταξύ δύο κειμένων (δηλ. εκείνου στο οποίο έχουμε ελάχιστο μήκος λέξεων 4 και εκείνου που έχουμε ελάχιστο μήκος λέξεων 6), χρησιμοποιήθηκε μια απλή έκδοση του SVM αλγορίθμου [167].

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

$$x = a * b \tag{9.1.1}$$

$$y = |a| * |b| \tag{9.1.2}$$

$$z = x/y \tag{9.1.3}$$

$$r = sin(z) \tag{9.1.4}$$

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.

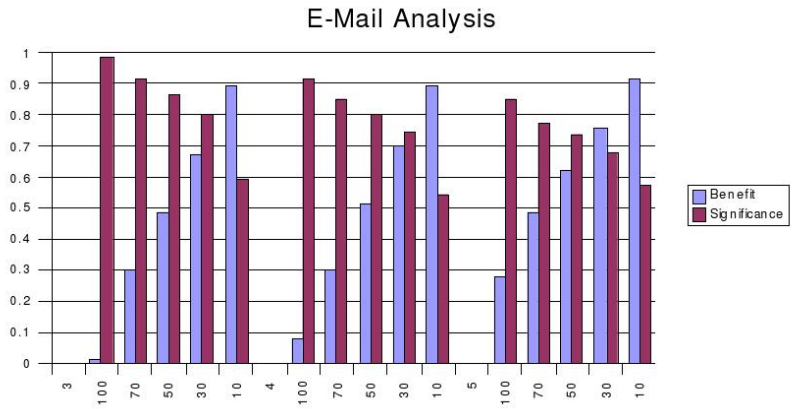
Με σκοπό να περιοριστεί ακόμη περισσότερο ο αριθμός των keywords του κειμένου, κρατήσαμε μόνο ένα ποσοστό αυτών και επανυπολογίσαμε από τη σχέση 9.1.4 την συσχέτιση μεταξύ των keywords του αρχικού κειμένου και του ποσοστού των keywords που κρατήθηκε.

9.1.1 Πειραματισμός με $e-mails$

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα που προέκυψαν από την πειραματική διαδικασία με κείμενα ηλεκτρονικού ταχυδρομείου. Κατά τη διάρκεια της πειραματικής διαδικασίας χρησιμοποιήθηκε ελάχιστο μήκος λέξεων τριών, τεσσάρων και πέντε γραμμάτων. Τα αποτελέσματα συνοψίζονται στην γραφική απεικόνιση του Σχήματος 9.1

Όπως φαίνεται και στο Σχήμα 9.1, 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

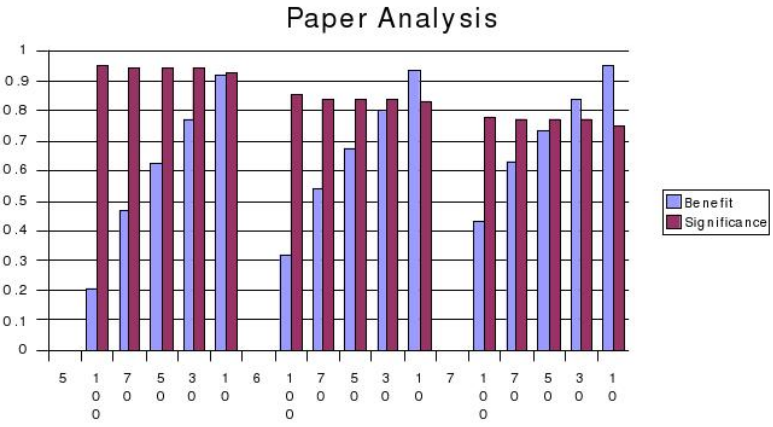


Σχήμα 9.1: Ανάλυση κειμένων ηλεκτρονικού ταχυδρομείου.

την ανάλυση κειμένων ηλεκτρονικού ταχυδρομείου, μειώνουμε δηλαδή το ελάχιστο μήκος λέξεων σε 3 γράμματα και κρατάμε τις μισές από τις κωδικολέξεις που προκύπτουν από την ανάλυση. Πρέπει να αναφερθεί επίσης ότι τα keywords βρίσκονται σε φθίνουσα σειρά διάταξης σε σχέση με τη συχνότητα εμφάνισης, πριν κρατηθεί το κατάλληλο ποσοστό.

9.1.2 Πειραματισμός με *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 Σχήμα 9.2 παρουσιάζονται τα αποτελέσματα που προέκυψαν μέσω της πειραματικής διαδικασίας.



Σχήμα 9.2: Ανάλυση κειμένων δημοσιεύσεων.

Όπως μπορούμε να δούμε από τη γραφική παράσταση του Σχήματος 9.2, κρατήθηκε ελάχιστο μήκος λέξεων 5, 6, 7 και περισσότεροι χαρακτήρες και στη συνέχεια κρατήθηκε ένα ποσοστό των keywords για καθένα από τον περιορισμό μήκους λέξεων. Όπως μπορούμε να δούμε, τα αποτελέσματα δεν επηρεάζονται (σημαντικά) από τον παράγοντα ποσοστού κράτησης των λέξεων. Αυτό μπορεί να εξηγηθεί ως εξής: τα κείμενα που επεξεργάζεται ο μηχανισμός εξαγωγής keywords σε αυτή την περίπτωση, περιέχουν περισσότερες από 900 μοναδικές λέξεις οι οποίες εμφανίζονται

πολλές φορές μέσα στο κείμενο και αυτό γιατί τα papers έχουν ένα συγκεκριμένο θεματικό πεδίο, με αποτέλεσμα, η επαναληπτικότητα των όρων είναι αναπόφευκτη. Το όριο της συσχέτισης ώστε να θεωρηθεί ότι το κείμενο δεν έχει χάσει το νόημά του, επιλέχθηκε να είναι το 80%. Αυτό σημαίνει ότι ο περιορισμός μήκους λέξεων για 7 ή περισσότερους χαρακτήρες μοιάζει να μην επιτυγχάνει το στόχο. Αντίθετα, με ελάχιστο μήκος λέξεων 5 ή 6 χαρακτήρων, το κείμενο που προκύπτει ξεπερνά σε συσχέτιση με το αρχικό κείμενο το όριο του 80% για την ομοιότητα.

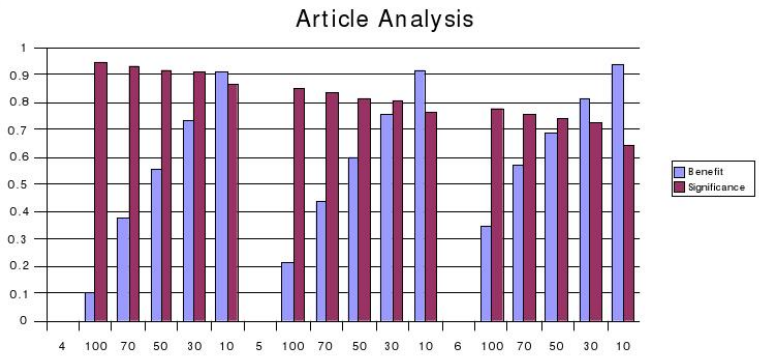
Το ζεύγος που αξιολογήθηκε ως βέλτιστο για να κρατηθεί, είναι το 6/10%, δηλαδή 6 χαρακτήρες ως ελάχιστο μήκος λέξεων και 10% των εξαγόμενων keywords, το οποίο μας οδηγεί σε 83% ομοιότητα και πάνω από 90% όφελος.

9.1.3 Πειραματισμός με άρθρα

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

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

Τα ζεύγη που μπορούν να περάσουν το όριο του 85%, μπορούν να βρεθούν μόνο στις περιπτώσεις που κρατούνται 4 και 5 χαρακτήρες ως ελάχιστο μήκος λέξεων. Πιο συγκεκριμένα, όλα τα ζεύγη που προκύπτουν από την χρήση 4 χαρακτήρων και η πρώτη επιλογή από τη χρήση 5 χαρακτήρων ικανοποιούν το όριο που αναφέρθηκε. Η πρώτη επιλογή από τη χρήση 5 χαρακτήρων, έχει πολύ μικρό όφελος (21%). Αντίθετα το ζεύγος 4/10% μας δίνει ομοιότητα πάνω από 85% και όφελος που ξεπερνάει το 90%. Αυτό σημαίνει ότι κόβουμε το 90% των μοναδικών keywords και αποθηκεύουμε μόνο ένα 10% αυτών που μας δίνουν πάνω από 85% ομοιότητα του τελικού κειμένου σε σχέση με το αρχικό. Τα παραπάνω συνοψίζονται και στο διάγραμμα του Σχήματος 9.3



Σχήμα 9.3: Ανάλυση άρθρων ειδήσεων από το διαδίκτυο.

9.1.4 Γενικά αποτελέσματα

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

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

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

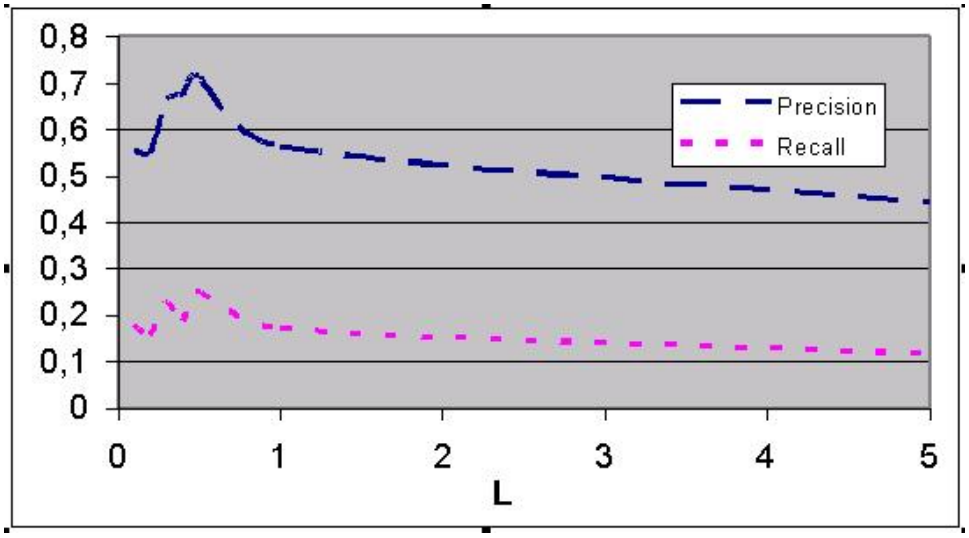
9.2 Πειραματισμός με το υποσύστημα εξαγωγής ουσιαστικών

Το τμήμα εξαγωγής ουσιαστικών, όπως ήδη αναφέρθηκε, αποτελεί μία σημαντική προσθήκη στο σύστημα PeRSSonal που έλαβε χώρα κατά την διάρκεια εκπόνησης της παρούσας εργασίας. Οι τεχνικές που περιγράφηκαν σε προηγούμενο κεφάλαιο αφορούν στο υποσύστημα εξαγωγής κωδικολέξεων του μηχανισμού και έχουν να κάνουν με την πιο αποτελεσματικότερη βαθμολόγηση αυτόν κάτι που αναμένεται να οδηγεί σε καλύτερα αποτελέσματα όσον αφορά στις διαδικασίες ανάκτησης πληροφορίας που ακολουθούν έπειτα από το μηχανισμό εξαγωγής κωδικολέξεων.

Προκειμένου να αξιολογήσουμε την απόδοση του συστήματος PeRSSonal όσον αφορά στην περίληψη των άρθρων που παράγετε, με την προσθήκη του υποσυστήματος εξαγωγής ουσιαστικών από το κείμενο διενεργήσαμε δύο σύνολα πειραματικών διαδικασιών. Πρώτα, προσπαθήσαμε να καθορίσουμε την καλύτερη δυνατή τιμή του παράγοντα L της σχέσης 7.1.5, ο οποίος εκφράζει το επιθυμητό επιπλέον βάρος που θέλουμε να δώσουμε σε ένα ουσιαστικό που περιλαμβάνεται σε μία πρόταση. Παράλληλα, προσπαθήσαμε να αξιολογήσουμε το αποτέλεσμα που έχει η εφαρμογή της τεχνικής ανάκτησης ουσιαστικών στη συνολική επίδοση του συστήματος χρησιμοποιώντας κλασικές μετρικές ακρίβειας - ανάκλησης. Για τα πειράματα χρησιμοποιήσαμε ένα σύνολο 3000 άρθρων από διάφορες πηγές (corpus). Τα άρθρα αυτά ανήκουν με μεγάλη συσχέτιση σε μία από τις επτά βασικές κατηγορίες του συστήματος και η πληροφορία αυτή χρησιμοποιήθηκε κατά την εξαγωγή των περιλήψεων με την εφαρμογή της σχέσης 7.1.10 (προκατηγοριοποιημένα άρθρα). Οι περιλήψεις των άρθρων είναι επομένως 'οι καλύτερες δυνατές' που μπορούν να εξαγχθούν αυτή τη στιγμή από τον μηχανισμό.

Όπως αναφέρθηκε και στο κεφάλαιο των αλγοριθμικών θεμάτων, η παράμετρος L χρησιμοποιείται για τον χειρισμό της επίπτωσης που έχει η διαδικασία ανάκτησης ουσιαστικών στην διαδικασία

περίληψης του συστήματος. Κατά την πειραματική διαδικασία δώσαμε διάφορες τιμές στην παράμετρο L ώστε να εντοπίσουμε την καλύτερη τιμή που πρέπει να εφαρμοστεί όσον αφορά πάντα στα άρθρα νέων με τα οποία καταπιάνεται το σύστημα PeRSSonal. Τα αποτελέσματα παρουσιάζονται στο σχήμα 9.4.



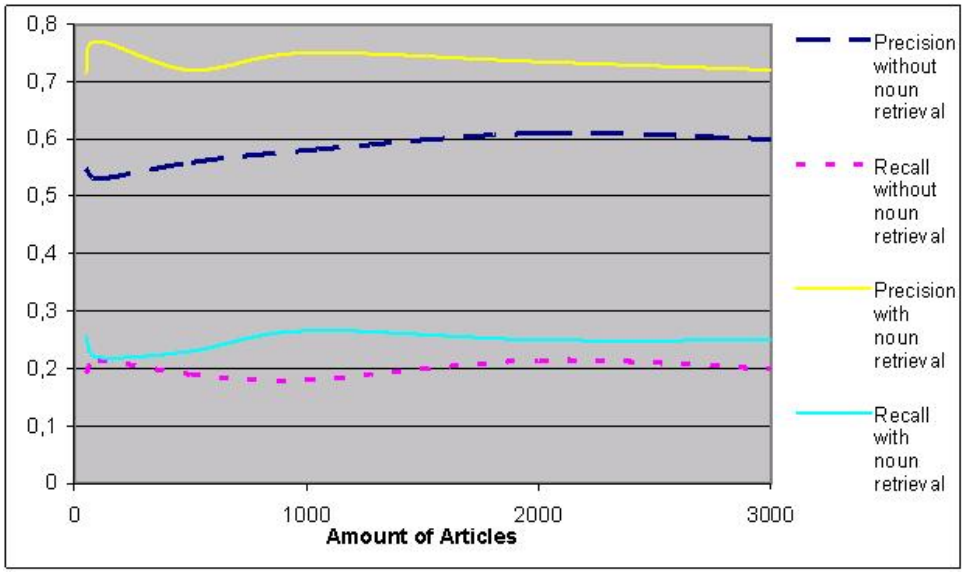
Σχήμα 9.4: Αποτελέσματα ακρίβειας / ανάκλησης για την περίληψη κειμένου μεταβάλλοντας τον παράγοντα L.

Από την γραφική παράσταση του σχήματος 9.4 είναι σαφές ότι μία τιμή μεταξύ 0.5 και 0.6 για την μεταβλητή L δίνει τα καλύτερα αποτελέσματα. Τιμές για το L πάνω από 1, φαίνεται να χειροτερεύουν τόσο την ακρίβεια όσο και την ανάκληση της διαδικασίας περίληψης σε σύγκριση με την περίπτωση όπου $L = 0$, δηλαδή όταν η τεχνική ανάκτησης ουσιαστικών δεν αξιοποιείται. Το παραπάνω έχει ως άμεση συνέπεια το γεγονός ότι όταν οι προτάσεις που περιέχουν ως επί το πλείστον ουσιαστικά κρατούνται για την διαδικασία περίληψης, αγνοώντας τις υπόλοιπες προτάσεις, η αποτελεσματικότητα της διαδικασίας ελαφρώς χειροτερεύει. Η εύρεση επομένως μιας χρυσής τομής για την παράμετρο L, που εξαρτάται πάντα και από τα κείμενα προς επεξεργασία, μπορεί να βελτιώσει αρκετά την διαδικασία της περίληψης. Το παραπάνω είναι ορατό επίσης και στο γράφημα που ακολουθεί (9.5) όπου η ακρίβεια και η ανάκληση αποτυπωνονται (χρησιμοποιόντας $L = 0.6$) όταν η διαδικασία περίληψης προχωρά με και χωρίς την αξιοποίηση των ουσιαστικών του κειμένου.

Από τη γραφική παράσταση του σχήματος 9.5 συμπεραίνουμε ότι η διαδικασία ανάκτησης ουσιαστικών μπορεί να δώσει σημαντικά καλύτερα αποτελέσματα στις περιλήψεις των κειμένων που προκύπτουν καθώς αυτές είναι πιο ακριβείς. Όσον αφορά στην ανάκληση, η βελτίωση είναι μικρή μεν, σημαντική δε λαμβάνοντας υπ΄ όψιν το γεγονός ότι από τη φύση της μια περίληψη αναπαριστάνει ένα επίπεδο αφαίρεσης, επομένως μια χαμηλής ανάκλησης αναπαράσταση, του αρχικού κειμένου.

9.3 Μηχανισμοί κατηγοριοποίησης και περίληψης

Κάθε μια από τις εξισώσεις (7.1.1),(7.1.6) και (7.1.9) για την βαθμολόγηση των προτάσεων ελέγχθηκε σε κάποια προκατηγοριοποιημένα (από ανθρώπους) κείμενα. Τα αποτελέσματα του μηχανισμού δείχνουν να είναι επαρκή σε σύγκριση με ήδη υπάρχοντα συστήματα. Ο βασικός μας στόχος είναι να παρουσιάζουμε μια προσωποποιημένη περίληψη άρθρων στον τελικό χρήστη και επομένως οι περιλήψεις που προκύπτουν βάσει των σχέσεων (7.1.1) και (7.1.6) δεν θα πρέπει να



Σχήμα 9.5: Επίπτωση της ανάκτησης ουσιαστικών στην ακρίβεια και την ανάκληση του συστήματος για την περίληψη κειμένου.

παράγουν περιλήψεις που διαφέρουν πολύ από ήδη υπάρχοντες αλγορίθμους. Η διαδικασία προσωποποίησης στην περίληψη δεν μπορεί να αξιολογηθεί σε σχέση με μια πρωτότυπη, ανθρώπινα παραγόμενη περίληψη αφού κάθε τέτοια εμπεριέχει τον υποκειμενικό ανθρώπινο παράγοντα. Ο μόνος πραγματικός εκτιμητής του συστήματος είναι ο τελικός χρήστης ο οποίος διαβάζει τις περιλήψεις.

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

9.3.1 Αξιολόγηση του μηχανισμού αυτόματης εξαγωγής 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 (Πίνακας 9.1) συνεπάγεται ότι ο μηχανισμός περίληψης που υλοποιήθηκε παράγει επαρκή αποτελέσματα συγκρινόμενος με δοκιμές που έγιναν με τον MEAD περιλήπτη, και σαφώς καλύτερα αποτελέσματα από τον περιλήπτη του MS Word. Προσθέτοντας τον παράγοντα κατηγοριοποίησης στη διαδικασία περίληψης, καταφέρνουμε να λάβουμε λίγο καλύτερα αποτελέσματα. Παρατηρούμε ότι η συνολική αύξηση είναι περίπου 10% σε σχέση με τα προηγούμενα αποτελέσματα όσον αφορά τις μετρικές της ακρίβειας και ανάκλησης. Η διαφορά οφείλεται στην διαδικασία κατηγοριοποίησης και, πιο συγκεκριμένα, στην προσθήκη της παραμέτρου k_3 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 Πίνακα 9.2 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 με τυχαίο ρυθμό 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.

9.3.2 Αξιολόγηση του μηχανισμού εξαγωγής προσωποποιημένης περίληψης

Η αξιολόγηση μιας δυναμικά εξαγόμενης προσωποποιημένης περίληψης κειμένου δεν είναι μια διαδικασία που μπορεί να γίνει με χρήση μέτρων σύγκρισης. Το μέτρο που χρησιμοποιείται για να αξιολογηθούν οι εξαγόμενες περιλήψεις είναι η συσχέτιση μεταξύ της περίληψης και του άρθρου που παρατηρείται από τους χρήστες του μηχανισμού. Η διαδικασία που ακολουθήθηκε για να αξιο-

Πίνακας 9.1: Σύγκριση του αλγορίθμου περίληψης του συστήματος με τον περιλήπτη του *MS Word*

	MS Word		**Κατασκευασμένος Μηχανισμός**	
	Ακρίβεια	**Ανάκληση**	**Ακρίβεια**	**Ανάκληση**
Άρθρο 1	0,33	0,12	0,66	0,75
Άρθρο 2	0,12	0,25	0,75	0,66
Άρθρο 3	0,25	0,12	0,5	0,66
Άρθρο 4	0,25	0,12	0,75	0,5
Άρθρο 5	0,33	0,5	0,66	1
Άρθρο 6	0,33	0,25	0,66	0,75
Άρθρο 7	0,25	0,33	0,75	0,66

Πίνακας 9.2: Αλλαγές στην ακρίβεια και την ανάκληση για την περίληψη ενός άρθρου ύστερα από την προσθήκη πιο αντιπροσωπευτικών *keywords* για την κατηγορία στην οποία το άρθρο ανήκει.

Χρόνος	**Άρθρα που προστέθηκαν στην κατηγορία**	**Υλοποιημένος μηχανισμός**	
		Ακρίβεια	**Ανάκληση**
10 λεπτά	0	0,5	0,66
8 ώρες	8	0,5	0,66
24 ώρες	31	0,66	0,5
36 ώρες	43	0,66	0,66
48 ώρες	59	0,66	0,66
62 ώρες	88	0,75	0,75
78 ώρες	103	0,75	0,8

λογηθούν 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 (7.1.6), όσο και από την εξίσωση (7.1.9), και (γ) άφησε τους χρήστες να επιλέξουν ποια περίληψη θεωρούν ως περισσότερο αντιπροσωπευτική για το άρθρο που διάβασαν. Η αντίστροφη διαδικασία εξετάστηκε επίσης, δόθηκαν δηλαδή πρώτα οι περιλήψεις στους χρήστες, στη συνέχεια το κείμενο και τέλος οι χρήστες αποφάνθηκαν για το ποια περίληψη θεωρούν ως περισσότερο αντιπροσωπευτική για το πλήρες άρθρο που διάβασαν. Και στις δύο περιπτώσεις που αναφέρθηκαν οι απαντήσεις ήταν οι ίδιες.

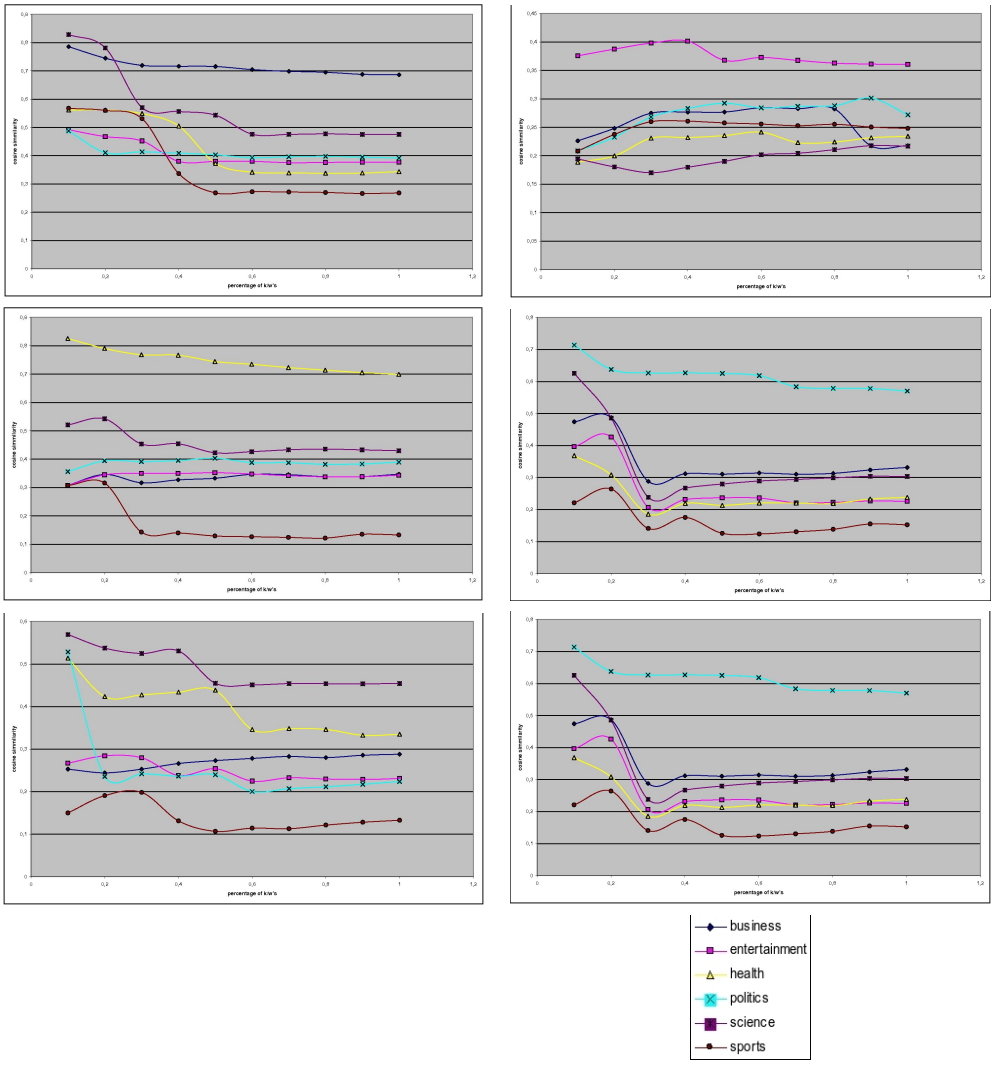
Οι χρήστες που έλαβαν μέρος στην πειραματική διαδικασία μπορούν να χωριστούν σε τρεις ομάδες: (α) νέοι χρήστες του συστήματος, (β) παλιοί χρήστες του συστήματος αλλά με μικρή δραστηριότητα (το οποίο σημαίνει λίγα δεδομένα για προσωποποίηση), και (γ) προχωρημένοι χρήστες του συστήματος με υψηλή καθημερινή δραστηριότητα (το οποίο σημαίνει πολλά δεδομένα για προσωποποίηση). Σύμφωνα με αυτές τις κατηγορίες, τρεις διαφορετικές καταστάσεις παρατηρήθηκαν. Οι νέοι χρήστες του συστήματος εξέφρασαν την άποψη ότι οι περιλήψεις που τους δόθηκαν ήταν όμοιες, κάτι που είναι μια λογική παρατήρηση εφόσον το σύστημα δεν έχει αρκετή πληροφορία για την διαδικασία προσωποποίησης και επομένως, η βαθμολόγηση των προτάσεων για την περίληψη δεν επηρεάζεται από τον παράγοντα k_4 (που χρησιμοποιείται για την προσωποποίηση της περίληψης). Οι χρήστες της δεύτερης ομάδας επέλεξαν, με ποσοστό μεγαλύτερο του 80% των άρθρων, την περίληψη που εξήχθη από την εξίσωση (7.1.6) (χωρίς τον παράγοντα προσωποποίησης). Αυτό ήταν επίσης αναμενόμενο αφού το προφίλ των χρηστών αυτών (με μικρή συμμετοχή) δεν ήταν πλήρες και περιείχε πολλά keywords που στην πραγματικότητα ήταν χαμηλής σημασίας τόσο για το άρθρο όσο και για την κατηγορία. Τα πλέον σημαντικότερα αποτελέσματα πηγάζουν από την τρίτη ομάδα χρηστών, τα μέλη της οποίας θεωρούνται από τους πιο 'έμπειρους' στη χρήση του συστήματος με σχεδόν σταθεροποιημένα προφίλ ύστερα από χρήση του συστήματος για μακρύ χρονικό διάστημα. Η σταθερότητα και η πληρότητα του προφίλ των χρηστών αυτών δίνει τη δυνατότητα προσωποποίησης στο μηχανισμό εξαγωγής περίληψης. Τα μέλη αυτής της ομάδας επέλεξαν σε ποσοστό μεγαλύτερο του 90% των άρθρων, την προσωποποιημένη περίληψη ως πιο αντιπροσωπευτική του άρθρου και μόνο 3% των περιλήψεων αξιολογήθηκαν ως 'όμοιες'. Είναι σημαντικό να τονιστεί ότι τα περισσότερα από τα υπολειπόμενα άρθρα (7%), αξιολογήθηκαν από τον μηχανισμό κατηγοριοποίησης του συστήματος ως 'ανήκοντα σε κάποια κατηγορία αλλά με ασθενή συσχέτιση'. Αυτό σημαίνει ότι αυτά ήταν άρθρα τα οποία προστέθηκαν στη συγκεκριμένη κατηγορία με την 'υποσημείωση' ότι το σύστημα δεν μπόρεσε με απόλυτη βεβαιότητα να τα κατατάξει σε κάποια κατηγορία, αλλά η κατηγορία στην οποία τελικά εισήχθηκαν είναι η πιο 'κοντινή' για αυτά τα άρθρα.

9.3.3 Αλληλεπίδραση μεταξύ της διαδικασίας περίληψης και κατηγοριοποίησης

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

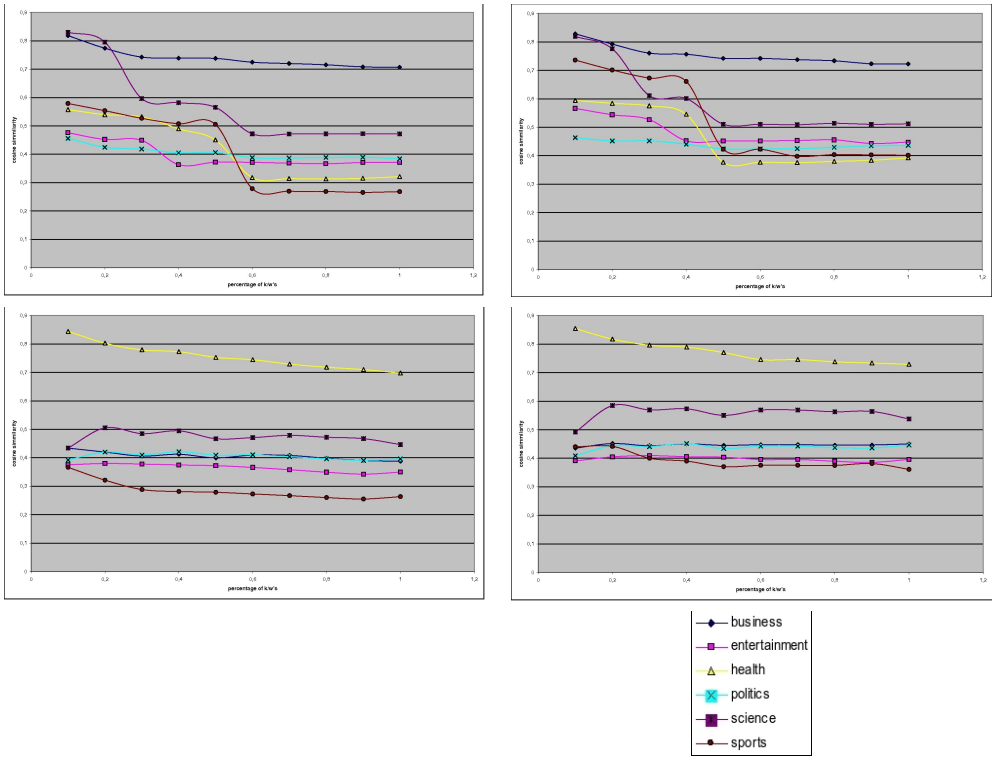
Αρχικά, χρειαζόταν να καθοριστεί το ποσοστό από keywords του κειμένου το οποίο πρέπει να κρατηθεί ουτοσώστε ο μηχανισμός κατηγοριοποίησης να έχει την μεγαλύτερη αποτελεσματικότητα. Προς αυτή την κατεύθυνση, μεταβάλαμε το ποσοστό των keywords που κρατούνται από

0,1 (δηλ. 10% των keywords) σε 1 (δηλ. όλα τα keywords) με βήμα 0,1, κάνοντας χρήση ενός αντιπροσωπευτικού κειμένου για κάθε μια από τις προαναφερθέντες κατηγορίες, και το κατηγοριοποιήσαμε. Το κείμενο που επιλέχθηκε για είσοδο στον μηχανισμό κατηγοριοποίησης δεν ήταν μέρος των κειμένων που χρησιμοποιήθηκαν για την κατασκευή της βάσης γνώσης (δεν ήταν μέρος του training set). Για κάθε ποσοστό από keywords μετρήθηκε η ομοιότητα συνημιτόνου μεταξύ του κειμένου και της κάθε κατηγορίας που υπάρχει στη βάση γνώσης. Εκτελέστηκαν πειράματα χρησιμοποιώντας ελάχιστο μήκος keywords 5 και 6 γράμματα, τόσο για την βάση γνώσης, όσο και για το κείμενο που εισήχθηκε στον μηχανισμό κατηγοριοποίησης. Ακολουθούν ορισμένα διαγράμματα που αποτυπώνουν τα αποτελέσματα.



Σχήμα 9.6: Ομοιότητα συνημιτόνου των κειμένων σε σχέση με τις κατηγορίες. Το *Training set* κατασκευάζεται με χρήση του 50% των *keywords* (διαδικασία προεπεξεργασίας).

Από το Σχήμα 9.6 (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

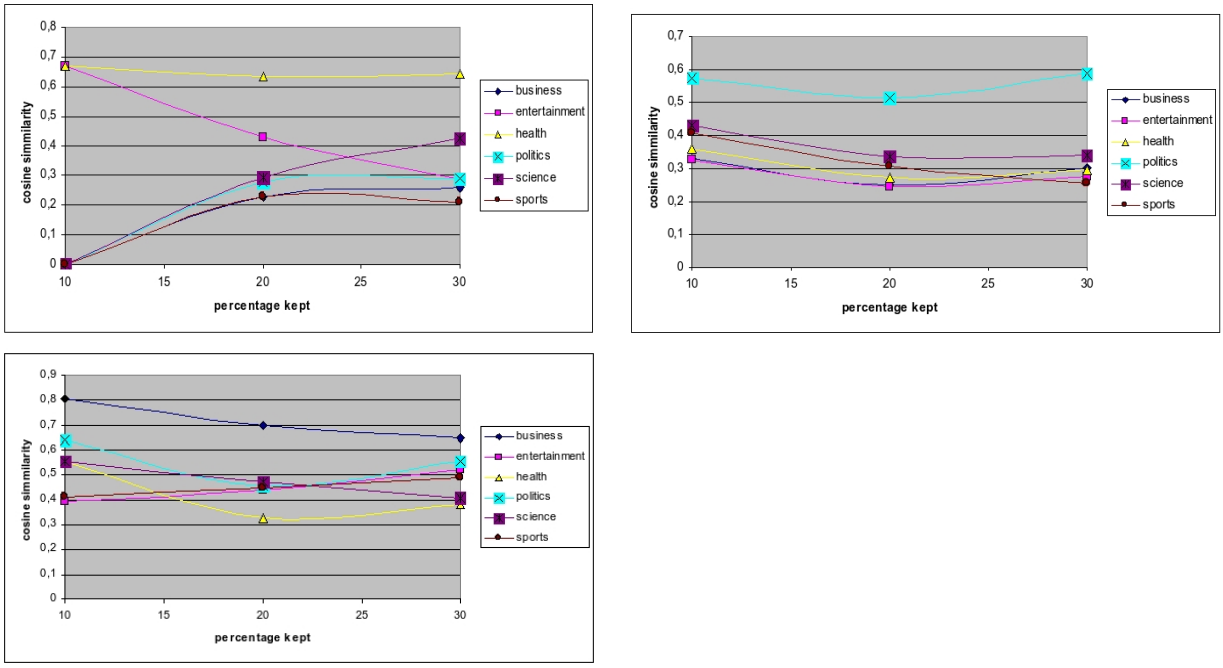


Σχήμα 9.7: Η πρώτη στήλη δείχνει την ομοιότητα συνημιτόνου μετρημένη χρησιμοποιώντας το 50% των *keywords* από το *training set*. Η δεύτερη στήλη δείχνει την ίδια ομοιότητα συνημιτόνου μετρημένη χρησιμοποιώντας το 100% των *keywords* του *training set*.

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 το Σχήμα 9.7, ένα κείμενο μπορεί να επιτύχει καλύτερο σκορ χρησιμοποιώντας ένα ελάχιστο μήκος 5 γραμμάτων για τα keywords και κρατώντας 50% των keyowrds που προκύπτουν. Με αυτό τον τρόπο, η βάση γνώσης είναι πιο φιλτραρισμένη, ενώ δεν μένουν έξω από τη διαδικασία keywords σημαντικά για κάποια/ες κατηγορία/ες.

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



Σχήμα 9.8: Ομοιότητα συνημιτόνου που μετρήθηκε για την κατηγοριοποίηση περιλήψεων χρησιμοποιώντας διάφορα ποσοστά για την δημιουργία των περιλήψεων

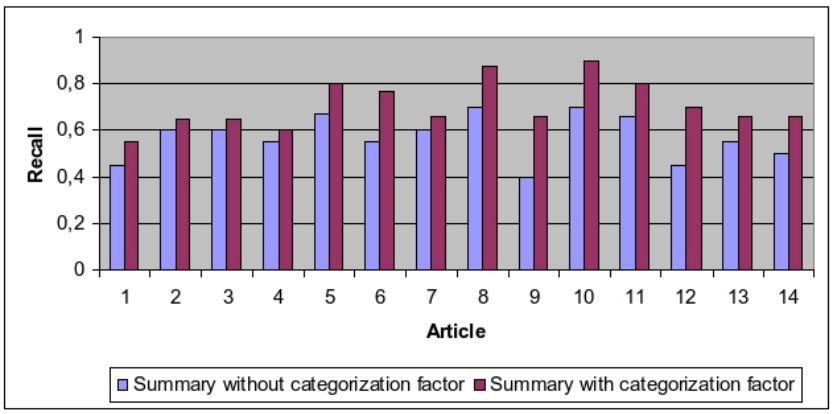
Χρησιμοποιήθηκαν διάφορα μεγέθη περιλήψεων με σκοπό να εντοπιστεί η επίδραση που έχουν στην κατηγοριοποίηση της περίληψης. Ακολουθούν ορισμένα διαγράμματα της πειραματικής διαδικασίας χρησιμοποιώντας κείμενα που ανήκουν σε διαφορετικές κατηγορίες, τα οποία αποκαλύπτουν το ιδανικό ποσοστό των προτάσεων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

Από αυτού του είδους την πειραματική διαδικασία καταλήξαμε στο συμπέρασμα ότι κρατώντας

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

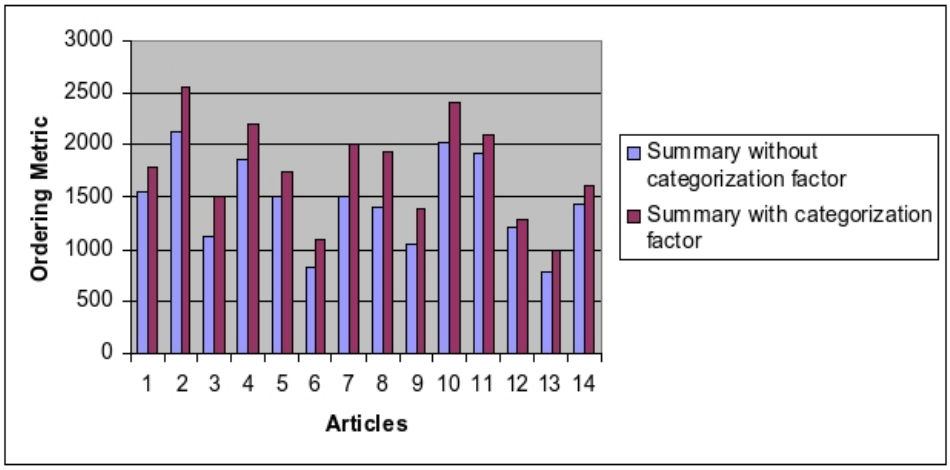
Ένα επιπλέον πεδίο στο οποίο έγινε πειραματισμός αφορούσε τη διερεύνηση της επίπτωσης που έχει η κατηγοριοποίησης στην διαδικασία της περίληψης. Για να αποκαλυφθεί η πιθανή συσχέτιση, κατασκευάσαμε τον μηχανισμό περίληψης ενσωματώνοντας σε αυτόν την δυνατότητα κατηγοριοποίησης. Αυτό σημαίνει πως, όταν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

Χρησιμοποιώντας κείμενα από συλλογές κειμένων (corpus texts), αρχικά παραγάγαμε την περίληψη του κειμένου χωρίς την χρήση του παράγοντα κατηγοριοποίησης (δηλ. k_3=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'.



Σχήμα 9.9: Σύγκριση της ανάκλησης των περιλήψεων οι οποίες εξήχθηκαν με και χωρίς την χρήση του παράγοντα κατηγοριοποίησης.

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



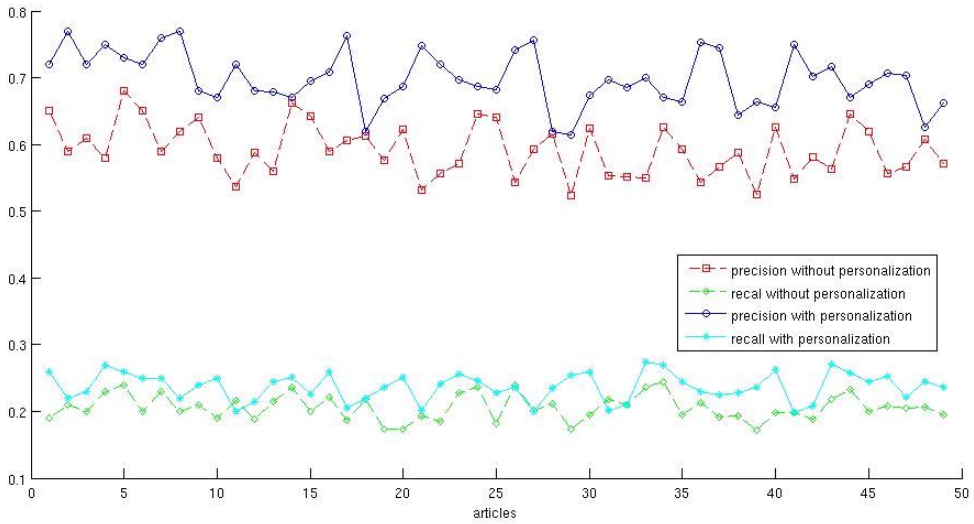
Σχήμα 9.10: Σύγκριση της μετρικής σειράς από περιλήψεις που εξήχθηκαν με και χωρίς τον παράγοντα κατηγοριοποίησης.

9.4 Μηχανισμός προσωποποιημένης περίληψης

Μια ακόμη επέκταση του συστήματος PeRSSonal που υλοποιήθηκε στα πλαίσια της παρούσας εργασίας είναι ο αλγόριθμος προσωποποίησης που περιγράφηκε στις ενότητες 7.1.5 και 7.2.6. Η αξιοποίηση πληθώρας παραμέτρων και πληροφοριών που προκύπτουν από την προσεκτική μελέτη των κινήσεων ενός συνδεδεμένου χρήστη, μας φανερώνουν πολλά στοιχεία για τις προτιμήσεις που έχει και που γνωρίζουμε ότι αλλάζουν με το πέρασμα του χρόνου. Καταλήξαμε επομένως στην παραγωγή ενός δυναμικού προφίλ χρήστη στο οποίο συμμετέχει ελάχιστα άμεσα ο ίδιος ο χρήστης και ως επί το πλείστον έμμεσα.

Προκειμένου να αξιολογήσουμε την βελτίωση που προσφέρει ο μηχανισμός προσωποποίησης που δημιουργήθηκε στις διαδικασίες του συστήματος, και πιο συγκεκριμένα στο υποσύστημα περίληψης, εκτελέσαμε πειραματική διαδικασία κατά την οποία χρησιμοποιήθηκαν οι κλασικές μετρικές ακρίβειας - ανάκλησης και 15 πραγματικοί εγγεγραμμένοι χρήστες του συστήματος PeRSSonal. Αρχικά ζητήσαμε από τους χρήστες να εγγραφούν και να χρησιμοποιήσουν το σύστημα όπως επιθυμούν (είτε μέσω του web interface είτε μέσω της εφαρμογής desktop) για χρονική περίοδο ενός μήνα. Με αυτόν τον τρόπο ο αλγόριθμος προσωποποίησης είχε αρκετό χρόνο να προσαρμόσει το προφίλ τους στις προτιμήσεις που παρουσιάζουν. Στη συνέχεια, δώσαμε 50 πλήρη άρθρα στους χρήστες και τους ζητήσαμε να βαθμολογήσουν μερικές προτάσεις των άρθρων αυτών τις οποίες εκτιμούν ότι είναι σημαντικές και επομένως θα έπρεπε να ανήκουν στην περίληψη του καθενός άρθρου. Επίσης παραγάγαμε και τις 'γενικές' περιλήψεις αυτών των άρθρων οι οποίες όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο δεν αξιοποιούν τον παράγοντα προσωποποίησης. Στη συνέχεια μετρήσαμε τις μετρικές ανάκλησης και ακρίβειας που έχει κάθε μία από τις δύο περιπτώσεις περιλήψεων (σε σχέση φυσικά με τις επιλογές που θεωρούν οι χρήστες ως ιδανικές) και τις συγκρίναμε. Τα αποτελέσματα παρουσιάζονται στη γραφική παράσταση του σχήματος 9.11.

Από το σχήμα 9.11 συνάγεται ότι η εφαρμογή του νέου αλγορίθμου προσωποποίησης έχει δώσει μία σημαντικά βελτιωμένη απόδοση στις δυνατότητες περίληψης του μηχανισμού τόσο για την ακρίβεια όσο και για την ανάκλησή του. Η βελτίωση είναι της τάξης του 17% για την ακρίβεια και 14% για την ανάκληση. Είναι σημαντικό όμως να τονίσουμε ότι οι στατιστικές βασίζονται στις επιλογές των χρηστών προκειμένου να εξαχθούν τα αποτελέσματα ακρίβειας και ανάκλησης και επομένως είναι υποκειμενικά 'προδιαθετιμένες' από τη φύση τους. Από την άλλη μεριά όμως,



Σχήμα 9.11: Ακρίβεια και ανάκληση του αλγορίθμου προσωποποίησης πάνω στην περίληψη κειμένου.

δεν υπάρχουν απόλυτα αντικειμενικά κριτήρια για την εξαγωγή της περίληψης από ένα κείμενο και επίσης είναι αυτή η 'υποκειμενικότητα' που ο αλγόριθμος προσωποποίησης προσπαθεί να υπολογίσει για κάθε χρήστη.

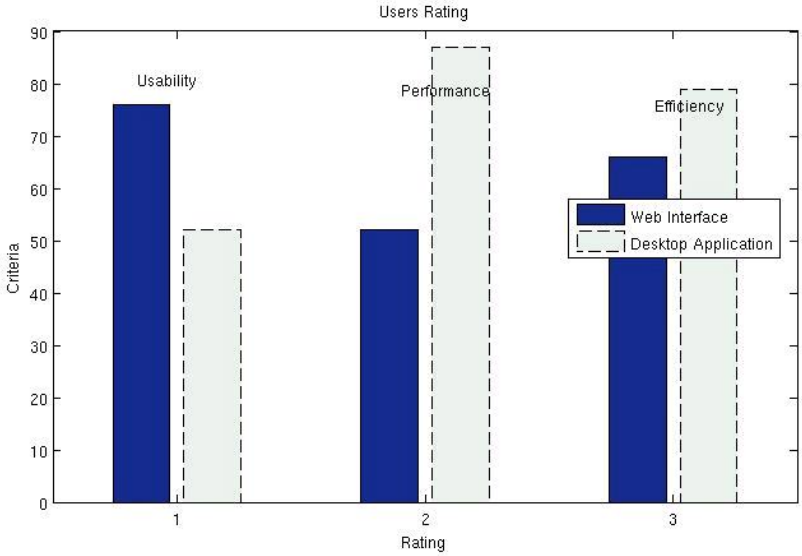
9.5 Εφαρμογή επιφάνειας εργασίας

Για την ανάπτυξη της εφαρμογής client side που αναπτύχθηκε στο επίπεδο παρουσίασης της πληροφορίας για το σύστημα PeRSSonal δόθηκε ιδιαίτερη έμφαση στα θέματα απόδοσης και εμφάνισης της πληροφορίας στο user interface. Στην παρούσα ενότητα παρουσιάζονται ορισμένα πειράματα που έγιναν για την αξιολόγηση της εφαρμογής καθώς και μερικά screenshots από την ίδια την εφαρμογή.

9.5.1 Αξιολόγηση

Προκειμένου να αξιολογήσουμε την εφαρμογή, ζητήσαμε από τους ίδιους 15 χρήστες του συστήματος στους οποίους αναφερθήκαμε και προηγουμένως να χρησιμοποιήσουν τόσο το web interface του PeRSSonal όσο και μία beta έκδοση της εφαρμογής για μία χρονική περίοδο ορισμένων ημερών. Στη συνέχεια ζητήσαμε από τους χρήστες να βαθμολογήσουν τα δύο συστήματα παρουσίασης της πληροφορίας σε ότι έχει να κάνει με: α) την χρηστικότητα και τη φιλικότητα ως προς τον χρήστη - το σύστημα δίνει αρκετές και καλές περιλήψεις;, β) την απόδοση και την διαδραστικότητα στις επιλογές - είναι οι χρόνοι απόκρισης ικανοποιητικοί; και γ) την αποδοτικότητα και την συντομία στον αναπαράσταση του περιεχομένου. Η βαθμολόγηση έγινε μια κλίματα από το ένα στο δέκα, με το δέκα να είναι το καλύτερο. Τα αποτελέσματα που προέκυψαν φαίνονται στη γραφική παράσταση του σχήματος 9.12.

Τα αποτελέσματα που παρουσιάζονται στο σχήμα 9.12 εκφράζουν την προτίμηση των χρηστών για την εφαρμογή επιφάνειας εργασίας σε ότι έχει να κάνει με τους τρόπους παρουσίασης της πληροφορίας που προσφέρει το PeRSSonal. Είναι σαφές όμως ότι η εφαρμογή υστερεί σε ότι



Σχήμα 9.12: Η αξιολόγηση των ίδιων των χρηστών για τα σύστηματα παρουσίασης του *PeRSSonal*.

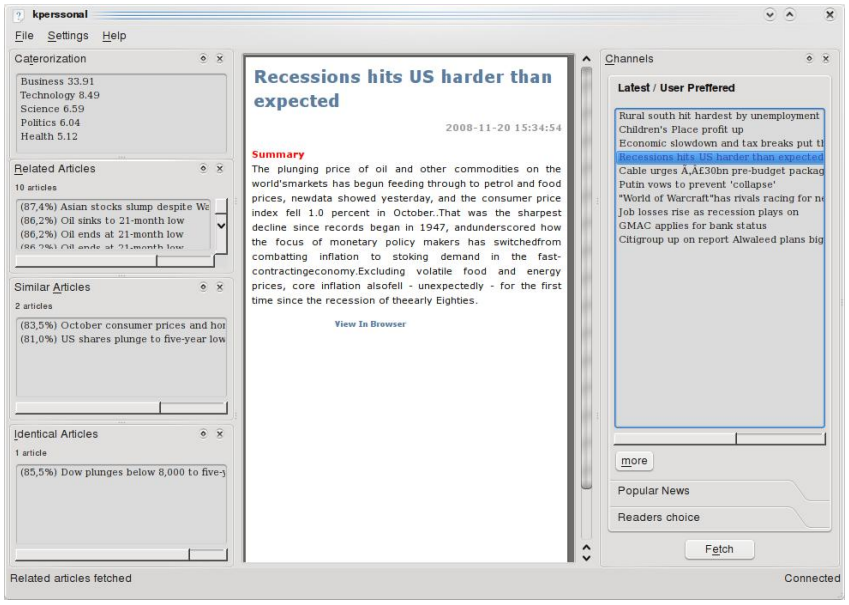
έχει να κάνει με την χρηστικότητά της κάτι που εν΄ μέρει οφείλεται στο ότι οι χρήστες συχνά δυσκολεύονται να συνηθίσουν μια νέα πλατφόρμα παρουσίασης και να εντοπίσουν γρήγορα τις δυνατότητες που προσφέρει. Αυτό είναι όμως αναμενόμενο καθότι α) απ΄ όσο γνωρίζουμε πρόκειται για την πρώτη ίσως εφαρμογή επιφάνεια εργασίας που είναι τόσο εστιασμένη στα θέματα της ανάκτησης πληροφορίας και τις δυνατότητες προσωποποίησης, και β) λαμβάνοντας υπ΄ όψιν την συμπυκνωμένη αναπαράσταση που έχει για ένα σχετικά μεγάλο όγκο πληροφορίας.

Πέρα από αυτά, σε ότι έχει να κάνει με την απόδοση και την διαδραστικότητα του συστήματος παρουσίασης, η desktop εφαρμογή υπερτερεί του web interface. Το γεγονός αυτό οφείλεται στις δυνατότητες caching και pre-fetching που η εφαρμογή desktop έχει αρχίσει να αξιοποιεί (παρότι δεν αναλύονται στην παρούσα εργασία). Επίσης, η αποδοτικότητα και η συντομία στην αναπαράσταση, ένα βασικό συστατικό του συστήματος παρουσίασης PeRSSonal αποτελεί ουσιαστικά συνάθροιση των προηγούμενων παραγόντων και δείχνει ότι παρά το γεγονός ότι η εφαρμογή βρίσκεται σε φάση ανάπτυξης, τα καινοτόμα χαρακτηριστικά που παρέχει θεωρούνται ιδιαίτερα σημαντικά από τους χρήστες ώστε να την προτιμούν.

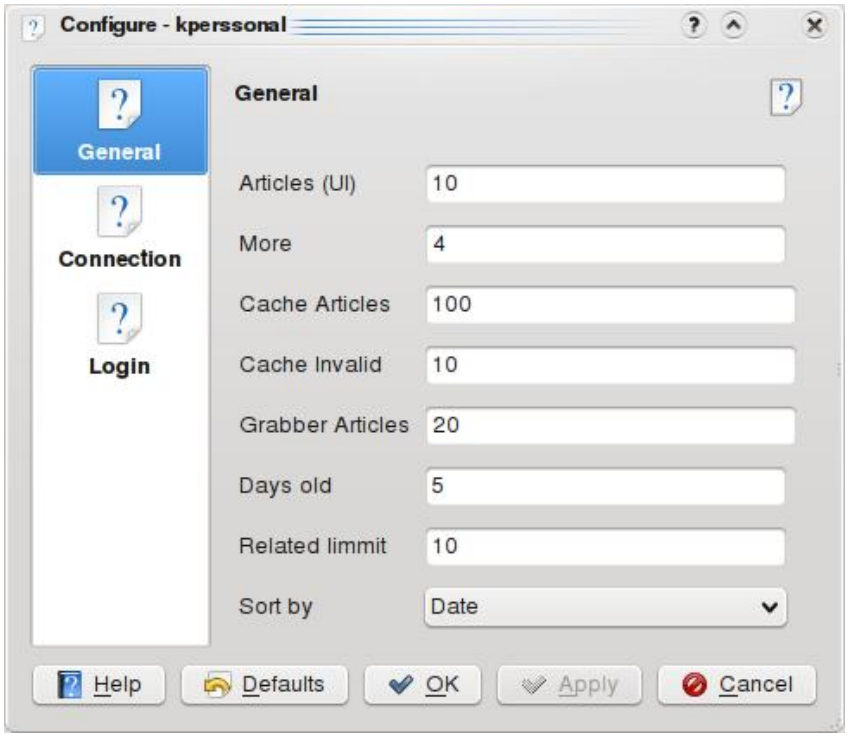
9.5.2 Παρουσίαση

Ακολουθούν ορισμένες ενδεικτικές οθόνες παρουσίασης της εφαρμογής. Στην εικόνα 9.13 φαίνεται το βασικό παράθυρο της εφαρμογής καθώς και τα κανάλια που προσφέρει. Αριστερά φαίνονται τα κανάλια των related, similar, identical άρθρων καθώς και η κατηγοριοποίηση που κάνει το σύστημα για το τρέχον άρθρο το οποίο φαίνεται στο κεντρικό μέρος της οθόνης. Δεξιά βρίσκονται τα κανάλια που μπορεί να δει ο χρήστης και που είναι τα τελευταία νέα, τα δημοφιλή και τα πιο συχνά αναγνωσμένα από τους χρήστες.

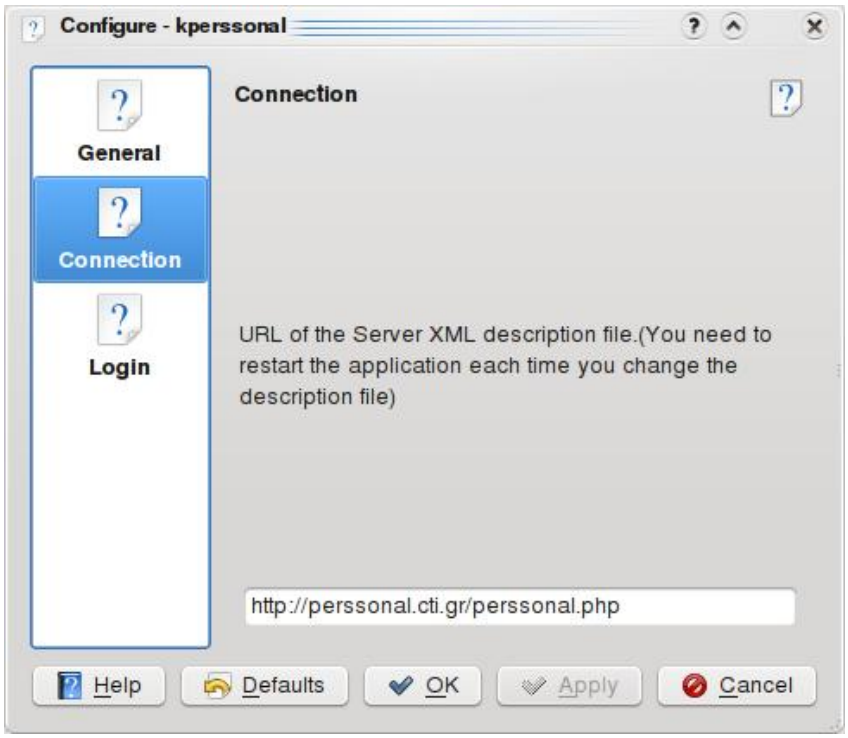
Στις εικόνες 9.14, 9.15 και 9.17 φαίνονται οι ρυθμίσεις της εφαρμογής που μπορεί να κάνει ο χρήστης.



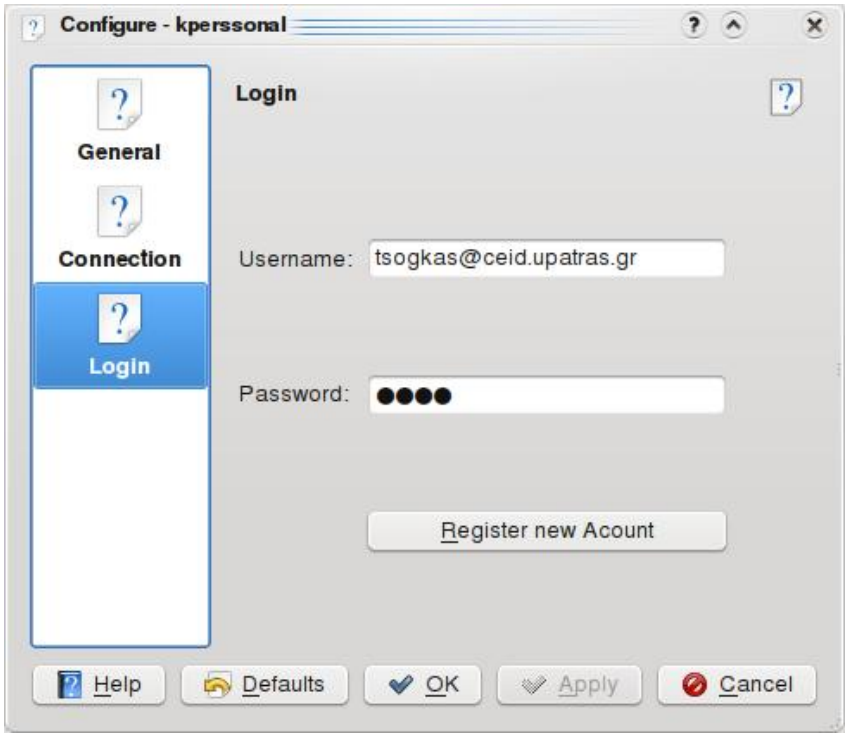
Σχήμα 9.13: Βασικό παράθυρο εφαρμογής.



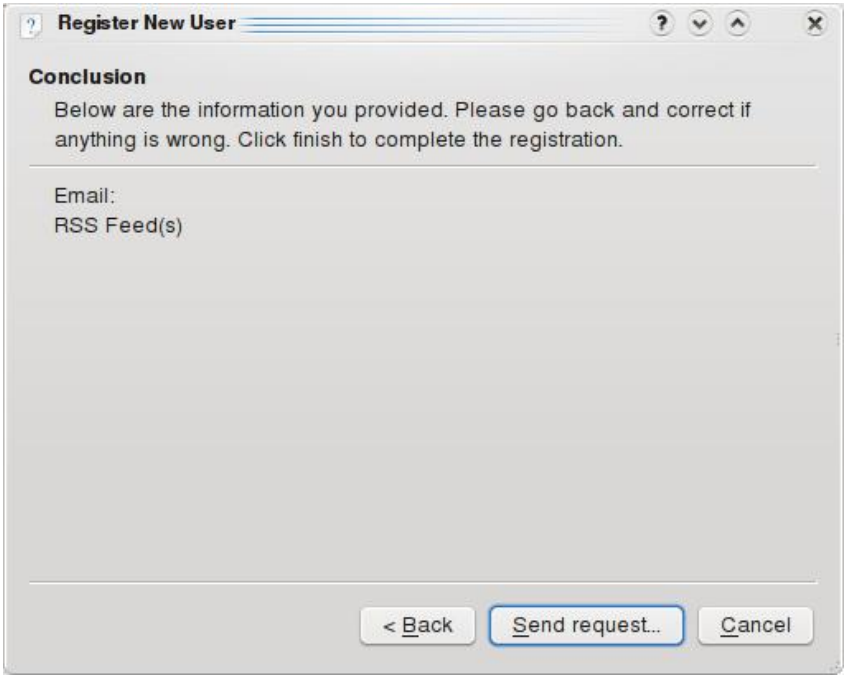
Σχήμα 9.14: Γενικές ρυθμίσεις εφαρμογής.



Σχήμα 9.15: Ρυθμίσεις σύνδεσης.



Σχήμα 9.16: Ρυθμίσεις χρήστη.



Σχήμα 9.17: Εγγραφή νέου χρήστη.

ΚΕΦΑΛΑΙΟ 10

Συμπεράσματα

> It is no measure of health to be well adjusted to a profoundly sick society.
>
> *Jiddu Krishnamurti, Indian Philosopher, 1986*

Στο παρόν κεφάλαιο παρουσιάζονται τα συμπεράσματα που προέκυψαν κατά τη διάρκεια εκπόνησης της παρούσας μεταπτυχιακής εργασίας.

Το υπερμέγεθες διαδίκτυο με την υπέρογκη πληροφορία που διακινείται σε αυτό κάνει την καθημερινή του χρήση δύσκολη για τον χρήστη. Στην εποχή μας και με τα μέσα που διαθέτει ακόμα και ο απλός χρήστης, η προσθήκη περιεχομένου στο διαδίκτυο από τον καθένα, είναι μια διαδικασία το ίδιο εύκολη με την απλή περιαγωγή στο χώρο του παγκόσμιου ιστού (π. χ. φαινόμενο blogging, το Web 2.0, κ.ο.κ.). Το πρόβλημα που δημιουργεί αυτή η ανεξέλεγκτη κατάσταση είναι ότι ακόμα οι πιο έμπειροι χρήστες καταναλώνουν πολύ χρόνο στην προσπάθεια εύρεσης πληροφορίας και συγκεκριμένα πηγών ενημέρωσης για τα θέματα που τους ενδιαφέρουν αφού κατακλύζονται από την πληροφορία.

Εστιάζοντας στο πρόβλημα του 'κατακλυσμού της πληροφορίας', επικεντρωνόμαστε σε αυτή που διακινείται στο διαδίκτυο και αφορά νέα και γεγονότα. Αυτό που θέλουμε ουσιαστικά να δημιουργήσουμε είναι ένα σύστημα το οποίο θα είναι σε θέση να παρουσιάζει, ειδήσεις που δημοσιεύονται στο διαδίκτυο, με τρόπο απλό και έχοντας στο νου μας τον παράγοντα άνθρωπο. Για να το επιτύχουμε αυτό πρέπει να παρέχουμε στο χρήστη μία υπηρεσία η οποία θα μπορεί να προσαρμόζεται σε αυτόν και να του παρέχει ποιοτικό και πλήρες περιεχόμενο για τις εξελίξεις που τον ενδιαφέρουν. Δεν στοχεύουμε στην ανάπτυξη ενός ακόμα portal νέων αφού κάτι τέτοιο δεν θα αντιμετώπιζε το πρόβλημα. Στοχεύουμε αντίθετα στο φιλτράρισμα της πληροφορίας και στην παρουσίασή της με την ελάχιστη αναγκαία ανάδραση από τον χρήστη στην επιφάνεια εργασίας του.

Το PeRSSonal είναι ένα σύστημα που δημιουργήθηκε ώστε να περνάει την πληροφορία μέσα από διάφορα στάδια επεξεργασίας. Αρχικά γίνεται το διαπέρασμα των ιστοσελίδων γνωστών news portals με χρήση των RSS Feeds που διαθέτουν και αποθηκεύεται ο html κώδικάς τους μέσω της διαδικασίας του crawling. Στη συνέχεια, από τον html κώδικα αναγνωρίζεται η χρήσιμη πληροφορία που αφορά το συγκεκριμένο άρθρο. Είναι σημαντικό σε αυτό το σημείο να απομακρύνονται

όσο το δυνατόν περισσότερα περιττά στοιχεία της σελίδας και να κρατούνται μόνο τα απαραίτητα. Ο μηχανισμός που αναπτύχθηκε γι΄ αυτό το φιλτράρισμα, βασίζεται στους αλγορίθμους που περιγράφηκαν σε προηγούμενα κεφάλαια και κάνει αρκετά καλή δουλεία. Έχει όμως αρκετά περιθώρια βελτίωσης ώστε τόσο να απομακρύνει επιβαρυντικά στοιχεία της σελίδας για το μηχανισμό (π. χ. ορισμένα scripts καταφέρνουν να περνάνε από τη διαδικασία), όσο και να μην χάνει σημαντικά τμήματα του άρθρου θεωρώντας τα ως άχρηστη πληροφορία. Σημαντικό ρόλο σε αυτό παίζει και η διαμόρφωση της σελίδας που χρησιμοποιούν τα news portal καθώς αν αυτή είναι εντελώς άναρχη και δίχως δομή είναι προφανές ότι η δουλειά του φιλτραρίσματος γίνεται δυσκολότερη.

Ακολουθώντας το φιλτράρισμα του κειμένου ο μηχανισμός προχωρά με τη διαδικασία της προεπεξεργασίας κειμένου και εξαγωγής των κωδικολέξεων. Πρόκειται για τη θεμελιώδη διεργασία σχεδόν όλων των μηχανισμών ανάκτησης πληροφορίας και επομένως επιθυμούμε τα καλύτερα δυνατά αποτελέσματα. Ο μηχανισμός προεπεξεργασίας που κατασκευάστηκε δοκιμάστηκε διεξοδικά ώστε να παράγει σωστές εξόδους και να κρατά τον πλεονασμό σε χαμηλά επίπεδα. Περιλαμβάνει τις διαδικασίες αφαίρεσης σημείων στίξης και αριθμών, ορθογραφικού ελέγχου και διόρθωσης λαθών, εύρεσης των ουσιαστικών του κειμένου, αφαίρεσης των λέξεων που ανήκουν στη λίστα των stopwords, το stemming των λέξεων και φυσικά την αντιστοίχηση των keywords που προκύπτουν με τις προτάσεις όπου αυτά εμφανίζονται. Η όλη διαδικασία κάνει εκτεταμένη χρήση regular expressions της βιβλιοθήκης boost-regex της C++ και είναι υλοποιημένη ώστε να αποφεύγονται περιττοί έλεγχοι ή επανάληψης με στόχο τη βελτίωση της απόδοσης.

Ο μηχανισμός που αναπτύχθηκε συνεχίζει με την κατηγοριοποίηση και περίληψη των άρθρων κάνοντας χρήση πολλών παραμέτρων οι οποίες μπορούν να προσαρμόζονται από το σύστημα δυναμικά ώστε να ανταποκρίνονται στο κείμενο. Τα αποτελέσματα σε κάθε περίπτωση είναι δύο ειδών: στα προσωποποιημένα που προκύπτουν δυναμικά όταν ένας χρήστης συνδέεται και τα ζητάει, και στα γενικά που αποτελούν τις εκδοχές για προβολή περιεχομένου μη-προσωποποιημένου για τον χρήστη που δεν επιθυμεί να εγγραφεί στο σύστημα.

Η παρουσίαση των αποτελεσμάτων αποτελεί ουσιαστικά και την τελική φάση στη ροή της πληροφορίας του συστήματος όπου τα αποτελέσματα μεταφέρονται χρησιμοποιώντας ανοιχτά πρωτόκολλα στην εφαρμογή χρήστη. Αυτή είναι σχεδιασμένη ώστε να παραμετροποιείται εύκολα και να δείχνει την πληροφορία για την οποία εκφράζει επιθυμία ο χρήστης. Θεωρούμε ότι μία τέτοια αντιμετώπιση θα είναι ο de facto τρόπος για την παρουσίαση της πληροφορίας στο χρήστη για τα χρόνια που ακολουθούν λαμβάνοντας υπ΄ όψιν τη συνδυαστική έκρηξη του διαδικτύου.

ΚΕΦΑΛΑΙΟ 11

Μελλοντική εργασία

> Ignorance, the root and the stem of every evil.
>
> *Plato, Greek Philosopher, 347 BC*

Στο τελευταίο αυτό κεφάλαιο, δίνονται ορισμένες μελλοντικές κατευθύνσεις για έρευνα στα πεδία ενδιαφέροντος καθώς και επεκτάσεις που δέχεται το σύστημα που αναπτύχθηκε. Πολλά από αυτά που παρουσιάζονται έχουν ήδη ξεκινήσει να ερευνόνται και να ενσωματώνονται στο PeRSSonal.

Όπως έχει αναφερθεί ήδη, το σύστημα PeRSSonal αναπτύχθηκε με γνώμονα την ευκολία στη βελτίωση καθώς κάθε τμήμα του μπορεί να λειτουργεί ανεξάρτητα από τα υπόλοιπα αρκεί να ακολουθούνται οι παραδοχές επικοινωνίας τους μέσω της κεντρικοποιημένης βάσης δεδομένων. Ως εκ' τούτου, καθένα υποσύστημα μπορεί να βελτιωθεί ή ακόμα και να αντικατασταθεί με κάποια καλύτερη τεχνική η οποία εκτιμάται ότι μπορεί να δώσει καλύτερα αποτελέσματα. Με αυτή την λογική είναι εφικτή και η εκτεταμένη αξιολόγηση καθενός τμήματος σε σύγκριση με κάτι προηγούμενο.

Το τμήμα της εξαγωγής κωδικολέξεων αναμένεται να βελτιωθεί μελλοντικά στα παρακάτω σημεία:

- Χρήση καλύτερων λίστών από stopwords οι οποίες ούτε θα απορρίπτουν σημαντικές λέξεις ούτε και θα δέχονται άλλες μη χρήσιμες. Μπορεί π. χ. να χρησιμοποιηθεί μια δυναμική λίστα που προσαρμόζεται ανάλογα με τη θεματολογία του κειμένου.
- Επέκταση του όλου μηχανισμού σε ένα ενοποιημένο πολυγλωσσικό περιβάλλον τεχνικών ανάκτησης πληροφορίας.
- Υποστήριξη για πολυμεσικό περιεχόμενο

Όσον αφορά στην μετατροπή του μηχανισμού σε πολυγλωσσικό σύστημα επεξεργασίας κειμένων, αυτό θα πρέπει να περιλαμβάνει λίστες με stopwords σε διάφορες γλώσσες, κανόνες stemming για την εκάστοτε γλώσσα καθώς και τα λεξικά των γλωσσών. Οι πληροφορίες αυτές θα πρέπει να αποθηκεύονται κεντρικά στη ΒΔ και να είναι διαθέσιμες on the fly κατά την εκτέλεση του μηχανισμού και αφού αναγνωριστεί η γλώσσα του κειμένου. Η βάση για την πολύγλωσση υποστήριξη

του μηχανισμού έχει ήδη κατασκευαστεί, αυτό που απαιτείται είναι η εύρεση ή η δημιουργία των κατάλληλων: α) κανόνων stemming, β) ορθογραφικών λεξικών, γ) λιστών stopwords, δ) κανόνων αναγνώρισης μερών του λόγου. Προκειμένου τα σύστημα να μπορεί να δεικτοδοτήσει και ελληνικά κείμενα, κινούμαστε άμεσα στην επέκτασή του τουλάχιστον για την ελληνική γλώσσα.

Όσον αφορά στα υποσυστήματα κατηγοριοποίησης και αυτόματης εξαγωγής περίληψης, στοχεύουμε σε ανάπτυξη πολλαπλών διαφορετικών αλγορίθμων μια και αυτά τα υποσυστήματα είναι ανεξάρτητα από τον υπόλοιπο μηχανισμό (modules). Επίσης στοχεύουμε στην ανάπτυξη ενός γραφικού περιβάλλοντος διαχείρισης για τον όλο μηχανισμό με έφεση την εύκολη πρόσβαση και διαχείρισή του. Το server-side caching είναι επίσης μία ενδιαφέρουσα προέκταση για τη βελτιστοποίηση της απόκρισης του συστήματος, ένα χαρακτηριστικό που ήδη μελετάται.

Η εφαρμογή client side παρουσίασης που αναπτύχθηκε μπορεί να βελτιωθεί σημαντικά στους τομείς της απόκριση αξιοποιώντας δυνατότητες τοπικής αποθήκευσης αποτελεσμάτων (client-side caching) αλλά και πρόβλεψης των επιλογών του χρήστη και prefetching αποτελεσμάτων βάσει του προφίλ χρήστη. Ήδη σε αυτό το τμήμα η έρευνα βρίσκεται σε εξέλιξη και τα πρώτα αποτελέσματα είναι ενθαρυντικά.

Πέρα από τους παραπάνω επιμέρους τομείς επέκτασης, βασικός στόχος είναι η διαρκή επέκταση και βελτίωση του μηχανισμού με αξιοποίηση ενδιαφέροντων ερευνητικών αποτελεσμάτων που προκύπτουν είτε από την δική μας ερευνητική δραστηρίοτητα ή και γενικότερα.

Ευρετήριο

Βιβλιογραφία

[1] Amazon. Online shopping. http://www.amazon.com.

[2] The apache web server. Website. http://httpd.apache.org/.

[3] Bbc news. News portal. http://news.bbc.co.uk/.

[4] Boost libraries for c++. Website. http://www.boost.org/.

[5] Cgicc. a c++ class library for writing cgi applications. Website. http://www.cgicc.org/.

[6] Cnn. News portal. http://www.cnn.com/.

[7] Dick hardt. how sxip works (whitepaper). Website. https://sxip.org/docs/specs/how-sxip-works.pdf2004.

[8] Dig internet search engine software. Website. http://www.htdig.org/.

[9] Dom. Document Object Model. http://www.w3.org/DOM/.

[10] Expat. XML parsing library. http://expat.sourceforge.net/.

[11] foxnews.com. News portal. http://www.foxnews.com/.

[12] gartner.com. Website. http://www.gartner.com/.

[13] Gentoo linux. Website. http://www.gentoo.org/.

[14] The gnu compiler collection. Website. http://www.netbeans.org/.

[15] Gnu wget - gnu project - free software foundation. Website. http://www.gnu.org/software/wget/.

[16] Google. Search engine. http://www.google.com/.

[17] Google news. News Portal. http://news.google.com/.

[18] Google shopping. Website. http://froogle.google.com/.

[19] Gun aspell. spell checker. http://aspell.net/.

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

[21] Httrack website copier - offline browser. Website. http://www.httrack.com/.

[22] Internetworldstats.com. Internet Statistics. http://www.internetworldstats.com/stats.htm.

[23] Java, java history. http://ils.unc.edu/blaze/java/javahist.html.

[24] Kdevelop. integrated development environment for unix, supporting kde/qt, c/c++ and many other languages. Website. http://www.kdevelop.org/.

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

[26] libcurl. Curl grabber. http://curl.haxx.se/.

[27] libstemmer. Multi-language stemmers. http://snowball.tartarus.org/download.php.

[28] Methabot web crawler. Website. http://bithack.se/methabot/.

[29] M.k.bergman. the deep web: Surfacing hidden value. Website. http://www.press.umich.edu/jep/07-01/bergman.html.

[30] Msn shopping. Website. http://shopping.msn.com/.

[31] Mysq++ c++ api interface to the mysql database. Website. http://www.mysql.org/downloads/api-mysql++.html.

[32] Mysql, opensource database. http://www.mysql.com.

[33] Nutch open source web search engine. Website. http://lucene.apache.org/nutch/.

[34] Open directory project. Website. http://www.dmoz.org.

[35] passport.net. Website. http://www.passport.net.

[36] Personews. Website. http://news.csd.auth.gr/.

[37] The php language runtime engine: Cli, cgi and apache2 sapis. Website. http://www.php.net/.

[38] The porter stemmer algorithm. Website. http://www.tartarus.org/~martin/PorterStemmer/.

[39] Postgresql, opensource database. http://www.postrgresql.org.

[40] Reuters. News portal. http://www.reuters.com/.

[41] Rss - real simple syndication. Website. http://www.w3.org/WAI/highlights/about-rss.html.

[42] trolltech.com. Trolltech's Qt. http://trolltech.com/products/.

[43] W3c - xml protocol. Website. http://www.w3.org/XML/.

[44] Web information retrieval environment (wire). Website. http://www.cwr.cl/projects/WIRE/.

[45] Webcrawler, the world's search engines spun together. Website. http://www.webcrawler.com/.

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

[47] Www size. Website. http://www.worldwidewebsize.com/.

[48] Xerces. A validating XML parser. http://xerces.apache.org/xerces-c/.

[49] yahoo.com. Search engine. http://www.yahoo.com/.

[50] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, R. Hoberman, and D. Caputo. Topic-based novelty detection: 1999 summer workshop at CLSP, final report, 1999.

[51] E. Amitay. Using common hypertext links to identify the best phrasal description of target web documents. In *Proceedings of the SIGIR*, volume 98, 1998.

[52] C. Apté, F. Damerau, and S.M. Weiss. *Towards language independent automated learning of text categorization models.* Springer-Verlag New York, Inc. New York, NY, USA, 1994.

[53] A. Arasu and H. Garcia-Molina. Extracting structured data from Web pages. *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 337–348, 2003.

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

[55] E. Banos, I. Katakis, N. Bassiliades, G. Tsoumakas, and I. Vlahavas. PersoNews: A Personalized News Reader Enhanced by Machine Learning and Semantic Filtering. *LECTURE NOTES IN COMPUTER SCIENCE*, 4275:975, 2006.

[56] R. Barzilay and L. Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. *Proceedings of HLT-NAACL 2004*, pages 113–120, 2004.

[57] N.J. Belkin and W.B. Croft. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, 35(12):29–38, 1992.

[58] VD Belur. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. *IEEE Computer Society Press, New York: IEEE press*, 1991.

[59] H. Berger and D. Merkl. A Comparison of Text-Categorization Methods applied to N-Gram Frequency Statistics. *Proc. of the 17th Australian Joint Conf. on Artificial Intelligence*, 2004.

[60] D. Bergmark. Collection synthesis. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 253–262. ACM New York, NY, USA, 2002.

[61] D. Bergmark, C. Lagoze, and A. Sbityakov. Focused Crawls, Tunneling, and Digital Libraries. *LECTURE NOTES IN COMPUTER SCIENCE*, pages 91–106, 2002.

[62] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic Web. *Scientific American*, 284(5):28–37, 2001.

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

[64] C. Bouras, C. Dimitriou, V. Poulopoulos, and V. Tsogkas. The importance of the difference in text types to keyword extraction: Evaluating a mechanism. In Hamid R. Arabnia, editor, *International Conference on Internet Computing*, pages 43–49. CSREA Press, 2006.

[65] C. Bouras, V. Poulopoulos, and V. Tsogkas. Efficient summarization based on categorized keywords. In *DMIN*, pages 285–291, 2007.

[66] C. Bouras, V. Poulopoulos, and V. Tsogkas. Personalizing text summarization based on sentence weighting. In *IADIS European First International Conference Data Mining (ECDM 2007), Lisbon, Portugal*, pages 3 – 10, 2007.

[67] C. Bouras, V. Poulopoulos, and V. Tsogkas. Creating dymanic personalized rss summaries. In *8th Industrial Conference on Data Mining – ICDM 2008, , Leipzig, Germany*. Springer, 2008.

[68] C. Bouras, V. Poulopoulos, and V. Tsogkas. PeRSSonal, the Automatic Summarization, Text Categorization, Personalized Syndication, System. *Handbook of Research on Social Interaction Technologies and Collaboration Software: Concepts and Trends*, 2008.

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

[70] C. Bouras and V. Tsogkas. Improving text summarization using noun retrieval techniques. In *Advanced Knowledge – based Systems, Invited Session of the 12nd International Conference on Knowledge – based and Intelligent Information & Engineering Systems(KES 2008), Zagreb, Croatia*. LNCS, 2008.

[71] R. Brandow, K. Mitze, and L.F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management: an International Journal*, 31(5):675–685, 1995.

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

[73] G. Cao. Support Vector Machine Active Learning with Applications to Text Classification.

[74] M.F. Caropreso, S. Matwin, and F. Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Text Databases and Document Management: Theory and Practice*, pages 78–102, 2001.

[75] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases. *Proceedings of the 23rd VLDB Conference*, pages 446–455, 1997.

[76] P.K. Chan. Constructing Web User Profiles: A non-invasive Learning Approach. *KDD-99 Workshop on Web Usage Analysis and User Profiling*, pages 7–12, 1999.

[77] K.C.C. Chang, B. He, and Z. Zhang. Toward large scale integration: Building a metaquerier over databases on the web. *Proc. of CIDR*, pages 44–55, 2005.

[78] W.W. Cohen. Text categorization and relational learning. *Proceedings of ICML-95, 12th International Conference on Machine Learning*, pages 124–132, 1995.

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

[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] L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle. The Platform for Privacy Preferences 1.0 (P3P1. 0) Specification. *W3C Recommendation*, 16, 2002.

[82] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pages 133–140. Association for Computational Linguistics Morristown, NJ, USA, 1992.

[83] B.D. Davison. Topical locality in the Web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–279. ACM New York, NY, USA, 2000.

[84] R.L. Donaway, K.W. Drummey, and L.A. Mather. A comparison of rankings produced by summarization evaluation measures. *ANLP/NAACL Workshops*, pages 69–78, 2000.

[85] S. Dumais and H. Chen. Hierarchical classification of Web content. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263, 2000.

[86] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155, 1998.

[87] J. Edwards, K. McCurley, and J. Tomlin. An adaptive model for optimizing performance of an incremental web crawler. *Proceedings of the 10th international conference on World Wide Web*, pages 106–113, 2001.

[88] B. Endres-Niggemeyer et al. *Summarizing information*. Springer New York, 1998.

[89] 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.

[90] T. Firmin and M.J. Chrzanowski. An Evaluation of Automatic Text Summarization Systems. *Advances in Automatic Text Summarization*, pages 325–336, 1999.

[91] Gary Flake, Steve Lawrence, and Lee L. Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August February0–FebruaryMarch 2000.

[92] W.B. Frakes and R. Baeza-Yates. *Information retrieval: data structures and algorithms*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1992.

[93] W. Frawley, G. Piatesky-Shapiro, and C. Matheus. An Overview In Knowledge Discovery in Databases AAAI, 1991.

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

[95] D. Fum, G. Guida, and C. Tasso. Forward and backward reasoning in automatic abstracting. *Proceedings of the 9th conference on Computational linguistics-Volume 1*, pages 83–88, 1982.

[96] J. Furnkranz, T. Mitchell, and E. Riloff. A case study in using linguistic phrases for text categorization on the WWW. *Learning for Text Categorization: Proceedings of the 1998 AAAI/ICML Workshop*, pages 98–05, 1998.

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

[98] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web communities from link topology. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, pages 225–234. ACM Press New York, NY, USA, 1998.

[99] J. Gimenez and L. Marquez. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 43–46, 2004.

[100] J. Goecks and J. Shavlik. Automatically Labeling Web Pages Based on Normal User Actions. *Proc. of the Intl. Conf. on Intelligent User Interfaces*.

[101] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128, 1999.

[102] U. Hahn and U. Reimer. Semantic Parsing and Summarizing of Technical Texts in the TOPIC System. *Informationslinguistik*, pages 153–193, 1986.

[103] P. Hensley, M. Metral, U. Shardanand, D. Converse, and M. Myers. Proposal for an Open Profiling Standard. *Technical Note, World Wide Web Consortium, June*, 1997.

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

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

[106] PS Jacobs and L.F. Rau. SCISOR: extracting information from on-line news. *Communications of the ACM*, 33(11):88–97, 1990.

[107] C. Jacquemin. *Spotting and Discovering Terms Through Natural Language Processing*. MIT Press, 2001.

[108] T. Joachims. *Text Categorization with Suport Vector Machines: Learning with Many Relevant Features*. Springer-Verlag London, UK, 1998.

[109] T. Joachims, D. Freitag, and T. Mitchell. WebWatcher: A Tour Guide for the World Wide Web. *Proceedings of IJCAI97*, pages 1–7, 1997.

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

[111] F. Karlsson. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Walter de Gruyter, 1995.

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

[113] J.O.N.M. KLEINBERG. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[114] V. Kluev. Compiling document collections from the Internet. In *ACM SIGIR Forum*, volume 34, pages 9–14. ACM Press New York, NY, USA, 2000.

[115] G. Klyne, F. Reynolds, C. Woodrow, H. Ohto, J. Hjelm, M. Butler, L. Tran, et al. Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies. *W3C Working Draft*, 8, 2002.

[116] K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, 2002.

[117] Y. Koike, T. Kamba, and M. Langheinrich. PIDL-Personalized Information Description Language. *W3C Note, 09*.

[118] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 170–178, 1997.

[119] R. Krovetz. Viewing morphology as an inference process. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202, 1993.

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

[121] D.D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. *Third Annual Symposium on Document Analysis and Information Retrieval*, pages 81–93, 1994.

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

[123] A. Maedche and S. Staab. Ontology learning for the Semantic Web. *Intelligent Systems, IEEE*, 16(2):72–79, 2001.

[124] I. Mani and M. Maybury. *Advances in automatic text summarization*. The MIT Press, 1999.

[125] I. Mani and G. Wilson. Robust temporal processing of news. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 69–76, 2000.

[126] D. Marcu. The rhetorical parsing of natural language texts. *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, pages 96–103, 1997.

[127] B. Masand. Lino, G., & Waltz, D.(1992). Classifying news stories using memory based reasoning. *Proceedings of 506 15th ACM SIGIR international conference on research and development in information retrieval*, pages 59–65.

[128] L.A. Mather and J. Note. Discovering Encyclopedic Structure and Topics in Text. *Sixth ACM SIGKDD*.

[129] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*, 752, 1998.

[130] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *AAAI Spring Symposium on Intelligent Agents in Cyberspace 1999*, 1999.

[131] T. Mitchell, B. Buchanan, G. DeJong, T. Dietterich, P. Rosenbloom, and A. Waibel. Machine Learning. *Annual Review of Computer Science*, 4(1):417–433, 1990.

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

[133] B. Mobasher, R. Cooley, and J. Srivastava. Automatic Personalization Through Web Usage Mining. 2000.

[134] MC Mont, S. Pearson, and P. Bramhall. Towards accountable management of identity and privacy: sticky policies and enforceable tracing services. *Database and Expert Systems Applications, 2003. Proceedings. 14th International Workshop on*, pages 377–382, 2003.

[135] M. Montes-y Gómez, A. Gelbukh, and A. López-López. Mining the News: Trends, Associations, and Deviations. *Computación y Sistemas*, 5(1):14–24, 2001.

[136] C. Mooers. Information retrieval viewed as temporal signalling. *International Congress of Mathematicians. Cambridge, Mass., 1950. Proceedings*, 1951.

[137] A.H. Morris, G.M. Kasper, and D.A. Adams. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, 3(1):17–35, 1992.

[138] M. Najork and J.L. Wiener. Breadth-first crawling yields high-quality pages. In *Proceedings of the 10th international conference on World Wide Web*, pages 114–118. ACM Press New York, NY, USA, 2001.

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

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

[141] M. Oka and Y. Ueda. Evaluation of Phrase-representation Summarization based on Information Retrieval Task. *Proceedings of the Workshop on Automatic Summarization, ANLP-NAACL2000*, 2000.

[142] M.E. Okurowski, H. Wilson, J. Urbina, T. Taylor, R.C. Clark, and F. Krapcho. Text summarizer in use: Lessons learned from real world deployment and evaluation. *Proceedings of the Workshop on Automatic Summarization, ANLP-NAACL2000*, 2000.

[143] C.D. Paice. Another stemmer. *ACM SIGIR Forum*, 24(3):56–61, 1990.

[144] CD Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management: an International Journal*, 26(1):171–186, 1990.

[145] M. Pazzani, J. Muramatsu, and D. Billsus. Syskill & Webert: Identifying interesting web sites. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 5461, 1996.

[146] B. Pinkerton. Finding What People Want: Experiences with the WebCrawler. *Proceedings of the Second International World Wide Web Conference*, 1994.

[147] M. Porter. The Porter Stemming Algorithm. *Accessible at http://www. tartarus. org/-martin/PorterStemmer.*

[148] S. Raghavan and H. Garcia-Molina. Crawling the Hidden Web. *PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES*, pages 129–138, 2001.

[149] GJ Rath, A. Resnick, and TR Savage. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143, 1961.

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

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

[152] J. Rissanen. *Stochastic Complexity in Statistical Inquiry Theory.* World Scientific Publishing Co., Inc. River Edge, NJ, USA, 1989.

[153] H. Saggion and G. Lapalme. Concept identification and presentation in the context of technical text summarization. *ANLP/NAACL Workshops*, pages 1–10, 2000.

[154] G. Salton, J. Allan, C. Buckley, and A. Singhal. Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts. *Science*, 264(5164):1421, 1994.

[155] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc. New York, NY, USA, 1986.

[156] B. Sankaran. Tamil Search Engine.

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

[158] M. Saravanan, Pc Reghu Raj, and S. Raman. Summarization and Categorization of text data in high-level data cleaning for information retrieval. *Applied Artificial Intelligence*, 17(5):461–474, 2003.

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

[160] N. Shadbolt, T. Berners-Lee, and W. Hall. The Semantic Web Revisited. *IEEE INTELLIGENT SYSTEMS*, pages 96–101, 2006.

[161] J. Shavlik, S. Calcari, T. Eliassi-Rad, and J. Solock. An instructable, adaptive interface for discovering and monitoring information on the World-Wide Web. *Proceedings of the 4th international conference on Intelligent user interfaces*, pages 157–160, 1998.

[162] S. Sizov, M. Biwer, J. Graupmann, S. Siersdorfer, M. Theobald, G. Weikum, and P. Zimmer. The BINGO! System for Information Portal Generation and Expert Web Search. *Conference on Innovative Systems Research (CIDR)*, 2003.

[163] N. Slonim and N. Tishby. The power of word clusters for text classification. *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research*, 2001.

[164] R. Swan and J. Allan. Automatic generation of overview timelines. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, 2000.

[165] R.R. Trujillo and A. Ardö. *Simulation tool to study focused web crawling strategies.* 2006.

[166] K. Tzeras and S. Hartmann. Automatic indexing based on Bayesian inference networks. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 22–35, 1993.

[167] V.N. Vapnik. *Statistical learning theory.* Wiley New York, 1998.

[168] J. Verbeek. An information theoretic approach to finding word groups for text classification. *Institute for Lang uage, L og ic and Comp utation, University of Am sterdam*, 2000.

[169] J. Wang, J.R. Wen, F. Lochovsky, and W.Y. Ma. Instance-based schema matching for web databases by domain-specific query probing. *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 408–419, 2004.

[170] D.H. Widyantoro, T.R. Ioerger, and J. Yen. Learning user interest dynamics with a three-descriptor representation. *Journal of the American Society for Information Science and Technology*, 52(3):212–225, 2001.

[171] C. Wongchokprasitti and P. Brusilovsky. NewsMe: A Case Study for Adaptive News Systems with Open User Model. In *Proceedings of the Third International Conference on Autonomic and Autonomous Systems*. IEEE Computer Society Washington, DC, USA, 2007.

[172] WA Woods and JG Schmolze. The KL-ONE family. Semantic Networks in Artificial Intelligence, Pp133-178, 1992.

[173] P. Wu, J.R. Wen, H. Liu, and W.Y. Ma. Query selection techniques for efficient crawling of structured web sources. *Proc. of ICDE*, 2006.

[174] Y. Yang. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 13–22, 1994.

[175] Y. Yang and C.G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems (TOIS)*, 12(3):252–277, 1994.

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

[177] S.R. Young and P.J. Hayes. Automatic classification and summarization of banking telexes. *Proceedings of the Second Conference on Artificial Intelligence Applications*, pages 402–408, 1985.

[178] D. Zeinalipour-Yazti and M. Dikaiakos. Design and implementation of a distributed crawler and filtering processor. *Proc. of NGITS 2002*, 2382:58–74.

[179] H. Zhang. The optimality of naive Bayes. In *Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference*, pages 562–567, 2004.