Πανεπιστήμιο Κρήτης
Σχολή Θετικών και Τεχνολογικών Επιστημών
Τμήμα Επιστήμης Υπολογιστών

Σύστημα ημι-αυτόματης κατηγοριοποίησης του περιεχομένου που συγκεντρώνει ένας δικτυακός τόπος από διαφορετικές πηγές

Ζωή Πολιτοπούλου

Μεταπτυχιακή Εργασία

Ηράκλειο, Δεκέμβριος 2005

Σύστημα ημι-αυτόματης κατηγοριοποίησης του περιεχομένου που συγκεντρώνει ένας δικτυακός τόπος από διαφορετικές πηγές

Εργασία που υποβλήθηκε από την
Πολιτοπούλου Ζωή
ως μερική εκπλήρωση των απαιτήσεων για την απόκτηση
ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

Συγγραφέας:

Πολιτοπούλου Ζωή
Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Εισηγητική Επιτροπή:

Χρήστος Νικολάου
Καθηγητής, Επόπτης

Δημήτριος Πλεξουσάκης
Αναπληρωτής Καθηγητής, Μέλος

Ηλίας Χούστης
Καθηγητής Πανεπιστημίου Θεσσαλίας, Μέλος

Δεκτή:

Δημήτριος Πλεξουσάκης, Αναπλ. Καθηγητής
Πρόεδρος Επιτροπής Μεταπτυχιακών Σπουδών

Ηράκλειο, Δεκέμβριος 2005

Περίληψη

Η ανάπτυξη του διαδικτύου έχει προσφέρει στους χρήστες του μια πληθώρα πληροφοριών, η πλειονότητα των οποίων βρίσκεται διαθέσιμη σε μορφή κειμένου. Ο χρήστης είναι καθημερινά αντιμέτωπος με τα προβλήματα της αυτόματης λήψης αυτού του υλικού και της οργάνωσης του σε κατηγορίες με έναν αυτόματο ή ημι-αυτόματο τρόπο και φυσικά πριν το υλικό αυτό φτάσει να χρησιμοποιηθεί. Τα Συστήματα Διαχείρισης Περιεχομένου (ΣΔΠ) συνεισφέρουν σε αυτή την προσπάθεια δίνοντας στους χρήστες έναν ενιαίο τρόπο οργάνωσης και παρουσίασης του περιεχομένου τους. Κανένα όμως από τα υπάρχοντα ΣΔΠ δεν παρέχει έναν αυτόματο ή ημι-αυτόματο τρόπο για την λήψη και κατηγοριοποίηση των κειμένων. Σε αυτό το σημείο έρχονται να συνεισφέρουν οι γνωστές από παλιότερα Τεχνικές Κατηγοριοποίησης Κειμένου (Text Categorization – TC), οι οποίες μετεξελίσσονται και προσαρμόζονται στα δεδομένα του διαδικτύου.

Η παρούσα εργασία ασχολείται με όλο το πρόβλημα της ανάκτησης και κατηγοριοποίησης κειμένου, ξεκινώντας από την διαδικασία λήψης από την πηγή είτε αυτή είναι κείμενο, είτε RSS, είτε το αποτέλεσμα της αναζήτησης μέσω Google API. Περιγράφει καταρχήν την διαδικασία λήψης και διαχείρισης κειμένου από τις παραπάνω πηγές. Μετέπειτα ασχολείται με την λεξικογραφική ανάλυση του κειμένου και παρουσιάζει έναν αλγόριθμο ο οποίος χρησιμοποιείται για την συγκεκριμένη εργασία. Αφού παραχθούν οι λέξεις κλειδιά κάθε κειμένου μετά οι λέξεις αυτές δίνονται στον αλγόριθμο κατηγοριοποίησης, ο οποίος προσπαθεί να εντάξει το υπό εξέταση κείμενο κάτω από μία από τις διαθέσιμες κατηγορίες. Οι κατηγορίες αυτές προκύπτουν από τη χρήση του προτύπου καταλόγου DMOZ, ο οποίος αποτελεί και την βάση των περισσότερων μηχανών αναζήτησης. Οι κατηγορίες περιγράφονται σε αντίστοιχες ιεραρχίες και προσφέρουν έναν τυποποιημένο και καθολικό τρόπο για να τις αναφέρει κανείς ή να περιγράψει πραγματικά αντικείμενα, ενέργειες, έγγραφα κτλ. ενώ μπορούν να χρησιμοποιηθούν και για την περιγραφή των χαρακτηριστικών του περιεχομένου ενός αντικειμένου. Ο συνδυασμός ενός αλγορίθμου κατηγοριοποίησης για το διαδίκτυο και του προτύπου καταλόγου DMOZ αποτελεί την πρώτη προσπάθεια που αναφέρεται στη βιβλιογραφία.

Τέλος η διαδικασία της λήψης κειμένου από το διαδίκτυο και της μετέπειτα κατηγοριοποίησης του μελετάται στα πλαίσια λειτουργίας ενός ΣΔΠ και μάλιστα ενός ΣΔΠ Ελεύθερου Λογισμικού / Ανοικτού Κώδικα, κάτι το οποίο παρουσιάζεται για πρώτη φορά στη βιβλιογραφία. Παρουσιάζεται η δυνατότητα ενσωμάτωσης της διαδικασίας λήψης και κατηγοριοποίησης κειμένου στις διαθέσιμες προς τον τελικό χρήστη δυνατότητες του ΣΔΠ και η ανταλλαγή πληροφορίας με τα υπόλοιπα μέρη του συστήματος καθώς και τα προβλήματα και οι περιορισμοί που αντιμετωπίστηκαν εξαιτίας της ενσωμάτωσης στο ΣΔΠ. Τέλος παρουσιάζονται κάποια σενάρια

χρήσης από όπου κανείς μπορεί να συμπεράνει πως η διαδικασία παράγει ικανοποιητικά αποτελέσματα και μπορεί να χρησιμοποιηθεί σε ένα περιβάλλον παραγωγής. Η όλη εργασία μπορεί να επεκταθεί με την χρήση περισσότερο εμπλουτισμένων περιγραφών μεταδεδομένων για τις κατηγορίες αλλά και την χρήση της γνώσης που αποκτάται από το σύστημα από τις διαδοχικές κατηγοριοποιήσεις. Τέλος θα είχε ιδιαίτερο ενδιαφέρον η παροχή της συγκεκριμένης διαδικασίας σαν ηλεκτρονικής υπηρεσίας (web service) ώστε να είναι εφικτή η χρήση της και από άλλα ΣΔΠ.

Abstract

The Internet era of the information management is synonymous of huge amounts of knowledge available to everybody via a web browser. To manage the continuously increasing available information, businesses use the so-called Content Management Systems (CMS) that provide effective technology-based support to enterprise and web content management. Small enterprises and individual Internet users have the possibility to use similar systems, a short of "low power" CMS, for the same purposes (from open source content management systems to blogs). The planet of content management is however under stress due to its evolution towards a more decentralized and collaborative platform. Many analysts describe this evolution under the term "Web 2.0": "The Web is shifting from an international library of interlinked pages to an information ecosystem, where data circulate like nutrients in a rain forest". Strong drivers for this include the rise of personal and interpersonal content management (as a social process) and the increasing possibilities for information and content aggregation. This work builds on these trends to design and develop a back-end system for a Content Management System (Web Content Classifier) able to capture content from the Web and effectively categorize it. Such a system offers valuable "social software" functionality to personal and interpersonal CMS thus augmenting their attractiveness to users.

Ευχαριστίες

Με την παράδοση αυτής της μεταπτυχιακής εργασίας μια πορεία περίπου 8 χρόνων φτάνει στο τέλος της, τουλάχιστον με το ρόλο της φοιτήτριας, στο Πανεπιστήμιο Κρήτης. Με την ευκαιρία αυτή λοιπόν θα ήθελα να ευχαριστήσω ορισμένους ανθρώπους που με βοήθησαν να φτάσω μέχρι εδώ.

Πρώτον απ' όλους θα ήθελα να ευχαριστήσω τον κ. Πέτρο Καβάσσαλη ο οποίος με εμπιστεύτηκε πριν πέντε χρόνια και με ενέταξε στο δυναμικό της ομάδας ΑΤΛΑΝΤΙΔΑ του Πανεπιστημίου Κρήτης. Στο εσωτερικό της ομάδας έμαθα πολλά και κυρίως να ακολουθώ ένα δομημένο τρόπο σκέψης.

Στη συνέχεια θα ήθελα να ευχαριστήσω τα μέλη της επιτροπής που υποστήριξαν και ενέκριναν αυτή τη μεταπτυχιακή εργασία, τον κ. Χρήστο Νικολάου καθηγητή του Τμήματος Επιστήμης Υπολογιστών, τον κ. Δημήτρη Πλεξουσάκη αναπληρωτή καθηγητή του Τμήματος Επιστήμης Υπολογιστών καθώς και τον κ. Ηλία Χούστη καθηγητή στο Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας. Ο κ. Χρήστος Νικολάου αν και θα απουσίαζε στην Αμερική για ένα μεγάλο διάστημα με εμπιστεύτηκε και δέχτηκε να είναι ο επόπτης αυτής της μεταπτυχιακής εργασίας. Τον κ. Πλεξουσάκη τον ευχαριστώ όχι μόνο επειδή ήταν μέλος αυτής της επιτροπής αλλά και για όσα μου έχει διδάξει όλα αυτά τα χρόνια μέσα από τα μαθήματα του. Τέλος, ευχαριστώ πολύ τον κ. Ηλία Χούστη που έκανε όλο αυτό το ταξίδι για να παρευρεθεί απλώς στην παρουσίαση αυτής της εργασίας, αλλά και για τη συζήτηση που ακολούθησε μετά την παρουσίαση. Επίσης θα ήθελα να ευχαριστήσω τον κ. Δημήτρη Κοτζίνο, Εντεταλμένο Επίκουρο του Τμήματος Επιστήμης Υπολογιστών ο οποίος αν και δεν ήταν μέλος της επιτροπής είχε πολύ μεγάλη υπομονή, με βοήθησε και με στήριξε σε όλη τη διάρκεια της εργασίας με χρήσιμες συμβουλές αλλά και καθοριστικές παρεμβάσεις στα σημεία που «κολλούσα».

Στο σημείο αυτό θα ήθελα να ευχαριστήσω και τα υπόλοιπα μέλη της ομάδας ΑΤΛΑΝΤΙΔΑ που με στήριξαν ο καθένας με τον τρόπο του, τον Χαράλαμπο Σάμπαλη και τη Θέμις Ζαμάνη. Ιδιαίτερα με τη Θέμις δουλεύουμε στο ίδιο γραφείο και έχουμε περάσει μαζί ξενύχτια, πολλές καλές αλλά και δύσκολες στιγμές. Είναι ένας ιδιαίτερα υπομονετικός άνθρωπος και πάντα πρόθυμη να βοηθήσει ακόμα και όταν πολλές φορές πάθαινα κρίσεις πανικού.

Επίσης, θα ήθελα να ευχαριστήσω όλους τους φίλους που γνώρισα στην Κρήτη μερικοί από τους οποίους δεν είναι εδώ σήμερα αλλά έχουμε κάνει πολλές εκδρομές και έχουμε περάσει πολύ όμορφες στιγμές, το Δημήτρη και το Σωτήρη Μπαλκούρα, τη Λένα Κανέλλου, την Τίνα Κατελανή,

το Στέλιο Γάσπαρη, το Γιώργο Μανεσσιώτη, το Δημήτρη Καπανίδη, τον Απόστολο Φώτη, τον Δημήτρη Μπαλουκίδη. Αλλά και την Κική Καραδήμου την οποία γνώρισα τα τελευταία δύο χρόνια και μου έδειξε ότι όταν κάτι το θες πολύ το πετυχαίνεις καθώς και το Γιάννη Ρουσίδη τον οποίο γνώρισα καλύτερα στη διάρκεια αυτής της εργασίας με τα ξενύχτια που κάναμε ιδιαίτερα τους τελευταίους τρεις μήνες.

Τέλος, ένα μεγάλο ευχαριστώ στους γονείς μου Παύλο και Παναγιώτα και την αδερφή μου Βασιλική οι οποίοι με στηρίζουν και με βοηθούν σε ότι και αν κάνω.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 Εισαγωγή

1.1 Καθορισμός του προβλήματος

Η ανάπτυξη του διαδικτύου τα τελευταία χρόνια έχει φέρει αλλαγές τόσο στην ποιότητα και το μέγεθος όσο και στην ταχύτητα πρόσβασης του διαθέσιμου περιεχομένου. Οι χρήστες κυριολεκτικά κατακλύζονται από πληροφορία την οποία δυσκολεύονται όχι μόνο να αφομοιώσουν αλλά πια και να φιλτράρουν. Σε αυτό πρέπει κανείς να προσθέσει τις διαφορετικές μορφές στις οποίες είναι διαθέσιμη η πληροφορία, κάτι που κάνει το μοντέλο διαχείρισης της να απέχει πολύ από τα παραδοσιακά μοντέλα βάσεων δεδομένων. Το μέγεθος, η ανομοιογένεια και η πολυπλοκότητα του διαδικτύου εισάγουν νέα ενδιαφέροντα ερευνητικά προβλήματα ξεπερνώντας υπάρχουσες τεχνικές, αρχιτεκτονικές και αλγορίθμους και προσφέροντας πρόσφορο έδαφος για την βελτίωσή τους ή τη δημιουργία καινούριων.

Στις μέρες μας πλέον το διαδίκτυο αντιμετωπίζεται σαν μια ολοκληρωμένη πλατφόρμα παροχής υπηρεσιών και όχι (μόνο) σαν ένα σύνολο από τυχαία ευρισκόμενα δεδομένα. Μάλιστα έχει ήδη ξεκινήσει και η συζήτηση για την επόμενη γενιά του διαδικτύου (Web 2.0), η οποία θα βασίζεται όχι πια στο λογισμικό αλλά στις παρεχόμενες υπηρεσίες στις οποίες θα μπορεί ο χρήστης να έχει πρόσβαση από οπουδήποτε και (σχεδόν) με οποιοδήποτε μέσο. Υπηρεσίες όπως τα επιτυχημένα παραδείγματα του Google[1] και του eBay[2] έχουν ήδη αντικαταστήσει σαν έννοιες το Netscape, με το οποίο ήταν άρρηκτα συνδεδεμένο το διαδίκτυο στις πρώτες μέρες του. Αντίστοιχα έννοιες όπως τα wikis[3] και τα blogs[4] έχουν αντικαταστήσει την έννοια της προσωπικής ιστοσελίδας και του ατομικού ιστοτόπου. Είμαστε δηλαδή στην αρχή της ολικής αλλαγής του παραδείγματος σύμφωνα με το οποίο ερμηνεύουμε και χρησιμοποιούμε το διαδίκτυο.

Δύο από τα βασικότερα προβλήματα της πληροφορίας στο διαδίκτυο είναι πως είναι κυρίως σε μορφή κειμένου και μάλιστα βρίσκεται σε μη δομημένη ή στην καλύτερη περίπτωση ημιδομημένη μορφή. Όλα αυτά έχουν δημιουργήσει την ανάγκη ύπαρξης Συστημάτων Διαχείρισης Περιεχομένου, τα οποία προσπαθούν να παρέχουν τα απαραίτητα εργαλεία τόσο για την ανάκτηση όσο και για την περαιτέρω διαχείριση του περιεχομένου αυτού. Αυτού του είδους τα συστήματα εισάγουν διαδοχικά μια καινούρια σειρά προβλημάτων που έχουν να κάνουν τόσο με

[1] www.google.com

[2] www.ebay.com

[3] http://en.wikipedia.org/wiki/Wiki

[4] http://en.wikipedia.org/wiki/Blog

τον τρόπο εισαγωγής του περιεχομένου που διαχειρίζονται (αν δηλαδή θα είναι αυτόματος, ημιαυτόματος ή από τον χρήστη) όσο και με τον τρόπο παρουσίασης αυτού του περιεχομένου στον τελικό χρήστη, μια και οι ανάγκες του ξεπερνούν πια την σειριακή παράθεση κειμένων (documents), συνδέσμων (links) και νέων (news feeds). Για αυτό το λόγο έχουν δημιουργηθεί και χρησιμοποιούνται όλο και περισσότερο σήμερα κατάλογοι περιεχομένου (content directories), βασική συνεισφορά των οποίων είναι η δυνατότητα κατηγοριοποίησης του περιεχομένου ώστε να είναι πιο εύκολη η αναζήτηση αλλά και η παρουσίασή του σε συνεπείς νοηματικές κατηγορίες. Εκτός των άλλων οι κατάλογοι αυτοί αποτελούν τη βάση πάνω στην οποία οι δημοφιλείς πλέον μηχανές αναζήτησης αναζητούν το περιεχόμενο που ενδιαφέρει το χρήστη. Στα πλαίσια ενός Συστήματος Διαχείρισης Περιεχομένου ένας τέτοιος κατάλογος μπορεί να χρησιμοποιηθεί για την παραγωγή των κατηγοριών στις οποίες ανήκει το υλικό του ΣΔΠ και κατ' επέκταση την παρουσίαση (αλλά και γενικότερη διαχείριση) του με βάση της κατηγορίες αυτές.

Η παρούσα μεταπτυχιακή εργασία αγγίζει τα δύο αυτά σοβαρά προβλήματα των Συστημάτων Διαχείρισης Περιεχομένου και προσπαθεί να κάνει εφικτή καταρχήν την αυτόματη ή ημιαυτόματη ανάκτηση συγκεκριμένων τύπων περιεχομένου και κατόπιν την αυτόματη ή ημιαυτόματη κατηγοριοποίησή τους. Πρέπει να σημειωθεί ότι τα υπάρχοντα ΣΔΠ δεν παρέχουν δυνατότητες κατηγοριοποίησης του περιεχομένου τους ενώ λίγα είναι αυτά που παρέχουν δυνατότητες αυτόματης ανάκτησης. Τα προβλήματα για τα οποία προτείνονται λύσεις στα πλαίσια ενός ΣΔΠ είναι:

- Το πρόβλημα της αυτόματης ανάκτησης περιεχομένου από το δίκτυο (σε ημιδομημένη μορφή: RSS, Google API results, HTML, απλά κείμενα)
- Το πρόβλημα της ημιαυτόματης κατηγοριοποίησης κειμένου που προέρχεται από τις πιο πάνω πηγές

Γενικότερα η δυνατότητα αυτοματοποίησης διαδικασιών όπως αυτές στα πλαίσια ενός ΣΔΠ κάνει την χρήση τέτοιων συστημάτων περισσότερη ελκυστική για τον τελικό χρήστη μια και του δίνει τη δυνατότητα όχι μόνο αυτοματισμών αλλά και καλύτερης διαχείρισης του υλικού του.

1.2 Κίνητρο και συνεισφορά της εργασίας

Η βασική συνεισφορά της εργασίες κινείται στους παρακάτω τρεις άξονες:

- Αυτόματη ανάκτηση περιεχομένου από πηγές που περιέχουν ημιδομημένη πληροφορία, όπως τα RSS feeds, τα αποτελέσματα του Google API και τα HTML κείμενα ή δομημένη πληροφορία όπως κείμενα που έχουν εισαχθεί στη βάση με ένα συγκεκριμένο τρόπο.

- Ημιαυτόματη κατηγοριοποίηση κειμένων που έχουν εισαχθεί σε βάση δεδομένων είτε αυτόματα είτε από το χρήστη, με δύο βασικά σημεία συνεισφοράς:
  - ⇒ Την λεξικογραφική ανάλυση του κειμένου και την εξαγωγή των χρήσιμων λέξεων για την μετέπειτα κατηγοριοποίησή του και την αναλυτική ποιοτική επισκόπηση των ερευνητικών αποτελεσμάτων στον συγκεκριμένο τομέα
  - ⇒ Την κατηγοριοποίηση του κειμένου που αναλύθηκε λεξικογραφικά με την χρήση ενός ευρεστικού αλγορίθμου που παρουσιάζεται στη συγκεκριμένη εργασία και βασίζεται στην χρήση του καταλόγου DMOZ, ο οποίος αποτελεί ένα αναγνωρισμένο πρότυπο (standard), για την παροχή των κατηγοριών. Και στην συγκεκριμένη περίπτωση παρέχεται μια αναλυτική επισκόπηση των προτεινόμενων μέχρι σήμερα αλγορίθμων κατηγοριοποίησης κειμένου (ανεξάρτητα από το αν προέρχεται το κείμενο αυτό από το διαδίκτυο ή όχι) και εξηγείται που διαφέρει ο προτεινόμενος αλγόριθμος από τους υπολοίπους.
- Παροχή των παραπάνω δυνατοτήτων στα πλαίσια ενός Συστήματος Διαχείρισης Περιεχομένου Ανοικτού Κώδικα, προσφέροντας με αυτόν τον τρόπο την βάση για την παροχή πραγματικών ημιαυτόματων υπηρεσιών στους χρήστες. Η δυνατότητα αυτή είναι σημαντικό να σημειωθεί πως απουσιάζει από όλα τα ΣΔΠ Ανοικτού Κώδικα και πως στα περισσότερα εμπορικά ΣΔΠ που συναντάται δεν είναι ολοκληρωμένη.

1.3 Οργάνωση του κειμένου

Στην εργασία αυτή παρουσιάζουμε τις λύσεις που προτείνονται στα παραπάνω προβλήματα, όπως αυτά αντιμετωπίζονται στα πλαίσια ενός ΣΔΠ. Οι λύσεις που δίνονται συζητούνται τόσο από την τεχνική τους πλευρά (στα πλαίσια της συγκεκριμένης πλατφόρμας πάνω στην οποία στηρίχτηκε η υλοποίηση, αλλά και στη δυνατότητα ενσωμάτωσής τους στην πλατφόρμα) όσο και στην αλγοριθμική τους πλευρά, όπου βέβαια αυτή είναι υπαρκτή.

Στο κεφάλαιο 2 παρουσιάζονται οι γενικές αρχές λειτουργίας ενός Συστήματος Διαχείρισης Περιεχομένου (ΣΔΠ) με έμφαση στην εφαρμογή τους από ΣΔΠ Ανοικτού Κώδικα, ενώ παρουσιάζεται και η υψηλού επιπέδου αρχιτεκτονική του ATL-CME που είναι και το ΣΔΠ Ανοικτού Κώδικα πάνω στο οποίο υλοποιήθηκε η παρούσα εργασία.

Στο κεφάλαιο 3 παρουσιάζεται μια σύνοψη των τάσεων στον τομέα της διαχείρισης περιεχομένου με έμφαση στην περιοχή της «προσωπικής διαχείρισης περιεχομένου». Καθορίζεται επίσης το θέμα που διαπραγματεύεται η εργασία και ο λόγος που κρίθηκε απαραίτητο να αναπτυχθεί.

Το κεφάλαιο 4 παρουσιάζει την αλγοριθμική συνεισφορά της εργασίας για: (1) την ανάκτηση περιεχομένου από το διαδίκτυο, (2) την ανάλυση κειμένου και εξαγωγή χρήσιμων λέξεων αλγοριθμικά από οποιοδήποτε κείμενο (μαζί με εκτεταμένη αναφορά στη σχετική βιβλιογραφία) και (3) την κατηγοριοποίηση κειμένου (μαζί πάλι με την σχετική βιβλιογραφία). Αναλύονται τα βήματα των αλγορίθμων αλλά και εξηγείται η σχεδιαστική ένταξή τους στο ΣΔΠ.

Στο κεφάλαιο 5 ο αναγνώστης μπορεί να βρει τις λεπτομέρειες τόσο για την υλοποίηση των αλγορίθμων που περιγράφηκαν στο προηγούμενο κεφάλαιο όσο και για την ροή της πληροφορίας κατά την διαδικασία της κατηγοριοποίησης. Επίσης αναλύονται οι περιορισμοί που τέθηκαν από την επιλογή του συγκεκριμένου ΣΔΠ ως πλατφόρμα πάνω στην οποία υλοποιήθηκαν οι προτεινόμενοι αλγόριθμοι αλλά τα οι απαραίτητες διεπαφές που απαιτήθηκαν με άλλα υποσυστήματα του ΣΔΠ ώστε η διαδικασία της κατηγοριοποίησης να ενταχθεί ομαλά στη ροή ενός τέτοιου συστήματος και να συνεργαστεί και να εκμεταλλευτεί τα υποσυστήματα που είναι ήδη σε λειτουργία.

Το κεφάλαιο 6 παρουσιάζει τα σενάρια χρήσης του υποσυστήματος της κατηγοριοποίησης που χρησιμοποιήθηκαν τόσο για την τεκμηρίωση του ευρεστικού αλγορίθμου όσο και για την καλύτερη επιλογή διάφορων αριθμητικών παραμέτρων που χρησιμοποιεί ο αλγόριθμος για να παίρνει αποφάσεις. Επίσης στο κεφάλαιο αυτό παρουσιάζονται τα συμπεράσματα που μπορεί κανείς να βγάλει από τα συγκεκριμένα σενάρια τόσο για την συνολική αποτελεσματικότητα του αλγορίθμου όσο και για την ποιοτική πλευρά των αποτελεσμάτων του.

Τέλος το κεφάλαιο 7 παρουσιάζει μια σύνοψη της εργασίας καθώς και κάποια γενικά συμπεράσματα που μπορούν να εξαχθούν από αυτή. Δίνει επίσης και κάποιες κατευθύνσεις μελλοντικής έρευνας στην συγκεκριμένη περιοχή μια και προφανώς το θέμα παρουσιάζει ερευνητικό ενδιαφέρον.

2 Διαχείριση Περιεχομένου και Ανοικτό Λογισμικό

2.1 Εισαγωγή

Αυτή η μεταπτυχιακή εργασία εστιάζεται στο σχεδιασμό και στην ανάπτυξη ενός συστήματος Ημι-αυτόματης κατηγοριοποίηση του περιεχομένου που συγκεντρώνει ένας δικτυακός τόπος από διαφορετικές πηγές (web, εσωτερική πληροφορία κλπ.). Ένα τέτοιο σύστημα συμπληρώνει τη λειτουργικότητα ενός Συστήματος Διαχείρισης Περιεχομένου (Content Management System), που είναι η τεχνολογική μηχανή την οποία, εκτεταμένα πλέον, χρησιμοποιούν οι δικτυακοί τόποι για τη διαχείριση και δημοσίευση περιεχομένου στο Web (ή σε άλλα μέσα, π.χ. κινητό τηλέφωνο, τηλεόραση κλπ). Στο κεφάλαιο αυτό παρουσιάζονται τα Συστήματα Διαχείρισης Περιεχομένου (ΣΠΔ), με έμφαση στα ΣΠΔ Ελεύθερου Λογισμικού / Ανοικτού Κώδικα. Περιγράφονται συνοπτικά οι λειτουργικές δυνατότητές τους και συγκρίνονται, στο πλαίσιο ενδιαφέροντος της εργασίας αυτής, ανάλογα με το αν υποστηρίζουν ή όχι δυνατότητες κατηγοριοποίησης περιεχομένου (content categorization). Παρουσιάζεται τέλος το περιβάλλον ATL Content Management Engine (ATL CME), στο οποίο υλοποιήθηκε, ως αυξητική λειτουργικότητα, το σύστημα Κατηγοριοποίησης Περιεχομένου από το διαδίκτυο (Web Content Classifier - WCC) που αποτελεί το αποτέλεσμα αυτής της μεταπτυχιακής εργασίας.

2.2 Τι είναι ένα Σύστημα Διαχείρισης Περιεχομένου

Ένα Σύστημα Διαχείρισης Περιεχομένου δεν είναι μόνο ένα προϊόν ή μια τεχνολογία. Ορίζεται ως ένας γενικός όρος που υποδεικνύει ένα τεχνολογικό σύστημα το οποίο περιλαμβάνει ένα ευρύ φάσμα διαδικασιών που αφορούν τη δημιουργία, αποθήκευση, τροποποίηση, ανάκτηση και παρουσίαση περιεχομένου, το οποίο με τη σειρά του μπορεί να είναι πολλών ειδών. Ουσιαστικά, ένα τέτοιο σύστημα συνδέει κανόνες, επιχειρησιακές διαδικασίες (ενδεχομένως και ροές εργασιών) με τη διαδικασία διαχείρισης και δημοσίευσης πληροφορίας στο Web, και ταυτόχρονα δίνει πρόσβαση σε εξουσιοδοτημένους χρήστες, βάσει κανόνων, διαδικασιών και πολιτικών που υλοποιεί ο οργανισμός που το χρησιμοποιεί.

Ένα Σύστημα Διαχείρισης Περιεχομένου καλύπτει τον κύκλο ζωής των σελίδων ενός δικτυακού τόπου παρέχοντας εργαλεία για τη δημιουργία του περιεχομένου, τη δημοσίευση και την αρχειοθέτησή του. Επίσης παρέχει τις δυνατότητες διαχείρισης της δομής, εμφάνισης των δημοσιευμένων σελίδων στους επισκέπτες / χρήστες του δικτυακού τόπου και οργανώνει πλαίσια πλοήγησης (navigation paths) από αυτούς [50].

Πρόκειται ουσιαστικά, για ένα ηλεκτρονικό σύστημα διαχείρισης και δημοσίευσης πληροφορίας το οποίο μπορεί να θεωρηθεί ως μια κοινή (shared) βάση δεδομένων που περιέχει όμως επιπλέον εργαλεία για την οργάνωση του ψηφιακού περιεχομένου και τη διαχείριση κανόνων ελεγχόμενης πρόσβασης στην όλη διαδικασία διαχείρισης-δημοσίευσης αυτού του περιεχομένου. Όταν λέμε «ψηφιακό περιεχόμενο» εννοούμε αρχεία, εικόνες, απλό κείμενο, γραφικά, ήχο, βίντεο.

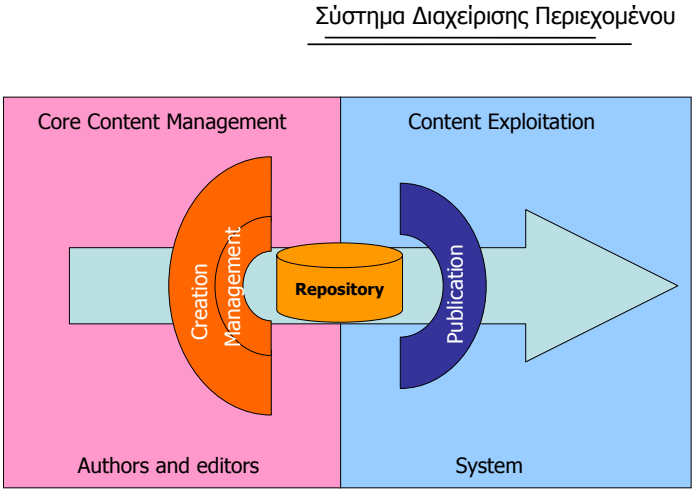
Ανάλογα με τις ανάγκες κάθε οργανισμού ένα ΣΔΠ μπορεί να υποστηρίζει τη διαχείριση μιας απλής βάσης δεδομένων, να παρέχει μηχανισμούς που προσδίδουν «σημασία» στο περιεχόμενο (μεταδεδομένα) αλλά και δυνατότητες δημιουργίας συνδέσμων μεταξύ των αρχείων και, φυσικά, πολύπλοκους μηχανισμούς που αφορούν σύνθετους κανόνες πρόσβασης και ανανέωσης του περιεχομένου [55] σε όσους εμπλέκονται στις σχετικές διαδικασίες (συγγραφείς, συντάκτες, οι υπεύθυνοι για τη διαχείριση του δικτυακού τόπου κλπ.). Τελικά, ένα ΣΔΠ μπορεί να θεωρηθεί ένα σύστημα που «δέχεται» περιεχόμενο (input), αυτό που γράφει ένας «συγγραφέας» ή αντλείται από μια πηγή, και «γεννά» (output) σελίδες με «δομή» (ένα βήμα πέρα από τις στατικές και δυναμικές σελίδες) είσοδο. Σε έναν οργανισμό, μια επιχείρηση, μια εφημερίδα,, ένα ΣΠΔ λειτουργεί ως ένα κοινό σημείο δημοσίευσης της πληροφορίας (single source publishing), πράγμα που υπονοεί τη «συγκέντρωση» της υπό δημοσίευση πληροφορίας σε ένα σημείο και την «αυτόματη» (στη βάση διαδικασιών και κανόνων) δημοσίευση της σε πολλά διαφορετικά μέσα, ενδεχομένως και με διαφορετικά πρότυπα (PDF, XML, κτλ.) [50].

Οι διαδικασίες που εφαρμόζονται στο περιεχόμενο από ένα Σύστημα Διαχείρισης Περιεχομένου μπορούν να χωριστούν σε τέσσερις μεγάλες κατηγορίες (Σχήμα 1) [63]:

- **Δημιουργία περιεχομένου:** Ένα Σύστημα Διαχείρισης Περιεχομένου περιλαμβάνει ένα περιβάλλον για τους δημιουργούς του περιεχομένου. Πρόκειται για έναν εύκολο τρόπο δημιουργίας νέων σελίδων ή ανανέωσης του περιεχομένου χωρίς τη γνώση της γλώσσας HTML ή άλλων προγραμματιστικών εργαλείων, απευθύνεται δηλαδή σε μη ειδικούς. Δίνει επίσης τη δυνατότητα διαχείρισης της δομής ενός δικτυακού τόπου καθορίζοντας τις κατηγορίες αλλά και τις συνδέσεις μεταξύ αυτών. Για παράδειγμα κατά την έκδοση ενός ηλεκτρονικού περιοδικού οι δημιουργοί του περιεχομένου είναι οι αρθρογράφοι των νέων και των άρθρων που δημοσιεύονται σε αυτό.
- **Διαχείριση περιεχομένου:** Η διαχείριση του περιεχομένου γίνεται συνήθως χρησιμοποιώντας μια κεντρική αποθήκη (βάση) δεδομένων. Εκεί αποθηκεύεται όλο το περιεχόμενο ενός δικτυακού τόπου καθώς και άλλες πληροφορίες για τη δομή του. Αυτή η αποθήκη δεδομένων επιτρέπει ακόμη τον έλεγχο εκδόσεων του περιεχομένου

(ποιος, πότε και τι άλλαξε) ενώ περιέχει και κανόνες πρόσβασης σ' αυτό. Τέλος, σε αυτό το στάδιο γίνεται η διαχείριση της ροής των εργασιών ενός εγγράφου (αν αυτή η δυνατότητα παρέχεται από το Σύστημα Διαχείρισης Περιεχομένου). Οι υπεύθυνοι για τη διαχείριση του περιεχομένου είναι οι διαχειριστές του συστήματος που φροντίζουν την ανάπτυξη και την αναβάθμιση αυτών των μηχανισμών. Στο παράδειγμα της έκδοσης ενός ηλεκτρονικού περιοδικού, είναι ο διαχειριστής του συστήματος (system administrator) μαζί με τον αρχισυντάκτη.

- **Δημοσίευση περιεχομένου:** Εφόσον το περιεχόμενο είναι αποθηκευμένο μπορεί να δημοσιευθεί είτε στο δικτυακό τόπο (εν προκειμένω αυτόν της εφημερίδας) είτε να χρησιμοποιηθεί από άλλους δικτυακούς τόπους σε διαφορετικά πρότυπα. Η διαδικασία της δημοσίευσης είναι συνυφασμένη με την παροχή μηχανισμών για την αυτόματη παρουσίασή του, με τη δυνατότητα δημιουργίας μεταδεδομένων για το περιεχόμενο που διαχειρίζεται, με την ύπαρξη διαφορετικών "templates" για κάθε σελίδα του δικτυακού τόπου, κλπ.. Φυσικά, η δημοσίευση περιεχομένου (ενός αρχείου για παράδειγμα) γίνεται εφόσον έχει δοθεί η σχετική άδεια από τον υπεύθυνο του οργανισμού (στο παράδειγμα της έκδοσης ενός ηλεκτρονικού περιοδικού, ο εκδότης / αρχισυντάκτης αυτού) σε έναν «συντάκτη» να αναρτήσει αυτή την πληροφορία.
- **Παρουσίαση περιεχομένου:** Τέλος, ένα σύστημα διαχείρισης περιεχομένου υποστηρίζει διαλειτουργικότητα μεταξύ των υπαρχόντων φυλλομετρητών, αποτελεσματική πλοήγηση στην ταξινόμηση των κατηγοριών περιεχομένου και παρέχει δυνατότητες «προσωποποίησης» (personalization) της εμφάνισης του περιεχομένου, με τρόπο που ο κάθε επισκέπτης / χρήστης να βλέπει (αν αυτό αποφασιστεί) ένα «διαφορετικό» δικτυακό τόπο, με το περιεχόμενο που αυτός επιθυμεί να περιέχει.



Σχήμα 1 Η ανατομία ενός συστήματος διαχείρισης περιεχομένου

Ο Πίνακας 1 παρουσιάζει συνοπτικά τις λειτουργικότητες ενός ΣΔΠ [49].

Core Content Management	**Δημιουργία περιεχομένου**	**Διαχείριση Περιεχομένου**
	Ενσωματωμένο περιβάλλον για τη σύνταξη του περιεχομένου (authoring environment) Διαχωρισμός κειμένου και παρουσίασης Δυνατότητα σύνταξης από πολλούς χρήστες Επαναχρησιμοποίηση περιεχομένου Δημιουργία μεταδεδομένων Τρόποι σύνταξης περιεχομένου για μη ειδικούς Ευκολία χρήσης	Έλεγχος εκδόσεων Αρχειοθέτηση Διαδικασίες ροής εργασίας (Workflow) Ασφάλεια Δυνατότητα επικοινωνίας με εξωτερικά συστήματα
Content Exploitation	**Δημοσιοποίηση**	**Παρουσίαση**
	Υποστήριξη Stylesheets Πρότυπα σελίδων (templates) Επεκτασιμότητα Υποστήριξη για πολλά είδη δεδομένων Προσωποποίηση Διαχωρισμός του περιβάλλοντος δοκιμών και του περιβάλλοντος παραγωγής Δυναμική δημοσιοποίηση περιεχομένου	Χρηστικότητα (Usability) Προσβασιμότητα Υποστήριξη για πολλούς φυλλομετρητές Περιορισμένη λειτουργικότητα από την πλευρά του χρήστη Αποτελεσματική πλοήγηση Μεταδεδομένα

Πίνακας 1 Τα βασικά στοιχεία ενός Συστήματος Διαχείρισης Περιεχομένου

2.3 Ανοικτό Λογισμικό και Συστήματα Διαχείρισης Περιεχομένου

2.3.1 Συστήματα Διαχείρισης Περιεχομένου Ανοικτού Λογισμικού

Τα Συστήματα Διαχείρισης Περιεχομένου Ανοικτού Λογισμικού είναι πακέτα λογισμικού με δυνατότητες διαχείρισης περιεχομένου και εργαλεία με λειτουργικότητες που περιγράφηκαν παραπάνω. Η βιομηχανία της διαχείρισης περιεχομένου περιλαμβάνει τρεις διακριτές στρατηγικές ομάδες: τις επιχειρήσεις που λέγονται "Big irons" (Art Technology Group, Broadvision, Interwoven, Vignette), τις "Verticals" (FileNet, Documentum) και τις μεγάλες επιχειρήσεις software (IBM,

Oracle, Microsoft). Ο κατάλογος αυτών των συστημάτων συμπληρώνεται με αυτά του Ανοικτού Λογισμικού που συνιστούν ένα νέο ρεύμα στη βιομηχανία αυτή και μια τάση σε μεγέθυνση, υποστηριζόμενη κυρίως από τις πολιτικές των κυβερνήσεων.

Για να αντιληφθεί κανείς τη μεγάλη απήχηση που έχουν γνωρίσει τα Συστήματα Διαχείρισης Περιεχομένου Ανοικτού Λογισμικού θα πρέπει να ανατρέξει στην ιστορία της αγοράς αυτών των προϊόντων. Πριν μερικά χρόνια οι συμβουλευτικές εταιρείες συμφωνούσαν ότι το μέλλον της διαχείρισης του περιεχομένου (content management) αφορούσε τις μεγάλες εταιρείες που προσέφεραν εμπορικά πακέτα λογισμικού. Πράγματι από το 1997 τέτοια συστήματα όπως αυτά που παρείχαν οι Vignette[5], Interwoven[6] και Documentum[7] θεωρούνταν αδιαμφισβήτητα οι κυρίαρχοι του χώρου της διαχείρισης περιεχομένου ανεξάρτητα από το μεγάλο κόστος αγοράς τους. Ωστόσο το 2001 ορισμένες αναλύσεις της Forrester[8] προέβλεπαν ότι ένα μεγάλο μέρος των εταιρειών που χρησιμοποιούσαν συστήματα διαχείρισης περιεχομένου θα επέστρεφαν σε λύσεις με χαμηλότερο κόστος που θα εξυπηρετούσαν τις ανάγκες τους καλύτερα στο πλαίσιο μιας συνδυαστικής εξέτασης κόστος / απόδοση. Την άποψη αυτή ενστερνίστηκε και η εταιρεία Jupiter Media Matrix[9] με το πρόσθετο επιχείρημα συνήθως οι λύσεις που αναπτύσσονται εντός ενός οργανισμού αποτελούν και τη σωστή επιλογή για την κάλυψη των αναγκών του. Πράγματι, η αγορά και εγκατάσταση των πιο δημοφιλών από τα εμπορικά πακέτα διαχείρισης περιεχομένου συνδέεται συχνά με [54]:

- Υψηλό κόστος για την αγορά και χρήση
- Δύσκολη ολοκλήρωση και συνεργασία με υπάρχουσες εφαρμογές

Αυτό που φαίνεται πια καθαρά είναι ότι τα πιο γνωστά και ευέλικτα συστήματα διαχείρισης περιεχομένου έχουν νόημα απόκτησης μόνο για πολύ μεγάλους οργανισμούς, δεν ανταποκρίνονται ακριβώς στις ανάγκες διαχείρισης περιεχομένου και δημοσίευσης στο Web των μικρότερων επιχειρήσεων και των δημοσίων και κοινωφελών οργανισμών. Για αυτές τις κατηγορίες, η βιομηχανία επεξεργάζεται κατάλληλα προϊόντα (Yahoo!, Microsoft), αλλά ίσως η πιο δυναμική πηγή τεχνολογικών λύσεων είναι το ρεύμα του Λογισμικού Ανοικτού Κώδικα. Μάλιστα, η αγορά των ΣΔΠ ΕΛ/ΑΚ αναπτύσσεται τόσο γρήγορα όσο και η αγορά των εμπορικών ΣΔΠ [56].

---

[5] Vignette, http://www.vignette.com/

[6] Interwoven, http://www.interwoven.com/

[7] Documentum, http://www.documentum.com/

[8] Forrester Research Inc., http://www.forrester.com/

[9] Jupiter Research, http://www.jupiterresearch.com/bin/item.pl/home

Στο Σχήμα 2 αποτυπώνεται ακριβώς η ωριμότητα των προϊόντων αυτών και δίνονται ορισμένα αντιπροσωπευτικά παραδείγματα για διάφορα segments της αγοράς εφαρμογών[10]. Διαπιστώνει κανείς η διαχείριση περιεχομένου είναι ένας τομέας που τα προϊόντα Ανοικτού Κώδικα (ΣΠΔ ΕΛ/ΑΚ, δικτυακές πύλες, συνεργατικοί μηχανισμοί) έχουν ελκυστικότητα (επιπλέον πολλές από τις άλλες τεχνολογίες που παρουσιάζονται στο σχήμα χρησιμοποιούνται στα πλαίσια ενός ΣΠΔ). Η ανάπτυξη και η ωριμότητά τους έγκειται στη συνεργασία με άλλες εφαρμογές ώστε να αποτελέσουν τα βασικά στοιχεία (building blocks) για την ανάπτυξη υπηρεσιών προσαρμοσμένες στις ανάγκες των χρηστών[11].

The Open-Source Stack: Growing and Expanding

Stack modules	Products	Maturity
Enterprise Applications	OpenCRX, SugarCRM, Compiere, Ohioedge	1/4
Content Management	Midgard, OpenCMS, Lenya, Typo3, Red Hat	3/4
Portals	Jetspeed, Zope/Plone, uPortal, JBoss Portal	2/4
Collaboration	Zope, Drupal, phpBB, Hula, Chandler, Mozilla	2/4
Search	Lucene, ht://Dig	3/4
Development Tools	Eclipse, NetBeans	4/4
Application Servers	JBoss Geronimo JOnAS	3/4
Application Integration	Openadaptor	1/4
Directory Services	OpenLDAP	3/4
RDBMS	MySQL, PostgreSQL	3/4
Operating System	Linux, FreeBSD	4/4

Σχήμα 2 Αποτύπωση της εικόνας των προϊόντων ΕΛ/ΑΚ

Η χρήση ενός εμπορικού συστήματος διαχείρισης περιεχομένου προσφέρει αναμφισβήτητα πολλά πλεονεκτήματα, κυρίως αυτό της τεκμηρίωσης, της εκπαίδευσης και της οργανωμένης υποστήριξης. Ζητούν όμως μια σημαντική επένδυση. Τα προϊόντα ΣΔΠ ΕΛ/ΑΚ μπορούν να δώσουν καλής απόδοσης και προσαρμοσμένες στις ανάγκες του χρήστη λύσεις, με ελεγχόμενο κόστος, και επιτρέπουν την καλή συνεργασία της διαχείρισης περιεχομένου με ήδη υπάρχουσες

[10] Content Management, Collaboration and Portals: Now and in the future, http://www.gartner.com

[11] Το ΕΛ/ΑΚ ακολουθεί δύο μοντέλα τα οποία εφαρμόζονται και στα ΣΔΠ. Το πρώτο αφορά ΣΠΔ που αναπτύσσονται στα πλαίσια μιας κοινότητας και το δεύτερο περιλαμβάνει ΣΠΔ που προσφέρουν υποστήριξη με το αντίστοιχο κόστος [49]. Η πρώτη κατηγορία ταιριάζει σε οργανισμούς που διαθέτουν ικανότητες για την προσαρμογή των ΣΠΔ στις ανάγκες τους ενώ η δεύτερη θα πρέπει να αξιολογηθεί σαν ένα εμπορικό προϊόν για το οποίο δεν απαιτείται κόστος για την άδεια χρήσης του.

εφαρμογές. Στα μειονεκτήματα ανάπτυξης ή χρήσης ενός συστήματος ανοικτού λογισμικού συγκαταλέγεται κυρίως η αβεβαιότητα για το μέλλον αυτού του λογισμικού. Η υποστήριξη του προϊόντος, η τεκμηρίωση και η εκπαίδευση των χρηστών αποτελούν επίσης βασικούς ανασταλτικούς παράγοντες. Ο Πίνακας 2 παρουσιάζει μια εποπτική εικόνα των διαφορών των εμπορικών ΣΠΔ από τα ΣΠΔ ΕΛ/ΑΚ.

Συστήματα Διαχείρισης Περιεχομένου Ανοιχτού Λογισμικού	**Εμπορικά Συστήματα Διαχείρισης Περιεχομένου**
Χαμηλό κόστος	Πολύ υψηλό κόστος: Κόστος αρχικής αγοράς / αδειών χρήσης / προσαρμογής
Κόστος ανά υπηρεσία, κανένα κόστος στην απόκτηση του λογισμικού	Το κόστος συνδέεται τόσο με το κόστος αγοράς όσο και με την εκπαίδευση των χρηστών
Εύκολη προσαρμογή	Δεν υπάρχει δυνατότητα προσαρμογής ή δημιουργίας μιας νέας λειτουργικότητας από τον οργανισμό
Εύκολη ενσωμάτωση σε υπάρχον λογισμικό	Δύσκολη συνεργασία με υπάρχοντα συστήματα
Υποστήριξη από κοινότητες χρηστών	Λιγότεροι προγραμματιστές σε σχέση με τους πολλούς χρήστες μιας κοινότητας
Η υποστήριξη παρέχεται από μεγάλες κοινότητες αλλά η αναζήτηση της λύσης κοστίζει το χρόνο αναζήτησης	Εξειδικευμένη υποστήριξη για κάθε οργανισμό
Μπορεί ανά πάσα στιγμή να σταματήσει να υποστηρίζεται από την κοινότητα χρηστών	Υπάρχει μέλλον
Η εκπαίδευση των χρηστών γίνεται με τη χρήση και από υπαλλήλους του οργανισμού που το χρησιμοποιεί	Εκπαίδευση του προσωπικού που θα το χρησιμοποιήσει
Μπορεί εύκολα να προσαρμόζει νέες τάσεις που εμφανίζονται στην αγορά	Είναι αρκετά δύσκολο να παρακολουθήσει τις εξελίξεις στο χώρο
Πολλές φορές η τεκμηρίωση είναι ελλιπής	Τεκμηρίωση για κάθε παρεχόμενη υπηρεσία

Πίνακας 2 Σύγκριση Συστημάτων Διαχείρισης Περιεχομένου

2.3.2 Πού σταματά ένα Σύστημα Διαχείρισης Περιεχομένου Ανοικτού Λογισμικού

Ο Πίνακας 3 παρουσιάζει τα πιο δημοφιλή συστήματα διαχείρισης περιεχομένου ανοικτού λογισμικού[12].

Συστήματα Διαχείρισης Περιεχομένου Ανοικτού Λογισμικού	
Red Hat CCM	http://ccm.redhat.com/
uPortal	http://www.uportal.org/
Midgard	http://www.midgard-project.org/
Zope	http://www.zope.org/
ApacheCocoon	http://cocoon.apache.org/
Drupal	http://www.drupal.org/
Mambo	http://www.mamboserver.com/

Πίνακας 3 Τα πιο δημοφιλή Συστήματα Διαχείρισης Περιεχομένου Ανοικτού Λογισμικού

Τα συστήματα αυτά παρέχουν αποτελεσματικούς μηχανισμούς και εργαλεία για την εύκολη και αποτελεσματική διαχείριση του περιεχομένου που διαθέτει ένας δικτυακός τόπος. Όλοι αυτοί οι μηχανισμοί αφορούν στην πλειονότητά τους την εισαγωγή, αποθήκευση και ανάκτηση της πληροφορίας όπως αυτοί περιγράψαμε στην ενότητα 2.2. Παρατηρείται ωστόσο ότι απουσιάζουν μηχανισμοί ανάκτησης και κατηγοριοποίησης του περιεχομένου που διαχειρίζονται. Εξαίρεση σε αυτόν τον κανόνα αποτελεί το Drupal το οποίο δίνει μια κάποια δυνατότητα στο διαχειριστή του δικτυακού τόπου να δημιουργεί ταξινομίες (taxonomy).

Η λειτουργικότητα του Drupal επιτρέπει την ταξινόμηση του περιεχομένου σε κατηγορίες και υποκατηγορίες. Υποστηρίζονται τρεις τύποι: λεξικά (πολλαπλές λίστες κατηγοριών), θησαυροί (λεξικά που επιτρέπουν τον καθορισμό σχέσεων μεταξύ των όρων) και ταξινομίες (λεξικά όπου οι σχέσεις μεταξύ των όρων που περιέχει το λεξικό δημιουργούν μια ιεραρχία). Η διαδικασία που ακολουθείται για την ταξινόμηση του περιεχομένου περιλαμβάνει σε πρώτη φάση τη δημιουργία

[12] Σύγκριση των δυνατοτήτων αλλά και μια πλήρης περιγραφή της λειτουργικότητάς τους μπορεί να βρεθεί στο δικτυακό τόπο, http://www.cmsmatrix.org. Στο [53] τα ΣΔΠ συγκρίνονται βάσει των εργαλείων και των τεχνολογιών που χρησιμοποιούν για να ικανοποιήσουν τις τέσσερις φάσεις της διαχείρισης περιεχομένου ενώ προτείνεται μια ταξινομία για την αξιολόγηση κάθε ΣΔΠ.

του λεξικού και στη συνέχεια την ανάθεση του περιεχομένου που εισάγεται στο δικτυακό τόπο σε κάποιον από τους όρους του λεξικού ή του θησαυρού.

Η προσέγγιση που ακολουθείται από το Drupal έχει τα ακόλουθα μειονεκτήματα:

- Η ταξινομία δημιουργείται από το διαχειριστή. Δεν ακολουθούνται πρότυπα ενώ η δημιουργία της βασίζεται στην ιδέα που έχει ο διαχειριστής για κάποιο θέμα
- Το είδος του περιεχομένου που μπορεί να κατηγοριοποιηθεί ακολουθώντας την παραπάνω προσέγγιση αφορά μόνο περιεχόμενο που δημιουργεί ο ίδιος ο χρήστης (ιστορίες, ψηφοφορίες, καταχωρήσεις). Πρόκειται ουσιαστικά για ένα τοπικό μηχανισμό όπου δεν υπάρχει πρόβλεψη κατηγοριοποίησης περιεχομένου που μπορεί να προέρχεται από άλλες πηγές π.χ. το διαδίκτυο
- Η κατηγοριοποίηση γίνεται εκ των υστέρων και όχι σε πραγματικό χρόνο. Η διαδικασία περιλαμβάνει την εισαγωγή του περιεχομένου και την αντιστοίχησή του στη συνέχεια σε κάποια από τις υπάρχουσες κατηγορίες.

Στην παρούσα μεταπτυχιακή εργασία υλοποιείται ένα περιβάλλον ημι-αυτόματης κατηγοριοποίησης του περιεχομένου που συγκεντρώνει ένας δικτυακός τόπος από διαφορετικές πηγές - web, εσωτερική πληροφορία, κλπ. (Web Content Classifier – WCC) το οποίο διευρύνει ακριβώς τη (κρίσιμη) λειτουργικότητα που απουσιάζει από τα υφιστάμενα Συστήματα Διαχείρισης Περιεχομένου, βελτιώνοντας την αποτελεσματικότητα τους. Το σύστημα που αναπτύχθηκε καλύπτει μάλιστα μια ανάγκη (κατηγοριοποίηση περιεχομένου) που γίνεται όλο και πιο σημαντική στις σημερινές συνθήκες ανάπτυξης του Web (Web 2.0, βλ. παρακάτω). Η υλοποίησή του ενσωματώθηκε το ΣΔΠ ΕΛ/ΑΚ ATL Content Management Engine (ATL CME) που έχει αναπτυχθεί στο Πανεπιστήμιο Κρήτης και χρησιμοποιείται σήμερα σε μια σειρά από δημόσιους οργανισμούς, κυρίως από το Πανεπιστημιακό Δίκτυο ΕΔΕΤ (http:///www.grnet.gr). Μια συνοπτική περιγραφή αυτού του περιβάλλοντος παρατίθεται στην ενότητα 2.4.

2.4 Το Σύστημα Διαχείρισης Περιεχομένου ATL CME

Το Σύστημα Διαχείρισης Περιεχομένου ATL CME έχει αναπτυχθεί στο πλαίσιο του Εθνικού Δικτύου Έρευνας και Τεχνολογίας – ΕΔΕΤ[13], υπό την ευθύνη της Ομάδας ΑΤΛΑΝΤΙΔΑ (ATLANTIS Group)

[13] Εθνικό Δίκτυο Έρευνας και Τεχνολογίας, http://www.grnet.gr

του Πανεπιστημίου Κρήτης. Πρόκειται για ένα περιβάλλον που αναπτύσσεται συνεχώς για να καλύψει ανάγκες οργανισμών που διαθέτουν μεγάλο όγκο πληροφορίας η οποία και πρέπει να δημοσιευθεί στο διαδίκτυο με μια εύκολα διαχειρίσιμη δομή. Προς αυτή την κατεύθυνση το ATL CME προσφέρει την απαραίτητη υποδομή και μεθοδολογία για τη δημιουργία μιας δικτυακής πύλης. Διατίθεται ως ένα προϊόν ελεύθερου λογισμικού και χρησιμοποιεί τεχνολογίες ανοικτού κώδικα (open source), τη γλώσσα προγραμματισμού PHP, XML και RSS πρότυπα.

2.4.1 Περιβάλλον λειτουργίας του ATL CME

Με την ενεργοποίηση της λειτουργίας του ATL CME από τον υπεύθυνο διαχειριστή (webmaster), δημιουργείται ο «σκελετός» της υπό κατασκευή δικτυακής πύλης με τα ακόλουθα στοιχεία:

- Βάση Δεδομένων (database)
- Μηχανισμός πυρήνα (core engine), που περιέχει τα προσαρτήματα λογισμικού για τη διαχείριση χρηστών, την κατασκευή του χάρτη ιστού, την υποστήριξη πολυγλωσσίας καθώς και την ασφάλεια του δικτυακού τόπου
- Προσαρτήματα λογισμικού ή λειτουργικοί μηχανισμοί (modules) – επιλογή από μία λίστα «available modules», όπως η διαχείριση αρχείων, εκδηλώσεων, συνδέσμων κ.α.
- Τμήματα πληροφορίας (blocks), που αναλαμβάνουν την παρουσίαση χρήσιμων πληροφοριών σε όλες τις σελίδες του δικτυακού τόπου.

Τα προσαρτήματα λογισμικού (modules) παρέχουν όλη τη λειτουργικότητα στο σύστημα διαχείρισης περιεχομένου ATL CME. Εκεί γίνεται η απόδοση των δικαιωμάτων πρόσβασης για τους χρήστες και τα γκρουπ των χρηστών του δικτυακού τόπου ενώ παρέχονται και οι μηχανισμοί για τη διαχείριση της πληροφορίας κάθε προσαρτήματος λογισμικού. Τα τμήματα πληροφορίας αποτελούν έναν τρόπο παρουσίασης της πληροφορίας που διαχειρίζεται ένα προσάρτημα λογισμικού. Είναι ορατά από όλες τις σελίδες του δικτυακού τόπου και παρέχουν έναν τρόπο ανάκτησης της πληροφορίας που βρίσκεται αποθηκευμένη στη βάση δεδομένων του δικτυακού τόπου.

Μετά την εγκατάσταση, η ανάπτυξη της δικτυακής πύλης συνίσταται πλέον στη διαχείριση των modules (και των «blocks») που υλοποιούν τη λειτουργικότητα που αυτή προσφέρει.

2.4.2 Λειτουργικότητα του Συστήματος Διαχείρισης Περιεχομένου ATL CME

Η λειτουργικότητα του Συστήματος Διαχείρισης Περιεχομένου ATL CME προσφέρει:

- Διαχείριση «εντός του κύκλου ζωής» της πληροφορίας που δημοσιοποιεί η δικτυακή πύλη (content lifecycle management / creation with desktop applications – publication in web formats).
- Εργαλεία για τη συλλογή πληροφορίας (content aggregation / content syndication).
- Ταξινόμηση και διαχείριση κατηγοριών περιεχομένου:
  - ⇒ Δημιουργία σύνθετων ιεραρχιών με μεγάλο βάθος και πλάτος, και (της απολύτως επαρκούς) μετα-πληροφορίας στους κόμβους της ιεραρχίας,
  - ⇒ Γραφική απεικόνιση των ιεραρχιών ώστε να είναι εύκολη η πρόσθεση νέων κατηγοριών και υποκατηγοριών,
  - ⇒ Ανακύκλωση χρήσης της πληροφορίας (από τη μία κατηγορία στη άλλη).
- Αναζήτηση και ανάκτηση πληροφορίας:
  - ⇒ Με τη χρήση της Μηχανής Αναζήτησης Google (που έχει ενταχθεί στη λειτουργία του Συστήματος ATL CME),
  - ⇒ Με τη μετατροπή desktop formats (doc, ppt, pdf, xls) σε format αναγνωρίσιμο από έναν web browser, ή άλλο κανάλι παρουσίασης (HTML, XML).
- Ενιαία εικαστική μορφή της πύλης και τρόπο πλοήγησης (unified look and feel).

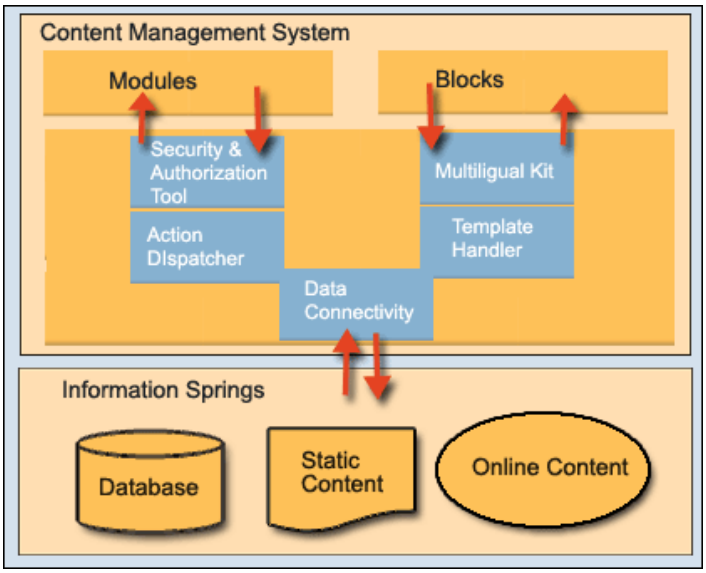
Ταυτόχρονα, υποστηρίζει την οργάνωση ad-hoc ομάδων εργασίας (groups συνεργασίας) και κοινοτήτων χρηστών, με:

- Διαδικασίες ανάθεσης ρόλων σε χρήστες (access control / role-based access control).
- Εργαλεία σύγχρονης και ασύγχρονης συνεργασίας και επικοινωνίας (file upload, link upload, calendar, forum, message and press release board, newsletter, IM & SMS interactive services).

2.4.3 Αρχιτεκτονική του Συστήματος Διαχείρισης Περιεχομένου ATL CME

Η φιλοσοφία σχεδιασμού του Συστήματος Διαχείρισης Περιεχομένου ATL CME βασίζεται σε μία έκδοση του λογισμικού προϊόντος Postnuke (έκδοση 0.64).

Η αρχιτεκτονική του Συστήματος Διαχείρισης Περιεχομένου ATL CME φαίνεται στο Σχήμα 3.



Σχήμα 3 Η αρχιτεκτονική του Συστήματος Διαχείρισης Περιεχομένου ATL CME

Στην καρδιά του συστήματος ATL CME, βρίσκεται εγκατεστημένο το Υποσύστημα «Μηχανισμός Πυρήνα» (core engine system), που παρέχει υπηρεσία, δηλαδή διευκολύνσεις (προγραμματιστικό περιβάλλον και επίπεδα αφαίρεσης) στα Προσαρτήματα λογισμικού (modules) και στα Τμήματα πληροφορίας (blocks) – που μπορούν να ενεργοποιούνται ή να απενεργοποιούνται όταν επιλέξει ο διαχειριστής. Το προγραμματιστικό περιβάλλον του πυρήνα προσομοιάζει αυτό των portlets, ένα πρότυπο για τη δημιουργία «συνιστωσών» (components) σε δικτυακές πύλες. Τα στοιχεία του πυρήνα αλληλοσυμπληρώνονται και ορίζουν ελέγχους ροής πληροφορίας για την ασφαλή λειτουργία της δικτυακής πύλης (security as access control) και την παροχή ολοκληρωμένης και δομημένης υπηρεσίας (integration).

Συνθήκες πρόσβασης και ασφάλειας: Στον πυρήνα (στοιχείο: Security and Authorization tool) εκτελείται ο έλεγχος πρόσβασης των χρηστών που χρησιμοποιούν τη δικτυακή πύλη με βάση την πολιτική του webmaster (access control policy) και η ανάθεση δικαιωμάτων σε αυτούς (role-based access control) – επί της πληροφορίας που περιέχει η πύλη (information-based access control model). Με αυτό τον τρόπο, η πληροφορία που περιέχει ο κόμβος συνίσταται σε ένα σύνολο από «αντικείμενα πληροφορίας» (information objects) με καταχωρημένα δικαιώματα πρόσβασης (ανάγνωσης / τροποποίησης: view, insert, delete, update) για χρήστες και ομάδες χρηστών. Πράγματι, το Σύστημα Διαχείρισης Περιεχομένου ATL CME χρησιμοποιεί το μοντέλο ελέγχου πρόσβασης που κάνει ταυτόχρονα χρήση της έννοιας «user» και «group». Κάθε χρήστης «ελέγχεται» με βάση την ταυτότητα που του αποδίδεται, αλλά και βάση της ομάδας χρηστών (π.χ.

ομάδα εργασίας, κοινότητα) στην οποία ανήκει, πράγμα που μπορεί να του αποδίδει συμπληρωματικά δικαιώματα πρόσβασης.

Η πληροφορία που συνδέεται με τους κανόνες πρόσβασης και τα σχετικά δικαιώματα αποθηκεύεται σε ειδικούς πίνακες, ανεξάρτητους από τα «αντικείμενα πληροφορίας» στα οποία αναφέρεται. Με αναφορά στον πίνακα χρηστών, κάθε module ορίζει, για τα «αντικείμενα πληροφορίας» που αυτό διαχειρίζεται (π.χ. ένα module του τύπου «Διαχείριση αρχείων», συνήθως διαχειρίζεται αρχεία που έχουν γεννηθεί από desktop εφαρμογές, doc, ppt, pdf, xls, jpeg κλπ.), τα αντίστοιχα δικαιώματα πρόσβασης και ενημερώνει σχετικώς το στοιχείο «Security and Authorization tool». Τέλος, η λογική των δικαιωμάτων πρόσβασης υποστηρίζει «κληρονομικότητα», με αποτέλεσμα δικαιώματα που ορίζονται σε ένα «information container» (ένα σύνολο από «αντικείμενα πληροφορίας», ένας «φάκελος αρχείων») να εφαρμόζονται και στους «φακέλους αρχείων» που έχουν ως «πρόγονο» (στο δένδρο της πληροφορίας) τον εν λόγω φάκελο. Την κληρονομικότητα έρχεται να συμπληρώσει η δυνατότητα «αρνητικών δικαιωμάτων» (negative rights), ώστε επιλεκτικά να απαγορεύεται η πρόσβαση κάποιων χρηστών σε «φακέλους απογόνους» ενός «φακέλου αρχείων».

Αποθήκευση πληροφορίας: Η αποθήκευση της πληροφορίας που περιέχει η δικτυακή πύλη βρίσκεται συγκεντρωμένη στο Υποσύστημα «Information Springs». Το σύστημα ATL CME μπορεί να διαχειρίζεται πληροφορία αποθηκευμένη σε πληθώρα μέσων, χωρίς να απασχολεί η μορφοποίηση που έχει λάβει (format). Μπορεί επίσης να συνδέεται με οποιαδήποτε βάση υποστηρίζει το πρωτόκολλο επικοινωνίας SQL. Στις δικτυακές πύλες που υποστηρίζει σήμερα το Σύστημα Διαχείρισης Περιεχομένου ATL CME, έχει κυρίως χρησιμοποιηθεί η βάση δεδομένων ανοικτού κώδικα (open source) PostgreSQL.

Παρουσίαση: Η οργάνωση και η παρουσίαση της αποθηκευμένης πληροφορίας στους επισκέπτες της δικτυακής πύλης, σε format αναγνωρίσιμο από ένα φυλλομετρητή (web browser), είναι αρμοδιότητα του Υποσυστήματος «Διαχείριση Περιεχομένου». Αυτό περιλαμβάνει τα Προσαρτήματα λογισμικού (ή λειτουργικοί μηχανισμοί, modules) και τα Τμήματα πληροφορίας (blocks). Τα προσαρτήματα λογισμικού καθορίζουν τον ακριβή τρόπο εμφάνισης των «αντικειμένων πληροφορίας» και τις συνθήκες πρόσβασης σε αυτά των χρηστών της δικτυακής πύλης (όπως απορρέουν από τα δικαιώματα που τους έχουν αναγνωρισθεί). Τα τμήματα πληροφορίας εμφανίζονται ως σταθερά τμήματα, σε όλες τις σελίδες. Η επιλογή του σχεδιαστικού προτύπου της κάθε σελίδας ορίζεται από το στοιχείο του Πυρήνα «Template handler».

Τα προσαρτήματα λογισμικού που είναι σήμερα διαθέσιμα μπορούν να κατηγοριοποιηθούν ως εξής (Σχήμα 4):

- **Δομικές Λειτουργικές Υπηρεσίες (core modules):** Modules Manager, Blocks Manager, Site map, Πολυγλωσσία, Διαχείριση Χρηστών.
- **Υποστηρικτικές Λειτουργικές Υπηρεσίες (support modules):** Διαχείριση Θεματικών Κατηγοριών, Παρουσίαση Ανακοινώσεων, Παρουσίαση Εκδηλώσεων, Κατάλογος, Διαχείριση Αρχείων, Διαχείριση Συνδέσμων, Διαχείριση Φωτογραφιών, Συλλογή περιεχομένου από τρίτες πηγές, FAQ, Newsletter, Forum, Get user feedback, banners, polls, e-mail alert, SMS alert, Αποστολή βλάβης.
- **Αυξημένες Λειτουργικές Υπηρεσίες (augmented modules):** Διαχείριση Ομάδων Εργασίας, Διαχείριση Ομάδων Εργασίας plus (εργαλεία προγραμματισμού / απολογισμού).
- **Βοηθητικές Λειτουργικές Υπηρεσίες (help modules):** Αναζήτηση, Στατιστικά.

Σχήμα 4 Λειτουργικότητα του Συστήματος Διαχείρισης Περιεχομένου ATL CME

3 Διαχείριση Περιεχομένου & "Γνώση των πολλών": Διαχείριση της Πληροφορίας στο τετράγωνο

3.1 Εισαγωγή

Το κεφάλαιο αυτό παρουσιάζει το πλαίσιο που οδήγησε στην σύλληψη και στον σχεδιασμό ενός συστήματος ημι-αυτόματης κατηγοριοποίησης του περιεχομένου που συγκεντρώνει ένας δικτυακός τόπος από διαφορετικές πηγές (web, εσωτερική πληροφορία κλ.π) – Web Content Classifier (WCC). Η τάση που δημιούργησε την ανάγκη ανάπτυξης αυτού του συστήματος είναι, σε ένα μακροσκοπικό επίπεδο, η διαφαινόμενη δυναμική μετασχηματισμού του Web που αποκαλείται «towards Web 2.0» (ο σταδιακός μετασχηματισμός του Web σε ένα περιβάλλον συνεργασίας και συνεργατικής διαχείρισης της πληροφορίας, και Web services σε εκτεταμένη κλίμακα). Σε ένα μικροσκοπικό επίπεδο, το σύστημα που αναπτύξαμε «εκμεταλλεύεται» δύο ισχυρά φαινόμενα στο Web, αυτό του «content aggregation» και αυτό της ανάδειξης συνεργατικών μοντέλων διαχείρισης περιεχομένου (personal & interpersonal content management) για να εμπλουτίσει την τεχνολογία διαχείρισης περιεχομένου ενός δικτυακού τόπου με λειτουργικότητα που επιτρέπει ημι-αυτόματη κατηγοριοποίηση των πηγών περιεχομένου και ως εκ τούτου να βελτιώσει την απόδοση ενός Συστήματος Διαχείρισης Περιεχομένου Ανοικτού Κώδικα (ΣΔΠ ΕΛ/ΑΚ).

3.2 Διαχείριση περιεχομένου «για όλους»

Οι οργανισμοί που παράγουν περιεχόμενο ζητούν να το αποθηκεύσουν και να το επαναχρησιμοποιήσουν στη συνέχεια. Το περιεχόμενο αυτό μπορεί να αποτελείται από πολύπλοκα δομημένα επιχειρησιακά έγγραφα και από άλλα, λιγότερο ή περισσότερο πολύπλοκα, που οι χρήστες τα μέλη ενός οργανισμού δημιουργούν, διαμοιράζονται, διαβιβάζουν σε άλλους και αποθηκεύουν. Ένα μέρος αυτού του περιεχομένου καταλήγει στο Web. Όλες αυτές οι ανάγκες διαχείρισης-δημοσίευσης πληροφορίας στο εσωτερικό ενός οργανισμού καλύπτονται από τα Συστήματα Διαχείρισης Περιεχομένου (ΣΠΔ) που γενικά περιλαμβάνουν:

- αποθήκες (databases) για το οποίο δημοσίευση περιεχόμενο (content repositories)

- μέσα για την διαχείριση εσωτερικών εγγράφων (e-documents) από τα οποία παράγεται, κατά ένα μεγάλο ποσοστό, το υπό δημοσίευση περιεχόμενο (document repositories).

Τα «content repositories» διαχειρίζονται αποτελεσματικά την πολυπλοκότητα της δομημένης πληροφορίας. Συγκεκριμένα, «ανακυκλώνουν» το υλικό των «document repositories» για να δημιουργήσουν πολύπλοκες δομές δεδομένων που συνιστούν «περιεχόμενο» (με την έννοια ότι περιέχουν σχέσεις και αντιστοιχίες μεταξύ δεδομένων / documents), ελέγχουν τη διαχείριση του περιεχομένου και τη διαδικασία δημοσίευσης του στη βάση κανόνων ροής εργασιών και εξουσιοδοτημένη πρόσβασης, και το μετατρέπουν σε διάφορες μορφές ώστε αυτό να είναι κατάλληλο για δημοσίευση στο Web, στις κινητές συσκευές και σε άλλες πλατφόρμες. Στο μεταξύ, αυτό που είναι «περιεχόμενο» προς διαχείριση για έναν οργανισμό διαφοροποιείται όλο και περισσότερο, με την έκρηξη των μέσων επικοινωνίας (e-mail, instant messaging κλπ), προς μορφές δομημένες βεβαίως (ένα e-mail είναι ένα δομημένο έγγραφο) αλλά όχι με τον τρόπο που δομούνται τα ηλεκτρονικά έγγραφα μιας επιχείρησης (e-documents) [82]. Το περιεχόμενο που μπορεί να αντλείται από το Web (π.χ. τιμοκατάλογοι, εμπορικοί κατάλογοι προϊόντων) κάνει τη διαφοροποίηση του υπό διαχείριση περιεχομένου ακόμη πιο εκτεταμένη.

Τα συστήματα διαχείρισης περιεχομένου που απευθύνονται στην αγορά των (μεγάλων) επιχειρήσεων διαθέτουν την ικανότητα να αντιμετωπίζουν αυτή την πολυπλοκότητα και κάνουν λειτουργικά διαφανή (transparent) την επικοινωνία ανάμεσα σε «document» & «content repositories». Λίγη προσοχή όμως είχε δοθεί στις ανάγκες των μικρότερων επιχειρήσεων, και των κυβερνητικών οργανισμών, που έχουν αντίστοιχες ανάγκες, και ακόμη λιγότερη στους απλούς χρήστες και στην αυξανόμενη ανάγκη οργάνωσης του όλο και πιο διαφοροποιημένου περιεχομένου που αυτοί διαθέτουν και πρέπει να διαχειριστούν. Πράγματι ακόμη και αυτοί οι χρήστες αρχίζουν να χρησιμοποιούν «τεχνολογία» για την ανάπτυξη προσωπικών δικτυακών τόπων οι οποίοι ξεφεύγουν από την παράθεση απλών σελίδων και διαθέτουν «συνεργατικούς μηχανισμούς» άντλησης πληροφορίας από το Web (π.χ. RSS) και αυτόματης δημοσίευσης σε έναν προσωπικό δικτυακό τόπο. Αυτή είναι μια τάση που έχει γίνει πολύ ισχυρή με τη διάδοση των blogs και των wikis. Με σημαντικά επακόλουθα.

Η χρήση των blogs από έναν απλό χρήστη του Ίντερνετ προσφέρει τη δυνατότητα επικοινωνίας με άλλους bloggers χρησιμοποιώντας συνδέσμους (reference) στα blogs τους. Χρειάζονται αποτελεσματικοί τρόποι τέτοιων «συνδέσεων». Η μία λύση, είναι η τοποθέτηση συνδέσμων στο εσωτερικό του κειμένου και μια άλλη συνιστά η μεταφορά του περιεχομένου στο εσωτερικό του

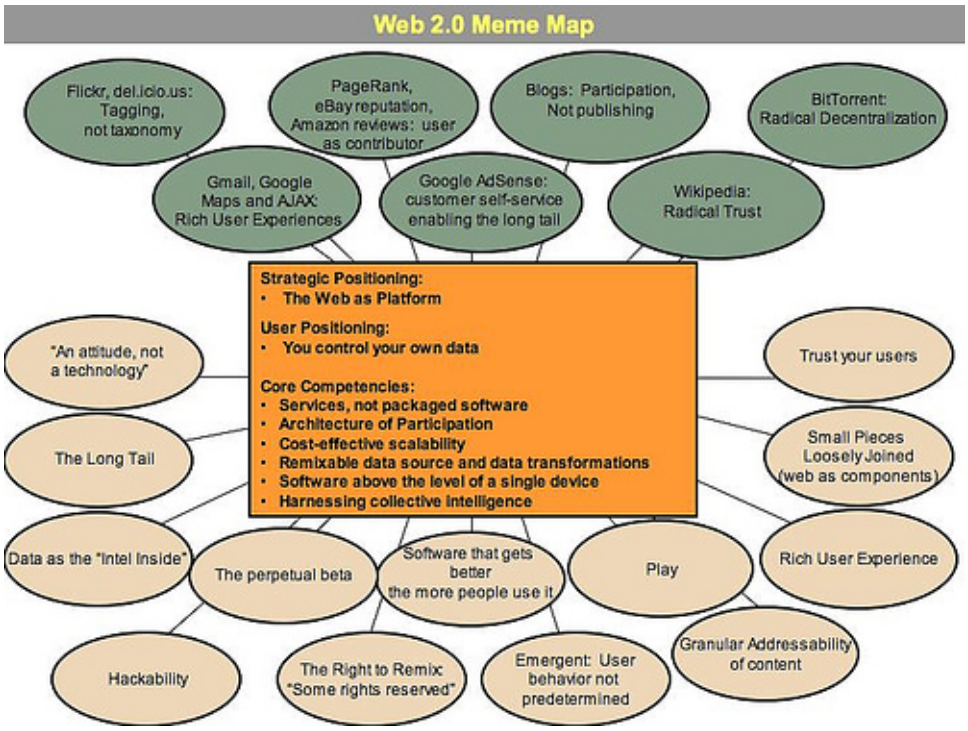
blog (aggregation). Όμως η ευκολία προσθήκης περιεχομένου στο blog ενός χρήστη του Ίντερνετ οδήγησε στην αύξηση του περιεχομένου που θα πρέπει αυτός να διαχειρίζεται. Για το λόγο αυτό απαιτούνται τρόποι οργάνωσης και παρουσίασης της πληροφορίας. Ορισμένα εργαλεία διαχείρισης blogs παρέχουν ήδη απλούς μηχανισμούς κατηγοριοποίησης με τη χρήση λέξεων – κλειδιών για το χαρακτηρισμό των κειμένων. Αυτό όμως δεν είναι αρκετό.

Τι χρειάζεται; Οι κυβερνητικοί οργανισμοί, οι μικρές επιχειρήσεις, οι απλοί χρήστες τους Ίντερνετ έχουν την ευκαιρία να αξιοποιήσουν πολλές διαφοροποιημένες πηγές περιεχομένου, από το Web και άλλες, στο πλαίσιο μιας πολιτικής διαχείρισης περιεχομένου της οποίας την πολυτέλεια μέχρι τώρα είχαν μόνο οι μεγάλες επιχειρήσεις. Ο όγκος και η διαφοροποίηση της πληροφορίας που μπορούν να διαχειριστούν μικροί οργανισμοί και οι απλοί χρήστες του Ίντερνετ είναι αυξανόμενοι. Πράγμα που δημιουργεί ανάγκες: α) αποτελεσματικών μηχανισμών άντλησης περιεχομένου από άλλους δικτυακούς τόπους (θεωρώντας ότι η άντληση «εσωτερικής» πληροφορίας για δημοσίευση στο Web είναι ένα θέμα σχετικό εύκολο σε μικρούς οργανισμούς) και β) οργάνωσης, με την έννοια της κατηγοριοποίησης (categorization), του περιεχομένου σε στάδια πριν την επιλογή δημοσίευσης, και βέβαια αφού δημοσιευθεί. Η συγκεκριμένη μεταπτυχιακή εργασία επιχειρεί να δώσει λύση στα δύο παραπάνω προβλήματα αυτό της συλλογής του περιεχομένου από το Web και στη συνέχεια της κατηγοριοποίησή του.

3.3 Από το Web 1.0 στο Web 2.0

Το θέμα της διαχείρισης ηλεκτρονικού περιεχομένου αποκτά μια πολυπλοκότητα που οφείλεται στην ίδια την αναδιάρθρωση του παγκόσμιου ιστού. Ο Παγκόσμιος Ιστός εξελίσσεται όλο και περισσότερο σε μια ολοκληρωμένη πλατφόρμα παροχής υπηρεσιών και όχι (μόνο) σαν ένα σύνολο από τυχαία ευρισκόμενα έγγραφα και περιεχόμενο. Μάλιστα έχει ήδη ξεκινήσει και η συζήτηση για την επόμενη γενιά του διαδικτύου (Web 2.0), η οποία θα βασίζεται όχι πια στο λογισμικό αλλά στις παρεχόμενες υπηρεσίες στις οποίες θα μπορεί ο χρήστης να έχει πρόσβαση από οπουδήποτε και (σχεδόν) με οποιοδήποτε μέσο. Υπηρεσίες όπως τα επιτυχημένα παραδείγματα του Google και του eBay έχουν ήδη αντικαταστήσει σαν έννοιες το Netscape, με το οποίο ήταν άρρηκτα συνδεδεμένο το διαδίκτυο στις πρώτες μέρες του. Αντίστοιχα έννοιες όπως τα wikis και τα blogs έχουν αντικαταστήσει την έννοια της προσωπικής ιστοσελίδας και του

ατομικού ιστοτόπου. Είμαστε δηλαδή στην αρχή της ολικής αλλαγής του παραδείγματος σύμφωνα με το οποίο ερμηνεύουμε και χρησιμοποιούμε το διαδίκτυο (Σχήμα 5).



Σχήμα 5 Web 2.0

Τρεις πολύ ισχυρές τάσεις σηματοδοτούν, σε ότι μας αφορά, την αναδιάρθρωση του παγκόσμιου ιστού, τις οποίες αναλύουμε παρακάτω:

1. Η εξέλιξη του σημασιολογικού ιστού για την υποστήριξη των υπηρεσιών του Web 2
2. Η μεγάλη διάδοση των τεχνολογιών προσωπικής και συνεργατικής διαχείρισης περιεχομένου (Personal και Interpersonal CM)
3. Η φαινομενική ανάπτυξη τεχνικών αυτόματης πολυσυλλογής περιεχομένου (Aggregation).

3.4 Η εξέλιξη του σημασιολογικού ιστού για την υποστήριξη του Web 2

Όπως αναφέρθηκε παραπάνω η αναδιάρθρωση του παγκόσμιου ιστού βασίζεται στην αρχή ότι πλέον ο παγκόσμιος ιστός είναι μια ολοκληρωμένη πλατφόρμα παροχής υπηρεσιών ενώ σημαντικό

ρόλο αποκτά ο χρήστης ο οποίος πλέον συμμετέχει σε συζητήσεις και συνεισφέρει / διαμοιράζεται περιεχόμενο. Το Web 2.0 σύμφωνα με τον Tim O' Reilly είναι μια «αρχιτεκτονική συμμετοχής» - ένα αστέρι του οποίου οι ακτίνες αποτελούν τις συνδέσεις μεταξύ δικτυακών εφαρμογών. Η αρχιτεκτονική αυτή βασίζεται στο «κοινωνικό λογισμικό» (social software) όπου οι χρήστες εκτός από το να καταναλώνουν περιεχόμενο δημιουργούν με τη σειρά τους νέο ενώ παρέχονται προγραμματιστικές διεπαφές που επιτρέπουν στους προγραμματιστές να προσθέτουν δικτυακές υπηρεσίες ή να διαχειρίζονται τα δεδομένα τους.

Οι τεχνολογίες που χρησιμοποιούνται από το Web 2.0 δεν είναι καινούριες αλλά στηρίζονται σε υπάρχουσες οι οποίες μάλιστα αναπτύχθηκαν κατά τη διάρκεια προσδιορισμού του σημασιολογικού ιστού. Όπως αναφέρεται: «Το Web 2.0 είναι η συνάντηση του σημασιολογικού ιστού του Tim Berners Lee με το κοινωνικό λογισμικό» (Web 2.0 is Tim Berners Lee's Semantic Web meets social software).

Η έννοια του Web 2.0 δεν περιγράφεται μόνο από την τεχνολογία αλλά από ένα πλαίσιο πολλών αρχών. Το Web 2.0 σχετίζεται με την απελευθέρωση των δεδομένων και την παροχή υπηρεσιών που μπορούν να δουλεύουν και να συνεργάζονται με δεδομένα άλλων ανθρώπων. Έτσι ενώ ο σημασιολογικός ιστός εισήγαγε έναν τρόπο αναπαράστασης των δεδομένων (XML, RDF) αλλά και ανταλλαγής τους (WSDL, SOAP) το Web 2.0 αξιοποιεί αυτά τα δεδομένα δημιουργώντας υπηρεσίες που μπορούν να αλληλεπιδρούν με πολλά είδη δεδομένων επιτρέποντας και στο χρήστη να συνεισφέρει και να δημιουργεί με τη σειρά του νέες υπηρεσίες [86].

Το πραγματικό Web 2.0 είναι η εξέλιξη του σημασιολογικού ιστού με την έννοια ενός δικτύου δεδομένων και υπηρεσιών που περιγράφονται με έναν κατανοητό τρόπο κατάλληλο να διαβαστεί και από τις μηχανές. Πρότυπα όπως η RDF αποτελούν τη βάση του Web 2.0 παρέχοντας έναν κατανεμημένο (distributed) τρόπο κατανόησης της έννοιας των δεδομένων. Τα δεδομένα εξάγονται με τέτοιο τρόπο ώστε οποιοσδήποτε και οπουδήποτε να μπορεί να χρησιμοποιήσει τα αποτελέσματα ενώ προστίθενται μεταδεδομένα προκειμένου ο παραγωγός και ο καταναλωτής να μπορεί να προσαρμόζει και να αναπτύσσει καινούριο περιεχόμενο καθώς το σύστημα μεγαλώνει [87].

3.5 Προσωπική και συνεργατική διαχείριση περιεχομένου (Personal και Interpersonal CM)

Δέκα χρόνια πριν, η διαδικασία ενημέρωσης των πολιτών ήταν μονόδρομη, το μέσο ήταν ο πομπός και ο πολίτης ο δέκτης των ειδήσεων. Η εμφάνιση όμως προσωπικών διαδικτυακών εφημερίδων (weblogs) επέτρεψε στους συντάκτες να παρακάμψουν τα παραδοσιακά μέσα ενημέρωσης εκφράζοντας πρακτικά οτιδήποτε μπορούσαν να σκεφτούν σε εκατομμύρια χρήστες σε όλο τον κόσμο [84].

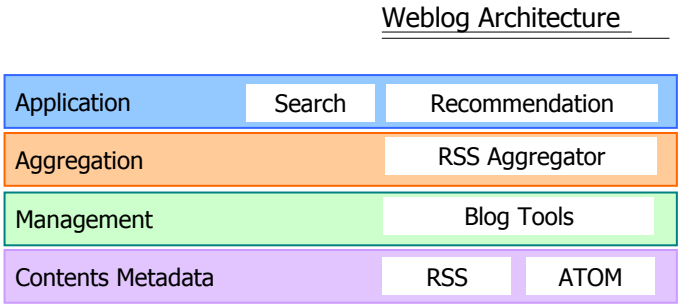
Για την έκφρασή τους οι χρήστες χρησιμοποίησαν τα blogs. Τα blogs ή weblogs από τους περισσότερους θεωρούνται ιστοσελίδες των οποίων το περιεχόμενο αποτελείται από καταχωρήσεις του χρήστη οργανωμένες σε αντίστροφη χρονολογική σειρά. Η δομή μιας σελίδας ενός weblog αποτελείται από ένα σύνολο καταχωρήσεων ενώ στο πλάι αυτών παρέχονται πληροφορίες για το προφίλ του ιδιοκτήτη. Η διαχείριση γίνεται από τον ίδιο το χρήστη χρησιμοποιώντας κάποιο εργαλείο. Τα blogs τείνουν να είναι προσωπικά και να διαβάζονται από τους ίδιους χρήστες σε τακτική βάση δημιουργώντας ένα δίκτυο από ενεργές κοινότητες. Η δημιουργία των κοινοτήτων αυτών έγκειται στη δημιουργία συνδέσμων από το ένα blog σε κάποιο άλλο. Αν και η πλειοψηφία των συντακτών αυτών των εφημερίδων εκφράζουν τις απόψεις τους απευθυνόμενοι ουσιαστικά σε ένα μικρό κύκλο συγγενών και φίλων πλέον έχουν δημιουργηθεί αντίστοιχες εφημερίδες και για πολιτικούς, οργανώσεις, ερευνητές. Οι τελευταίοι εκφράζοντας τα σχόλιά τους για διάφορα θέματα συνεισφέρουν στον τρόπο με τον οποίο συζητούνται και αναφέρονται αυτά. Επιπλέον πολλές επιχειρήσεις χρησιμοποιούν τα weblogs ως διαφημιστικά μέσα των προϊόντων τους, για τη δημοσίευση ελαττωμάτων και λύσεων σε διάφορα προβλήματα που μπορεί να προκύψουν κατά τη χρήστη των προϊόντων τους [83].

Αν κανείς αφαιρέσει το περιεχόμενο των weblogs παρατηρεί ότι το κοινό σημείο τους είναι το σχήμα, το οποίο παρέχει ένα δίκτυο συνδέσμων προς άλλα weblogs δημιουργώντας έτσι μια μορφή κοινωνικών δικτύων. Αυτό το σχήμα διαφοροποιεί το περιεχόμενο των weblogs από το περιεχόμενο που παράγεται γενικά στο διαδίκτυο. Οι εγγραφές στα blogs είναι μικρές, ανεπίσημες και αρκετές φορές προσωπικές ανεξάρτητα από το θέμα που προσεγγίζουν ενώ μπορούν να θεωρηθούν ως η αρχή μιας συζήτησης. Η ύπαρξη μηχανισμών που επιτρέπουν το σχολιασμό των εγγραφών επιτρέπουν στους αναγνώστες τη συμμετοχή τους στη συζήτηση ενώ η δυνατότητα αυτή διευρύνεται μέσω της δημοσίευσης του περιεχομένου του blog (η δημοσίευση γίνεται

συνήθως με κάποιο πρότυπο: RSS, ATOM). Ακόμη ένα χαρακτηριστικό των blogs είναι η συχνή ανανέωσή τους και η ταξινόμηση του κειμένου με χρονολογική σειρά [85].

Ένα blog είναι ένα σύνολο εγγραφών. Κάθε εγγραφή αποτελείται από συνδέσμους, ημερομηνία και ώρα δημιουργίας, ένα μόνιμο σύνδεσμο στον οποίο θα βρίσκεται η εγγραφή μετά την απομάκρυνσή της από την πρώτη σελίδα, το όνομα του συντάκτη καθώς και σχόλια που μπορεί να έχουν διατυπωθεί από άλλους χρήστες. Πιο συγκεκριμένα, οι σύνδεσμοι προς άλλα blogs καθώς και τα σχόλια που τους συνοδεύουν αποτελούν τις συνδέσεις με τους υπόλοιπους χρήστες του διαδικτύου και υποδηλώνουν την ύπαρξη ενός blog. Η ύπαρξη ημερομηνίας και ώρας για κάθε εγγραφή αποτυπώνει το χρόνο ανανέωσης ενός blog και αποτελεί το συνδετικό κρίκο του συντάκτη με τον επισκέπτη. Τέλος, η δημιουργία του μόνιμου συνδέσμου για κάθε εγγραφή συμβάλλει στη δημιουργία ακριβών αναφορών από το ένα blog στο άλλο.

Συνολικά, η αρχιτεκτονική των weblogs παρουσιάζεται στο Σχήμα 6 [88]. Στο πρώτο επίπεδο οι εγγραφές των weblogs εξάγονται μέσω κάποιου προτύπου RSS, ATOM. Η διαχείριση (προσθήκη, ανανέωση, σχολιασμός) του περιεχομένου πραγματοποιείται χρησιμοποιώντας ειδικά εργαλεία που παρέχουν οι πλατφόρμες διαχείρισης weblogs ενώ περιεχόμενο από εξωτερικές πηγές συλλέγεται μέσω εφαρμογών RSS. Η αρχιτεκτονική συμπληρώνεται από εφαρμογές αναζήτησης στο περιεχόμενο και προτάσεων από άλλους χρήστες (κάποιος χρήστης προτείνει σε άλλους περιεχόμενο σχετικό με κάποιο θέμα).



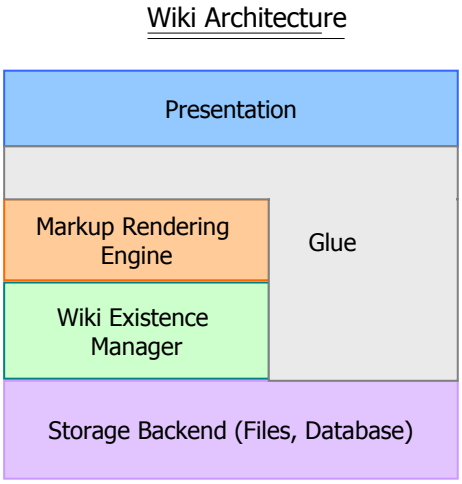
Σχήμα 6 Η αρχιτεκτονική των weblogs

Τα weblogs ουσιαστικά αποτελούν έναν τρόπο προσωπικής επικοινωνίας και έκφρασης ενός συγκεκριμένου συγγραφέα / συντάκτη με εκατομμύρια χρήστες του διαδικτύου. Η ύπαρξη συνδέσμων μεταξύ των blogs οδηγούν στη δημιουργία κοινοτήτων. Στην περίπτωση που υπάρχει ήδη μια κοινότητα που ενδιαφέρεται για κάποιο θέμα (π.χ. στα πλαίσια ενός έργου, στο

εσωτερικό ενός οργανισμού) χρησιμοποιούνται άλλες εφαρμογές συνεργασίας όπως είναι τα wikis ή οι μηχανισμοί instant messaging.

Τα wikis προσφέρουν ένα διαφορετικό τρόπο δημοσίευσης και παρουσίασης της πληροφορίας. Μερικές από τις εφαρμογές που ενσωματώνουν περιλαμβάνουν ένα περιβάλλον συγγραφής, και άλλες υπηρεσίες όπως λίστα των σελίδων που έχουν μεταβληθεί, ιστορικό των ενεργειών κάθε χρήστη και μηχανισμούς απόδοσης αδειών για τη μεταβολή των σελίδων. Επίσης, ορισμένα συστήματα περιέχουν μηχανισμούς διαχείρισης εκδόσεων και δυνατότητα επιστροφής στην προηγούμενη κατάσταση. Σε ένα τέτοιο περιβάλλον τα μέλη μιας κοινότητας προσθέτουν περιεχόμενο δημιουργώντας ή ενημερώνοντας τις σελίδες του wiki και στη συνέχεια δημιουργούν συνδέσμους προς αυτές. Υπάρχει ακόμη η δυνατότητα προσθήκης σχολίων στο περιεχόμενο άλλων χρηστών ή ανανέωσης των σελίδων που έχουν δημιουργήσει άλλοι χρήστες. Ο διαχειριστής ενός wiki έχει επιπρόσθετα δικαιώματα που αφορούν τη διαχείριση της πληροφορίας και των χρηστών [89].

Στο Σχήμα 7 παρουσιάζεται η αρχιτεκτονική των Wikis. Το περιεχόμενο που προστίθεται σε ένα wiki αποθηκεύεται είτε σε αρχεία είτε σε μια βάση δεδομένων. Το δεύτερο επίπεδο αποφασίζει την ύπαρξη μιας σελίδας στο wiki ενώ στη συνέχεια ένας μηχανισμός επιτρέπει την μετατροπή του περιεχομένου σε HTML/XML. Το κομμάτι "Glue" υλοποιεί μηχανισμούς προσθήκης περιεχομένου μέσω μιας φόρμας και επικοινωνεί με το επίπεδο της αποθήκευσης της πληροφορίας. Τέλος, το τελευταίο επίπεδο παρέχει μηχανισμούς παρουσίασης της πληροφορίας.



Σχήμα 7 Η αρχιτεκτονική των Wikis

Τα wikis χρησιμοποιούνται συνήθως στο εσωτερικό μιας εταιρείας για τη δημοσίευση εγγράφων σε ένα σημείο με δυνατότητες διαχείρισης από τους υπαλλήλους που σχετίζονται με αυτά καταβάλλοντας ελάχιστη προσπάθεια και με ελάχιστη καθυστέρηση. Ένα wiki μπορεί να χρησιμοποιηθεί ακόμη για την οργάνωση και διαχείριση σημειώσεων συναντήσεων, αλλά και ως ημερολόγιο των δραστηριοτήτων μιας επιχείρησης.

3.6 Αυτόματη πολυσυλλογή περιεχομένου (Aggregation)

Τα blogs αποτέλεσαν καταρχήν ένα σημείο όπου ο χρήστης μπορούσε να αποθηκεύσει τους συνδέσμους που τον ενδιέφεραν ενώ του δινόταν η δυνατότητα σύνδεσης με άλλα blogs. Η ύπαρξη προτύπων όπως τα RSS/ATOM έκαναν ακόμη πιο εύκολη αυτή την επικοινωνία μεταξύ των bloggers και ταυτόχρονα εισήγαγαν μια τάση συλλογής περιεχομένου. Το περιεχόμενο αυτό μπορεί να προέρχεται είτε από άλλα blogs είτε από δικτυακούς τόπους που εξάγουν το περιεχόμενό τους με κάποιο πρότυπο.

Η διαδικασία της αυτόματης (πολυ) συλλογής πληροφορίας ωστόσο δεν περιορίζεται σε αυτά που μόλις περιγράψαμε. Η δημιουργία του παγκόσμιου ιστού και η τάση της δημοσίευσης της διαθέσιμης πληροφορίας οδήγησαν στην ανάγκη δημιουργίας εργαλείων που να μπορούν να εξάγουν το περιεχόμενο που βρισκόταν αποθηκευμένο στις διάφορες σελίδες. Δημιουργήθηκαν λοιπόν, εργαλεία όπως αυτόματοι μηχανισμοί (robots), σαρωτές (crawlers) και μηχανές αναζήτησης που ουσιαστικά μετέφεραν το περιεχόμενο των δικτυακών τόπων σε άλλους. Τα robots χρησιμοποιούνται από τις μηχανές αναζήτησης και στόχος τους είναι ο ευρετηριασμός ενός δικτυακού τόπου προκειμένου το περιεχόμενό του να είναι διαθέσιμο στη βάση δεδομένων της μηχανής αναζήτησης. Οι web crawlers ακολουθούν μια διαδικασία σάρωσης του διαδικτύου και μεταφέρουν το περιεχόμενο που βρίσκουν σε κάποιο δικτυακό τόπο. Με αυτή την μέθοδο, το περιεχόμενο ανακτάται από τις τρίτες πηγές στο τρέχον σχήμα (περιεχόμενο και διάταξη). Δεδομένου όμως ότι ένας web crawler σαρώνει όλο το διαδίκτυο και άρα μπορεί να επιστρέφει και περιεχόμενο που δεν είναι σχετικό με κάποιο θέμα χρησιμοποιήθηκαν άλλα εργαλεία που διενεργούν εστιασμένη σάρωση (focused crawling). Τα αποτελέσματα είναι λιγότερα και περισσότερο ακριβή. Η λειτουργία των παραπάνω εργαλείων έγκειται στην εξαγωγή της χρήσιμης πληροφορίας αφού αναλυθεί η δομή των ιστοσελίδων ενώ η διαδικασία περιλαμβάνει επίσης τον καθορισμό αρκετών παραμέτρων που αφορούν τη διαδικασία μεταφοράς του περιεχομένου [58].

Στη συνέχεια δημιουργήθηκαν εργαλεία που συλλέγουν πληροφορία από δικτυακούς τόπους και αναλύουν το περιεχόμενο τους χωρίς να είναι απαραίτητη η γνώση της δομής των ιστοσελίδων. Τα εργαλεία αυτά ονομάζονται συναθροιστές (aggregators). Οι συναθροιστές μπορούν να κατασκευάζουν συλλογές περιεχομένου που προέρχεται είτε από το εσωτερικό ενός οργανισμού, είτε από πολλούς οργανισμούς, είτε χρησιμοποιώντας πληροφορία και από τους δύο. Η χρήση τους αποσκοπεί είτε στην αξιολόγηση και σύγκριση συγκεκριμένων αγαθών και υπηρεσιών είτε στη δημιουργία νέων συλλογών περιεχομένου ενώ η λειτουργία τους βασίζεται στις σχέσεις που δημιουργούνται με τους δικτυακούς τόπους προς συνάθροιση [34].

Ένας άλλος τρόπος για την άντληση περιεχομένου από έναν δικτυακό τόπο περιλαμβάνει τη χρήση προτύπων όπως τα RSS / ATOM. Η μέθοδος αυτή ονομάζεται «syndication». Η διαδικασία περιλαμβάνει την διανομή του περιεχομένου ενός δικτυακού τόπου προς όλους τους ενδιαφερομένους μέσω κάποιου προτύπου. Ο ενδιαφερόμενος δικτυακός τόπος αναλαμβάνει τη δημιουργία μιας εφαρμογής για τη μεταφορά του περιεχομένου σε αυτόν. Το πλεονέκτημα αυτής της μεθόδου είναι η αυτόματη ενημέρωση του περιεχομένου στην περίπτωση αλλαγών.

Ουσιαστικά βρισκόμαστε μπροστά σε ένα πολύ ισχυρό φαινόμενο, μια σειρά από καινούργιες διαδικασίες άντλησης της πληροφορίας που όπως παρατηρεί ο S. Madnick εφαρμόζονται τόσο στο εσωτερικό ενός οργανισμού όσο και μεταξύ οργανισμών (αλλά και στο επίπεδο του απλού χρήστη του Ίντερνετ) και προσφέρουν τη δυνατότητα συγκρότησης «βιβλιοθηκών πληροφορίας» (information collections): «Aggregators are used to build integrated information collections for many purposes, such as forming comparisons and managing relationships. These new integrated information collections can be built from information sources inside an organization (i.e., intra-organizational), between organizations (i.e., interorganizational) or both. Comparison type aggregators are focused on collecting information about specific goods and services for evaluation. Shopbots such as for those for purchasing books, music and electronics are good examples of this capability. Relationship type aggregators form new information collections based on relationships with the aggregatees. For example, financial account aggregators (e.g., Yodlee, VerticalOne, CashEdge) are being adopted by major financial (e.g., Chase, Citibank, Merrill Lynch) and non-financial institutions (e.g., CNBC, AOL)».

	Comparison	**Relationship**
Inter-Organizational	Comparison of book prices or shipping costs from alternative suppliers	Consolidation of all one's frequent flyer or financial accounts
Intra-Organizational	Comparison of manufacturing costs from multiple plants	Consolidation of all information about each customer from the company's separately maintained web sites across functions (e.g., accounting, service) and geography (e.g., domestic and international)

Πίνακας 4 Παραδείγματα τύπων πολυσυλλογής [34]

3.7 Επίσημη και ανεπίσημη διαχείριση περιεχομένου (Formal and informal CM)

Η διαθεσιμότητα όλων των παραπάνω εργαλείων αλλά και η ευκολία δημιουργίας δικτυακών τόπων συνέβαλλε στη συγκέντρωση μεγάλου όγκου περιεχομένου το οποίο χρίζει μεθόδων οργάνωσης και κατηγοριοποίησης.

Η διαδικασία της οργάνωσης του περιεχομένου προσεγγίζεται διαφορετικά ανάλογα με τον οργανισμό στον οποίο απευθύνεται. Οι οργανισμοί που παράγουν περιεχόμενο είναι είτε οι επιχειρήσεις είτε άλλοι φορείς. Στην πρώτη περίπτωση το περιεχόμενο υπόκεινται σε περιορισμούς όσον αφορά τη δημοσίευσή του. Στη δεύτερη συνήθως το παραγόμενο περιεχόμενο δημοσιεύεται στο διαδίκτυο και μάλιστα χρησιμοποιούνται κάποιοι από τους τρόπους που περιγράφηκαν παραπάνω για τον εμπλουτισμό του.

Η διαχείριση του περιεχομένου στο περιβάλλον μιας επιχείρησης εξαρτάται από τις τεχνικές που θα εφαρμοστούν σε όλες τις φάσεις του κύκλου ζωής του (Σχήμα 8). Οι μέθοδοι που εφαρμόζονται διαφέρουν ανάλογα με την εταιρεία, ωστόσο υπάρχουν κοινά χαρακτηριστικά που αφορούν τα στάδια προετοιμασίας του περιεχομένου. Τα στάδια αυτά είναι [90]:

- Δημιουργία και συλλογή του περιεχομένου

Στο πρώτο στάδιο καταρχήν αναγνωρίζεται το περιεχόμενο που θα δημιουργηθεί καθώς και εκείνο που υπάρχει διαθέσιμο. Το περιεχόμενο παράγεται είτε στο εσωτερικό μιας επιχείρησης είτε συλλέγεται από κάποιο συνεργάτη της. Ένα ΣΔΠ παρέχει εργαλεία για την δημιουργία του περιεχομένου ενσωματώνοντας πρότυπα όπως το XML. Στην περίπτωση που αυτό υπάρχει διαθέσιμο σε έντυπη μορφή μετατρέπεται σε ηλεκτρονική μορφή με τα αντίστοιχα μεταδεδομένα. Το στάδιο αυτό περιλαμβάνει επίσης διαδικασίες αποθήκευσης των αναφορών που παράγονται στα διάφορα κομμάτια της επιχείρησης.

- Διαχείριση του περιεχομένου

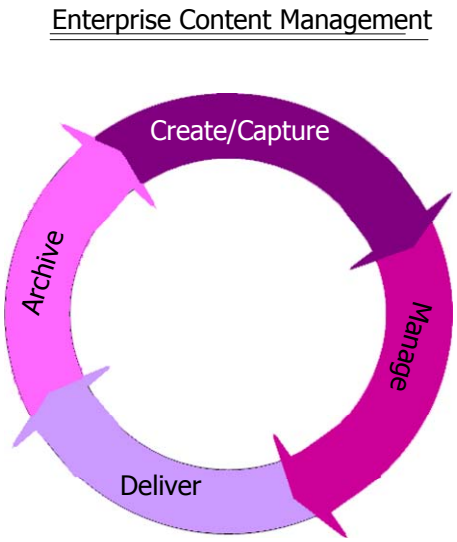
Η διαχείριση περιλαμβάνει τον ευρετηριασμό, την ομαδοποίηση καθώς και τη δημιουργία συνδέσμων μεταξύ του περιεχομένου και των βάσεων δεδομένων στις οποίες αποθηκεύεται. Στο σημείο αυτό εισάγονται επίσης διαδικασίες κατηγοριοποίησης του περιεχομένου βάσει λέξεων – κλειδιών ή όρων. Επίσης προσδιορίζεται η ταυτότητα του συγγραφέα και όλα τα απαραίτητα γλωσσολογικά και νομικά πρότυπα της επιχείρησης. Ορισμένες υπηρεσίες περιλαμβάνουν ακόμη τον έλεγχο εκδόσεων, την διαδικασία ελέγχου των σταδίων της έγκρισης, επισκόπησης και δημοσίευσης του περιεχομένου καθώς και την απόδοση αδειών πρόσβασης. Τέλος, στο στάδιο αυτό αποφασίζεται το πρότυπο παρουσίασης που θα χρησιμοποιηθεί για το έγγραφο.

- Παράδοση του περιεχομένου

Το τρίτο στάδιο περιλαμβάνει τον διαχωρισμό της πληροφορίας από το πρότυπο (layout) που χρησιμοποιείται ώστε να είναι δυνατή η διανομή του περιεχομένου σε διαφορετικά είδη π.χ. XML. Παρέχονται επίσης, υπηρεσίες παράδοσης του περιεχομένου σε προκαθορισμένο χρόνο καθώς και υπηρεσίες πολυγλωσσίας.

- Συντήρηση – Αρχειοθέτηση του περιεχομένου

Οι προσφερόμενες υπηρεσίες σε αυτό το στάδιο αποσκοπούν στη συνεχή ενημέρωση (up-to-date) του περιεχομένου αλλά και την αρχειοθέτησή του με βάση τους επιχειρησιακούς κανόνες που ενσωματώνει μια επιχείρηση.



Σχήμα 8 Στρατηγική της διαχείρισης περιεχομένου των επιχειρήσεων

Όπως φαίνεται παραπάνω η διαχείριση του περιεχομένου μιας επιχείρησης περιλαμβάνει ένα σύνολο από πολύπλοκες διαδικασίες που εφαρμόζονται σε όλα τα στάδια της ζωής ενός εγγράφου. Στην περίπτωση ενός μικρού οργανισμού η διαδικασία είναι απλούστερη αφού το μεγαλύτερο μέρος του περιεχομένου συλλέγεται από τρίτες πηγές ενώ η διανομή και η αρχειοθέτηση δεν ακολουθούν συγκεκριμένα πρότυπα. Προκύπτουν λοιπόν δύο ανάγκες οι οποίες αφορούν την δημιουργία μηχανισμών συλλογής περιεχομένου από διάφορες πηγές καθώς και τις μεθόδους διαχείρισης και οργάνωσης του περιεχομένου αυτού.

Η συλλογή του περιεχομένου πραγματοποιείται χρησιμοποιώντας μηχανισμούς συνάθροισης (aggregators) / σάρωσης (crawlers) καθώς και μηχανισμούς διαχείρισης του περιεχομένου των δικτυακών τόπων που το παρέχουν με κάποιο πρότυπο.

Η δεύτερη μέθοδος, η διαχείριση του περιεχομένου περιλαμβάνει μεθόδους οργάνωσης και κατηγοριοποίησης. Ένα παράδειγμα αυτής της οργάνωσης αποτελεί ο δικτυακός τόπος CiteULike[14]. Το CiteULike προσφέρει έναν τρόπο οργάνωσης των δημοσιεύσεων ενός χρήστη δημιουργώντας ουσιαστικά μια βιβλιοθήκη στο διαδίκτυο. Η διαδικασία περιλαμβάνει την εγγραφή του χρήστη και την προσθήκη των δημοσιεύσεων που τον ενδιαφέρουν στο σύστημα. Η υπηρεσία εξάγει αυτόματα τις πληροφορίες της δημοσίευσης (citation) διευκολύνοντας και άλλους χρήστες στην

[14] CiteULike, http://www.citeulike.org/

αναζήτησή της. Η αποθήκευση των δημοσιεύσεων πραγματοποιείται σε ένα μηχάνημα κεντρικά οπότε ο χρήστης έχει πρόσβαση στη βιβλιοθήκη ανεξάρτητα από τη φυσική του θέση. Ανάλογες προσπάθειες στοχεύουν στην οργάνωση της συλλογής των αγαπημένων δικτυακών τόπων ενός χρήστη (bookmarks)[15]. Και σε αυτή την περίπτωση ο χρήστης καλείται να προσθέσει το σύνολο των δικτυακών τόπων σε ένα σύστημα οπότε στη συνέχεια δημιουργούνται βιβλιοθήκες συνδέσμων για κάποιο θέμα.

Μια άλλη τάση στην οργάνωση του περιεχομένου βασίζεται στο περιεχόμενο που έχει αποθηκευμένο κάποιος χρήστης στον υπολογιστή του (desktop). Ένα παράδειγμα αυτής της προσέγγισης αποτελεί το προϊόν «έξυπνος κατάλογος» (smart folder) που υλοποιείται από την εταιρεία Blinkx[16]. Η εφαρμογή αυτή δημιουργεί καταλόγους οι οποίοι ανανεώνονται όταν προστίθεται νέα πληροφορία. Η δημιουργία των καταλόγων βασίζεται στις έννοιες που περιγράφει το περιεχόμενο. Το είδος του περιεχομένου που αποθηκεύεται σε αυτούς τους καταλόγους ποικίλει και μπορεί να περιλαμβάνει εκτός από δεδομένα που είναι τοπικά αποθηκευμένα στο δίσκο, περιεχόμενο από δικτυακούς τόπους, νέα κτλ. Ο χρήστης μπορεί να πραγματοποιήσει αναζητήσεις στο περιεχόμενο χρησιμοποιώντας λέξη – φράση. Ανεξάρτητα από αυτό όμως ο κατάλογος μπορεί να προτείνει επίσης πληροφορία βασισμένη στο περιεχόμενο (context) των εγγράφων που διαχειρίζεται ο χρήστης εκείνη τη στιγμή. Μια άλλη δυνατότητα που παρέχεται από την υπηρεσία αυτή περιλαμβάνει το ιστορικό των αρχείων που έχει προσπελάσει ο χρήστης.

Διαφορετική προσέγγιση για το ίδιο πρόβλημα αποτελεί η αναζήτηση περιεχομένου στον υπολογιστή ενός χρήστη (desktop search). Η προσέγγιση αυτή υλοποιείται από το MSN και το Google και περιλαμβάνει τον ευρετηριασμό του περιεχομένου που βρίσκεται αποθηκευμένο στον υπολογιστή ενός χρήστη με δυνατότητες αναζήτησης μέσα σε αυτό.

3.8 Υποδομή της ανεπίσημης διαχείρισης περιεχομένου (Back end στο informal CM)

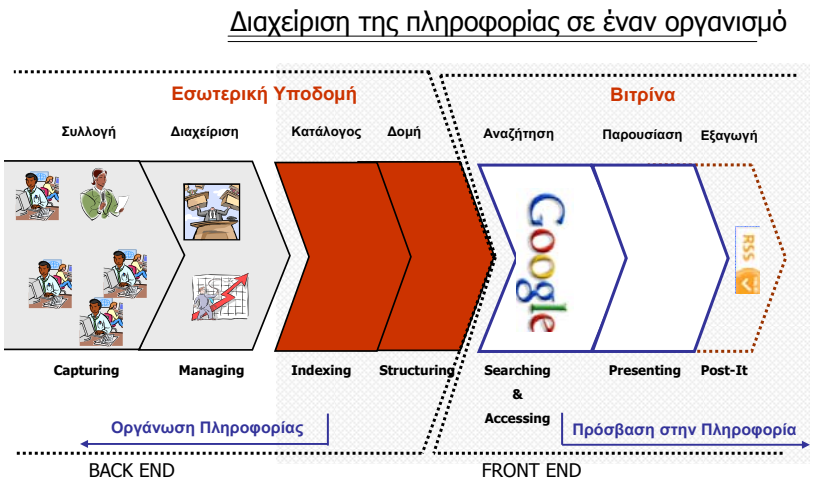
Η διαχείριση του περιεχομένου σε οργανισμούς που δεν παράγουν μεγάλο όγκο πληροφορίας έγκειται, όπως αναφέρθηκε στην προηγούμενη ενότητα, στην ανάπτυξη μεθόδων συλλογής

[15] Scraping Google Directory to do Semantic Clustering of Tags from intersp.icio.us, http://jonaquino.blogspot.com/2005/03/scraping-google-directory-to-do.html

[16] Blinkx, http://www.blinkx.com/overview.php

περιεχομένου από τρίτες πηγές και οργάνωσης / κατηγοριοποίησης του περιεχομένου αυτού. Τα στάδια του κύκλου ζωής της πληροφορίας στα πλαίσια ενός συστήματος διαχείρισης παρουσιάζονται στο Σχήμα 9. Το σύστημα αποτελείται από δύο μέρη: την εσωτερική υποδομή και τη βιτρίνα.

Η εσωτερική υποδομή περιλαμβάνει διαδικασίες συλλογής και διαχείρισης της πληροφορίας καθώς και μεθόδους διαμόρφωσης καταλόγων πληροφορίας και δομής (indexing, structuring). Από την άλλη πλευρά η βιτρίνα περιλαμβάνει διαδικασίες πρόσβασης στο περιεχόμενο δηλαδή μεθόδους αναζήτησης και ανάκτησης της πληροφορίας καθώς και εξαγωγή της με κάποιο πρότυπο.



Σχήμα 9 Ο κύκλος ζωής της πληροφορίας

Στα πλαίσια της παρούσας μεταπτυχιακής εργασίας ασχολούμαστε με την εσωτερική υποδομή ενός ΣΔΠ και συγκεκριμένα με τις μεθόδους συλλογής και διαχείρισης του περιεχομένου.

Το είδος του περιεχομένου που διαχειρίζεται το σύστημα αφορά όχι μόνο περιεχόμενο που προέρχεται από το εσωτερικό ενός οργανισμού αλλά και περιεχόμενο που προέρχεται από το διαδίκτυο με διαφορετικούς τρόπους. Το εσωτερικό περιεχόμενο αφορά περιεχόμενο που εισάγεται στο ΣΔΠ χρησιμοποιώντας μηχανισμούς εισαγωγής περιεχομένου που παρέχονται από αυτό. Το περιεχόμενο από το διαδίκτυο προέρχεται από δύο πηγές: στην πρώτη περίπτωση προέρχεται από τις μηχανές αναζήτησης και στη δεύτερη από άλλους δικτυακούς τόπους που παρέχουν το περιεχόμενό τους μέσω κάποιου προτύπου. Η διαδικασία συλλογής αφορά λοιπόν

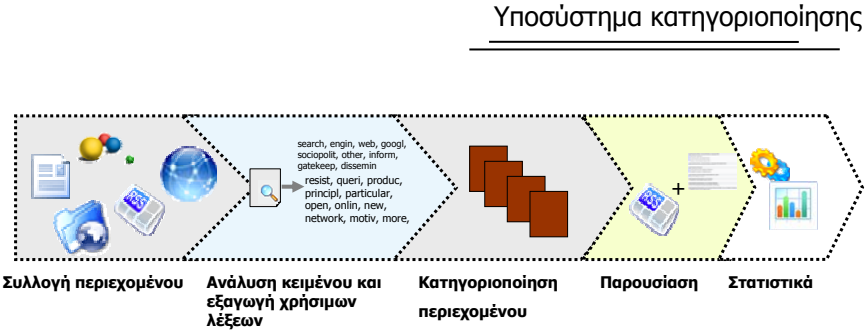
τον εμπλουτισμό των πηγών περιεχομένου ενός ΣΔΠ δίνοντας έμφαση στο περιεχόμενο που μπορεί να συγκεντρωθεί από το διαδίκτυο.

Η συλλογή του περιεχομένου και η αποθήκευσή του σε μια βάση δεδομένων ωστόσο δεν έχει αξία εάν δεν εφαρμόζονται κάποιοι κανόνες για τη δόμηση αλλά και την κατηγοριοποίηση του με τρόπο που να είναι κατανοητός από το χρήστη του συστήματος. Προς αυτή την κατεύθυνση στο πλαίσιο της παρούσας εργασίας υλοποιήθηκαν μηχανισμοί για την κατηγοριοποίηση και την παρουσίαση του περιεχομένου που βρίσκεται αποθηκευμένο στη βάση δεδομένων ενός δικτυακού τόπου. Η διαδικασία χρησιμοποιεί μια ταξινομία που είναι διαθέσιμη στο διαδίκτυο. Εφόσον το περιεχόμενο κατηγοριοποιηθεί παρέχεται δυνατότητα πλοήγησης στο αποτέλεσμα της κατηγοριοποίησης. Τέλος, παρέχονται στατιστικά στοιχεία για την διαδικασία.

4 Σχεδίαση υποσυστήματος ημιαυτόματης κατηγοριοποίησης κειμένων

4.1 Εισαγωγή

Στο κεφάλαιο αυτό περιγράφονται τα βασικά στοιχεία που συνθέτουν το υποσύστημα ημιαυτόματης κατηγοριοποίησης κειμένων το οποίο προσαρμόστηκε στο ΣΔΠ ATL CME. Οι διαφορετικές διαδικασίες που περιλαμβάνει το υποσύστημα συνδέονται μεταξύ τους όπως φαίνεται στο Σχήμα 10.



Σχήμα 10 Εποπτική εικόνα του υποσυστήματος κατηγοριοποίησης περιεχομένου

Τα στάδια της διαδικασίας κατηγοριοποίησης ενός κειμένου είναι τα εξής:

- **Συλλογή:** Η διαδικασία της συλλογής περιλαμβάνει τη διαδικασία εισαγωγής του περιεχομένου στο υποσύστημα κατηγοριοποίησης και την αποθήκευσή του (ενότητα 4.2 με αναδρομή στη σχετική βιβλιογραφία).
- **Ανάλυση κειμένου:** Η ανάλυση του κειμένου και η εξαγωγή χρήσιμων λέξεων περιλαμβάνει διαδικασίες επεξεργασίας και μετατροπής του περιεχομένου σε κατάλληλη μορφή προκειμένου να κατηγοριοποιηθεί (ενότητα 4.3 με αναδρομή στη σχετική βιβλιογραφία).
- **Κατηγοριοποίηση:** η κατηγοριοποίηση του περιεχομένου ολοκληρώνεται με την εφαρμογή ενός αλγορίθμου κατηγοριοποίησης (ενότητα 4.4 με αναδρομή στη σχετική βιβλιογραφία).

4.2 Συλλογή κειμένων από το διαδίκτυο: αυτόματη, ημιαυτόματη, δια χειρός

Το πρώτο στάδιο στην διαδικασία κατηγοριοποίησης του περιεχομένου αποτελεί η συλλογή και αποθήκευσή του. Ανατρέχοντας στη βιβλιογραφία προτείνονται διάφοροι τρόποι συλλογής με πιο διαδεδομένη την προσέγγιση που στηρίζεται στη δημιουργία «συσκευαστών» (wrappers) και «συναθροιστών» (aggregators). Μια δεύτερη προσέγγιση αφορά τη συλλογή περιεχομένου από δικτυακούς τόπους που εξάγουν το περιεχόμενο τους μέσω κάποιου προτύπου. Στην ενότητα που ακολουθεί συνοψίζονται οι κύριες προσεγγίσεις στο θέμα της συλλογής περιεχομένου από το διαδίκτυο, όπως παρουσιάζονται στη βιβλιογραφία, και περιγράφονται οι τρόποι συλλογής περιεχομένου στο υποσύστημα κατηγοριοποίησης.

4.2.1 Το πρόβλημα της συλλογής περιεχομένου από το διαδίκτυο στη βιβλιογραφία

Η διαθεσιμότητα εγγράφων στο διαδίκτυο δημιούργησε την ανάγκη εύρεσης αποτελεσματικών τρόπων για την «εκμετάλλευση» του περιεχομένου τους. Οι απλές σελίδες αποτελούν κυρίως έναν τρόπο παρουσίασης της πληροφορίας που περιέχουν χωρίς να δίνουν επιπλέον δυνατότητες εκμετάλλευσης και διαχείρισης της πληροφορίας. Για το λόγο αυτό στα μέσα της δεκαετίας του ’90 προτάθηκε η αρχιτεκτονική της «μεσολάβησης» (mediation architecture). Η αρχιτεκτονική αυτή διακρίνει τρία είδη διαδικασιών: αυτούς που διευκολύνουν (facilitators) την ολοκλήρωση των εφαρμογών, τους μεσολαβητές (mediators) οι οποίοι παρέχουν εφαρμογές για την επικοινωνία με διαφορετικές πηγές και τους «συσκευαστές» (wrappers) που παρέχουν μια αντιστοίχηση για την αναπαράσταση των δεδομένων, του μοντέλου των δεδομένων καθώς και της γλώσσας επερωτήσεων που χρησιμοποιείται. Η παραπάνω αρχιτεκτονική υιοθετήθηκε από το μεγαλύτερο μέρος της επιστημονικής κοινότητας και χρησιμοποιήθηκε ως οδηγός για την εξαγωγή δομημένης πληροφορίας από τα διαθέσιμα έγγραφα. Για να γίνει αυτό, δημιουργήθηκαν εφαρμογές που εκμεταλλεύονται την ομοιότητα πολλών εγγράφων εξάγοντας πρότυπα (templates) που περιγράφουν τη δομή τους. Τα πρότυπα αυτά χρησιμοποιούνται για την αυτόματη αναγνώριση και εξαγωγή των δεδομένων και μπορούν να εφαρμοστούν σε HTML σελίδες ή αρχεία δεδομένων (όπως Word, Excel) [71], [72], [76].

Οι wrappers ταξινομούνται ανάλογα με τον τρόπο επεξεργασίας των εγγράφων (ιστοσελίδων), τις προδιαγραφές που χρησιμοποιούν για την αντιστοίχηση των εγγράφων και με το πόσο δηλωτικοί είναι. Οι wrappers αντιμετωπίζουν τις ιστοσελίδες είτε ως δέντρα είτε ως ροές χαρακτήρων. Στην

πρώτη περίπτωση τα έγγραφα κωδικοποιούνται με κάποιο μοντέλο αναπαράστασης (π.χ. DOM) ενώ στη δεύτερη χρησιμοποιούνται κανονικές εκφράσεις. Η δεύτερη κατηγοριοποίηση έγκειται στον τρόπο που δημιουργούνται οι κανόνες για την εξαγωγή του κειμένου (χειρωνακτικός, ημι-αυτόματος, αυτόματος). Στη χειρωνακτική προσέγγιση ο χρήστης είναι υπεύθυνος για την δημιουργία κανόνων εξαγωγής αναλύοντας ένα αντιπροσωπευτικό σύνολο σελίδων καθώς και για την ανανέωσή τους. Στην αυτόματη προσέγγιση οι χρήστες θα πρέπει πρώτα να σχολιάσουν ένα σύνολο εκπαιδευτικών παραδειγμάτων χρησιμοποιώντας την αντίστοιχη διεπαφή (interface). Στη συνέχεια αλγόριθμοι μηχανικής μάθησης εφαρμόζονται για την εξαγωγή των αντίστοιχων κανόνων. Στην ημι-αυτόματη προσέγγιση δε χρησιμοποιούνται αλγόριθμοι μηχανικής μάθησης αλλά η δημιουργία του αρχείου προδιαγραφών γίνεται ευκολότερη αφού το σύστημα προτείνει πρότυπα και ο χρήστης θα πρέπει να τα εγκρίνει ή να τα αλλάξει. Ο τρίτος τρόπος κατηγοριοποίησης γίνεται βάσει της αυτονομίας των κανόνων εξαγωγής του κειμένου από τη γλώσσα προγραμματισμού που χρησιμοποιείται. Ορισμένοι wrappers παρέχουν τη δική τους γλώσσα ενώ άλλοι χρησιμοποιούν μια μικτή προσέγγιση (συνδυασμός της γλώσσας που χρησιμοποιούν μαζί με κανόνες εξαγωγής) [80].

Σε ορισμένες περιπτώσεις οι wrappers χρησιμοποιούνται ως βοηθοί στην επιλογή και το συνδυασμό πηγών περιεχομένου με στόχο τη δημιουργία νέων. Οι Huffman και Steier [77] για παράδειγμα, προτείνουν τη δημιουργία ενός βοηθού που επιτρέπει στο χρήστη τη συλλογή περιεχομένου από διάφορες πηγές. Η επιλογή γίνεται βάσει φίλτρων που εφαρμόζει ο χρήστης. Ορισμένα από αυτά τα φίλτρα αφορούν τις πηγές από όπου θα αντληθεί το περιεχόμενο, τους περιορισμούς που θα έχει (πιθανές πηγές περιεχομένου), τη δομή του καθώς και τις ιδιότητες της υπηρεσίας που θα δημιουργηθεί (ακρίβεια, κόστος, χρόνος). Η διαδικασία που ακολουθείται περιλαμβάνει την ύπαρξη ενός λεξικού / οντολογίας για το περιεχόμενο, τον καθορισμό του περιεχομένου των πηγών καθώς και τη δυνατότητα επερωτήσεων σ' αυτό.

Το σύστημα Wiccap [75] περιέχει επίσης εργαλεία για τη συλλογή του περιεχομένου από διάφορες πηγές και επιτρέπει τη δημιουργία διαφορετικών όψεων (views) αυτού του υλικού από τους χρήστες βάσει συγκεκριμένων κριτηρίων. Το περιεχόμενο που συλλέγεται αποθηκεύεται σε XML αρχεία η δομή των οποίων ακολουθεί συγκεκριμένα XML σχήματα. Με αυτό τον τρόπο τα αποτελέσματα του συστήματος μπορούν να χρησιμοποιηθούν από άλλα συστήματα που υποστηρίζουν XML.

Οι προσεγγίσεις που περιγράφηκαν παραπάνω χρησιμοποιούν τη δομή των HTML σελίδων προκειμένου να εξάγουν την πληροφορία που χρειάζεται και να δημιουργήσουν συλλογές

περιεχομένου. Ωστόσο, σε καμία δεν προτείνονται τρόποι για την αυτόματη ή ημι-αυτόματη ενημέρωση των συλλογών στην περίπτωση ανανέωσης του περιεχομένου. Λύση στο πρόβλημα επιχειρούν να δώσουν οι Dave, Bogen, Karadkar, Francisco-Revilla, Furuta, και Shipman [35] οι οποίοι δημιουργούν δυναμικές συλλογές περιεχομένου από δικτυακούς τόπους που παρέχουν το περιεχόμενό τους μέσω του προτύπου RSS (RSS feeds). Στην προσέγγιση αυτή ο χρήστης επιλέγει τους δικτυακούς τόπους που τον ενδιαφέρουν χρησιμοποιώντας διάφορα φίλτρα όπως την ημερομηνία δημιουργίας ή ενημέρωσης μιας ιστοσελίδας. Το σύστημα ελέγχει την ύπαρξη νέας έκδοσης της ιστοσελίδας και στην περίπτωση που υπάρχει τη «φέρνει» στο χρήστη. Για κάθε νέα σελίδα γίνεται εξαγωγή λέξεων – κλειδιών προκειμένου να παρέχεται και η δυνατότητα αναζήτησης στο συγκεντρωμένο περιεχόμενο. Στη συνέχεια η ανανέωση του περιεχομένου από τη συγκεκριμένη πηγή γίνεται αυτόματα χωρίς τη συμβολή του χρήστη.

Ανατρέχοντας στη βιβλιογραφία το πρόβλημα της συλλογής περιεχομένου από ιστοσελίδες επιλύεται με τη δημιουργία wrappers οι οποίοι εκμεταλλεύονται τη δομή τους και αναλαμβάνουν την παρουσίαση του περιεχομένου στο χρήστη. Ωστόσο τα καινούρια πρότυπα που χρησιμοποιούνται πλέον (RSS, XML), απαιτούν νέες προσεγγίσεις στο πρόβλημα της συλλογής χωρίς ωστόσο να καταργούν και υπάρχουσες υλοποιήσεις. Μερικά από τα μειονεκτήματα των προσεγγίσεων που έχουν ακολουθηθεί έως τώρα αφορούν τα παρακάτω:

- Οι χρήστες πρέπει να γνωρίζουν τη γλώσσα που χρησιμοποιούν οι wrappers
- Οι χρήστες πρέπει να έχουν τεχνικές γνώσεις για τη δημιουργία των προτύπων αλλά και να μπορούν να αξιολογούν τον κώδικα των σελίδων που θα συλλέξουν
- Οι wrappers δε μπορούν να χειριστούν σελίδες που είναι κωδικοποιημένες σε XML μορφή (π.χ. νέα με RSS)
- Ανεξάρτητα με το είδος του wrapper (αυτόματος, ημιαυτόματος, χειρωνακτικός) απαιτείται η συμβολή του χρήστη
- Εφόσον γίνει αλλαγή του προτύπου (template) του κώδικα μιας σελίδας ο χρήστης πρέπει να παρέμβει και να μοντελοποιήσει το νέο πρότυπο
- Το αποτέλεσμα έχει μικρή ακρίβεια ειδικά σε σελίδες που έχουν πολλές εξαιρέσεις στον κώδικά τους.

Στη συγκεκριμένη εργασία δε χρησιμοποιήθηκε η προσέγγιση των wrappers αφού το αποτέλεσμα που παράγεται έχει μικρή ακρίβεια. Η προσέγγιση που υιοθετήθηκε συλλέγει περιεχόμενο από δικτυακούς τόπους που χρησιμοποιούν κάποιο πρότυπο (RSS, XML) ή είναι ήδη δομημένο με συγκεκριμένη μορφή (εσωτερικό περιεχόμενο).

4.2.2 Τρόποι συλλογής περιεχομένου στο υποσύστημα κατηγοριοποίησης

Το περιεχόμενο που συλλέγεται στο υποσύστημα κατηγοριοποίησης κειμένων προέρχεται τόσο από εξωτερικές πηγές όσο και από το Σύστημα Διαχείρισης Περιεχομένου ATL CME (Σχήμα 11). Οι διαφορετικές πηγές περιεχομένου λοιπόν είναι οι εξής:

- Content Syndication: πρόκειται για περιεχόμενο (κυρίως νέα) που προέρχονται από εξωτερικούς δικτυακούς τόπους μέσω των δύο προτύπων (standards): RSS[17,18] και Atom[19].
- GoogleAPI[20]: περιεχόμενο που προέρχεται με χρήση της υπηρεσίας GoogleAPI. Η μηχανή αναζήτησης Google προσφέρει μια δικτυακή υπηρεσία (web service) που επιτρέπει την αναζήτηση στη βάση δεδομένων της καθώς και την ενσωμάτωση της λειτουργικότητάς της σε οποιονδήποτε δικτυακό τόπο.
- Εσωτερικό περιεχόμενο: στην κατηγορία αυτή ανήκει το περιεχόμενο που έχει ήδη καταχωρηθεί στο Σύστημα Διαχείρισης Περιεχομένου μέσω των μηχανισμών που διαθέτει. Στην παρούσα φάση το υποσύστημα κατηγοριοποίησης ενσωματώνει περιεχόμενο που προέρχεται από άρθρα, συνδέσμους και αρχεία.

Όπως φαίνεται παραπάνω το υποσύστημα κατηγοριοποίησης διαχειρίζεται περιεχόμενο που είναι ημι-δομημένο. Το περιεχόμενο που προέρχεται από δικτυακούς τόπους με RSS είναι κωδικοποιημένο ως XML αρχείο και περιγράφεται με ένα συγκεκριμένο XML σχήμα ανάλογα με το πρότυπο που χρησιμοποιείται. Το περιεχόμενο που εισάγεται στο δικτυακό τόπο μέσω της λειτουργικότητας GoogleAPI είναι ένα XML αρχείο που ακολουθεί ένα XML σχήμα γνωστό εκ των προτέρων στο χρήστη. Τέλος, το εσωτερικό περιεχόμενο είναι ήδη αποθηκευμένο στη βάση δεδομένων του δικτυακού τόπου. Τα χαρακτηριστικά των πινάκων στους οποίους καταχωρείται το περιεχόμενο είναι γνωστά στο διαχειριστή του δικτυακού τόπου.

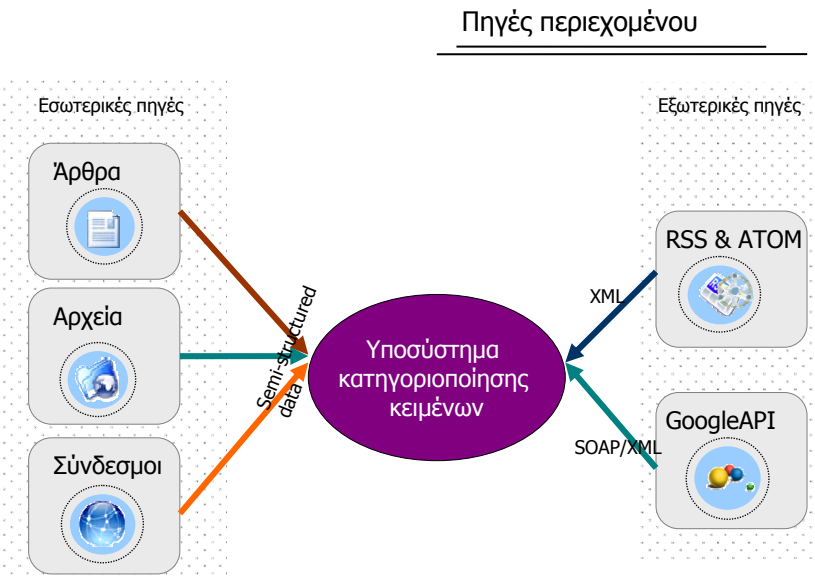
---

[17] RSS 1.0 Specification Modules, http://web.resource.org/rss/1.0/

[18] RSS 2.0 Specification Modules, http://blogs.law.harvard.edu/tech/rss

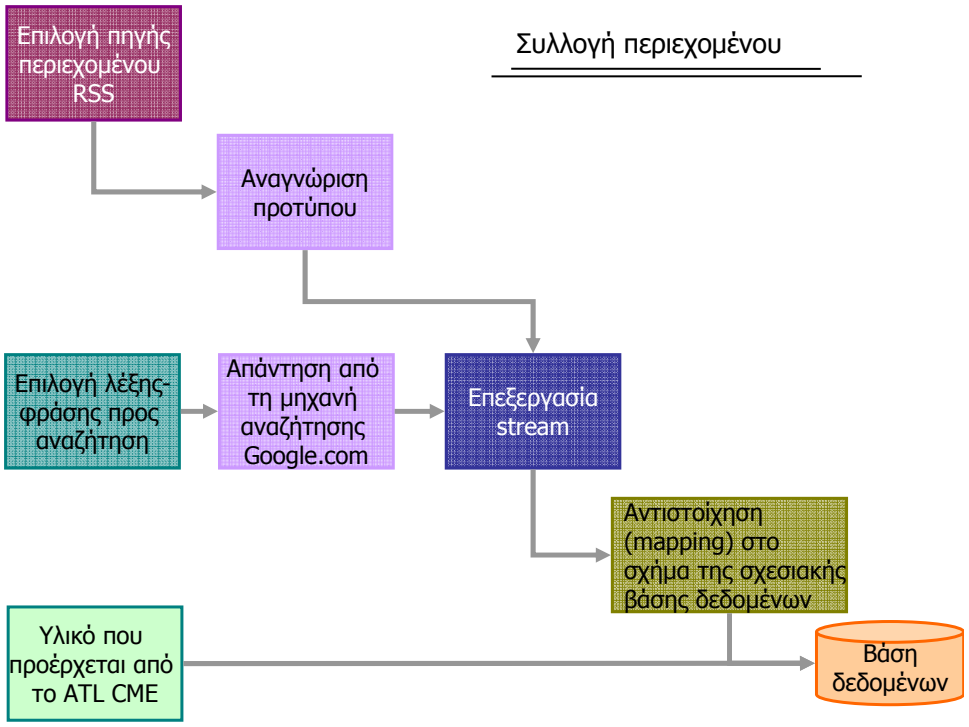
[19] Atom Specification, http://ietfreport.isoc.org/idref/draft-ietf-atompub-format/

[20] Google API, http://www.google.com/apis/



Σχήμα 11 Πηγές περιεχομένου για το υποσύστημα κατηγοριοποίησης περιεχομένου

Στο Σχήμα 12 φαίνονται αναλυτικά τα στάδια της διαδικασίας συλλογής για κάθε μία από τις παραπάνω πηγές περιεχομένου. Για το περιεχόμενο που ακολουθεί τα πρότυπα RSS/ATOM η διαδικασία περιλαμβάνει την αναγνώριση του προτύπου, την επεξεργασία του αρχείου και την δημιουργία ενός προτύπου (template) για την αντιστοίχηση του XML δέντρου στο σχήμα της σχεσιακής βάσης δεδομένων του υποσυστήματος κατηγοριοποίησης. Στην περίπτωση που το περιεχόμενο προέρχεται από την υπηρεσία GoogleAPI το βήμα της αναγνώρισης προτύπου παραλείπεται και η διαδικασία ολοκληρώνεται με την αποθήκευση του υλικού στη σχεσιακή βάση δεδομένων. Τέλος, όταν το υλικό βρίσκεται ήδη στο Σύστημα Διαχείρισης Περιεχομένου ATL CME αρκεί να γίνει ανάκτηση της πληροφορίας από τη βάση δεδομένων. Περισσότερες λεπτομέρειες για τη διαδικασία και τα στάδια υλοποίησης της συλλογής περιεχομένου δίνονται στο αντίστοιχο Κεφάλαιο (Κεφάλαιο 5).



Σχήμα 12 Τα στάδια της διαδικασίας συλλογής περιεχομένου

4.3 Αλγόριθμοι ανάλυσης κειμένου και εξαγωγής χρήσιμων λέξεων

Η εφαρμογή ενός αλγορίθμου κατηγοριοποίησης προϋποθέτει καταρχήν την ανάλυση του κειμένου προς κατηγοριοποίηση. Η ανάλυση αυτή αφορά τα εξής θέματα: την αναπαράσταση του κειμένου σε κάποια συγκεκριμένη μορφή, την εξαγωγή των χρήσιμων λέξεων από αυτό καθώς και την αφαίρεση καταλήξεων και προθεμάτων από τις λέξεις. Στην ενότητα αυτή, παρουσιάζονται οι διάφορες προσεγγίσεις που συναντώνται στη βιβλιογραφία για κάθε ένα από αυτά τα θέματα και περιγράφονται τα βασικά σημεία της προσέγγισης που ακολουθήθηκε στην παρούσα εργασία.

4.3.1 Αλγόριθμοι ανάλυσης που υπάρχουν στη βιβλιογραφία

4.3.1.1 Προ-επεξεργασία κειμένου

Η ανάλυση του κειμένου προς κατηγοριοποίηση περιλαμβάνει μια σειρά από βήματα για την προ-επεξεργασία του κειμένου. Αυτά τα βήματα είναι τα εξής:

- Μετατροπή των χαρακτήρων των λέξεων σε πεζά ή σε κεφαλαία (Case folding)
- Αφαίρεση των συχνά χρησιμοποιούμενων λέξεων (Stop words)

- Εξαγωγή των ριζών των λέξεων του κειμένου (Stemming).

Μετατροπή των χαρακτήρων των λέξεων σε πεζά: είναι η διαδικασία μετατροπής όλων των χαρακτήρων των λέξεων ενός κειμένου σε κεφαλαία ή μικρά γράμματα. Για παράδειγμα οι λέξεις "Did", "DiD" και "dID" θα μετατραπούν όλες στη μορφή "did" ή "DID" αντίστοιχα. Η διαδικασία αυτή έχει το πλεονέκτημα της μείωσης των συγκρίσεων στην φάση της εξαγωγής των χρήσιμων λέξεων του κειμένου.

Αφαίρεση των συχνά χρησιμοποιούμενων λέξεων: πρόκειται για λέξεις που δεν έχουν σημασιολογική σχέση με το κείμενο στο οποίο περιέχονται. Οι λέξεις αυτές μπορεί να είναι λέξεις που εμφανίζονται συχνά σε κείμενα (π.χ. άρθρα, προθέσεις) ή λέξεις που θεωρούνται κοινές για το συγκεκριμένο πεδίο στο οποίο ανήκει το κείμενο προς κατηγοριοποίηση (π.χ. η λέξη "Computers" σε ένα κείμενο που αφορά υπολογιστές).

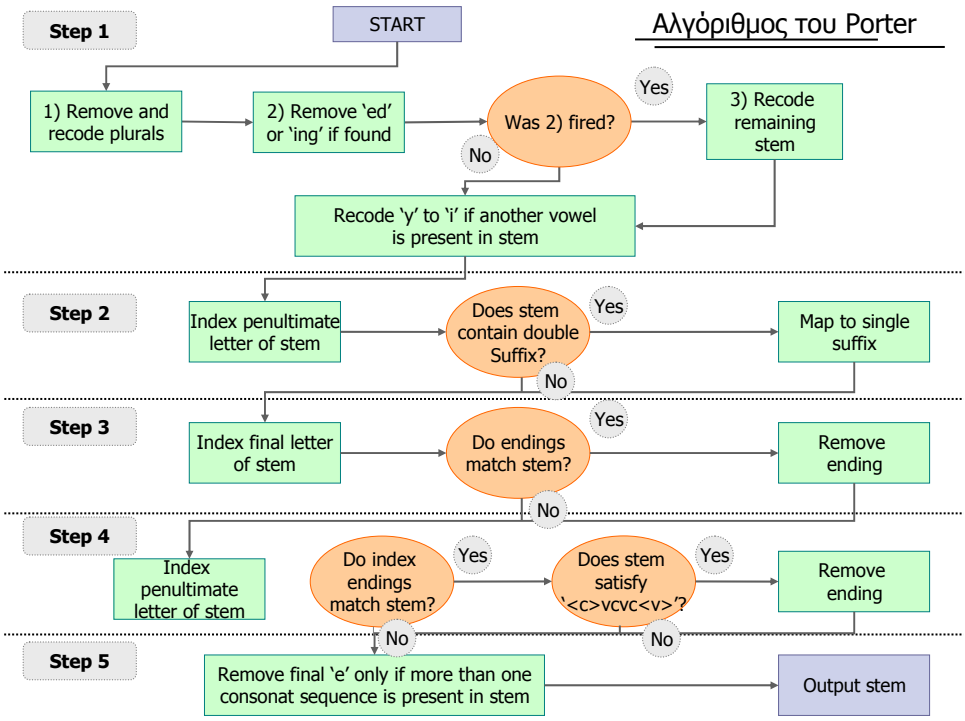
Αφαίρεση συνώνυμων λέξεων: οι συνώνυμες λέξεις δεν προσφέρουν στην σημασιολογία του κειμένου και επομένως μπορούν να αφαιρεθούν. Η εύρεση των συνώνυμων πραγματοποιείται χρησιμοποιώντας λεξικά ή έτοιμες λίστες από λέξεις. Ένα από τα λεξικά που χρησιμοποιούνται ευρέως είναι το WordNet[21], μια ηλεκτρονική βιβλιοθήκη με λέξεις η οποία κατηγοριοποιεί ιεραρχικά τους όρους της αγγλικής γλώσσας (ουσιαστικά, ρήματα, επίθετα, επιρρήματα) δημιουργώντας σύνολα συνώνυμων (synsets). Τα σύνολα αυτά αναπαριστούν έναν όρο ή μια έννοια και συνδέονται ανάλογα με τη συνάφειά τους. Οι σχέσεις μεταξύ των συνόλων με άλλα υψηλότερα ή χαμηλότερα στην ιεραρχία καθορίζονται με διαφορετικούς τύπους σχέσεων (Is-A, Part-Of). Το WordNet επεκτείνει τις δυνατότητές του και στη σημασιολογική ομοιότητα δηλαδή στην ανακάλυψη λέξεων που έχουν το ίδιο νόημα αλλά δεν είναι όμοιες λεξικογραφικά. Τέλος, περιέχει υπο-σύνολα των synsets που ονομάζονται senses. Τα senses ομαδοποιούν διαφορετικές έννοιες του ίδιου όρου [40].

Εξαγωγή των ριζών των λέξεων του κειμένου: είναι η διαδικασία αφαίρεσης προθεμάτων και καταλήξεων από τις λέξεις του κειμένου ώστε να παραμείνει μόνο η ρίζα κάθε λέξης. Έτσι λέξεις όπως οι "Computing", "Computer" και "Computational" μετατρέπονται στον όρο "Compute". Στη βιβλιογραφία υπάρχουν δύο προσεγγίσεις για την εύρεση των ριζών των λέξεων. Οι προσεγγίσεις αυτές χρησιμοποιούνται ευρέως και αφορούν την αγγλική γλώσσα. Αναπτύχθηκαν δε, από τους Porter (1980) και Lovins (1968).

[21] WordNet, http://wordnet.princeton.edu/

Ο αλγόριθμος του Porter [41] παρουσιάστηκε το 1980. Βασίζεται στην ιδέα ότι οι καταλήξεις στην αγγλική γλώσσα (περίπου 1200) δημιουργούνται από συνδυασμούς μικρότερων και απλούστερων καταλήξεων. Αποτελείται από 5 ή 6 βήματα (ανάλογα με τον ορισμό του βήματος) κάθε ένα από τα οποία εκτελείται γραμμικά (Σχήμα 13). Στον αλγόριθμο γίνονται ορισμένες παραδοχές. Ένα σύμφωνο είναι ένα γράμμα εκτός από τα A, E, I, O, U και Y. Ακόμα ως σύμφωνο θεωρείται ένα φωνήεν του οποίου προηγείται ένα φωνήεν. Για παράδειγμα στη λέξη "boy" τα σύμφωνα είναι τα B και Y ενώ στη λέξη "try" είναι τα T και R. Ένα φωνήεν είναι ένα γράμμα που δεν είναι σύμφωνο. Μία ακολουθία συμφώνων με μέγεθος μεγαλύτερο ή ίσο με ένα αποτυπώνεται σαν C ενώ η αντίστοιχη ακολουθία από φωνήεντα αναπαρίσταται από το γράμμα V. Έτσι μια λέξη μπορεί να αναπαρασταθεί σαν [C] $(VC)^m$ [V] όπου ο δείκτης m δείχνει m επαναλήψεις του VC και οι αγκύλες [] ορίζουν την προαιρετική εμφάνιση των περιεχομένων τους. Η τιμή m ονομάζεται μέτρο μιας λέξης και μπορεί να πάρει οποιαδήποτε τιμή μεγαλύτερη ή ίση με το μηδέν. Χρησιμοποιείται για να αποφασιστεί εάν μια κατάληξη θα πρέπει να αφαιρεθεί. Γενικά, οι κανόνες που χρησιμοποιούνται είναι της μορφής: (συνθήκη) S1 -> S2 που σημαίνει ότι η κατάληξη S1 θα αντικατασταθεί από S2 εάν τα γράμματα που απομένουν από το S1 ικανοποιούν τη συνθήκη.

Στο πρώτο βήμα του αλγορίθμου γίνεται χειρισμός των πληθυντικών και των αορίστων. Το βήμα αυτό λόγω πολυπλοκότητας χωρίζεται σε τρία υπο-βήματα. Το πρώτο χειρίζεται τους πληθυντικούς (π.χ. kisses -> kiss και αφαίρεση του es). Το δεύτερο αφαιρεί τις καταλήξεις ed και ing ή μετατρέπει την κατάληξη eed σε ee όπου αυτό απαιτείται. Η διαδικασία συνεχίζεται και αν έχει αφαιρεθεί η κατάληξη ed ή η ing η ρίζα που απομένει μετασχηματίζεται ακολουθώντας συγκεκριμένους κανόνες. Το τρίτο κομμάτι απλώς μετατρέπει το τελικό y σε i. Τα βήματα 2-5 ασχολούνται κυρίως με τη διαφορετική σειρά στις ομάδες καταλήξεων. Για το λόγο αυτό μετατρέπουν τις διπλές καταλήξεις σε μια κατάληξη ενώ αφαιρούν και καταλήξεις που πληρούν ορισμένα κριτήρια όπως φαίνονται και στο ακόλουθο σχήμα.



Σχήμα 13 Περιγραφή του αλγορίθμου του Porter

Ο δεύτερος πιο δημοφιλής αλγόριθμος εξαγωγής ριζών λέξεων παρουσιάστηκε από την Lovins το 1968 [42]. Πρόκειται για έναν αλγόριθμο που παράγει το αποτέλεσμά του με ένα πέρασμα (αφαιρεί μια κατάληξη τη φορά) και αφαιρεί τις καταλήξεις βασιζόμενος στην αρχή του πιο μεγάλου ταιριάσματος. Ο αλγόριθμος χρησιμοποιεί μια λίστα από 297 καταλήξεις οι οποίες συνδέονται με έναν περιορισμό από μια διαθέσιμη λίστα περιορισμών. Οι περιορισμοί αυτοί αποτρέπουν την αφαίρεση της κατάληξης μιας λέξης εφόσον πληρούνται κάποιες προϋποθέσεις. Επίσης, χρησιμοποιούνται αρκετοί κανόνες που αντιμετωπίζουν τις πιο κοινές εξαιρέσεις στην αγγλική γλώσσα. Κάθε κατάληξη συνδέεται με την εξαίρεση ότι η παραγόμενη ρίζα θα πρέπει να έχει τουλάχιστον δύο γράμματα ενώ κάποιοι άλλοι κανόνες ακολουθούν έναν από τους παρακάτω όρους:

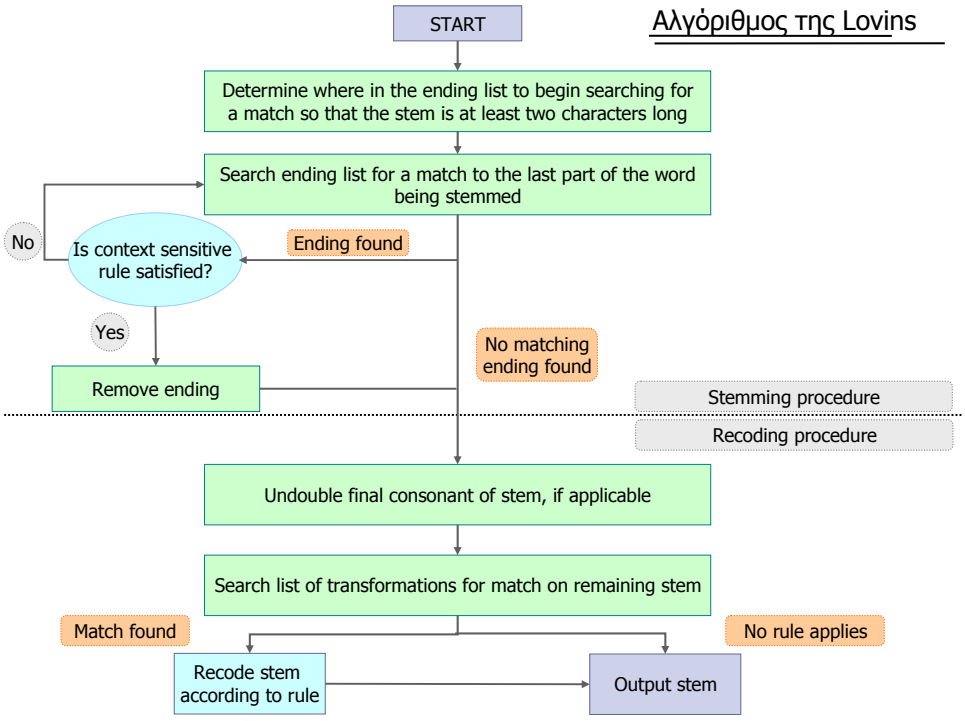
- Αυξάνεται το ελάχιστο μέγεθος της παραγόμενης ρίζας αφού αφαιρεθεί η κατάληξη
- Μια κατάληξη δεν αφαιρείται όταν συγκεκριμένα γράμματα εμφανίζονται στην παραγόμενη ρίζα
- Συνδυασμό των παραπάνω όρων.

Όπως φαίνεται και στο Σχήμα 14 ο αλγόριθμος περιλαμβάνει δύο στάδια. Το πρώτο, είναι το στάδιο εξαγωγής των ριζών (stemming phase) όπου αφαιρούνται οι καταλήξεις και ελέγχεται η εφαρμογή τυχόν εξαιρέσεων μεταξύ των βημάτων. Το δεύτερο στάδιο χρησιμοποιεί κανόνες για

την ανασυγκρότηση των λέξεων από τις καταλήξεις. Με τον τρόπο αυτό εξασφαλίζεται ότι οι παραγόμενες ρίζες ταιριάζουν με καταλήξεις άλλων παρεμφερών λέξεων. Ένα από τα βασικά προβλήματα αυτής της διαδικασίας είναι η αναξιοπιστία και το υψηλό ποσοστό αποτυχίας είτε στην ανασυγκρότηση μιας λέξης από την ρίζα της είτε στην αντιστοίχηση των ριζών των λέξεων με αυτές που έχουν παρόμοιο νόημα.

Τα προβλήματα που εντοπίζονται στην χρήση αυτού του αλγορίθμου είναι τα ακόλουθα:

- Αδυναμία να εξάγει τις ρίζες ορισμένων λέξεων που δεν περιλαμβάνονται στη λίστα κανόνων
- Αναξιοπιστία και υψηλό ποσοστό αποτυχίας είτε στην ανασυγκρότηση μιας λέξης από την ρίζα της είτε στην αντιστοίχηση των ριζών των λέξεων που έχουν παρόμοιο νόημα
- Χαμηλή ταχύτητα δεδομένης της μεγάλης λίστας κανόνων που χρησιμοποιούνται καθώς και της δεύτερης φάσης.



Σχήμα 14 Περιγραφή του αλγορίθμου της Lovins

4.3.1.2 Μοντέλα αναπαράστασης του κειμένου

Η εξόρυξη πληροφορίας συνήθως χειρίζεται δεδομένα τα οποία είναι μη δομημένα ή ημι-δεδομένα (XML) οπότε η κατηγοριοποίησή τους αποτελεί μια επίπονη διαδικασία. Για να ξεπεραστεί αυτό το

πρόβλημα εφαρμόζονται διάφορες τεχνικές ευρετηριασμού των κειμένων προς κατηγοριοποίηση. Ο ευρετηριασμός ενός κειμένου περιγράφει τη διαδικασία αντιστοίχησης του κειμένου σε μια δομημένη διάταξη η οποία αναπαριστά το περιεχόμενό που περιέχει. Η αναπαράσταση πραγματοποιείται χρησιμοποιώντας τους όρους (λέξεις) που περιέχει το κείμενο. Στη βιβλιογραφία έχουν περιγραφεί αρκετές προσπάθειες που δεν βασίζονται μόνο στους όρους του κειμένου αλλά σε φράσεις [23] ή ακολουθίες ομάδων αλφαριθμητικών (kernel strings) [78].

Στη συγκεκριμένη ενότητα παρουσιάζονται τα βασικά στοιχεία των κλασσικών μοντέλων αναπαράστασης ενός κειμένου. Αυτά είναι τα:

- Δυαδικό μοντέλο (Boolean Model)
- Διανυσματικό μοντέλο (Vector Space Model)
- Πιθανοκρατικό μοντέλο (Probabilistic Model).

Δυαδικό μοντέλο: Στο μοντέλο αυτό ένα έγγραφο (d_j) μοντελοποιείται ως ένα σύνολο κλειδιών (keywords) και παριστάνεται με το διάνυσμα $d_j = \left(w_{1,j}, ..., w_{t,j}\right)$ όπου $w_{i,j} = 1$ αν η λέξη k_i εμφανίζεται στο κείμενο d_j (αλλιώς $w_{i,j} = 0$). Το σύνολο όλων των λέξεων ευρετηριασμού αναπαρίσταται ως $K = \{k_1, ..., k_t\}$. Δηλαδή ένα κείμενο d είναι μια σύζευξη όρων όπου όρος είναι μια λέξη σε θετική ή αρνητική μορφή.

Πλεονεκτήματα	Μειονεκτήματα
Εύκολο στην κατανόηση	Δεν είναι ευέλικτο. AND σημαίνει όλα, OR σημαίνει οποιοδήποτε
Αποτελεσματικό όταν κανείς γνωρίζει τι ψάχνει και τι περιέχει η συλλογή	Αδυναμία ελέγχου του μεγέθους της απάντησης
Αποδοτική υλοποίηση	Δεν επιτρέπει την έκφραση σύνθετων πληροφοριακών αναγκών
	Δεν υπάρχει διάταξη των αποτελεσμάτων όσων αφορά τη συνάφειά τους (λιγότερο, περισσότερο συναφές)
	Εάν στην ερώτηση για τη συνάφεια ενός εγγράφου κάποιος χρήστης απαντήσει θετικά (ή αρνητικά αντίστοιχα) δεν υπάρχει τρόπος τροποποίησης της ερώτησης στο δυαδικό μοντέλο

Πίνακας 5 Πλεονεκτήματα και μειονεκτήματα του δυαδικού μοντέλου

Διανυσματικό μοντέλο: Στο μοντέλο αυτό κάθε έγγραφο αναπαρίσταται από ένα διάνυσμα. Κάθε συνιστώσα του διανύσματος είναι η τιμή ενός όρου που δείχνει την συνάφεια του όρου αυτού με το έγγραφο. Συνήθως οι υψηλές τιμές υποδηλώνουν και μεγάλη συνάφεια. Ένα έγγραφο d_j εκφράζεται ως ένα διάνυσμα $d_j = \left(w_{1,j}, ..., w_{t,j}\right)$ όπου $w_{i,j} = tf_{ij} idf_i$. Στον τύπο αυτό $tf_{ij} = freq_{ij} / \max_k \left\{freq_{kj}\right\}$ με $freq_{ij}$ το πλήθος των εμφανίσεων του όρου i στο έγγραφο j και $\max_k \left\{freq_{kj}\right\}$ το μεγαλύτερο πλήθος εμφανίσεων ενός όρου στο έγγραφο j. Η ποσότητα idf_i ορίζεται ως η αντίστροφη συχνότητα ενός όρου i και ισούται με $\log_2\left(N / df_i\right)$ όπου df_i είναι το πλήθος των εγγράφων που περιέχουν τον όρο i. Για να υπολογιστεί ο βαθμός συσχέτισης ενός εγγράφου με μια επερώτηση θα πρέπει να υπολογιστεί ο ακόλουθος τύπος:

$$sim\left(d_j, q\right) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}.$$

Πλεονεκτήματα	Μειονεκτήματα
Απλό στην κατανόηση	Χάνει σημαντική πληροφορία από το αρχικό κείμενο όπως τη σειρά των όρων
Αποδοτική υλοποίηση για μεγάλες συλλογές εγγράφων	Απουσία σημασιολογίας (π.χ. σημασία λέξεων)
Λαμβάνει υπόψη τις τοπικές (tf) και καθολικές (idf) συχνότητες όρων	Υποθέτει την ανεξαρτησία των όρων (π.χ. άγνοια συνωνύμων)

Πίνακας 6 Πλεονεκτήματα και μειονεκτήματα του διανυσματικού μοντέλου

Πιθανοκρατικό μοντέλο: Στο μοντέλο αυτό το πρόβλημα της ανάκτησης πληροφορίας επιλύεται με χρήση πιθανοτήτων. Συγκεκριμένα, το μοντέλο αυτό προσπαθεί να εκτιμήσει την πιθανότητα να βρει ο χρήστης το έγγραφο d_i συναφές με μια επερώτηση q ενώ γίνεται η υπόθεση ότι η πιθανότητα αυτή είναι ανεξάρτητη από τυχόν άλλα έγγραφα που υπάρχουν σε μια συλλογή εγγράφων. Έτσι για μια δεδομένη επερώτηση q υπάρχει ένα σύνολο εγγράφων R που είναι σχετικό με αυτή. Το πιθανοκρατικό μοντέλο ταξινομεί τα έγγραφα με βάση την πιθανότητα να είναι σχετικά με την παραπάνω επερώτηση δηλαδή $P(R \mid q, d_i)$. Η ομοιότητα υπολογίζεται από τον τύπο: $sim(d_j, q) = \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left\langle \log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})} \right\rangle$ όπου $P(k_i \mid R)$ είναι η πιθανότητα ότι ένας όρος k_i είναι παρών σε ένα έγγραφο που έχει επιλεγεί τυχαία από κάποιο σύνολο εγγράφων ενώ $w_{i,j} \in \{0,1\}$ και $w_{i,q} \in \{0,1\}$. Εάν V είναι ένα σύνολο εγγράφων και V_i είναι ένα σύνολο εγγράφων το οποίο περιέχει τον όρο k_i τότε $P(k_i \mid R) = \frac{V_i + 0.5}{V + 1}$ και $P(k_i \mid \overline{R}) = \frac{n_i - V_i + 0.5}{n - V + 1}$.

Πλεονεκτήματα	Μειονεκτήματα
Τα έγγραφα ταξινομούνται βάσει της πιθανότητάς τους να είναι σχετικά με μια επερώτηση	Δεν λαμβάνεται υπόψη η συχνότητα εμφάνισης των όρων ενός εγγράφου
	Για κάθε όρο θα πρέπει να υπολογίζεται η πιθανότητα $P(k_i \mid R)$

Πίνακας 7 Πλεονεκτήματα και μειονεκτήματα του πιθανοκρατικού μοντέλου

4.3.2 Η προσέγγιση που χρησιμοποιήθηκε στην εργασία

Η προσέγγιση που χρησιμοποιήθηκε στην παρούσα εργασία έχει ως στόχο την ανάλυση του κειμένου και την εύρεση ενός μικρού συνόλου λέξεων που το χαρακτηρίζει (εύρεση λεξικού). Η διαδικασία περιλαμβάνει όλα τα βήματα που περιγράφηκαν στην υπο-ενότητα 4.3.1.1 ενώ το κείμενο αναπαρίσταται από ένα διάνυσμα με συνιστώσες τις ρίζες των λέξεων του κειμένου. Τα βήματα που ακολουθούνται είναι τα εξής:

- Μετατροπή όλων των γραμμάτων των λέξεων σε πεζά
- Αφαίρεση των συχνά χρησιμοποιούμενων λέξεων
- Εξαγωγή των ριζών των λέξεων
- Αναπαράσταση του κειμένου προς κατηγοριοποίηση.

Αναλυτικότερα:

Βήμα 1: Μετατροπή όλων των γραμμάτων των λέξεων σε πεζά.

Βήμα 2: Αφαίρεση όσων λέξεων θεωρείται ότι δεν προσφέρουν στην σημασιολογία του κειμένου. Η αφαίρεση γίνεται χρησιμοποιώντας τρεις λίστες: μια λίστα με συχνά χρησιμοποιούμενες λέξεις (άρθρα, προθέσεις), μια λίστα με τα πιο συνήθη ομαλά ρήματα (περιλαμβάνονται οι καταλήξεις σε –ing και –ed) και τέλος μια λίστα με τα πιο συνηθισμένα ανώμαλα ρήματα (περιλαμβάνονται ο αόριστος, ο παρακείμενος καθώς και η κατάληξη –ing για κάθε ένα από τα ρήματα).

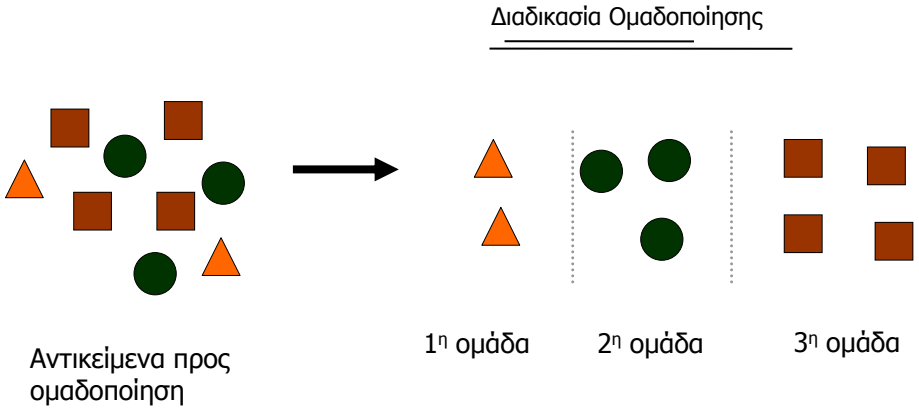
Βήμα 3: Εφαρμογή του αλγορίθμου του Porter για την εξαγωγή των ριζών των λέξεων.

Βήμα 4: Το μοντέλο αναπαράστασης που χρησιμοποιείται προσομοιώνει το διανυσματικό μοντέλο. Στο διανυσματικό μοντέλο ένα έγγραφο αναπαρίσταται ως διάνυσμα με συνιστώσες τις τιμές συνάφειας ενός όρου με το συγκεκριμένο έγγραφο. Στη συγκεκριμένη προσέγγιση το κείμενο αναπαρίσταται ως ένα διάνυσμα αλλά οι συνιστώσες του αποτελούνται από τις ρίζες των λέξεων που συνιστούν το κείμενο. Στο διάνυσμα αυτό αποθηκεύεται επίσης η συχνότητα εμφάνισης κάθε όρου στο κείμενο. Η συχνότητα εμφάνισης, στο συγκεκριμένο μοντέλο, ορίζεται ως ο αριθμός των εμφανίσεών του στο κείμενο. Η αναπαράσταση αυτή επιλέχθηκε λόγω της ευκολίας της ταχύτητας στην υλοποίησή της. Επειδή βασίζεται κυρίως στη στατιστική ανάλυση των λέξεων του κειμένου έχει όλα τα μειονεκτήματα του διανυσματικού μοντέλου.

4.4 Αλγόριθμοι κατηγοριοποίησης

Η κατηγοριοποίηση κειμένων έχει γίνει αντικείμενο έρευνας τόσο στην περιοχή της ανάκτησης (retrieval) όσο και της εξόρυξης (mining) πληροφορίας. Σε πρώτη φάση, χρησιμοποιήθηκε για την βελτίωση της ακρίβειας της πληροφορίας που παράγεται από τα συστήματα ανάκτησης. Ωστόσο, τα τελευταία χρόνια η κατηγοριοποίηση χρησιμοποιείται για την καλύτερη οργάνωση συλλογών αρχείων ή για την οργάνωση των αποτελεσμάτων που επιστρέφονται από τις μηχανές αναζήτησης [10]. Τέλος, η κατηγοριοποίηση κειμένων έχει χρησιμοποιηθεί για τη δημιουργία ιεραρχιών των ομάδων στις οποίες αντιστοιχούνται τα κείμενα [1]. Μια διαφορετική προσέγγιση ανακαλύπτει τις ομάδες σε μια υπάρχουσα ταξινομία (π.χ. Yahoo!) και στη συνέχεια χρησιμοποιεί την ιεραρχία που προκύπτει για την κατηγοριοποίηση νέων εγγράφων [81]. Μπορεί κανείς να διακρίνει δύο είδη κατηγοριοποίησης [27]:

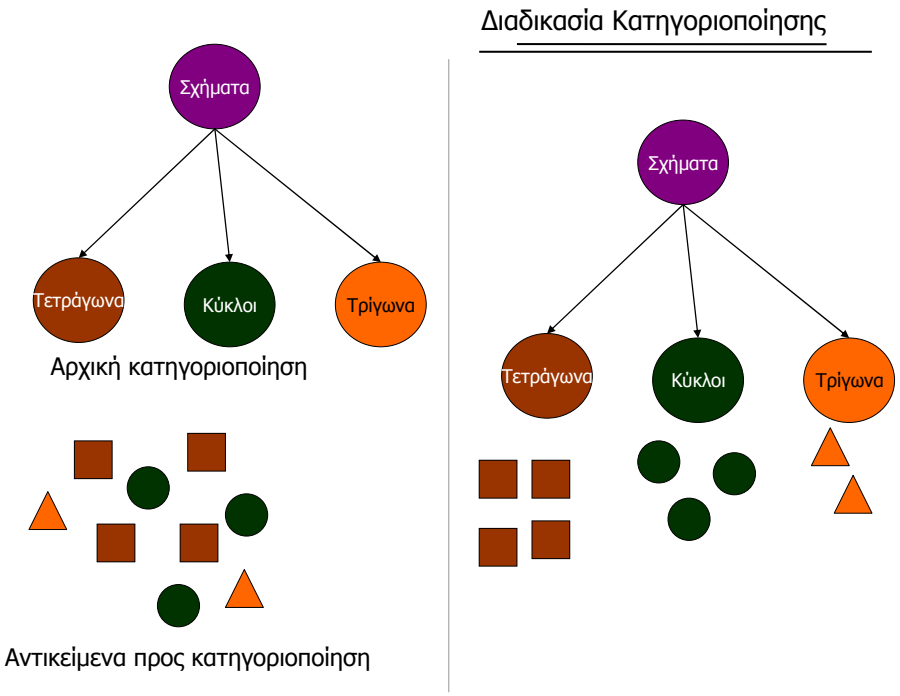
1. **Κατηγοριοποίηση χωρίς εκ προτέρων γνώση των κατηγοριών (text clustering):** Η διαδικασία αυτή ονομάζεται ομαδοποίηση (clustering) και μπορεί να θεωρηθεί ως η εργασία εύρεσης μιας δομής κατηγοριών χωρίς να είναι γνωστή από την αρχή καμία πληροφορία γι' αυτή τη δομή (Σχήμα 15). Ένα αντικείμενο προς κατηγοριοποίηση ανήκει σε μια κατηγορία εφόσον η ομοιότητα μεταξύ αυτού και της κατηγορίας είναι υψηλή [64]. Συνήθως οι ομάδες διακρίνονται από μεγάλη ομοιότητα μεταξύ των μελών τους και χαμηλή ομοιότητα μεταξύ αντικειμένων που ανήκουν σε διαφορετικές ομάδες.



Σχήμα 15 Η διαδικασία της ομαδοποίησης

2. **Κατηγοριοποίηση με εκ των προτέρων γνώση των κατηγοριών (text categorization):** Πρόκειται για τη διαδικασία αντιστοίχησης των εγγράφων σε προκαθορισμένες κατηγορίες οι οποίες δημιουργούνται βασιζόμενες στην πληροφορία που εξάγεται από ένα σύνολο εκπαιδευτικών εγγράφων. Η διαδικασία αυτοματοποιείται

χρησιμοποιώντας τη μέθοδο της επαγωγικής μάθησης όπου η κατηγοριοποίηση πραγματοποιείται με βάση τα χαρακτηριστικά ήδη κατηγοριοποιημένων εγγράφων. [25] (Σχήμα 16).



Σχήμα 16 Η διαδικασία κατηγοριοποίησης

4.4.1 Αλγόριθμοι κατηγοριοποίησης χωρίς γνώση των κατηγοριών

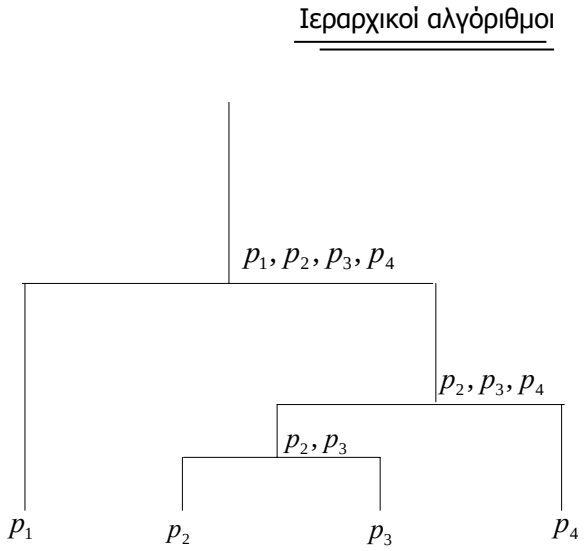
Η πιο γνωστή μέθοδος κατηγοριοποίησης είναι η μέθοδος της ομαδοποίησης. Στη μέθοδο αυτή τα στοιχεία προς κατηγοριοποίηση χωρίζονται σε ομάδες (clusters) με τέτοιο τρόπο ώστε τα μέλη της ίδιας ομάδας να έχουν όσο το δυνατόν μεγαλύτερη ομοιότητα ενώ μέλη διαφορετικών ομάδων να έχουν όσο το δυνατόν μεγαλύτερη ανομοιογένεια [66].

Υπάρχουν δύο είδη αλγορίθμων ομαδοποίησης:

- Οι ιεραρχικοί αλγόριθμοι και
- Οι αλγόριθμοι κατάτμησης

4.4.1.1 Ιεραρχικοί αλγόριθμοι (Hierarchical clustering)

Κάθε ιεραρχικός αλγόριθμος δημιουργεί μια ακολουθία από διαμερίσεις τμημάτων με μία μοναδική ομάδα στην κορυφή της δενδρικής ακολουθίας (Σχήμα 17). Κάθε επίπεδο δημιουργείται από τη συγχώνευση δύο ομάδων του κατώτερου επιπέδου (από κάτω προς τα πάνω) ή την διαίρεση μιας μεγαλύτερης ομάδας σε μικρότερες (από πάνω προς τα κάτω). Στο Σχήμα 17 φαίνεται η διαδικασία που ακολουθείται σε έναν ιεραρχικό αλγόριθμο με 4 αντικείμενα στην εκκίνηση.



Σχήμα 17 Δενδρόγραμμα ενός ιεραρχικού αλγορίθμου

Υπάρχουν δύο προσεγγίσεις που μπορούν να ακολουθηθούν στους ιεραρχικούς αλγορίθμους [1].

- **Συσσώρευση (Agglomerative):** Σε πρώτη φάση κάθε αντικείμενο συνιστά μια ξεχωριστή ομάδα. Στόχος του αλγορίθμου είναι η συγχώνευση αυτών των ομάδων των αντικειμένων προκειμένου να σχηματίσουν μεγαλύτερες ομάδες με όσο το δυνατό μεγαλύτερη ομοιότητα. Η διαδικασία αυτή πραγματοποιείται ελέγχοντας τους γειτονικούς κόμβους των αντικειμένων και καταλήγει σε μια και μοναδική ομάδα.
- **Διαίρεση (Divisive):** Σε αυτή την περίπτωση υπάρχει μόνο μια αρχική ομάδα αντικειμένων. Σε κάθε βήμα η ομάδα αυτή διαιρείται σε μικρότερες ομάδες μέχρις ότου κάθε αντικείμενο να αποτελεί μια δική του ομάδα ή πλέον να μη μπορεί να γίνει άλλη διαίρεση των ομάδων.

Η διαδικασία που ακολουθείται στους αλγορίθμους συσσώρευσης περιλαμβάνει τα παρακάτω βήματα (αντίστοιχα στους ιεραρχικούς αλγορίθμους διαίρεσης):

- Υπολογισμός της ομοιότητας μεταξύ δύο ομάδων με αντικείμενα (clusters).

- Συγχώνευση των δύο ομάδων που έχουν τη μεγαλύτερη ομοιότητα μεταξύ τους.
- Ενημέρωση του πίνακα όπου αποθηκεύονται οι ομοιότητες των ομάδων.
- Επανάληψη των δύο παραπάνω βημάτων έως ότου προκύψει μόνο μια ομάδα.

Πλεονεκτήματα	**Μειονεκτήματα**
Παράγουν ομάδες με μεγάλη ποιότητα	Πολυπλοκότητα $O(n^2)$
Είναι κατάλληλοι για μεγάλο όγκο δεδομένων	Μεγάλη ανομοιογένεια μεταξύ των παραγόμενων ομάδων
	Απαιτούν μεγάλο χώρο για την αποθήκευσή τους
	Δε μπορούν να χειριστούν δεδομένα με πολύ θόρυβο επειδή οι αποφάσεις για την συγχώνευση δύο ομάδων είναι τελικές (δεν υπάρχει επιστροφή σε προηγούμενη κατάσταση)

Πίνακας 8 Πλεονεκτήματα και μειονεκτήματα των ιεραρχικών αλγορίθμων

Οι μετρικές που χρησιμοποιούνται για τον υπολογισμό της ομοιότητας δύο ομάδων είναι οι εξής:

Single-linkage function (SL): η μέγιστη ομοιότητα μεταξύ δύο αντικειμένων (υ_a, υ_b) στις δύο ομάδες (C_i, C_j). Υπολογίζεται από τον ακόλουθο τύπο: $sim_{SL}(C_i, C_j) = \max_{\upsilon_a \in C_i, \upsilon_b \in C_j} sim(\upsilon_\alpha, \upsilon_b)$.

Complete-linkage function (CL): η μικρότερη ομοιότητα μεταξύ δύο αντικειμένων στις δύο ομάδες. Υπολογίζεται από τον ακόλουθο τύπο: $sim_{CL}(C_i, C_j) = \min_{\upsilon_a \in C_i, \upsilon_b \in C_j} sim(\upsilon_\alpha, \upsilon_b)$.

Average-linkage function (AL): ο μέσος όρος όλων των ομοιοτήτων μεταξύ των αντικειμένων στις δύο ομάδες. Υπολογίζεται από τον ακόλουθο τύπο: $sim_{AL}(C_i, C_j) = sim_A(C_i, C_j)$.

4.4.1.2 Αλγόριθμοι κατάτμησης (Partitional clustering)

Οι αλγόριθμοι κατάτμησης σε αντίθεση με τους ιεραρχικούς αλγόριθμους διαμερίζουν τα δεδομένα μόνο σε ένα σημείο. Έτσι εάν πρέπει να δημιουργηθούν K ομάδες με αντικείμενα ο αλγόριθμος κατάτμησης παράγει αυτά τα αντικείμενα αμέσως. Η μέθοδος που ακολουθείται αναζητά ένα σύνολο δεδομένων το οποίο να ικανοποιεί μια συνάρτηση αξιολόγησης η οποία και βασίζεται σε ορισμένα κριτήρια βελτιστοποίησης. Ο μηχανισμός αναζήτησης χρησιμοποιείται για

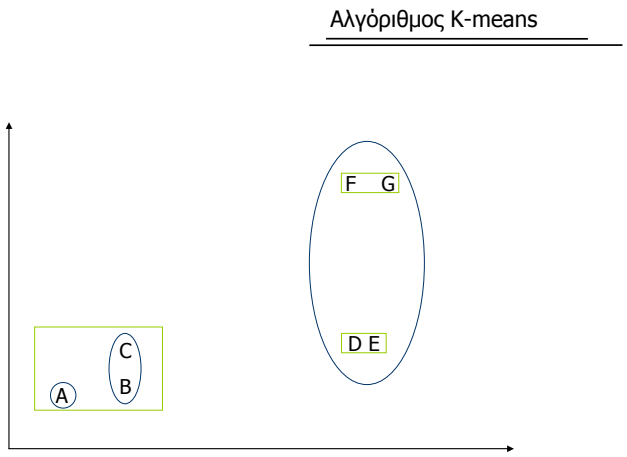
την εύρεση ομάδων αντικειμένων υψηλής ποιότητας [22] χρησιμοποιώντας την συνάρτηση αξιολόγησης ως οδηγό. Τα διαφορετικά είδη αλγορίθμων διαίρεσης συνίστανται στην επιλογή των κριτηρίων βελτιστοποίησης. Στην πράξη, ο αλγόριθμος τρέχει πολλές φορές στα δεδομένα με διαφορετικές αρχικές καταστάσεις οπότε τελικά επιλέγεται η καλύτερη διαμόρφωση.

Οι πιο συχνές προσεγγίσεις στους αλγορίθμους κατάτμησης είναι οι αλγόριθμοι τετραγωνικού λάθους (Squared Error) και οι γραφο-θεωρητικοί (Graph-Theoretic) [8].

4.4.1.2.1 Αλγόριθμοι τετραγωνικού λάθους (Squared Error Algorithms)

Ένας από τους πιο γνωστούς και συχνά χρησιμοποιούμενους αλγορίθμους είναι ο αλγόριθμος K-means [3] (πολυπλοκότητα $O(tkn)$ με n τον αριθμό των αντικειμένων προς κατηγοριοποίηση, k τον αριθμό των ζητούμενων ομάδων και t τον αριθμό των επαναλήψεων) και οι διάφορες εκδοχές του. Ο αλγόριθμος αυτός θεωρεί μια αρχική διαμέριση των δεδομένων και εκχωρεί τα δεδομένα σε ομάδες βάσει της ομοιότητας που έχει το αντικείμενο και η ομάδα στην οποία ανήκει. Η διαδικασία συνεχίζεται μέχρις ότου ικανοποιηθεί μια συνθήκη (π.χ. δε μπορεί να γίνει ανάθεση των αντικειμένων σε άλλες ομάδες). Στο Σχήμα 18 φαίνεται ένα παράδειγμα εφαρμογής του αλγορίθμου K-means. Θεωρώντας αρχική κατάσταση του αλγορίθμου τα στοιχεία A, B, C δημιουργούνται τρεις ομάδες. Αυτές είναι οι {A}, {B, C} και {F, G, D, E}. Υπολογίζοντας το κριτήριο τετραγωνικού λάθους αποδεικνύεται ότι αυτό είναι μεγαλύτερο από αυτό της επιλογής των ομάδων {A, B, C}, {F, G} και {D, E}. Το κριτήριο τετραγωνικού λάθους (ή η ανομοιογένεια στο εσωτερικό μιας ομάδας) για μια ομάδα L και ένα σετ αντικειμένων X υπολογίζεται ως: $e^2(X,L)=\sum_{j=1}^{K}\sum_{i=1}^{n_j}\left\|x_i^{(j)}-c_j\right\|^2$ με $x_i^{(j)}$ τον i-στό όρος που ανήκει στην j-στή ομάδα ενώ c_j είναι το κεντρικό στοιχείο της ομάδας.

[22] Η υψηλή ποιότητα μιας ομάδας αντικειμένων εκφράζεται από τον αριθμό των αντικειμένων τα οποία ανήκουν σε αυτή την ομάδα και έχουν μικρή πιθανότητα να ανήκουν και σε κάποια άλλη γειτονική.



Σχήμα 18 Παράδειγμα εφαρμογής αλγορίθμου K-means

Η διαδικασία που ακολουθείται για την εύρεση K ομάδων στον αλγόριθμο K-means είναι η εξής:

- Επιλογή K σημείων σαν αρχικά κεντρικά σημεία (centroids)
- Αντιστοίχηση όλων των σημείων στο πιο κοντινό κεντρικό σημείο
- Υπολογισμός του κεντρικού σημείου κάθε ομάδας
- Επανάληψη των δύο παραπάνω βημάτων μέχρι να σταθεροποιηθούν οι ομάδες.

Η απόδοση του αλγορίθμου K-means εξαρτάται από την επιλογή της αρχικής κατάτμησης. Εάν η αρχική κατάτμηση δεν είναι σωστή μπορεί να προκύψει το πρόβλημα του τοπικού ελαχίστου. Γι' αυτό το λόγο έχουν προταθεί στη βιβλιογραφία διάφορες εκδοχές [5], [11]. Μερικές από αυτές προσπαθούν να επιλέξουν μια σωστή αρχική κατάτμηση έτσι ώστε ο αλγόριθμος να επιτύχει την εύρεση του καθολικού ελαχίστου, ενώ άλλες επιτρέπουν την διάσπαση και τη συγχώνευση των ομάδων που δημιουργούνται.

Πλεονεκτήματα	Μειονεκτήματα
Απλός, κατανοητός	Επιλογή του αριθμού των ομάδων από την αρχή
Τα αντικείμενα αυτόματα αποδίδονται σε ομάδες	Όλα τα αντικείμενα εισάγονται σε μια ομάδα
Γρήγορος $O(tkn)$	Δε μπορεί να χειριστεί ομάδες διαφορετικού μεγέθους και πυκνότητας
Αποτελεσματικός αν και απαιτείται να εφαρμοστεί πολλές φορές στα δεδομένα	Μπορεί να εφαρμοστεί μόνο σε δεδομένα που έχουν την έννοια του κέντρου. Παραλλαγές του αλγορίθμου μπορεί να μην έχουν αυτόν τον περιορισμό όμως είναι αρκετά ακριβές (K-medoids)

Πίνακας 9 Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου K-means

4.4.1.2.2 Γραφο-θεωρητικοί αλγόριθμοι (Graph-Theoretic Clustering)

Κάποιες φορές οι ομάδες έχουν ασυνήθιστα σχήματα οπότε ο αλγόριθμος K-means αποτυγχάνει. Η εναλλακτική προσέγγιση περιλαμβάνει την κωδικοποίηση της ομοιότητας των χαρακτηριστικών των αντικειμένων αντί των απόλυτων ιδιοτήτων τους [79]. Η ομοιότητα αναπαρίσταται χρησιμοποιώντας έναν γράφο συγγένειας (affinity) με βάρη. Στο γράφο αυτό τα δεδομένα αναπαρίστανται ως κόμβοι ενώ η ανομοιότητα (dissimilarity) μεταξύ δύο κόμβων υπολογίζεται από το μήκος της ακμής που τους συνδέει. Συνήθως οι ομάδες σχηματίζονται από υπογράφους ισχυρά συνδεδεμένους. Οι υπογράφοι δημιουργούνται από το κόψιμο των γράφων εφόσον αφαιρεθούν οι μακρύτερες ακμές τους. Η συγγένεια μεταξύ δύο κόμβων ενός γράφου υπολογίζεται βρίσκοντας την τιμή τριών μετρικών:

- Ένταση: $aff(x,y) = \exp\left\{-\left(\frac{1}{2\sigma_i^2}\right)\left(\|I(x)-I(y)\|^2\right)\right\}$
- Απόσταση: $aff(x,y) = \exp\left\{-\left(\frac{1}{2\sigma_d^2}\right)\left(\|x-y\|^2\right)\right\}$
- Σύσταση: $aff(x,y) = \exp\left\{-\left(\frac{1}{2\sigma_t^2}\right)\left(\|c(x)-c(y)\|^2\right)\right\}$

4.4.1.3 Σύγκριση των παραπάνω αλγορίθμων

Οι ιεραρχικοί αλγόριθμοι θεωρούνται καλύτεροι όσον αφορά την ποιότητα των σχηματιζόμενων ομάδων ωστόσο απαιτούν τετραγωνικό χρόνο για την ολοκλήρωσή τους σε σχέση με τον αριθμό του αρχικού συνόλου αντικειμένων. Αντίθετα, οι αλγόριθμοι κατάτμησης απαιτούν γραμμικό χρόνο αλλά παράγουν ομάδες με χαμηλότερη ποιότητα. Επιπλέον οι ιεραρχικοί αλγόριθμοι χρησιμοποιούν μια συγκεκριμένη συνάρτηση προκειμένου να υπολογίσουν την ομοιότητα μεταξύ δύο επιμέρους ομάδων. Στην περίπτωση μάλιστα που τα δεδομένα είναι πολλά οι ομάδες που δημιουργούνται είναι ανομοιογενείς και δεν είναι δυνατή η απόδοση χαρακτηριστικών στα αντικείμενα που ανήκουν σε μια ομάδα. Οι αλγόριθμοι κατάτμησης θεωρούν γνωστό τον αριθμό των ομάδων εκ των προτέρων ενώ κάθε ομάδα έχει το δικό της μέτρο σύγκρισης της ομοιότητας (κέντρο). Ο Πίνακας 10 συνοψίζει τα βασικά στοιχεία των ιεραρχικών αλγορίθμων και των αλγορίθμων κατάτμησης.

Ιεραρχικοί αλγόριθμοι	**Αλγόριθμοι κατάτμησης**
Η ομαδοποίηση των αντικειμένων πραγματοποιείται χρησιμοποιώντας ήδη υπάρχουσες ομάδες	Τα αντικείμενα αποδίδονται αυτόματα σε ομάδες
Πολυπλοκότητα τετραγωνικού χρόνου	Πολυπλοκότητα γραμμικού χρόνου
Δε μπορούν να χειριστούν δεδομένα με πολύ θόρυβο επειδή οι αποφάσεις για την συγχώνευση δύο ομάδων είναι τελικές (δεν υπάρχει επιστροφή σε προηγούμενη κατάσταση)	Εφαρμόζονται πολλές φορές στα δεδομένα οπότε πολλές αποφάσεις μπορεί να ανακληθούν
Εφαρμόζονται σε δεδομένα χωρίς κανένα περιορισμό	Εφαρμόζονται κυρίως σε δεδομένα που έχουν την έννοια της διαμέρισης
Παράγονται ομάδες με υψηλή ποιότητα	Η ποιότητα εξαρτάται από το αρχικό σύνολο των αντικειμένων
Είναι κατάλληλοι για μεγάλο όγκο δεδομένων	Αντιμετωπίζουν το πρόβλημα του τοπικού ελαχίστου
Ο χρήστης επιλέγει σε ποιο σημείο θα κόψει το δέντρο	Ο αριθμός των ομάδων είναι προκαθορισμένος από την αρχή
Μεγάλη ανομοιογένεια μεταξύ των παραγόμενων ομάδων	Συνήθως οι ομάδες βρίσκονται κοντά όσον αφορά την ομοιότητά τους

Πίνακας 10 Σύγκριση των ιεραρχικών αλγορίθμων και των αλγορίθμων ομαδοποίησης

4.4.2 Αλγόριθμοι κατηγοριοποίησης με εκ των προτέρων γνώση των κατηγοριών

Στόχος της κατηγοριοποίησης κειμένου είναι η ταξινόμηση των δεδομένων σε έναν προκαθορισμένο αριθμό κατηγοριών. Ανάλογα με τον αριθμό κατηγοριών στις οποίες θα ανατεθεί ένα αντικείμενο διακρίνονται δύο είδη κατηγοριοποίησης. Στην πρώτη περίπτωση κάθε αντικείμενο μπορεί να ανήκει σε μία (single-label) κατηγορία, ενώ στη δεύτερη περίπτωση κάθε αντικείμενο μπορεί να ανήκει σε καμία ή περισσότερες από μία (multi-label) κατηγορίες. Τέλος, ανάλογα με τον τρόπο κατηγοριοποίησης υπάρχουν δύο προσεγγίσεις: βασισμένη στις κατηγορίες και βασισμένη στα αντικείμενα προς κατηγοριοποίηση. Στην πρώτη περίπτωση στόχος είναι η εύρεση των δεδομένων που ανήκουν σε μια δεδομένη κατηγορία ενώ στη δεύτερη περίπτωση στόχος είναι η εύρεση όλων των κατηγοριών στις οποίες ανήκει ένα αντικείμενο [25].

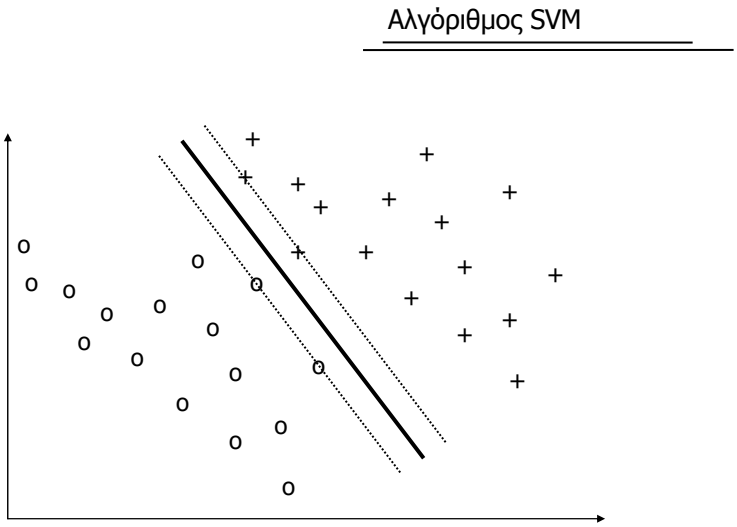
Ανατρέχοντας στη βιβλιογραφία μπορεί κανείς να βρει πολλούς αλγορίθμους κατηγοριοποίησης. Στην παρούσα εργασία γίνεται αναφορά μόνο στους πιο γνωστούς από αυτούς. Αυτοί είναι:

- Support Vector Machines - SVM
- K-Nearest Neighbor – kNN
- Neural Network – Nnet
- Linear Least-squares Fit (LLSF) mapping
- Naïve Bayes classifier – NB.

4.4.2.1 Διανυσματικοί μηχανισμοί υποστήριξης (Support Vector Machines - SVM)

Η προσέγγιση αυτή βασίζεται στα διανύσματα υποστήριξης (support vectors). Το πρόβλημα έγκειται στην εύρεση μιας επιφάνειας η οποία χωρίζει τα δεδομένα σε δύο κλάσεις με τον καλύτερο δυνατό τρόπο. Η επιφάνεια αυτή ονομάζεται επιφάνεια απόφασης (best decision surface). Ένα παράδειγμα εφαρμογής του αλγορίθμου SVM παρουσιάζεται στο Σχήμα 19 όπου ο γίνεται γραμμικός διαχωρισμός των δεδομένων. Η επιφάνεια με τους σταυρούς και με τους κύκλους περιγράφει θετικά και αρνητικά παραδείγματα αντίστοιχα, ενώ οι γραμμές αναπαριστούν τις επιφάνειες απόφασης. Οι δύο διακεκομμένες γραμμές δείχνουν τα δύο άκρα στα οποία μπορεί κανείς να μεταφέρει τις επιφάνειες απόφασης χωρίς να αλλάξει η κατηγοριοποίηση. Τα στοιχεία που βρίσκονται πάνω στις διακεκομμένες γραμμές ονομάζονται υποστηρικτικά διανύσματα (support vectors). Η επιφάνεια απόφασης που αναπαρίσταται με την μεσαία γραμμή είναι η καλύτερη δυνατή δεδομένου ότι είναι το μέσο στοιχείο μεταξύ των προηγούμενων επιφανειών απόφασης (στο περιθώριο μεταξύ των δύο επιφανειών απόφασης δεν υπάρχει κανένα παράδειγμα

ενώ η απόσταση της μεσαίας επιφάνειες από κάθε παράδειγμα και των δύο συνόλων είναι η μέγιστη) [24], [17]. Αξίζει να αναφερθεί ότι η επιφάνεια απόφασης που επιλέγεται κάθε φορά είναι η καλύτερη από ένα μικρό σύνολο εκπαιδευτικών παραδειγμάτων.



Σχήμα 19 Γραμμική προσέγγιση διανυσματικών μηχανισμών υποστήριξης

Πλεονεκτήματα	**Μειονεκτήματα**
Καλά αποτελέσματα σε γραμμικά και μη-γραμμικά προβλήματα	Τα διανύσματα υποστήριξης (support vectors) για μεγάλο εκπαιδευτικό σύνολο παραδειγμάτων μπορεί να έχουν μεγάλο μέγεθος οπότε μειώνεται η ταχύτητα κατηγοριοποίησης
Δεν απαιτείται επιλογή όλων των όρων του εγγράφου προς κατηγοριοποίηση παρά μόνο των διανυσμάτων υποστήριξης	Η επιλογή της περιοχής του πυρήνα στην οποία θα αναζητηθούν τα διανύσματα υποστήριξης απαιτεί χρόνο
Η ικανότητα του αλγορίθμου SVM να μάθει είναι ανεξάρτητη της διάστασης του χώρου	

Πίνακας 11 Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου SVM

4.4.2.2 Κ-πιο κοντινός γείτονας (K-Nearest Neighbor – kNN)

Στόχος του αλγορίθμου αυτού είναι να αποφασίσει εάν ένα έγγραφο d_i ανήκει σε μια κατηγορία c_j. Η διαδικασία που ακολουθείται περιλαμβάνει την αναζήτηση του k πιο κοντινού (όσον

αφορά την ομοιότητα) γείτονα ανάμεσα στο αρχικό σετ εγγράφων. Συγκεκριμένα, ελέγχεται εάν τα k πιο κοντινά έγγραφα ανήκουν επίσης στην κατηγορία c_j. Εάν η απάντηση είναι θετική για ένα αρκετά μεγάλο ποσοστό αυτών τότε όντως το έγγραφο ανήκει σ' αυτήν την κατηγορία. Για να αποφασιστεί εάν ένα έγγραφο ανήκει σε μια κατηγορία αρκεί να υπολογιστεί η παρακάτω σχέση: $y\left(\vec{x}, c_j\right) = \sum_{\vec{d_i} \in kNN} sim\left(\vec{x}, \vec{d_i}\right) y\left(\vec{d_i}, c_j\right) - b_j$ όπου $y\left(\vec{x}, c_j\right) \in \{0,1\}$ είναι το αποτέλεσμα της κατηγοριοποίησης του εγγράφου d_i στην κατηγορία c_j, $sim\left(\vec{x}, \vec{d_i}\right)$ είναι η ομοιότητα μεταξύ του εγγράφου $\vec{x}$ (το έγγραφο ανήκει στους «γείτονες» του αρχικού εγγράφου) και του d_i και ο όρος b_j είναι ένας αριθμός που χαρακτηρίζει μια κατηγορία. Η εύρεση του όρου b_j γίνεται χρησιμοποιώντας τη μετρική F_1 (εξηγείται στην παράγραφο 4.4.2.6). Τέλος, η επιτυχία του αλγορίθμου έγκειται στον υπολογισμό του αριθμού των εγγράφων (k) για τα οποία θα πρέπει να ελεγχθεί αν ανήκουν στην κατηγορία c_i. Πειράματα έδειξαν ότι οι τιμές του k θα πρέπει να κυμαίνονται από 30 έως 45.

Πλεονεκτήματα	**Μειονεκτήματα**
Αποτελεσματικότητα	Δυσκολία στην εύρεση μιας καλής μετρικής για την απόσταση των δύο κατηγοριών
Δυνατός (robust) σε δεδομένα με πολύ θόρυβο υπολογίζοντας το μέσο όρο των k-πιο κοντινών γειτόνων	Απαιτεί πολύ χρόνο για την κατηγοριοποίηση αφού πρέπει να πάρουν μια βαθμολογία όλα τα έγγραφα με τα οποία θα συγκριθεί το έγγραφο προς κατηγοριοποίηση
	Στην περίπτωση που υπάρχουν πολλές διαστάσεις (dimensionality) στο πρόβλημα η απόσταση μεταξύ των γειτόνων μπορεί να κυριαρχηθεί από άσχετα χαρακτηριστικά

Πίνακας 12 Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου kNN

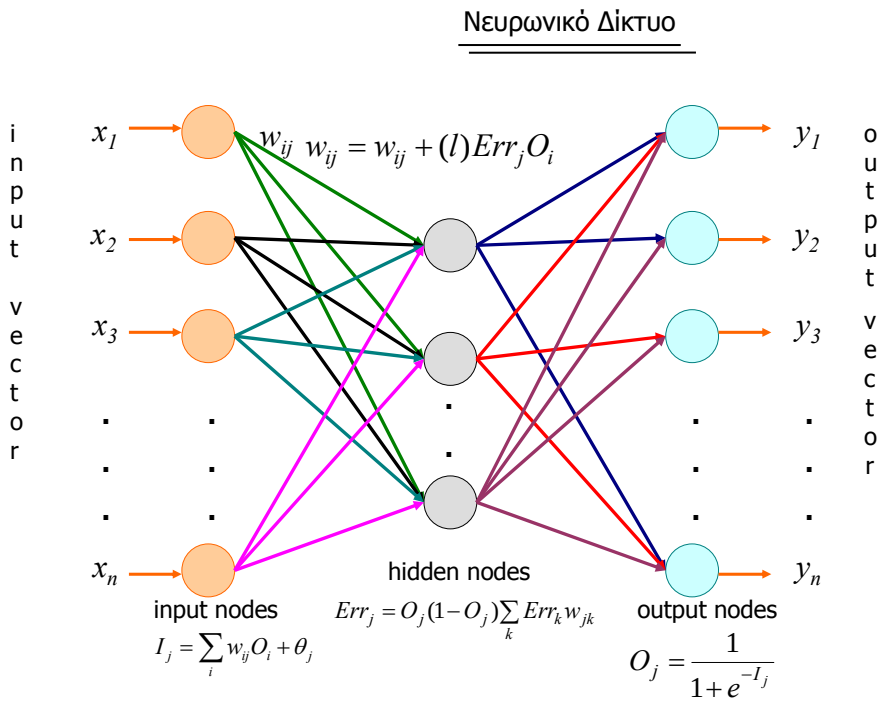
4.4.2.3 Νευρωνικό Δίκτυο (Neural Network - NNet)

Οι τεχνικές που βασίζονται στα νευρωνικά δίκτυα χρησιμοποιούνται κυρίως στον τομέα της Τεχνητής Νοημοσύνης. Η κατηγοριοποίηση με νευρωνικά δίκτυα αποτελείται από ένα δίκτυο ομάδων όπου οι μονάδες εισόδου αναπαριστούν όρους ενώ οι μονάδες εξόδου αναπαριστούν την

κατηγορία ή τις κατηγορίες στις οποίες ανήκει ένα αντικείμενο. Τα βάρη που αποδίδονται στις ακμές του δικτύου, και συνδέουν τις μονάδες, αναπαριστούν σχέσεις εξάρτησης. Εάν για παράδειγμα, ζητείται να κατηγοριοποιηθεί ένα έγγραφο d_j η διαδικασία που ακολουθείται περιλαμβάνει την φόρτωση των βαρών των όρων του σαν είσοδο, την ενεργοποίηση τους και τη διάδοσή τους μέσα στο δίκτυο. Η έξοδος του δικτύου αποτελεί τις τελικές κατηγορίες. Ένας τρόπος για την εκπαίδευση ενός νευρωνικού δικτύου είναι η «προς τα πίσω διάδοση των λαθών» (backpropagation) όπου τα βάρη των όρων φορτώνονται στην είσοδο και τα λάθη που προκύπτουν διαδίδονται προς τα πίσω προκειμένου να γίνουν αλλαγές στις παραμέτρους του δικτύου με στόχο την ελαχιστοποίησή τους [24].

Στο Σχήμα 20 παρουσιάζεται ένα παράδειγμα νευρωνικού δικτύου. Ο αντικειμενικός στόχος της διαδικασίας που περιγράφηκε παραπάνω είναι η δημιουργία ενός συνόλου βαρών που θα κατηγοριοποιεί την πλειοψηφία των πλειάδων του εκπαιδευτικού συνόλου δεδομένων σωστά. Τα βήματα είναι τα εξής:

- Αρχικοποίηση των βαρών με τυχαίες τιμές
- Προσθήκη των πλειάδων εισόδου στο δίκτυο μία προς μία
- Για κάθε μονάδα υπολογισμός της τιμής εισόδου του δικτύου για μια μονάδα ως γραμμικός συνδυασμός όλων των εισόδων, υπολογισμός της τιμής εξόδου χρησιμοποιώντας την συνάρτηση ενεργοποίησης, υπολογισμός του λάθους, ανανέωση των βαρών και της στατιστικής απόκλισης (bias).



Σχήμα 20 Παράδειγμα ενός νευρωνικού δικτύου

4.4.2.4 Linear Least-squares Fit (LLSF) mapping

Η προσέγγιση LLSF είναι μια μέθοδος χαρτογράφησης που αναπτύχθηκε από τον Yang το 1992. Στην προσέγγιση αυτή ένα μοντέλο οπισθοδρόμησης πολλών μεταβλητών μαθαίνει αυτόματα από ένα σύνολο κατάρτισης εγγράφων και των κατηγοριών στις οποίες ανήκουν. Τα δεδομένα αναπαριστώνται από ζεύγη διανυσμάτων εισόδου / εξόδου με το διάνυσμα εισόδου να αποτελείται από τους όρους και τα βάρη τους για κάθε έγγραφο (όπως στον αλγόριθμο SVM 4.4.2.1) ενώ το διάνυσμα εξόδου αποτελείται από τις κατηγορίες στις οποίες κατηγοριοποιήθηκε το έγγραφο. Λύνοντας το πρόβλημα LLSF παράγεται ένας πίνακας με τις λέξεις και τις κατηγορίες στις οποίες αυτές αντιστοιχούν. Ο τύπος που υπολογίζει αυτόν τον πίνακα είναι ο: $F_{LS} = \arg\min_F \|FA - B\|^2$ με A και B τα δεδομένα εκπαίδευσης ενώ ο πίνακας F_{LS} είναι ο ζητούμενος πίνακας ο οποίος και ορίζει την αντιστοίχηση ενός τυχαίου εγγράφου με το διάνυσμα που αναπαριστά τις κατηγορίες και τα βάρη τους. Ταξινομώντας τις κατηγορίες αυτές βάσει του βάρους τους σχηματίζεται μια λίστα από όλες τις κατηγορίες στις οποίες μπορεί να ανήκει ένα συγκεκριμένο έγγραφο [19].

4.4.2.5 Κατηγοριοποίηση με τη μέθοδο Naïve Bayes (Naïve Bayes classifier - NB)

Η βασική ιδέα της κατηγοριοποίησης με τη μέθοδο Naïve Bayes στηρίζεται στον υπολογισμό των συνδυασμένων πιθανοτήτων των λέξεων και των κατηγοριών προκειμένου να εκτιμηθεί η πιθανότητα ένα δοθέν έγγραφο να ανήκει σε μια κατηγορία. Στη μέθοδο αυτή οι λέξεις θεωρούνται ανεξάρτητες. Αυτό σημαίνει ότι η πιθανότητα υπό συνθήκη μιας λέξης δεδομένης μιας κατηγορίας είναι ανεξάρτητη από τις πιθανότητες υπό συνθήκη των υπολοίπων λέξεων δεδομένης της συγκεκριμένης κατηγορίας. Η υπόθεση αυτή μειώνει την πολυπλοκότητα του αλγορίθμου σε σχέση με άλλες μεθόδους αφού δε χρησιμοποιούνται συνδυασμοί λέξεων για την πρόβλεψη της κατηγοριοποίησης ενός εγγράφου σε μια δεδομένη κατηγορία.

Πλεονεκτήματα	**Μειονεκτήματα**
Αποτελεσματικότητα αφού δεν χρησιμοποιούνται συνδυασμοί λέξεων για την πρόβλεψη	Η υπόθεση της υπό όρους ανεξαρτησίας των κατηγοριών (class conditional independence) οδηγεί σε απώλεια της ακρίβειας των αποτελεσμάτων
Εύκολη υλοποίηση	Υπάρχουν εξαρτήσεις μεταξύ των μεταβλητών του προβλήματος οι οποίες δεν μοντελοποιούνται με τη μέθοδο Naïve Bayes
Καλά αποτελέσματα στις περισσότερες περιπτώσεις	

Πίνακας 13 Πλεονεκτήματα και μειονεκτήματα της μεθόδου Naïve Bayes

4.4.2.6 Σύγκριση των παραπάνω αλγορίθμων

Η αξιολόγηση των παραπάνω αλγορίθμων γίνεται με στόχο τη μέτρηση της αποτελεσματικότητάς τους δηλαδή της ικανότητας να παίρνουν σωστές αποφάσεις σχετικά με την κατηγοριοποίηση. Οι μετρικές που χρησιμοποιούνται είναι η ακρίβεια (precision), η ανάκληση (recall) και η μετρική F (F-measure).

Η ακρίβεια ορίζεται ως η πιθανότητα ότι αν ένα τυχαίο έγγραφο κατηγοριοποιηθεί κάτω από μια κατηγορία, τότε αυτή η απόφαση είναι σωστή. Η ανάκληση ορίζεται ως η πιθανότητα να παρθεί η απόφαση, ότι ένα τυχαίο αντικείμενο θα κατηγοριοποιηθεί σωστά σε μια κατηγορία στην οποία και έπρεπε να κατηγοριοποιηθεί. Τυπικά, η ακρίβεια και η ανάκληση μπορούν να ορισθούν ως:

$\widehat{\pi}_i = \frac{TP_i}{TP_i + FP_i}$ και $\widehat{\rho}_i = \frac{TP_i}{TP_i + FN_i}$ αντίστοιχα. Στις προηγούμενες σχέσεις FP_i είναι ο αριθμός των εγγράφων που έχουν κατηγοριοποιηθεί λανθασμένα στην κατηγορία c_i, TP_i είναι ο αριθμός των εγγράφων που έχουν κατηγοριοποιηθεί σωστά και FN_i είναι ο αριθμός των εγγράφων τα οποία λανθασμένα δεν κατηγοριοποιήθηκαν κάτω από την κατηγορία c_i. Η μετρική F υπολογίζεται ορίζεται ως ο αρμονικός μέσος της ακρίβειας και της ανάκλησης ($F(r, p) = \frac{2rp}{r + p}$) με p την ακρίβεια και r την ανάκληση.

Από τα πειράματα που έχουν διεξαχθεί με όλες τις παραπάνω μεθόδους ([19], [24]) παρατηρείται ότι ο αλγόριθμος που βασίζεται σε διανυσματικούς μηχανισμούς υποστήριξης (SVM) υπερτερεί από όλους τους υπόλοιπους ενώ η προσέγγιση βασισμένη σε Naïve Bayes έχει τη χειρότερη απόδοση.

4.4.3 Συλλογή και κατηγοριοποίηση κειμένου που προέρχεται από εξόρυξη από το διαδίκτυο

Οι αλγόριθμοι που παρουσιάστηκαν παραπάνω κατηγοριοποιούν τα έγγραφα αναλύοντας τις σχέσεις μεταξύ τους. Ωστόσο όταν οι τεχνικές αυτές εφαρμόζονται σε πεδία (domains) με πολλές κατηγορίες η ακρίβεια στην κατηγοριοποίηση μπορεί να εξασφαλιστεί μόνο όταν χρησιμοποιηθεί μεγάλο πλήθος εκπαιδευτικών παραδειγμάτων (training examples). Για το λόγο αυτό, αρκετές προσεγγίσεις στη βιβλιογραφία χρησιμοποιούν το διαδίκτυο σαν επιπλέον πηγή γνώσης σε συνδυασμό με κάποια από τις τεχνικές κατηγοριοποίησης που αναλύθηκε παραπάνω.

Στο [14] ως επιπλέον πηγή γνώσης χρησιμοποιείται ο κατάλογος του Yahoo!. Στόχος είναι η κατηγοριοποίηση ιστοσελίδων βάσει της περιγραφής που παρέχεται για κάθε ιστοσελίδα από τον κατάλογο του Yahoo!. Στο [37] προτείνεται η εξόρυξη πληροφορίας από τις μηχανές αναζήτησης για την κατηγοριοποίηση τμημάτων κειμένου με όσο το δυνατό πιο αυτοματοποιημένο τρόπο. Η προτεινόμενη διαδικασία εκμεταλλεύεται τα αποτελέσματα που επιστρέφονται από τις μηχανές αναζήτησης χρησιμοποιώντας το περιεχόμενό τους ως πηγή εκπαίδευσης τόσο της αρχικής ταξινόμησης των κατηγοριών (οι κατηγορίες έχουν εξαχθεί από τις λέξεις – κλειδιά του περιεχομένου προς κατηγοριοποίηση και συνιστούν ένα δέντρο. Κάθε κατηγορία περιγράφεται με μια λέξη ενώ η έννοιά της συντίθεται από την ένωση όλων των «παιδιών» της) όσο και των

κομματιών περιεχομένου που θα πρέπει να κατηγοριοποιηθούν. Η τελική κατηγοριοποίηση πραγματοποιείται χρησιμοποιώντας τον αλγόριθμο kNN (K-Nearest Neighbor - 4.4.2.2).

Οι Chuang και Chien [38] χρησιμοποιούν το διαδίκτυο για να ανακαλύψουν τη σχετικότητα (όσον αφορά την ομοιότητα) μικρών κομματιών περιεχομένου (text segments) που μπορεί να προκύπτουν από την επεξεργασία εγγράφων (λέξεις – κλειδιά, τίτλοι) και ερωτήσεων φυσικής γλώσσας. Κάθε μικρό κείμενο που εξάγεται από την επεξεργασία των αρχικών κειμένων μετατρέπεται σε ένα σύνολο όρων. Το σύνολο όρων εμπλουτίζεται με τα αποτελέσματα των μηχανών αναζήτησης χρησιμοποιώντας τα μικρά κείμενα ως επερωτήσεις. Στη συνέχεια το σύνολο όρων που δημιουργείται συγκρίνεται (όσον αφορά την ομοιότητα) με τα αντικείμενα που ανήκουν ήδη σε κάποια ομάδα εφαρμόζοντας έναν ιεραρχικό αλγόριθμο κατηγοριοποίησης (4.4.1.1). Το δυαδικό δέντρο των κατηγοριών που προκύπτει μετατρέπεται σε ένα φυσικό και περιεκτικό δέντρο (ρηχό και με μεγάλο εύρος).

Τέλος, οι Huang, Chuang και Chien [18], [39] χρησιμοποιούν μια κατηγοριοποίηση που βασίζεται στο διαδίκτυο για να παράγουν αυτόματα θεματικά μεταδεδομένα (thematic metadata) που υπάρχουν σε κείμενα. Στόχος είναι η κατηγοριοποίηση κειμένων, δεδομένης μιας θεματικής ιεραρχίας κατηγοριών. Η προσέγγιση που προτείνεται εκμεταλλεύεται τη διαθέσιμη πληροφορία στο διαδίκτυο για να δημιουργήσει εκπαιδευτικά παραδείγματα τόσο για τις θεματικές κατηγορίες όσο και για τα μεταδεδομένα που προέρχονται από τα κείμενα. Η μέθοδος περιλαμβάνει τρία λειτουργικά μέρη: εξαγωγή χαρακτηριστικών για τα κείμενα μέσω του διαδικτύου, εκπαίδευση των στατιστικών μοντέλων για τις θεματικές κατηγορίες και κατηγοριοποίηση των κειμένων στις κατάλληλες κατηγορίες. Η εξαγωγή χαρακτηριστικών πραγματοποιείται στέλνοντας τα κομμάτια κειμένου (text segments) στις μηχανές αναζήτησης αφού πρώτα δημιουργηθούν οι κατάλληλες επερωτήσεις προς αυτές. Το αποτέλεσμα είναι η παραγωγή ενός συνόλου αντικειμένων για κάθε θεματική κατηγορία. Τα αντικείμενα αυτά σε συνδυασμό με τα κομμάτια κειμένου μετατρέπονται σε διανύσματα για τον υπολογισμό της ομοιότητας μεταξύ τους. Ο αλγόριθμος HCQF αναλαμβάνει τη δημιουργία ενός εκπαιδευτικού σετ για κάθε κατηγορία ενώ στο τελικό στάδιο εφαρμόζεται ο αλγόριθμος kNN για την κατηγοριοποίηση των κειμένων.

Οι παραπάνω προσεγγίσεις αποτέλεσαν τη βάση του αλγορίθμου που χρησιμοποιήθηκε και υλοποιήθηκε στα πλαίσια της εργασίας μας.

4.4.4 Ο αλγόριθμος κατηγοριοποίησης που χρησιμοποιήθηκε

Όπως αναφέραμε και στην προηγούμενη ενότητα οι συνηθισμένες τεχνικές κατηγοριοποίησης κειμένων χρησιμοποιούν τις σχέσεις μεταξύ των εγγράφων για την κατηγοριοποίησή τους. Ωστόσο για την κατηγοριοποίηση μικρών τμημάτων περιεχομένου (στη συγκεκριμένη περίπτωση ένα νέο, ένας σύνδεσμος, ένα κείμενο) είναι ανάγκη να ακολουθηθεί μια διαφορετική προσέγγιση. Αυτή περιλαμβάνει τη χρήση μιας εξωτερικής πηγής η οποία και θα αποδώσει περισσότερα σημασιολογικά χαρακτηριστικά σε κάθε ένα από τα μικρά τμήματα πληροφορίας.

Στην προτεινόμενη προσέγγιση θεωρούμε ότι τα τμήματα περιεχομένου μπορούν να αποδοθούν σε μια από τις κατηγορίες των ιεραρχιών που υπάρχουν στο διαδίκτυο. Παραδείγματα αυτών των ιεραρχιών είναι το Yahoo![23], το Google directory[24], το DMOZ[25]. Η ιεραρχία των καταλόγων αυτών προσφέρει έναν ενιαίο τρόπο περιγραφής και πρόσβασης στα δεδομένα τους και επιπλέον αναπαριστά τις περισσότερες εκφάνσεις της ανθρώπινης δραστηριότητας.

Στη συνέχεια της ενότητας αυτής παρουσιάζεται ο αλγόριθμος που εφαρμόστηκε για την κατηγοριοποίηση του περιεχομένου ενώ για την καλύτερη κατανόησή του παρέχεται ένα πλήρες παράδειγμα.

4.4.4.1 Η δική μας προσέγγιση στο πρόβλημα της κατηγοριοποίησης κειμένων

Η μέθοδος που ακολουθείται «εκμεταλλεύεται» την ιεραρχία κατηγοριών του DMOZ (Open Directory Project). Τα κύρια μέρη της διαδικασίας φαίνονται στο Σχήμα 21. Διακρίνονται τρία επιμέρους στάδια στην εξέλιξη της κατηγοριοποίησης:

- Συγκέντρωση και επεξεργασία του περιεχομένου από διαφορετικές πηγές
- Συνδυασμός μιας τεχνικής κατηγοριοποίησης και της ιεραρχίας του DMOZ
- Εφαρμογή του αλγορίθμου της κατηγοριοποίησης και ανάθεση των κειμένων σε κάποια από τις διαθέσιμες κατηγορίες.

Στην ενότητα αυτή δεν θα ασχοληθούμε με την συγκέντρωση του περιεχομένου η οποία θεωρείται δεδομένη. Θέματα υλοποίησης περιγράφονται εκτενέστερα στο Κεφάλαιο 5 ενώ οι πηγές άντλησης του περιεχομένου αναλύονται στην υπο-ενότητα 4.2.2 του παρόντος. Παρομοίως

[23] Ο κατάλογος του Yahoo!: http://dir.yahoo.com

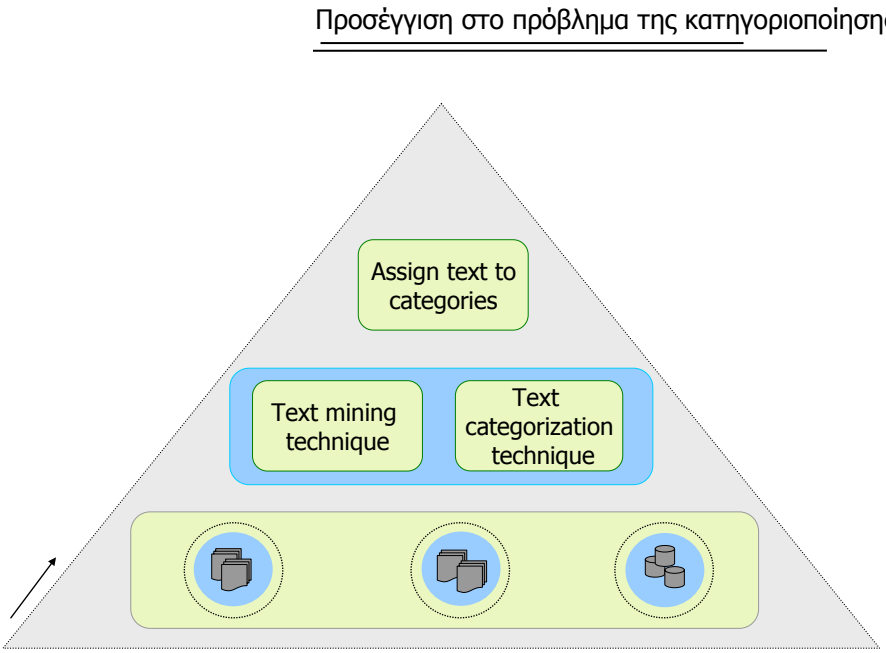
[24] Ο κατάλογος του Google: http://directory.google.com

[25] Ο κατάλογος του DMOZ: http://dmoz.org

ο αλγόριθμος που εφαρμόστηκε για την εξαγωγή των χρήσιμων λέξεων του κειμένου αναλύεται εκτενώς στην υπο-ενότητα 4.3.2.

Στο δεύτερο στάδιο το σύνολο των χρήσιμων λέξεων του κειμένου προς κατηγοριοποίηση εφαρμόζεται πάνω στον κατάλογο του DMOZ και δημιουργείται μια ιεραρχία με πιθανές κατηγορίες. Οι κατηγορίες αυτές ταξινομούνται ανά βάρος και το κείμενο ταξινομείται κάτω από την κατηγορία με το μεγαλύτερο βάρος.

Στην ακόλουθη υπο-ενότητα παρουσιάζονται αναλυτικά τα κύρια σημεία του αλγορίθμου.

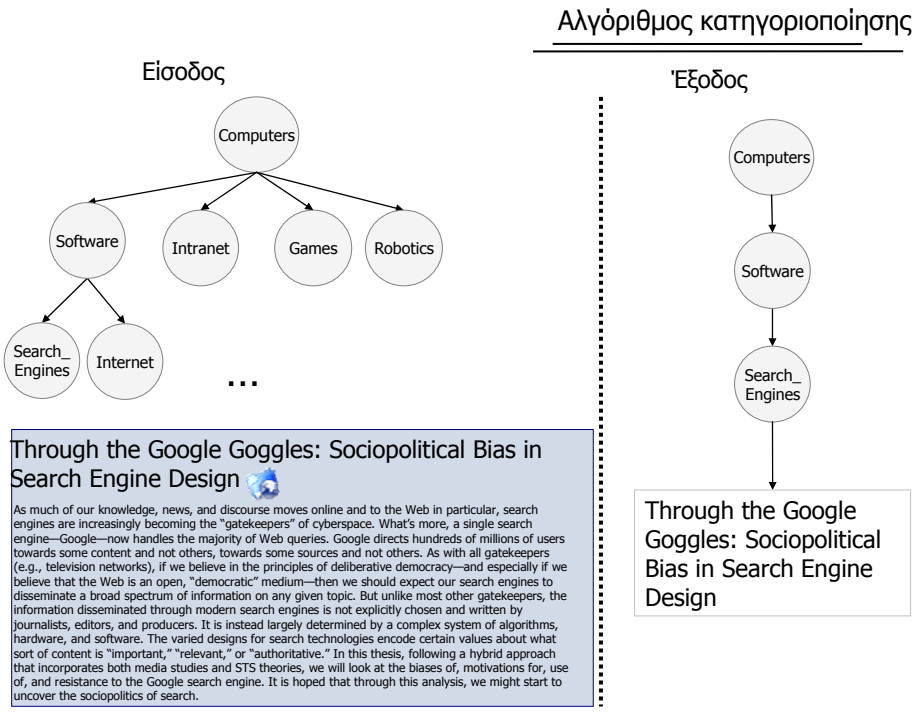


Σχήμα 21 Προσέγγιση που ακολουθείται στο υποσύστημα κατηγοριοποίησης

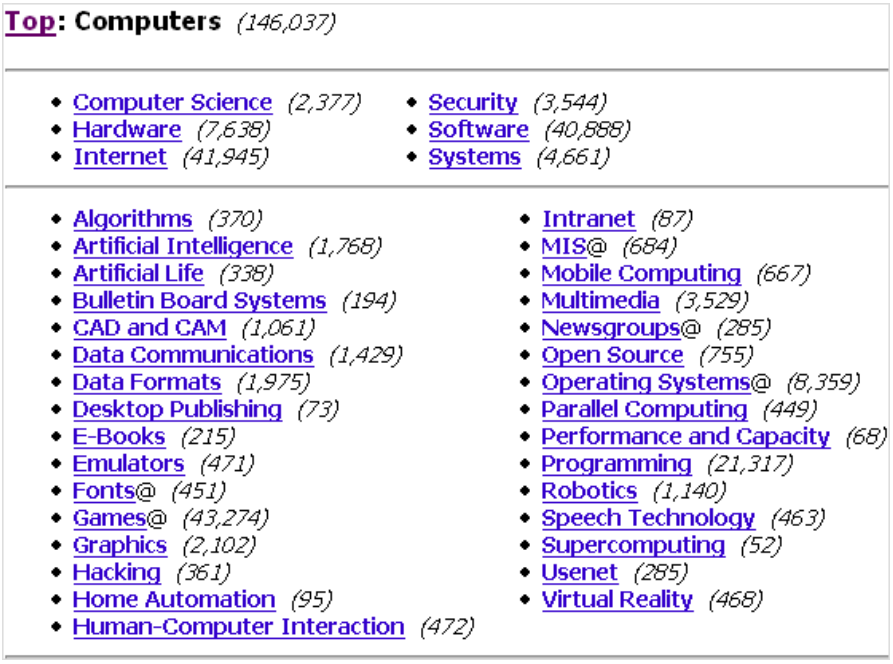
4.4.4.2 Περιγραφή του προτεινόμενου αλγορίθμου κατηγοριοποίησης

Στο Σχήμα 22 φαίνεται η είσοδος και η έξοδος του αλγορίθμου. Στόχος είναι η κατηγοριοποίηση ενός κειμένου δεδομένης της ιεραρχίας κατηγοριών που παρέχει ο κατάλογος DMOZ. Η είσοδος του αλγορίθμου αποτελείται από το κείμενο της κατηγοριοποίησης καθώς και το δέντρο των κατηγοριών. Η έξοδος είναι το μονοπάτι της κατηγορίας (ξεκινώντας από την αρχή της ιεραρχίας) στην οποία αποδίδεται το κείμενο.

Το DMOZ δεικτοδοτεί ένα σύνολο κατηγοριών (εννοιών) που ακολουθούν ιεραρχική οργάνωση σε μορφή κατευθυνόμενου ακυκλικού γράφου. Είναι ο μεγαλύτερος κατάλογος που δημιουργείται από ανθρώπους στο διαδίκτυο ενώ οικοδομείται και συντηρείται από μία κοινότητα εθελοντών συντακτών. Στο Σχήμα 23 παρουσιάζεται το περιεχόμενο της κατηγορίας Top: Computers που μπορεί να βρεθεί στο σύνδεσμο http://dmoz.org/Computers/. Οι κατηγορίες που περιέχουν το σύμβολο '@' στο τέλος του ονόματός τους αποτελούν συνδέσμους προς άλλες κατηγορίες του καταλόγου και βρίσκονται σε διαφορετικό μονοπάτι από αυτό της περιγραφόμενης κατηγορίας. Όλες οι άλλες κατηγορίες αποτελούν υπο-κατηγορίες της συγκεκριμένης κατηγορίας. Στο τελευταίο επίπεδο περιέχονται οι δικτυακοί τόποι που έχουν καταχωρηθεί στον κατάλογο. Τα στοιχεία που αποθηκεύονται γι' αυτούς είναι ο τίτλος και μια σύντομη περιγραφή τους. Στη συγκεκριμένη εργασία χρησιμοποιείται μόνο η δομή καταλόγου χωρίς καμία αναφορά στους δικτυακούς τόπους.



Σχήμα 22 Η είσοδος και η έξοδος του αλγορίθμου κατηγοριοποίησης



Σχήμα 23 Το περιεχόμενο της κατηγορίας Top: Computers

Πριν την έναρξη του αλγορίθμου κατηγοριοποίησης υπάρχουν ορισμένες εργασίες που θα πρέπει να πραγματοποιηθούν και αφορούν την επεξεργασία του καταλόγου του DMOZ. Αυτές είναι:

Επιλογή της κατηγορίας του καταλόγου του DMOZ που θα χρησιμοποιηθεί στο δικτυακό τόπο: Ο κατάλογος του DMOZ περιλαμβάνει έναν πολύ μεγάλο αριθμό κατηγοριών. Για το λόγο αυτό οι απαιτήσεις σε αποθηκευτικό χώρο είναι μεγάλες. Στην αρχή της διαδικασίας λοιπόν, επιλέγεται ποια από τις κεντρικές κατηγορίες αποτυπώνει καλύτερα το θέμα του δικτυακού τόπου (Computers, Arts, Games, Business, News, Society, κλπ.). Η κατηγορία αυτή καθώς και όλες οι υποκατηγορίες της αποθηκεύονται στη βάση δεδομένων που υποστηρίζει το δικτυακό τόπο.

Ταξινόμηση του καταλόγου με αλφαβητική σειρά: Για κάθε κατηγορία στον κατάλογο του DMOZ αποθηκεύονται δύο στοιχεία. Αυτά είναι ο τίτλος της κατηγορίας καθώς και το μονοπάτι που ακολουθείται για να φτάσει κανείς σε αυτή την κατηγορία. Ο τίτλος είναι συνήθως μια λέξη που περιγράφει την κατηγορία ενώ το μονοπάτι είναι μια ακολουθία λέξεων με αρχή τη ρίζα του δέντρου. Σε αυτό το βήμα η διαδικασία περιλαμβάνει την ταξινόμηση του δέντρου των κατηγοριών με βάση τον τίτλο τους.

Ευρετηριασμός (Indexing): Η διαδικασία του ευρετηριασμού βασίζεται στην αντιστοίχηση κάθε γράμματος του αγγλικού αλφάβητου με έναν αριθμό. Ο αριθμός αυτός δείχνει τη θέση της πρώτης κατηγορίας που έχει σαν πρώτο γράμμα το συγκεκριμένο γράμμα της αλφαβήτου. Με τον

τρόπο αυτό η αναζήτηση μιας κατηγορίας περιορίζεται μόνο σε ένα μικρό κομμάτι του καταλόγου του DMOZ.

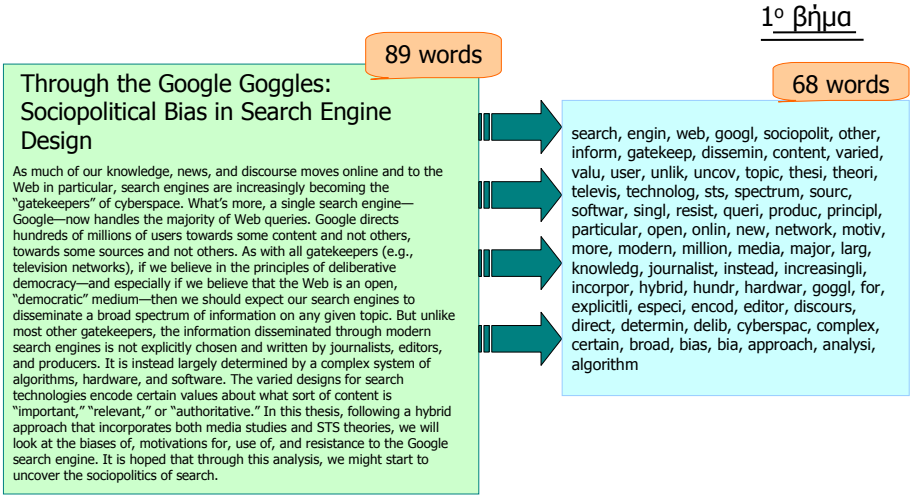
Επεξεργασία κατηγοριών με πολλές λέξεις: Το τελευταίο βήμα της επεξεργασίας του καταλόγου του DMOZ περιλαμβάνει το χειρισμό κατηγοριών των οποίων οι τίτλοι αποτελούνται από περισσότερες από μια λέξεις. Στην περίπτωση αυτή, η κατηγορία χωρίζεται σε ίσο αριθμό κατηγοριών όσες είναι και οι λέξεις του τίτλου της ενώ το μονοπάτι παραμένει το ίδιο. Με τη μέθοδο αυτή παρέχεται η δυνατότητα ανακατασκευής του καταλόγου των κατηγοριών.

Αφού ολοκληρωθούν οι παραπάνω εργασίες μπορεί να εφαρμοστεί ο προτεινόμενος αλγόριθμος κατηγοριοποίησης.

Αναλυτικά τα βήματα που ακολουθούνται για την ολοκλήρωση του αλγορίθμου είναι τα εξής:

Βήμα 1: Δημιουργία ενός συνόλου με τις σημαντικότερες λέξεις του κειμένου προς κατηγοριοποίηση.

Για κάθε κείμενο προς κατηγοριοποίηση (νέο, σύνδεσμος, αρχείο) εξάγεται ένα σύνολο με τις σημαντικότερες λέξεις, με τον τρόπο που αναλύθηκε στην υπο-ενότητα 4.3.2. Κάθε λέξη αποκτά ένα βάρος ανάλογα με τη συχνότητα εμφάνισής της στο κείμενο. Οι λέξεις που απαρτίζουν το παραπάνω σύνολο αποτελούν τις ρίζες των αρχικών λέξεων. Στο Σχήμα 24 δίνεται ένα παράδειγμα του λεξικού που παράγεται από το κείμενο «Through the Google Goggles: Sociopolitical Bias in Search Engine Design».



Σχήμα 24 Εξαγωγή λεξικού για ένα κείμενο

Βήμα 2: Αναζήτηση των λέξεων του συνόλου του βήματος 1 μέσα στον κατάλογο με τις κατηγορίες του DMOZ.

Για κάθε μία λέξη, που ανήκει στο σύνολο που δημιουργήθηκε στο βήμα 1, εκτελείται αναζήτηση στο κομμάτι της ιεραρχίας των κατηγοριών βάσει του αρχικού γράμματος αυτής.

Αν η λέξη που αναζητείται αποτελεί μέρος μιας κατηγορίας τότε η αντίστοιχη κατηγορία (καθώς και το μονοπάτι της) αποθηκεύεται σε μία λίστα κατηγοριών. Τα στοιχεία που αποθηκεύονται για κάθε κατηγορία, είναι το «μονοπάτι» της καθώς και μια τιμή που αποτελεί το βάρος της κατηγορίας και εξαρτάται από τη συχνότητα εμφάνισης της ρίζας της λέξης που αναζητούνταν. Στον αλγόριθμο κατηγοριοποίησης καθορίστηκε (ευρεστικά) ότι το βάρος της κατηγορίας που αντιστοιχείται σε έναν όρο του λεξικού θα ισούται με το μισό της συχνότητας εμφάνισης της ρίζας της λέξης στο κείμενο. Δηλαδή, εάν k είναι η συχνότητα εμφάνισης της ρίζας της λέξης – κλειδί τότε το βάρος της κατηγορίας c στην οποία αντιστοιχίζεται υπολογίζεται ως $weight_c = \alpha\kappa\acute{\varepsilon}\rho\alpha\iota o\ \mu\acute{\varepsilon}\rho o\varsigma(\frac{k}{2})$.

Βήμα 3: Δημιουργία ταξινομημένης λίστας με τις ευρεθείσες κατηγορίες.

Στο βήμα αυτό ταξινομούνται οι κατηγορίες που βρέθηκαν ως σχετικές ανάλογα με το βάρος τους.

Βήμα 4: Επιστροφή της πιο σχετικής κατηγορίας

Εάν το πρώτο στοιχείο της λίστας (κατηγορία) έχει βάρος μεγαλύτερο από μια τιμή που έχει οριστεί ευρεστικά (μετά από τη διεξαγωγή μιας σειράς πειραμάτων) τότε το αποτέλεσμα του αλγορίθμου είναι αυτή η κατηγορία. Εάν όμως έχει μικρότερο βάρος από αυτή την τιμή τότε το δέντρο που δημιουργούν οι κατηγορίες διατρέχεται από τα φύλλα προς τη ρίζα. Η διαδικασία που ακολουθείται περιγράφεται στο Σχήμα 25 και περιλαμβάνει την μεταφορά των βαρών των κατηγοριών στις πατρικές τους κατηγορίες. Εάν και η πατρική κατηγορία έχει κάποιο βάρος (επιστράφηκε ως σχετική) τότε το βάρος του απογόνου της προστίθεται στο δικό της βάρος. Η διαδικασία συνεχίζεται μέχρι να βρεθεί η κατηγορία που θα πληροί τον παραπάνω περιορισμό ή μέχρι να φτάσουμε στο τέλος του δέντρου.

Ειδικές περιπτώσεις

- Εάν δύο ή περισσότερες κατηγορίες έχουν τιμή μεγαλύτερη από την τιμή που έχει οριστεί επιστρέφονται και όλες στο χρήστη και αποφασίζει εκείνος την πιο σχετική

- Εάν δεν βρεθεί σχετική κατηγορία το κείμενο εισάγεται κάτω από μια κατηγορία με την αντίστοιχη ετικέτα.

Για να γίνει κατανοητή η διαδικασία της κατηγοριοποίησης περιγράφουμε το παρακάτω παράδειγμα (Σχήμα 25). Έστω ότι η τιμή του threshold είναι 5 και οι σχετικές κατηγορίες περιέχονται στον πίνακα του βήματος 1. Τα βήματα που ακολουθούνται είναι:

1ο βήμα

Top/computers/hardware/devices/processors	1
Top/computers/hardware/devices/memory	2
Top/computers/hardware/systems	2
Top/computers/internet	1
Top/computers	1

Εφόσον η μέγιστη τιμή του βάρους είναι 2 τα βάρη αποδίδονται στις πατρικές κατηγορίες.

2ο βήμα

Top/computers/hardware/devices	3
Top/computers/hardware	2
Top/computers	2

Στη συγκεκριμένη περίπτωση το βάρος της κατηγορίας Top/Computers δεν μεταφέρεται στην πατρική κατηγορία Top αφού η Top/Computers περιέχει ακόμη απογόνους στο δέντρο (αρκεί να ελεγχθεί η αμέσως επόμενη κατηγορία αλφαβητικά που είναι η Top/computers/hardware).

3ο βήμα

Top/computers/hardware	5
Top/computers	2

Εδώ η μέγιστη τιμή του βάρους στο δέντρο είναι 5 οπότε η πιο σχετική κατηγορία είναι η Top/computers/hardware.

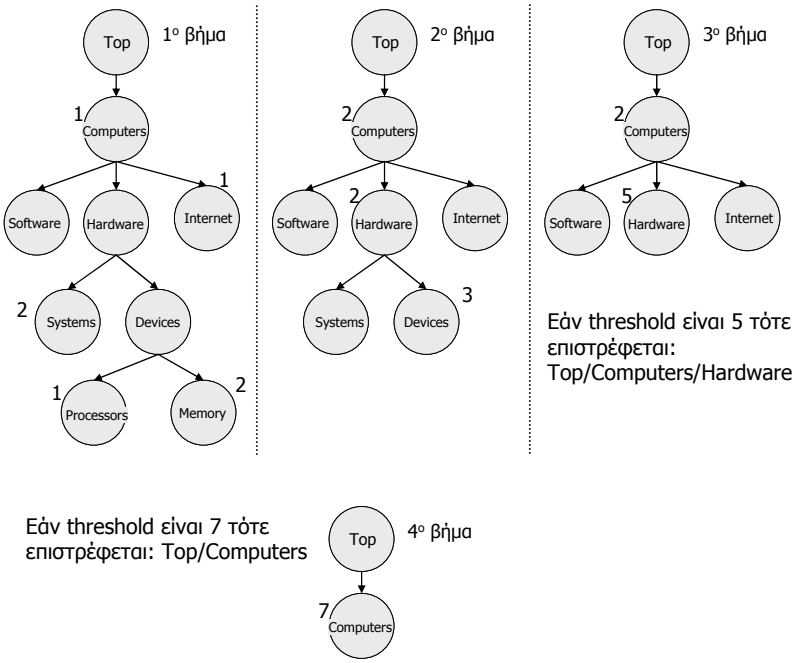
Εάν η τιμή της σταθεράς ήταν 7 τότε ο αλγόριθμος θα είχε ένα ακόμη βήμα και η πιο σχετική κατηγορία θα ήταν η Top/computers.

4ο βήμα

Top/computers	7

Εάν η τιμή της σταθεράς ήταν 8 ο αλγόριθμος δεν θα επέστρεφε καμία τιμή.

Στην περίπτωση που η τιμή της σταθεράς ήταν 2 τότε ο αλγόριθμος θα σταματούσε στο βήμα 2 και θα επέστρεφε και τις δύο κατηγορίες: Top/computers/hardware και Top/computers.



Σχήμα 25 Εύρεση της τελικής κατηγορίας διατρέχοντας το δέντρο των κατηγοριών ανά επίπεδο

4.5 Εποπτική εικόνα του υποσυστήματος κατηγοριοποίησης

Μια εποπτική εικόνα του υποσυστήματος κατηγοριοποίησης καθώς και τα λειτουργικά του μέρη παρατίθενται στο Σχήμα 26.

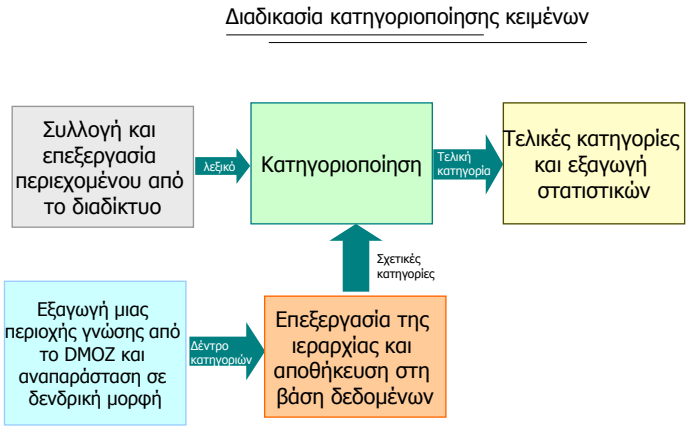
Συλλογή και επεξεργασία περιεχομένου από το διαδίκτυο: Περιλαμβάνει διαδικασίες συλλογής περιεχομένου από άλλους δικτυακούς τόπους ή από το ΣΔΠ ATL CME. Στο στάδιο αυτό γίνεται η επεξεργασία του κειμένου με στόχο την εξαγωγή ένα σύνολο με τις σημαντικότερες λέξεις του κειμένου (λεξικό).

Εξαγωγή μιας περιοχής γνώσης από το DMOZ και αναπαράσταση της σε δενδρική μορφή: Το RDF αρχείο που παρέχει το DMOZ φορτώνεται στο δικτυακό τόπο στον οποίο πραγματοποιείται η κατηγοριοποίηση. Από εκεί γίνεται εξαγωγή της περιοχής γνώσης που περιγράφει καλύτερα το περιεχόμενο του δικτυακού τόπου. Το RDF αρχείο αναλύεται και δημιουργείται ένα δέντρο με κατηγορίες.

Επεξεργασία της ιεραρχίας και αποθήκευση στη βάση δεδομένων: Το δέντρο με τις κατηγορίες ελέγχεται για τυχόν διπλές εγγραφές και αποθηκεύεται στη βάση δεδομένων.

Κατηγοριοποίηση: Στο σημείο αυτό εφαρμόζεται ο αλγόριθμος κατηγοριοποίησης. Οι όροι του λεξικού ελέγχονται στην ιεραρχία και παράγουν ένα σύνολο με σχετικές κατηγορίες. Ο αλγόριθμος κατηγοριοποίησης εξάγει την τελική κατηγορία στην οποία αντιστοιχίζεται το κείμενο.

Τελικές κατηγορίες και εξαγωγή στατιστικών: Το προϊόν της κατηγοριοποίησης αποτυπώνεται σε μια σελίδα όπου παρουσιάζονται οι κατηγορίες και τα κείμενα που ανήκουν σε κάθε μία. Το προϊόν της κατηγοριοποίησης εξάγεται ακόμη και σε μορφή RSS. Για το περιεχόμενο εξάγονται στατιστικά στοιχεία που αποτυπώνουν το ποσοστό της κατηγοριοποίησης καθώς και στοιχεία για το είδος και των αριθμό των κειμένων που κατηγοριοποιούνται.



Σχήμα 26 Συνοπτική παρουσίαση της διαδικασίας κατηγοριοποίησης περιεχομένου

5 Υλοποίηση

5.1 Εισαγωγή

Στο κεφάλαιο 5 παρουσιάζεται η υλοποίηση του υποσυστήματος κατηγοριοποίησης, στο ΣΔΠ ATL CME. Σε πρώτη φάση αναλύονται οι περιορισμοί που προκύπτουν από το περιβάλλον υλοποίησης που χρησιμοποιήθηκε, η επίδρασή τους στην δική μας εργασία και οι τρόποι αντιμετώπισής τους. Στη συνέχεια αναλύεται η αρχιτεκτονική του υποσυστήματος και η υλοποίηση των κομματιών (components) του. Τέλος, παρουσιάζονται οι λειτουργικότητες που παρέχονται από το σύστημα για το χρήστη.

5.2 Περιορισμοί και προκλήσεις της υλοποίησης

Το υποσύστημα κατηγοριοποίησης υλοποιήθηκε και εντάχθηκε στο ΣΔΠ ATL CME. Αυτό είχε ως αποτέλεσμα να κληρονομήσει όλους τους περιορισμούς που εισάγει το ίδιο το σύστημα αλλά και οι τεχνολογίες που χρησιμοποιεί. Οι περιορισμοί αυτοί είναι:

1. Η γλώσσα προγραμματισμού PHP

Η PHP είναι μια γλώσσα που εφαρμόζεται κυρίως στο επίπεδο της παρουσίασης του περιεχομένου και χρησιμοποιείται για την ανάπτυξη εφαρμογών του διαδικτύου. Εστιάζει δε, στη δημιουργία ενός επιπέδου αλληλεπίδρασης με διάφορες βάσεις δεδομένων για την παραγωγή δυναμικού περιεχομένου στο διαδίκτυο. Το γεγονός αυτό επηρεάζει την ταχύτητα της γλώσσας και η απόδοσή της δεν πλησιάζει κάποια από τις γνωστές γλώσσες προγραμματισμού (C ή C++).

2. Διαχείριση XML αρχείων με την PHP

Η γλώσσα PHP διαθέτει δύο βιβλιοθήκες για την ανάλυση XML αρχείων: την DOM και τη SAX. Οι βιβλιοθήκες αυτές υποστηρίζουν τρία είδη κωδικοποίησης: US-ASCII, ISO-8859-1 και UTF-8. Επομένως περιεχόμενο που είναι κωδικοποιημένο σε κάποια άλλη μορφή δε μπορεί να αναγνωριστεί. Το πρόβλημα αυτό λύθηκε με τη δημιουργία μιας βιβλιοθήκης που περιέχει τις πιο γνωστές κωδικοποιήσεις. Κατά τη συλλογή περιεχομένου λοιπόν χρησιμοποιείται μια εφαρμογή η οποία αντιστοιχίζει όλους τους χαρακτήρες της αλφαβήτου στην κωδικοποίηση ISO-8859-1. Στην περίπτωση που ο δικτυακός τόπος χρησιμοποιεί άλλο πρότυπο για την παρουσίαση του περιεχομένου ενεργοποιείται μια άλλη εφαρμογή που αναλαμβάνει την αντιστοίχηση των χαρακτήρων από ISO-8859-1 στην χρησιμοποιούμενη κωδικοποίηση.

3. Έλλειψη αποθήκευσης δεδομένων σε χώρο διαμοιραζόμενο από όλους τους χρήστες μιας web εφαρμογής

Η υλοποίηση της γλώσσας προγραμματισμού PHP δεν περιλαμβάνει υποστήριξη για την αποθήκευση δεδομένων σε χώρο διαμοιραζόμενο από όλους τους χρήστες μιας web εφαρμογής (application level support). Αν κάποια δεδομένα πρέπει να είναι προσβάσιμα από όλα τα κομμάτια του κώδικα θα πρέπει είτε να είναι αποθηκευμένα στη βάση δεδομένων, με όλη την καθυστέρηση που μπορεί να επιφέρει η σύνδεση και η ανάκτηση τους, είτε σε μεταβλητές που ορίζονται κατά την ενεργοποίηση ενός προσαρτήματος λογισμικού. Το μειονέκτημα στη δεύτερη περίπτωση αφορά τη δυσκολία ανανέωσης του περιεχομένου που κρατάνε οι μεταβλητές αυτές. Για την υλοποίηση της παρούσας εργασίας ακολουθήθηκε η πρώτη προσέγγιση.

4. Μεγάλο μέγεθος του αρχείου της ιεραρχίας του DMOZ

Το μέγεθος του αρχείου της ιεραρχίας του DMOZ είναι περίπου 500MB. Αυτό καθιστά απαγορευτική την απευθείας σύνδεση με τον κατάλογο και την κατηγοριοποίηση του περιεχομένου στο διαδίκτυο. Για το λόγο αυτό αποφασίστηκε το αρχείο να αποθηκεύεται τοπικά στον εξυπηρετητή και να χωρίζεται σε μικρότερα κομμάτια που είναι πιο εύκολα στη διαχείριση και αντιστοιχούν στο θέμα του δικτυακού τόπου. Το κλαδί του δέντρου που θα χρησιμοποιηθεί το αποφασίζει ο διαχειριστής του δικτυακού τόπου κατά την εγκατάσταση του συγκεκριμένου προσαρτήματος λογισμικού.

5. Το σύστημα ATL CME χρησιμοποιεί σχεσιακή βάση δεδομένων

Η σχεσιακή βάση δεδομένων που χρησιμοποιείται από το ATL CME περιορίζει τις δυνατότητες διαχείρισης αρχείων που ακολουθούν XML πρότυπα και δεν επιτρέπει / υποστηρίζει γλώσσες αναζήτησης πάνω σε XML δεδομένα. Για την αντιμετώπιση του συγκεκριμένου προβλήματος επιλέχθηκε η αντιστοίχηση του XML σχήματος που υποστηρίζει το RSS στο σχεσιακό μοντέλο της βάσης δεδομένων και της αποθήκευσης μόνο του περιεχομένου, μια και η βάση δεδομένων δεν υποστηρίζει αποθήκευση XML.

6. Ύπαρξη πολλών διαφορετικών εκδόσεων και προτύπων RSS

Το συγκεκριμένο πρόβλημα εντοπίζεται στο χειρισμό του περιεχομένου που συλλέγεται με το πρότυπο RSS. Συγκεκριμένα, υπάρχουν πολλές εκδόσεις του RSS και αρκετές υλοποιήσεις που έχουν προταθεί στην κοινότητα και χρησιμοποιούνται αν και δεν έχουν προτυποποιηθεί. Η λύση που ακολουθήθηκε για την επίλυση αυτού του περιορισμού αφορά τη δημιουργία ενός ενδιάμεσου επιπέδου για την αναγνώριση και το χειρισμό των διαφορετικών προτύπων. Συγκεκριμένα, για κάθε πρότυπο αποθηκεύθηκαν στοιχεία όπως το namespace που χρησιμοποιεί,

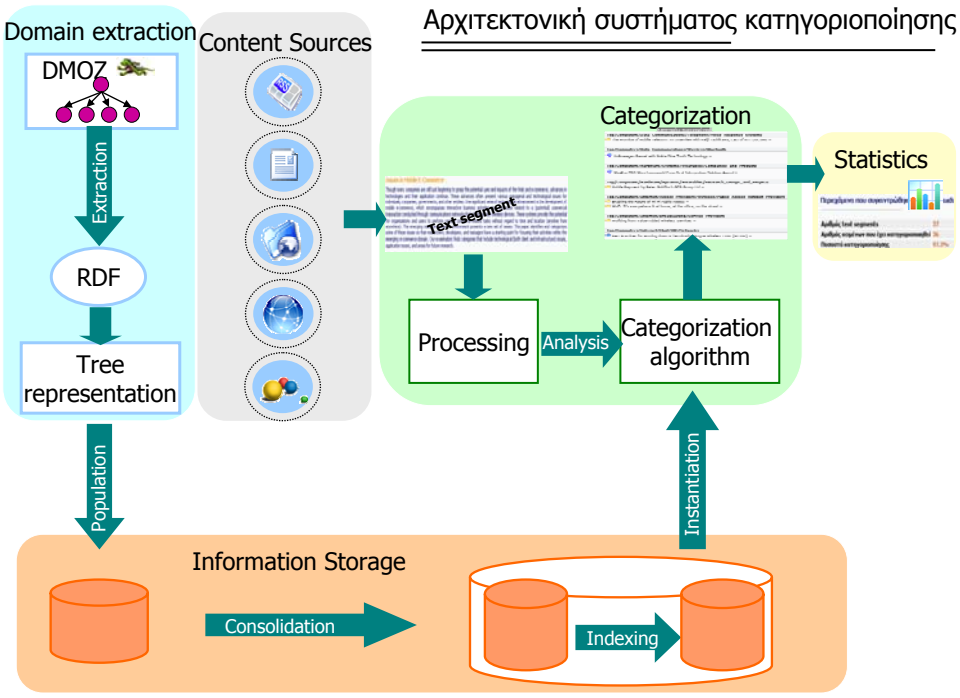
επιπρόσθετες ετικέτες (tags) καθώς και η αντίστοιχη περιγραφή τους η οποία χρησιμοποιείται για την παρουσίαση του περιεχομένου που περιγράφουν.

5.3 Υποσύστημα κατηγοριοποίησης

Το υποσύστημα κατηγοριοποίησης όπως περιγράφηκε στο κεφάλαιο 4 αποτελείται από τα εξής μέρη:

- **Συλλογή περιεχομένου από διάφορες πηγές:** Το κομμάτι αυτό αναλαμβάνει τη συλλογή του περιεχομένου που θα κατηγοριοποιηθεί. Οι πηγές περιεχομένου αφορούν τόσο εσωτερικές πηγές (περιεχόμενο που είναι αποθηκευμένο ήδη στο ATL CME) όσο και εξωτερικές πηγές (RSS, GoogleAPI)
- **Εξαγωγή κατηγορίας από τον κατάλογο του DMOZ και αποθήκευση της:** Πρόκειται για τη διαδικασία εξαγωγής μιας θεματικής ενότητας από τον κατάλογο του DMOZ (ανάλογα με το θέμα του δικτυακού τόπου) και την αποθήκευση της στη βάση δεδομένων του δικτυακού τόπου
- **Κατηγοριοποίηση του περιεχομένου που συλλέχθηκε:** Το κομμάτι αυτό αναλαμβάνει την κατηγοριοποίηση του περιεχομένου που έχει συλλεχθεί αντιστοιχίζοντας το σε κάποια από τις κατηγορίες του DMOZ
- **Εξαγωγή στατιστικών:** Τα στατιστικά που εξάγονται από το υποσύστημα κατηγοριοποίησης αποτυπώνουν το ποσοστό της κατηγοριοποίησης ενώ παρέχονται επιπλέον στοιχεία που βασίζονται στις μετρικές που αναφέρθηκαν στην ενότητα 4.4.2.6.

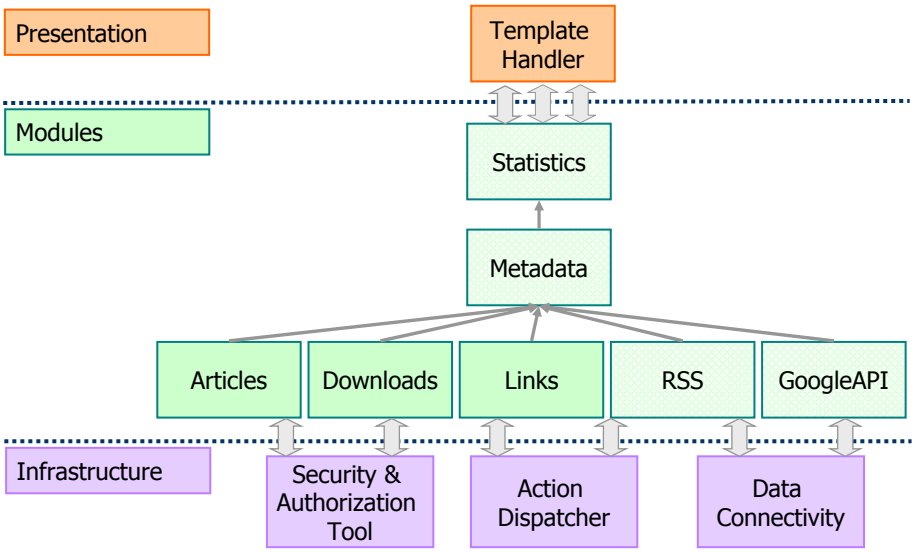
Ο τρόπος με τον οποίο συνεργάζονται τα παραπάνω μέρη αναλύεται στις ακόλουθες ενότητες. Η αρχιτεκτονική του υποσυστήματος παρουσιάζεται στο Σχήμα 27.



Σχήμα 27 Αρχιτεκτονική υποσυστήματος κατηγοριοποίησης

5.3.1 Ένταξη του υποσυστήματος κατηγοριοποίησης στο Σύστημα ATL CME

Η αύξηση της λειτουργικότητας του ΣΔΠ ATL CME, όπως αναφέρθηκε στην ενότητα 2.4, έγκειται στην προσθήκη νέων προσαρτημάτων λογισμικού. Η ολοκλήρωση του υποσυστήματος κατηγοριοποίησης λοιπόν, βασίστηκε τόσο στην ανάπτυξη καινούριων προσαρτημάτων λογισμικού όσο και στη συνεργασία με ορισμένα από τα υπάρχοντα. Στο Σχήμα 28 παρουσιάζονται τα προσαρτήματα λογισμικού που αναπτύχθηκαν στα πλαίσια αυτής της εργασίας (γραμμοσκιασμένα κουτιά) καθώς και η συνεργασία τους με τα διάφορα τμήματα του ΣΔΠ ATL CME. Το τελευταίο επίπεδο αποτελεί τη βασική υποδομή η οποία εξασφαλίζει την αλληλεπίδραση με τη βάση δεδομένων του δικτυακού τόπου, το μοντέλο ασφάλειας καθώς και τη διαχείριση των ενεργειών που γίνονται στο εσωτερικό των προσαρτημάτων λογισμικού. Στο μεσσαίο επίπεδο βρίσκονται τα modules που αναπτύχθηκαν αλλά και όσα από τα υπάρχοντα χρησιμοποιούνται στο υποσύστημα κατηγοριοποίησης ενώ το τελευταίο επίπεδο αναλαμβάνει την παρουσίαση της πληροφορίας.



Σχήμα 28 Τα προσαρτήματα λογισμικού που υλοποιήθηκαν και η αλληλεπίδρασή τους με τα άλλα μέρη του ΣΔΠ ATL CME

Το Σύστημα Διαχείρισης Περιεχομένου ATL CME έχει αναπτυχθεί σε μια αρχιτεκτονική «τριών επιπέδων» - πηγές περιεχομένου, μηχανισμός πυρήνα, προσαρτήματα λογισμικού και τμήματα πληροφορίας (information springs, core engine, modules & blocks). Το υποσύστημα κατηγοριοποίησης ενσωματώθηκε σε αυτή την αρχιτεκτονική προσφέροντας νέα λειτουργικότητα αλλά χρησιμοποιώντας και μέρος της υπάρχουσας υποδομής. Η ένταξη του υποσυστήματος κατηγοριοποίησης έγινε ως εξής:

Πηγές περιεχομένου

Οι πηγές περιεχομένου που υποστηρίζονταν από το Σύστημα Διαχείρισης Περιεχομένου ATL CME αφορούσαν περιεχόμενο που προέρχεται από το διαχειριστή του δικτυακού τόπου μέσω της προσθήκης στατικών σελίδων καθώς και από το περιεχόμενο που διαχειρίζονται τα προσαρτήματα λογισμικού το οποίο αποθηκεύεται στη βάση δεδομένων. Οι πηγές περιεχομένου εμπλουτίστηκαν εισάγοντας το διαδίκτυο ως επιπλέον πηγή πληροφορίας. Αυτό έγινε μέσω δύο νέων προσαρτημάτων λογισμικού: GoogleAPI και RSS/ATOM. Το προσάρτημα λογισμικού RSS/ATOM χρησιμοποιείται για τη συλλογή περιεχομένου από τρίτους δικτυακούς τόπους που παράγουν RSS feeds ενώ το GoogleAPI επιτρέπει την αναζήτηση πληροφορίας στη βάση δεδομένων της μηχανής αναζήτησης Google. Τα αποτελέσματα που επιστρέφονται από την αναζήτηση εισάγονται στο σύστημα και κατηγοριοποιούνται με βάση τη δική μας ιεραρχία.

Βασικός πυρήνας συστήματος διαχείρισης

Το υποσύστημα κατηγοριοποίησης δεν προσέφερε νέα λειτουργικότητα στο κομμάτι αυτό. Ωστόσο η προσαρμογή των νέων προσαρτημάτων λογισμικού πραγματοποιήθηκε χρησιμοποιώντας τη λειτουργικότητα που προσφέρουν τα διάφορα μέρη του πυρήνα. Συγκεκριμένα, η δημιουργία και ανάθεση δικαιωμάτων πρόσβασης στους χρήστες και τις ομάδες χρηστών για κάθε νέο προσάρτημα λογισμικού έγινε χρησιμοποιώντας το μοντέλο ασφάλειας (Security and Authorization tool) του συστήματος. Η μετάβαση του ελέγχου της πληροφορίας από το ένα κομμάτι στο άλλο υλοποιήθηκε χρησιμοποιώντας τον αποστολέα ενεργειών (Action Dispatcher) ενώ η σύνδεση με τη βάση δεδομένων (αποθήκευση και ανάκτηση πληροφορίας) πραγματοποιήθηκε από τη διεπαφή σύνδεσης (Data Connectivity). Τέλος, η παρουσίαση των δεδομένων πραγματοποιήθηκε με τη χρήση των υπαρχόντων προτύπων (templates, stylesheets) αλλά και τη δημιουργία καινούριων. Το κομμάτι «Διαχείριση προτύπων» (template handler) αναλαμβάνει τη διασύνδεση των προσαρτημάτων λογισμικού με τα πρότυπα.

Προσαρτήματα λογισμικού και τμήματα πληροφορίας

Στα πλαίσια αυτής της εργασίας αναπτύχθηκε μια σειρά από προσαρτήματα λογισμικού τα οποία υποστηρίζουν τη λειτουργικότητα του υποσυστήματος κατηγοριοποίησης αλλά και κάποια τμήματα πληροφορίας. Συγκεκριμένα τα προσαρτήματα λογισμικού που υλοποιήθηκαν (Σχήμα 28) αφορούν την διαχείριση του περιεχομένου που προέρχεται από το εξωτερικό περιβάλλον του συστήματος, την διαδικασία της κατηγοριοποίησης καθώς και την εξαγωγή στατιστικών στοιχείων για τα αποτελέσματα της κατηγοριοποίησης. Αντίστοιχα τα τμήματα πληροφορίας αναλαμβάνουν την παρουσίαση της πληροφορίας στις σελίδες του δικτυακού τόπου (η πληροφορία αφορά τόσο το αποτέλεσμα της κατηγοριοποίησης όσο και την παρουσίαση της πληροφορίας των προσαρτημάτων λογισμικού RSS και GoogleAPI). Η λειτουργικότητα των προσαρτημάτων λογισμικού αναλύεται στην ενότητα 5.5.

Συμπερασματικά, το υποσύστημα κατηγοριοποίησης προσέδωσε στο σύστημα ATL CME μηχανισμούς οι οποίοι επιτρέπουν τη συλλογή επιπρόσθετων μορφών περιεχομένου με την επιπλέον δυνατότητα της κατηγοριοποίησης τόσο του υπάρχοντος όσο και του καινούριου περιεχομένου.

5.3.2 Συλλογή περιεχομένου

Το περιεχόμενο που κατηγοριοποιείται μέσω του υποσυστήματος κατηγοριοποίησης αφορά νέα, άρθρα, συνδέσμους, κείμενα καθώς και τα αποτελέσματα της αναζήτησης που επιστρέφονται από

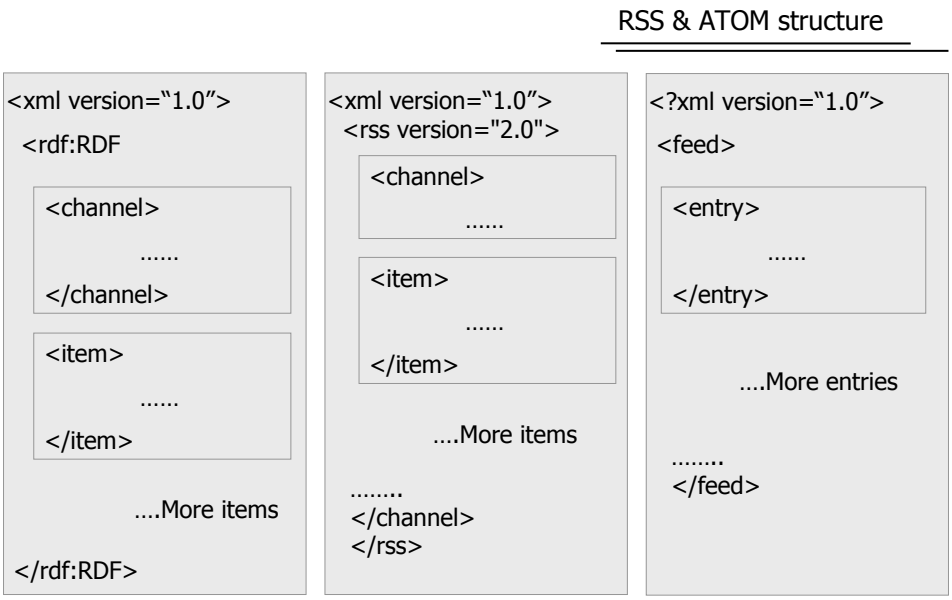
το Google. Η συλλογή του περιεχομένου πραγματοποιείται με δύο τρόπους (όταν πρόκειται για εξωτερικό περιεχόμενο): RSS & ATOM και GoogleAPI.

5.3.2.1 RSS & ATOM

Το προσάρτημα λογισμικού RSS έχει ως κύριο στόχο την συλλογή περιεχομένου από δικτυακούς τόπους που παρέχουν το υλικό τους μέσω των προτύπων RSS και ATOM. Οι εκδόσεις που είναι διαθέσιμες για τα πρότυπα αυτά είναι οι: RSS 0.91, RSS 0.92, RSS 2.0, RSS 1.0 και ATOM 0.3, ATOM 0.5. Στις παραπάνω εκδόσεις προστίθενται επιπλέον υλοποιήσεις (modules) που επιτρέπουν τη δημοσίευση πρόσθετων ειδών περιεχομένου (βιβλία, εκδηλώσεις, ταξινομίες). Τα "modules" διακρίνονται σε δύο είδη: τα προτεινόμενα (proposed) και τα καθιερωμένα (standard). Ο διαχωρισμός αυτός υποδεικνύει εάν έχουν υιοθετηθεί από την ομάδα εργασίας του προτύπου ή όχι.

Στο Σχήμα 29 παρουσιάζεται η δομή που ακολουθούν οι πιο συχνά χρησιμοποιούμενες εκδόσεις των παραπάνω προτύπων. Όπως φαίνεται πρόκειται για XML / RDF αρχεία το καθένα από τα οποία μπορεί να περιέχει συγκεκριμένα στοιχεία. Το περιεχόμενο που εξάγεται με RSS 1.0 είναι ένα RDF αρχείο που αποτελείται από δύο μεγάλα μέρη: το channel και το item. Το channel παρέχει γενικές πληροφορίες για τον τίτλο και το περιεχόμενο του αρχείου. Το στοιχείο item περιγράφει ένα νέο (τίτλο, περιγραφή, σύνδεσμο). Ανάλογα συντάσσονται και τα άλλα αρχεία. Στο πρότυπο ATOM το στοιχείο channel αντικαθίσταται από το feed ενώ η περιγραφή του νέου γίνεται χρησιμοποιώντας το στοιχείο entry αντί για το item[26].

[26] Σύγκριση των προτύπων RSS 2.0 και ATOM 1.0 γίνεται στην ακόλουθη σελίδα: http://www.tbray.org/atom/RSS-and-Atom ενώ παραδείγματα αρχείων που ακολουθούν τα πρότυπα RSS 0.91 και 0.92 υπάρχουν στη σελίδα, http://blogs.law.harvard.edu/tech/rss#sampleFiles.



Σχήμα 29 Η δομή των αρχείων που ακολουθούν τα πρότυπα RSS 1.0 & 2.0 και ATOM 0.3

Η συλλογή περιεχομένου από δικτυακούς τόπους που προσφέρουν το υλικό τους μέσω RSS / ATOM ενεργοποιεί μια διαδικασία φόρτωσης του συγκεκριμένου αρχείου, επεξεργασίας, αποθήκευσης και παρουσίασης στο δικτυακό τόπο. Αναλυτικά τα βήματα περιγράφονται στο παρακάτω (Σχήμα 31).

Βήμα 1: Επιλογή πηγής και μεταφορά του περιεχομένου στο δικτυακό τόπο

Ο χρήστης του συστήματος είναι υπεύθυνος για την επιλογή της πηγής παρέχοντας το σύνδεσμό της. Η μεταφορά των περιεχομένων της σελίδας στο δικτυακό τόπο πραγματοποιείται χρησιμοποιώντας μια αίτηση μέσω του πρωτοκόλλου HTTP. Εάν το περιεχόμενο που μεταφέρεται είναι κρυπτογραφημένο μέσω του πρωτοκόλλου SSL[27] χρησιμοποιείται η βιβλιοθήκη CURL[28] της γλώσσας PHP που παρέχει έναν τρόπο σύνδεσης και επικοινωνίας με τη συγκεκριμένη σελίδα. Εάν στο μηχανισμό έχει ενεργοποιηθεί η επιλογή κρυφής μνήμης (cache) τότε ελέγχεται εάν η συγκεκριμένη σελίδα έχει μεταφερθεί ξανά μέσα στο χρονικό διάστημα που έχει οριστεί στην μνήμη. Εάν έχει πραγματοποιηθεί, τότε τα περιεχόμενα της σελίδας δεν ανανεώνονται. Με τον τρόπο αυτό μειώνεται ο χρόνος μεταφοράς ενώ αποφεύγεται η ύπαρξη πολλών αντιγράφων του ίδιου αρχείου. Εφόσον η διαδικασία ολοκληρωθεί γίνεται η επεξεργασία του περιεχομένου της σελίδας καθώς αναλύεται (parsed) (βήμα 2).

[27] Ορισμός για το πρωτόκολλο SSL, http://www.webopedia.com/TERM/S/SSL.html

[28] Η βιβλιοθήκη CURL, http://gr.php.net/curl

Βήμα 2: Ανάλυση του περιεχομένου της σελίδας

Αντικείμενο αυτού του βήματος είναι η ανάλυση του περιεχομένου της σελίδας που μεταφέρθηκε προτού αποθηκευθεί στη βάση δεδομένων. Καταρχήν γίνεται αναγνώριση του προτύπου που χρησιμοποιείται. Ανάλογα με το πρότυπο ενεργοποιείται η κατάλληλη συνάρτηση για την ανάλυση του αρχείου. Στο σημείο αυτό γίνεται αναγνώριση και της κωδικοποίησης του αρχείου. Εάν η κωδικοποίηση του περιεχομένου διαφέρει από αυτή που χρησιμοποιεί ο δικτυακός τόπος τότε γίνεται η μετατροπή στην κωδικοποίηση του δικτυακού τόπου ώστε να εξασφαλιστεί η ορθή παρουσίαση του περιεχομένου. Η ανάλυση του αρχείου γίνεται χρησιμοποιώντας το μοντέλο SAX [29] δεδομένου ότι το περιεχόμενο που αναλύεται είναι δομημένο και δεν απαιτείται αποθήκευση του XML αρχείου[30]. Το μοντέλο αυτό επιτρέπει την ανάλυση ενός αρχείου XML ανεξαρτήτως μεγέθους αφού επεξεργάζεται τις ετικέτες (tags) μία προς μία και δεν απαιτεί τη φόρτωση ολόκληρου του αρχείου στη μνήμη. Ουσιαστικά η ανάλυση ξεκινά και όταν βρεθεί μια ετικέτα καλείται η κατάλληλη συνάρτηση που υποδεικνύει το χειρισμό της ετικέτας. Στη συγκεκριμένη περίπτωση η ανάλυση ενός αρχείου RSS/ATOM γίνεται χρησιμοποιώντας μια στοίβα στην οποία αποθηκεύονται τόσο το στοιχείο που αναλύεται όσο και το περιεχόμενό του. Έτσι για κάθε αρχείο δημιουργείται μια δενδρική δομή.

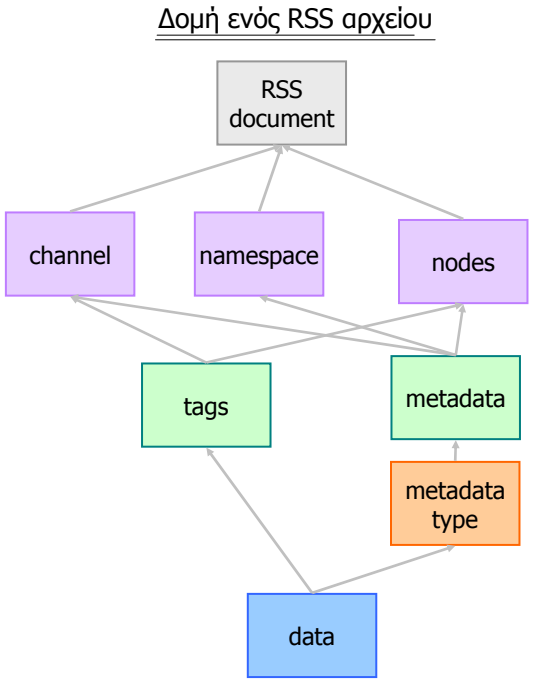
Βήμα 3: Αντιστοίχηση της δομής του αρχείου στο σχήμα της σχεσιακής βάσης δεδομένων

Η διαδικασία που περιγράφεται παραπάνω παράγει μια δενδρική δομή για κάθε αρχείο που αναλύεται. Η αποθήκευση στη βάση προϋποθέτει την μετατροπή αυτής της δομή στο σχήμα της σχεσιακής βάσης που χρησιμοποιείται στο δικτυακό τόπο. Μέχρι σήμερα έχουν προταθεί αρκετοί τρόποι προκειμένου να αποθηκευθούν XML δεδομένα σε μια σχεσιακή βάση ([43],[44],[45],[46],[47],[48]). Κοινό χαρακτηριστικό σε όλες τις προσεγγίσεις είναι η ανάλυση του αρχείου σε μικρά κομμάτια και η εύρεση χαρακτηριστικών (κανόνες, μοντέλα) τα οποία θα περιγράφουν αυτά τα κομμάτια. Κάθε κομμάτι αποθηκεύεται σε μια γραμμή στον αντίστοιχο πίνακα της βάσης δεδομένων. Η προσέγγιση που ακολουθείται στην παρούσα εργασία περιλαμβάνει την κατανομή της πληροφορίας σε διαφορετικούς πίνακες χρησιμοποιώντας αναφορές μεταξύ των πινάκων. Η δενδρική δομή του αρχείου αναπαρίσταται χρησιμοποιώντας πίνακες για τα δεδομένα και δείκτες από τα δεδομένα προς τους πατρικούς κόμβους του δένδρου. Συγκεκριμένα, η δομή των πινάκων ακολουθεί την ιεραρχία των στοιχείων του XML αρχείου. Η

[29] XML Parser Functions, http://gr.php.net/xml

[30] Το μοντέλο SAX προτιμάται συνήθως όταν το αρχείο που αναλύεται περιέχει ένα συγκεκριμένο σχήμα, όταν δεν απαιτείται αποθήκευση του αρχείου στη μνήμη και όταν το XML αρχείο που παράγεται δεν θα χρησιμοποιηθεί μετά την ανάλυση.

σύνδεση τους παρουσιάζεται στο Σχήμα 30. Ένα αρχείο RSS αποτελείται από το στοιχείο κανάλι (πληροφορίες για το αρχείο), από τα namespaces που περιγράφουν τυχόν RSS modules που μπορεί να περιέχονται στο αρχείο και τους κόμβους. Οι κόμβοι αποτελούνται από ετικέτες (tags) και μεταδεδομένα. Τα μεταδεδομένα μπορεί να έχουν συγκεκριμένους τύπους. Οι ετικέτες και τα μεταδεδομένα περιγράφουν το περιεχόμενο του RSS αρχείου.



Σχήμα 30 Η δομή των πινάκων που περιγράφουν ένα RSS αρχείο

Βήμα 4: Αποθήκευση στη βάση

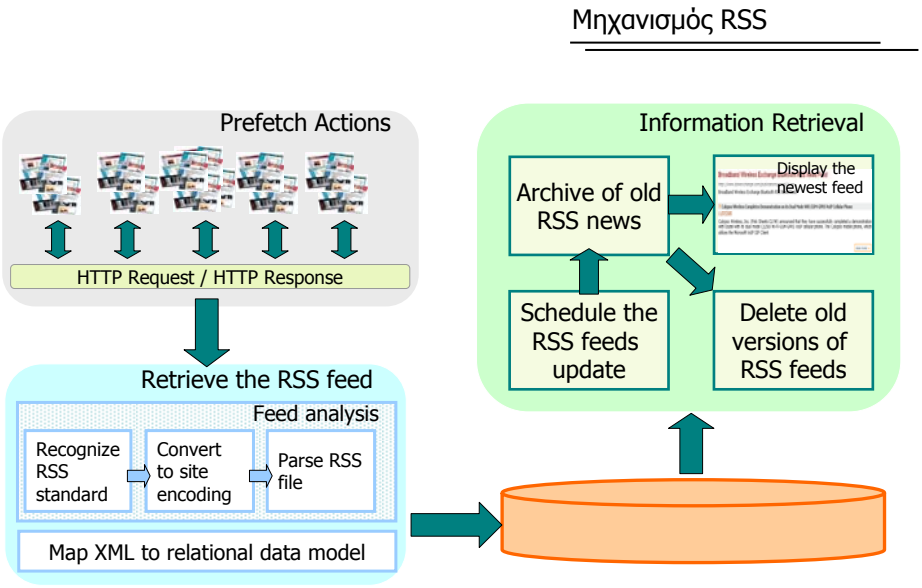
Εφόσον έχει γίνει η αναπαράσταση του δένδρου το περιεχόμενο αποθηκεύεται στη σχεσιακή βάση. Οι πίνακες που χρησιμοποιούνται ακολουθούν τη δομή του αρχείου. Για κάθε αρχείο αποθηκεύονται πληροφορίες για το ίδιο το αρχείο (σύνδεσμος, ημερομηνία, πρότυπο που ακολουθείται), για το στοιχείο "channel" (πρότυπο RSS) ή "entry" (πρότυπο ATOM). Στη συνέχεια κάθε νέο αποθηκεύεται σαν γραμμή σε έναν πίνακα ανάλογα με το στοιχείο (tag) στο οποίο ανήκει. Στην περίπτωση που υπάρχουν επιπλέον πληροφορίες για κάποιο νέο (στη μορφή μεταδεδομένων) αποθηκεύονται και αυτές σε έναν ειδικό πίνακα. Τα μεταδεδομένα προέρχονται συνήθως από τις επιπλέον λειτουργικότητες που έχουν αναπτυχθεί για τα πρότυπα RSS 1.0 και RSS 2.0 (Σχήμα 30). Περισσότερες πληροφορίες για τις λειτουργικότητες που υποστηρίζονται δίνονται στην ενότητα 5.5 του παρόντος.

Βήμα 5: Ανάκτηση των πληροφοριών

Εφόσον το αρχείο βρίσκεται αποθηκευμένο στη βάση δεδομένων μπορεί να παρουσιαστεί σε οποιαδήποτε σελίδα του δικτυακού τόπου. Η ανανέωση του αρχείου γίνεται είτε χειρωνακτικά στις στιγμές που επιθυμεί ο υπεύθυνος του δικτυακού τόπου είτε αυτόματα με τη δημιουργία ενός προγράμματος (scheduler). Εάν επιλεγεί ο δεύτερος τρόπος τότε το σύστημα αναλαμβάνει την μεταφορά των νέων εκδόσεων του αρχείου στο δικτυακό τόπο σε χρόνο που καθορίζει ο χρήστης. Στην περίπτωση που εμφανιστεί κάποιο λάθος κατά τη μεταφορά του αρχείου ο κωδικός του λάθους και μια μικρή περιγραφή αποθηκεύονται στη βάση δεδομένων προκειμένου να αποφασίσει ο χρήστης εάν επιθυμεί να συνεχίσει να παίρνει περιεχόμενο από την πηγή δεδομένων. Το συγκεκριμένο προσάρτημα λογισμικού κρατάει αρχείο των διαφορετικών εκδόσεων της πηγής ενώ υπάρχει η δυνατότητα διαγραφής κάποιων από αυτές με χειρωνακτικό ή αυτόματο τρόπο (μετά από συγκεκριμένο χρονικό διάστημα διαγράφονται όλες οι εκδόσεις του αρχείου).

Όπως έχει ήδη αναφερθεί το περιεχόμενο που συγκεντρώνεται μέσω RSS χρησιμοποιείται ως είσοδος στο υποσύστημα κατηγοριοποίησης. Συγκεκριμένα, είσοδο του υποσυστήματος αποτελεί μόνο η τελευταία έκδοση του αρχείου που υπάρχει διαθέσιμη στο δικτυακό τόπο και μάλιστα όχι το σύνολο του περιεχομένου που μπορεί να περιγράφει. Είσοδος είναι κάθε νέο χωριστά και μάλιστα συγκεκριμένα στοιχεία από κάθε νέο. Τα στοιχεία αυτά είναι:

- Τίτλος νέου
- Περιεχόμενο νέου
- Κατηγορία στην οποία ανήκει
- Θέμα στο οποίο ανήκει.



Σχήμα 31 Περιγραφή του μηχανισμού RSS

5.3.2.2 GoogleAPI

Ο δεύτερος μηχανισμός που έχει υλοποιηθεί για τη συλλογή του περιεχομένου εκμεταλλεύεται τη λειτουργικότητα της μηχανής αναζήτησης Google [15] χρησιμοποιώντας την υπηρεσία GoogleAPI. Ουσιαστικά το GoogleAPI είναι ένα πρόγραμμα που επιτρέπει την εύρεση και χρήση περιεχομένου από το διαδίκτυο [13]. Η διαδικασία που απαιτείται για τη χρήση αυτής της υπηρεσίας περιλαμβάνει τη δημιουργία ενός λογαριασμού στη μηχανή αναζήτησης Google, ο οποίος χρησιμοποιείται για την ταυτοποίηση του χρήστη της υπηρεσίας και η «φόρτωση» (download) του λογισμικού σε κάποιο μηχάνημα. Το όριο των αναζητήσεων περιορίζεται στις 1000 ανά ημέρα.

Τα Google Web APIs υλοποιούνται σαν υπηρεσία ιστού (web service). Η υπηρεσία υποστηρίζει μεθόδους σε SOAP[31] οι οποίες περιγράφονται σε ένα αρχείο WSDL[32]. Οι μέθοδοι αυτές μπορούν να χρησιμοποιηθούν όπως είναι σε ένα προγραμματιστικό περιβάλλον ή να δημιουργηθούν νέες. Η λειτουργικότητα της υπηρεσίας απαιτεί ακόμη τη συγγραφή προγραμμάτων που συνδέονται στην υπηρεσία Google Web APIs ανταλλάσσοντας μηνύματα γραμμένα σε SOAP.

Αναλυτικά τα βήματα της διαδικασίας είναι τα ακόλουθα:

Βήμα 1: Δημιουργία της αίτησης με τη λέξη – φράση προς αναζήτηση στο Google.com
Στο υποσύστημα κατηγοριοποίησης η διαδικασία ενεργοποιείται διατυπώνοντας μια λέξη – φράση προς αναζήτηση. Το πρόγραμμα που έχει υλοποιηθεί αναλαμβάνει τη σχηματοποίηση του μηνύματος SOAP με παραμέτρους που καθορίζονται από το Google. Η αίτηση αναζήτησης χρησιμοποιεί τη μέθοδο HTTP POST ακολουθούμενη από τις αντίστοιχες επικεφαλίδες ('Content-type':'text/xml', 'SOAPAction':'urn:GoogleSearchAction'). Η σύνδεση γίνεται στο api.google.com (Google cluster) στην πόρτα (port) 80 ενώ μια επιπλέον αίτηση γίνεται στον κατάλογο /search/beta2 περνώντας την αίτηση και τις επικεφαλίδες. Η απάντηση στην αίτηση αυτή είναι είτε ένα έγγραφο (απάντηση στην αναζήτηση) είτε ένας κωδικός λάθους (σε μορφή XML εγγράφου). Ένα παράδειγμα της αίτησης αλλά και της απάντησης που επιστρέφεται παρέχονται στο Σχήμα 32 ενώ μια εποπτική εικόνα της παραπάνω διαδικασίας παρουσιάζεται στο Σχήμα 33 (κομμάτι "BACK END").

[31] The SOAP (Simple Object Access Protocol) Specification, http://www.w3.org/TR/soap/

[32] The WSDL (Web Service Definition Language) Specification, http://www.w3.org/TR/wsdl

SOAP Request & Response

Σχήμα 32 Παράδειγμα αίτησης και της απάντησης που επιστρέφει το Google

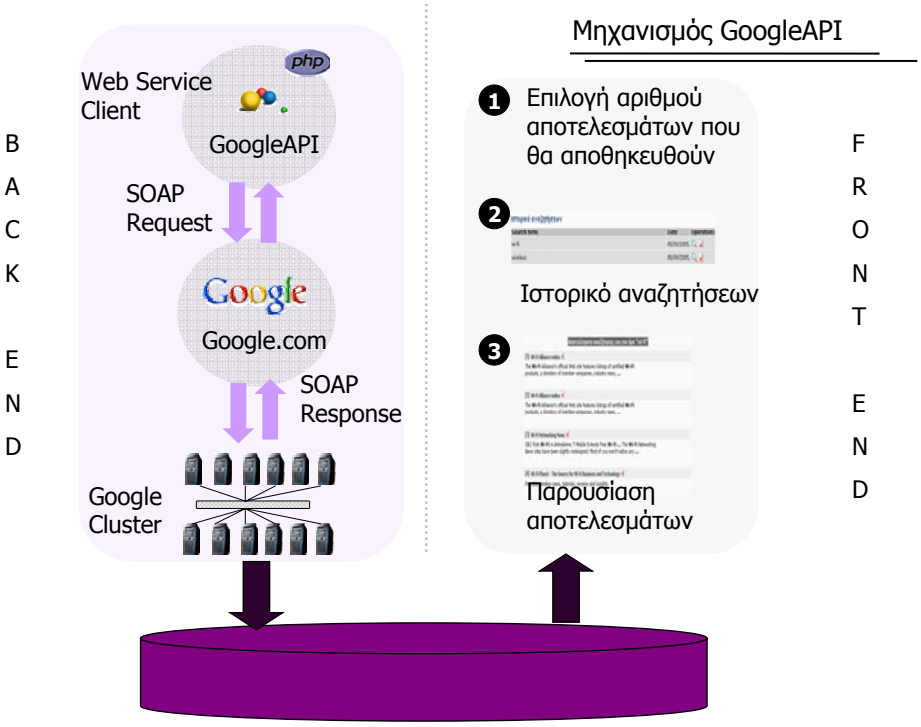
Βήμα 2: Μετατροπή του XML εγγράφου στο σχεσιακό σχήμα της βάσης δεδομένων

Όπως φαίνεται στο Σχήμα 32 η απάντηση στην αναζήτηση ακολουθεί τη δομή ενός XML εγγράφου. Για κάθε αποτέλεσμα οι πληροφορίες που επιστρέφονται αφορούν τον τίτλο, την περιγραφή, το σύνδεσμο καθώς και το μέγεθος της ιστοσελίδας. Τα στοιχεία που είναι χρήσιμα ως είσοδος για το υποσύστημα κατηγοριοποίησης και αποθηκεύονται είναι ο τίτλος και η περιγραφή (ο σύνδεσμος αποθηκεύεται στην περίπτωση που η λειτουργικότητα αυτή χρησιμοποιηθεί ανεξάρτητα από το υποσύστημα κατηγοριοποίησης).

Βήμα 3: Ανάκτηση της πληροφορίας

Η ανάκτηση της πληροφορίας αφορά το κομμάτι που αναφέρεται ως "FRONT END" στο Σχήμα 33. Οι δυνατότητες που παρέχονται αφορούν την παρουσίαση των αποτελεσμάτων που επιστρέφονται από μια λέξη – φράση προς αναζήτηση, την κράτηση ιστορικού των αναζητήσεων

καθώς και την επιλογή του αριθμού αποτελεσμάτων που θα αποθηκευθούν (το Google παρέχει σαν βάση την αποθήκευση 10 αποτελεσμάτων).



Σχήμα 33 Ο μηχανισμός GoogleAPI

5.3.2.3 Εσωτερικό Περιεχόμενο

Η τρίτη πηγή περιεχομένου αφορά περιεχόμενο που είναι ήδη αποθηκευμένο στη βάση δεδομένων του δικτυακού τόπου. Από τα διαθέσιμα προσαρτήματα λογισμικού του περιβάλλοντος υλοποίησης το περιεχόμενο που μπορεί να κατηγοριοποιηθεί αφορά άρθρα, ανακοινώσεις, κείμενα και συνδέσμους.

Η πληροφορία που αποθηκεύεται για τα κείμενα και τους συνδέσμους είναι:

- Τίτλος κειμένου ή συνδέσμου
- Σύνδεσμος ή αρχείο
- Τύπος (αρχείο ή σύνδεσμος / κατάλογος αρχείων ή συνδέσμων)
- Περιγραφή
- Βάρος (χρησιμοποιείται για τον καθορισμό της σειράς εμφάνισης του συνδέσμου ή του κειμένου)

Τα διαθέσιμα πεδία για τις ανακοινώσεις και τα άρθρα είναι:

- Τίτλος ανακοίνωσης / άρθρου
- Περιγραφή
- Πηγή
- Ημερομηνία εισαγωγής
- Βάρος (χρησιμοποιείται για τον καθορισμό της σειράς εμφάνισης της ανακοίνωσης ή του άρθρου)
- Τύπος (άρθρο / ανακοίνωση ή κατάλογος)
- Σύνδεσμος εφόσον πρόκειται για εξωτερική πηγή
- Αρχείο
- Ημερομηνία που θα σταματήσει να εμφανίζεται στο κοινό ένα άρθρο / ανακοίνωση.

Από όλα τα παραπάνω πεδία αυτά που χρειάζονται για την κατηγοριοποίηση είναι ο τίτλος και η περιγραφή για κάθε πόρο. Η ανάκτηση γίνεται μέσω μερικών απλών επερωτήσεων στη βάση δεδομένων του δικτυακού τόπου.

5.3.3 Επεξεργασία της ταξινομίας του DMOZ

Όπως αναφέρθηκε στην ενότητα 4.4.4 η κατηγοριοποίηση στο υποσύστημα κατηγοριοποίησης πραγματοποιείται με βάση την ιεραρχία των κατηγοριών του καταλόγου DMOZ. Το μεγάλο μέγεθος του καταλόγου καθιστά σχεδόν αδύνατη την απευθείας αναζήτηση των κατηγοριών στο δικτυακό τόπο του DMOZ. Η διαδικασία που περιγράφεται στην παρούσα ενότητα έχει ως στόχο την αποθήκευση ενός μέρους του καταλόγου στη βάση δεδομένων του δικτυακού τόπου.

Όπως φαίνεται στα ακόλουθα σχήματα (Σχήμα 34, Σχήμα 35) η διαδικασία χωρίζεται σε δύο μέρη. Το πρώτο αφορά την μεταφορά και αποθήκευση του καταλόγου στο δικτυακό τόπο και το δεύτερο την επεξεργασία του δέντρου προκειμένου να πραγματοποιείται γρήγορα η πρόσβαση και αναζήτηση σ' αυτό.

Βήμα 1: Φόρτωση της δομής της ιεραρχίας του καταλόγου DMOZ

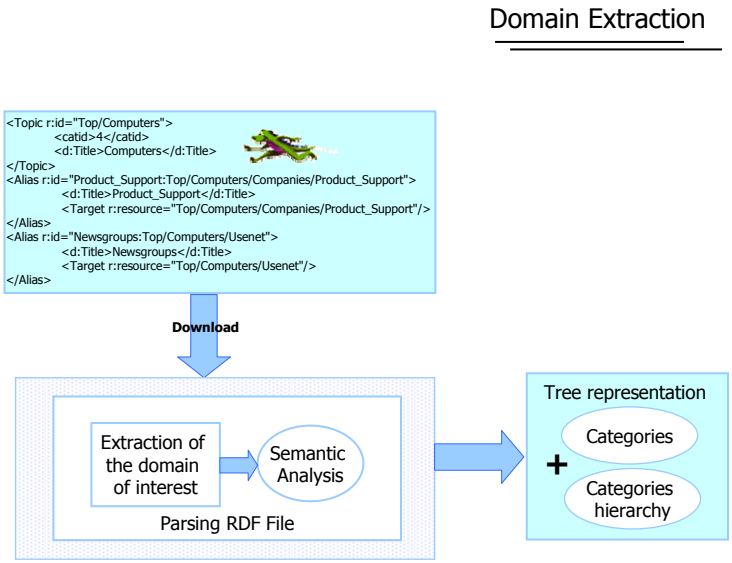
Η δομή της ιεραρχίας του DMOZ παρέχεται με τη μορφή RDF[33] αρχείου και περιγράφει περίπου 590.000 κατηγορίες. Το αρχείο αυτό χωρίζεται σε πολλά μέρη κάθε ένα από τα οποία περιγράφει μια συγκεκριμένη περιοχή γνώσης (εδώ ως βασικές περιοχές γνώσεις θεωρούνται όλες οι

[33] RDF (Resource Description Framework) specification, http://www.w3.org/RDF/

κατηγορίες που βρίσκονται στο πρώτο επίπεδο της ιεραρχίας). Η διαδικασία αυτοματοποιήθηκε ως ένα βαθμό δημιουργώντας ένα πρόγραμμα με είσοδο το αρχείο του DMOZ, το όνομα της κατηγορίας που εξάγεται καθώς και το όνομα του αρχείου στο οποίο θα αποθηκευθεί. Το αρχείο που χρησιμοποιείται για τον εκάστοτε δικτυακό τόπο σχετίζεται άμεσα με το θέμα που περιγράφει αυτός (π.χ. υπολογιστές, τέχνη, παιχνίδια, υγεία κτλ.).

Βήμα 2: Ανάλυση του αρχείου RDF

Το RDF αρχείο που παράγεται περιέχει αρκετές πληροφορίες σχετικά με τη δομή του αρχείου, τις συνδέσεις μεταξύ κατηγοριών διαφορετικών θεμάτων που μπορούν να αγνοηθούν[34]. Για το λόγο αυτό δημιουργήθηκε ένας συγκεκριμένος αναλυτής RDF (RDF parser) ο οποίος επεξεργάζεται το αρχείο και εξάγει μόνο την απαραίτητη πληροφορία. Για κάθε κατηγορία αποθηκεύονται τέσσερα στοιχεία: τίτλος, όνομα με το οποίο περιγράφεται, το μονοπάτι στο δέντρο των κατηγοριών καθώς και το αναγνωριστικό (id) που περιγράφει την πατρική κατηγορία.

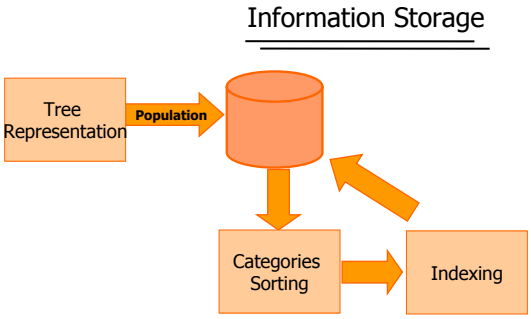


Σχήμα 34 Εξαγωγή θεματικής κατηγορίας από τον κατάλογο του DMOZ

Βήμα 3: Αποθήκευση και επεξεργασία της ιεραρχίας

[34] Μια κατηγορία που περιέχει υποκατηγορίες συνήθως παρέχει συνδέσμους προς αυτές. Αυτοί οι σύνδεσμοι μπορούν να αγνοηθούν εφόσον η αναλυτική περιγραφή των υποκατηγοριών δίνεται σε άλλο μέρος του αρχείου (με αυτό τον τρόπο αποφεύγεται επανάληψη της ίδιας πληροφορίας). Επίσης εάν μια κατηγορία ανήκει σε περισσότερους από έναν κλάδους του υποδένδρου τότε στον έναν από αυτούς παρέχεται σύνδεσμος προς την αναλυτική περιγραφή της. Εφόσον όμως στο συγκεκριμένο δικτυακό τόπο επιλέγεται μια συγκεκριμένη περιοχή γνώσης δεν έχει νόημα να αποθηκευθεί ο σύνδεσμος.

Η φάση αυτή περιλαμβάνει κάποιες ενέργειες για γρηγορότερη πρόσβαση και ανάκτηση πληροφοριών από την ιεραρχία. Το πρώτο βήμα αφορά τη διαγραφή διπλών εγγραφών στον κατάλογο[35]. Στη συνέχεια γίνεται ταξινόμηση και ευρετηριασμός της ιεραρχίας κατηγοριών. Η ταξινόμηση περιλαμβάνει την αποθήκευση των κατηγοριών με αλφαβητική σειρά ενώ ο ευρετηριασμός την χαρτογράφηση των γραμμάτων της αλφαβήτου με βάση τη θέση των κατηγοριών στον πίνακα.



Σχήμα 35 Αποθήκευση και επεξεργασία της ιεραρχίας κατηγοριών του DMOZ

5.3.4 Κατηγοριοποίηση

Η διαδικασία που ακολουθείται για την κατηγοριοποίηση του κειμένου παρουσιάζεται στο Σχήμα 36. Υπάρχουν δύο βασικά μέρη στη διαδικασία κατηγοριοποίησης: η εξαγωγή ενός λεξικού (ένα σύνολο με τις σημαντικότερες λέξεις) που περιγράφει το κείμενο που θα κατηγοριοποιηθεί και η εφαρμογή του αλγορίθμου κατηγοριοποίησης.

Εξαγωγή λεξικού

Η εξαγωγή του λεξικού από ένα κείμενο προϋποθέτει την εφαρμογή των αλγορίθμων που περιγράφηκαν στην ενότητα 4.3. Ως τμήμα κειμένου θεωρείται το σύνολο των λέξεων που εξάγεται από κάποια από τις διαθέσιμες πηγές περιεχομένου (Πίνακας 14).

---

[35] Μερικά παραδείγματα κατηγοριών που περιγράφονται από το ίδιο όνομα αλλά είναι διαφορετικές (διαφορετικό μονοπάτι) είναι οι παρακάτω:
Top/Computers/Programming/Languages/Java/User_Groups/Asia,
Top/Computers/Programming/Languages/Perl/User_Groups/Asia,
Top/Computers/Software/Operating_Systems/Linux/User_Groups/Asia
Σε αυτή την περίπτωση στη βάση δεδομένων αποθηκεύεται ένα σύνθετο όνομα που περιλαμβάνει τόσο τον τίτλο της κατηγορίας όσο και ένα κομμάτι από το μονοπάτι που ακολουθείται μέσα στο δέντρο. Με αυτό τον τρόπο εξασφαλίζεται η μοναδικότητα των κατηγοριών.

Πηγή περιεχομένου	Τμήματα κειμένου
RSS/ATOM	Κάθε νέο που περιέχεται στο αντίστοιχο αρχείο (τίτλος, περιγραφή, κατηγορία)
GoogleAPI	Σύνολο λέξεων που αποτελείται από τον τίτλο και την περιγραφή για κάθε δικτυακό τόπο που περιέχεται στα αποτελέσματα
Άρθρα	Τίτλος και περιγραφή άρθρου
Σύνδεσμοι	Τίτλος και περιγραφή συνδέσμου
Αρχεία	Τίτλος και περιγραφή αρχείου

Πίνακας 14 Πηγές περιεχομένου και τα τμήματα κειμένου (text segments) που παρέχουν στο υποσύστημα κατηγοριοποίησης

Το κείμενο που εξάγεται από τις πηγές περιεχομένου περιγράφεται με ένα σύνολο φράσεων. Το πρώτο βήμα περιλαμβάνει την αφαίρεση όλων των σημείων στίξης καθώς και τυχόν κώδικα html. Τα σημεία στίξης υπάρχουν αποθηκευμένα σε έναν πίνακα ενώ για την αφαίρεση του κώδικα HTML, χρησιμοποιείται συνάρτηση της PHP. Με το πέρας αυτών των ενεργειών για κάθε τμήμα κειμένου έχει δημιουργηθεί ένα σύνολο με τις λέξεις που περιέχει. Τα επόμενα βήματα περιέχουν ενέργειες προεπεξεργασίας του κειμένου όπως αυτές περιγράφηκαν στην ενότητα 4.3.2. Αυτές αφορούν τη μετατροπή των γραμμάτων των λέξεων σε πεζά, την αφαίρεση λέξεων που δεν προσφέρουν στη σημασιολογία του κειμένου και την εξαγωγή των ριζών των λέξεων.

Η μετατροπή των γραμμάτων των λέξεων σε πεζά εξασφαλίζεται από τη λειτουργικότητα της γλώσσας προγραμματισμού που χρησιμοποιείται. Η αφαίρεση των λέξεων που δεν προσφέρουν στη σημασιολογία του κειμένου πραγματοποιείται ελέγχοντας κάθε λέξη του συνόλου με τις λέξεις που υπάρχουν σε τρεις λίστες. Η πρώτη λίστα αποτελείται από τις 300 πιο συχνά χρησιμοποιούμενες λέξεις ενώ οι άλλες δύο από τα 330 πιο συχνά χρησιμοποιούμενα ομαλά και ανώμαλα ρήματα μαζί με το τρίτο πληθυντικό πρόσωπο, τον γερούνδιο, τον αόριστο και τον παρακείμενο τους. Στη συνέχεια για τις λέξεις που απομένουν εξάγονται οι ρίζες εφαρμόζοντας τον αλγόριθμο του Porter. Οι ρίζες των λέξεων αποτελούν και το λεξικό κάθε τμήματος κειμένου.

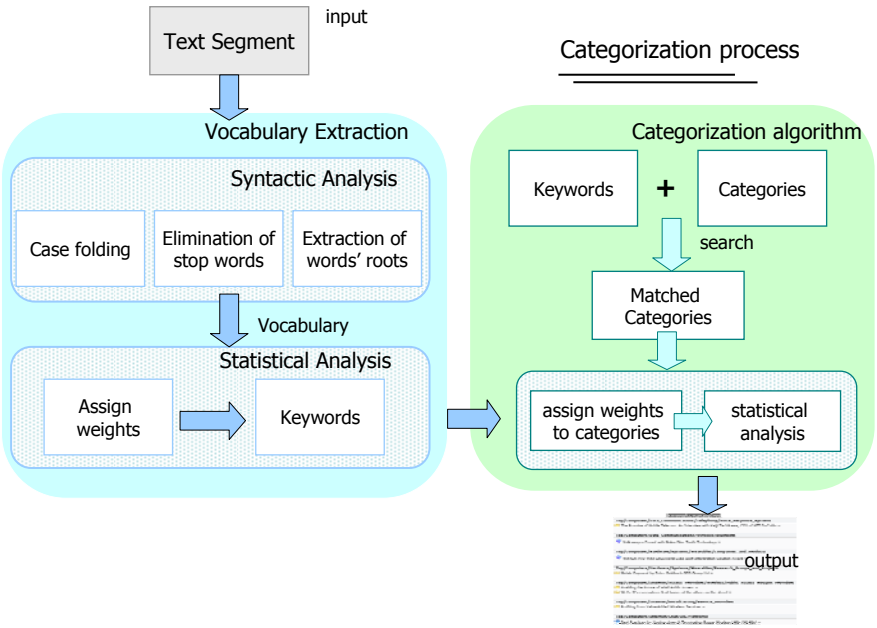
Οι διαδικασίες που περιγράφηκαν ουσιαστικά αποτελούν τη συντακτική ανάλυση του κειμένου. Το επόμενο βήμα είναι η στατιστική ανάλυση των ριζών των λέξεων που προέκυψαν. Η στατιστική ανάλυση περιλαμβάνει τον έλεγχο των λέξεων και την εύρεση της συχνότητας εμφάνισής τους.

Κατηγοριοποίηση του περιεχομένου

Στο βήμα αυτό είναι διαθέσιμο τόσο το λεξικό για κάθε τμήμα κειμένου όσο και η ιεραρχία κατηγοριών του καταλόγου που θα χρησιμοποιηθεί. Η διαδικασία της κατηγοριοποίησης ολοκληρώνεται με την εφαρμογή του αλγορίθμου που περιγράφηκε στην ενότητα 4.4.4.2. Ουσιαστικά, η εύρεση της πιο σχετικής κατηγορίας έγκειται στην αναζήτηση των λέξεων του λεξικού μέσα στην ιεραρχία και στην απόδοση βαρών στις κατηγορίες που επιστρέφονται ως σχετικές. Για να μειωθεί ο χρόνος αναζήτησης ο κατάλογος ευρετηριάζεται και η αναζήτηση πραγματοποιείται σειριακά μόνο σε ένα τμήμα της ιεραρχίας χρησιμοποιώντας την αντίστοιχη συνάρτηση της PHP. Το βάρος μιας κατηγορίας σχετίζεται με την συχνότητα εμφάνισης της λέξης στην οποία αντιστοιχείται. Παράλληλα, στην περίπτωση που ο αλγόριθμος επιστρέψει αρκετές κατηγορίες ως σχετικές έχει θεσπιστεί ευρεστικά μια σταθερά η οποία αποτελεί ένα κάτω όριο του βάρους των κατηγοριών. Εάν το βάρος μιας κατηγορίας είναι μεγαλύτερο από αυτή τη σταθερά τότε η κατηγορία επιστρέφεται ως η πιο σχετική. Εάν δε βρεθεί η πιο σχετική κατηγορία ακολουθώντας την παραπάνω διαδικασία η αναζήτηση πραγματοποιείται στην ιεραρχία ανά επίπεδο. Με τον τρόπο αυτό αποφεύγονται επιπλέον αναζητήσεις, εφόσον βρεθεί σχετική κατηγορία σε κάποιο επίπεδο.

Εμφάνιση των τελικών κατηγοριών

Η τελική φάση της κατηγοριοποίησης αφορά την παρουσίαση των κατηγοριών στις οποίες αντιστοιχήθηκαν τα κείμενα. Η παρουσίαση αυτή γίνεται είτε ως απλή σελίδα στο δικτυακό τόπο είτε μέσω του προτύπου RSS. Η εξαγωγή της κατηγοριοποίησης με RSS πραγματοποιείται προκειμένου να μπορεί να χρησιμοποιηθεί το υλικό και οι κατηγορίες από κάποιον άλλο δικτυακό τόπο.



Σχήμα 36 Διαδικασία κατηγοριοποίησης

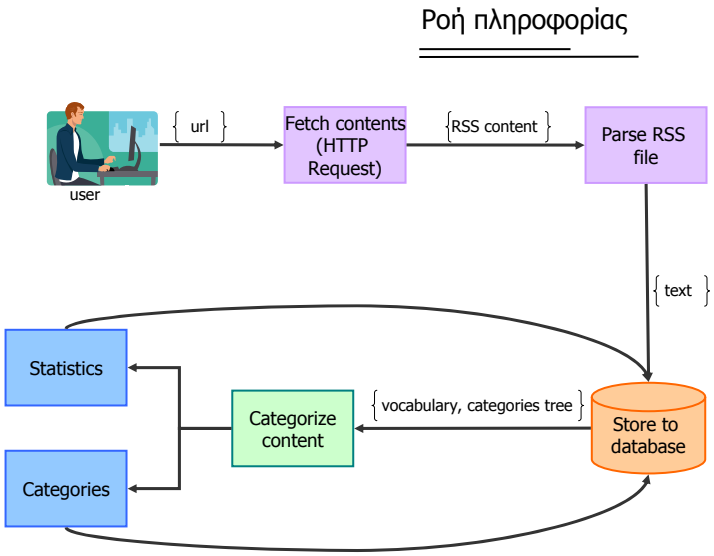
5.3.5 Στατιστικά

Τα στατιστικά που παρέχονται από το υποσύστημα κατηγοριοποίησης αφορούν το ποσοστό της κατηγοριοποίησης καθώς και ποιοτικά στοιχεία για την ακρίβεια της κατηγοριοποίησης. Συγκεκριμένα, για κάθε πηγή περιεχομένου παρέχεται ο αριθμός των τμημάτων κειμένου που είναι διαθέσιμα, ο αριθμός που τελικά κατηγοριοποιήθηκε καθώς και το ποσοστό της κατηγοριοποίησης. Κάποια επιπλέον στατιστικά παρέχονται στην περίπτωση που ο χρήστης έχει αντιστοιχήσει τα κείμενα σε κάποιες κατηγορίες οπότε έχει νόημα η σύγκριση της παραγόμενης κατηγοριοποίησης με αυτή του χρήστη.

5.4 Ροή της πληροφορίας στο σύστημα: από την συλλογή στην παρουσίαση

Για να αποκτήσουμε μια εποπτική ιδέα για τη ροή της πληροφορίας μεταξύ των διαφόρων τμημάτων του συστήματος, παρουσιάζουμε τη ροή της πληροφορίας για το σενάριο εισαγωγής περιεχομένου μέσω RSS και στη συνέχεια της κατηγοριοποίησης του. Ξεκινώντας από το σύνδεσμο στον οποίο υπάρχει η πηγή καταλήγουμε με το περιεχόμενο της πηγής κατηγοριοποιημένο. Τα βήματα που εκτελούνται είναι τα παρακάτω (Σχήμα 37):

- Ο χρήστης εισάγει το σύνδεσμο της RSS πηγής (url) από την οποία συλλέγεται το περιεχόμενο
- Ένα κομμάτι του συστήματος (fetcher) αναλαμβάνει τη μεταφορά του περιεχομένου της συγκεκριμένης σελίδας χρησιμοποιώντας το πρωτόκολλο HTTP
- Στο σύστημα μεταφέρεται ένα XML αρχείο το οποίο αναλύεται κατάλληλα από το κομμάτι του συστήματος RSS parser
- Η έξοδος της προηγούμενης διαδικασίας παράγει κείμενο με δομημένη μορφή το οποίο αποθηκεύεται στη βάση δεδομένων του δικτυακού τόπου
- Στη συνέχεια εφαρμόζεται ο αλγόριθμος κατηγοριοποίησης για το περιεχόμενο που έχει αποθηκευθεί. Την είσοδο αποτελεί ένα σύνολο λέξεων που εξάγεται από το κείμενο καθώς και η ιεραρχία κατηγοριών του DMOZ
- Το αποτέλεσμα της κατηγοριοποίησης είναι η αντιστοίχηση του υπάρχοντος περιεχομένου σε κάποια κατηγορία και η εξαγωγή στατιστικών που αφορούν το ποσοστό κατηγοριοποίησης.



Σχήμα 37 Ροή πληροφορίας

5.5 Λειτουργικότητες του συστήματος

Η λειτουργικότητα του υποσυστήματος κατηγοριοποίησης παρέχεται με μια σειρά προσαρτημάτων που έχουν υλοποιηθεί. Αυτά είναι:

- RSS
- GoogleAPI

- Metadata.

5.5.1 Προσάρτημα λογισμικού RSS

Είναι ένα από τα βασικά προσαρτήματα λογισμικού και αναλαμβάνει τη συλλογή του περιεχόμενου από δικτυακούς τόπους που το παρέχουν μέσω των προτύπων RSS και ATOM. Οι λειτουργίες που μπορούν να επιτελεστούν από το χρήστη είναι οι εξής:

5.5.1.1 Ανάθεση αδειών για τη διαχείριση του περιεχομένου

Ο διαχειριστής του δικτυακού τόπου είναι υπεύθυνος για την ανάθεση αδειών στους χρήστες και τα γκρουπ τους. Υπάρχουν τρία είδη αδειών:

- RSS_ADD: προσθήκη νέας πηγής περιεχομένου
- RSS_EDIT: ανανέωση των περιεχομένων της πηγής
- RSS_DELETE: διαγραφή της πηγής.

Ανάλογα με τις άδειες που αποδίδονται, οι χρήστες μπορούν να εκτελέσουν και τις αντίστοιχες ενέργειες.

5.5.1.2 Διαχείριση νέου module

Το πρότυπο RSS έχει αρκετές εκδόσεις οι οποίες χρησιμοποιούνται από δικτυακούς τόπους για την παροχή του υλικού τους. Ειδικά στις εκδόσεις 1.0 και 2.0 έχουν προταθεί αρκετές υλοποιήσεις με στόχο την περιγραφή και άλλων ειδών περιεχομένου εκτός από τα νέα. Οι υλοποιήσεις αυτές αναφέρονται ως "modules". Υπάρχουν τρία προτυποποιημένα (standard) modules και κάποια ακόμη που έχουν προταθεί [36] για αποδοχή. Από αυτά το παρόν προσάρτημα λογισμικού υποστηρίζει την πρώτη κατηγορία και κάποια από τη δεύτερη. Συγκεκριμένα αυτά είναι:

- Dublin Core: Στην υλοποίηση αυτή υιοθετούνται στοιχεία που υπάρχουν στο πρότυπο μεταδεδομένων Dublin Core[37].
- Syndication: Παρέχει πληροφορίες σχετικά με τη συχνότητα ανανέωσης του περιεχομένου ενός RSS αρχείου. Επίσης υιοθετεί στοιχεία από το πρότυπο «Open Content Syndication (OCS)[38]».

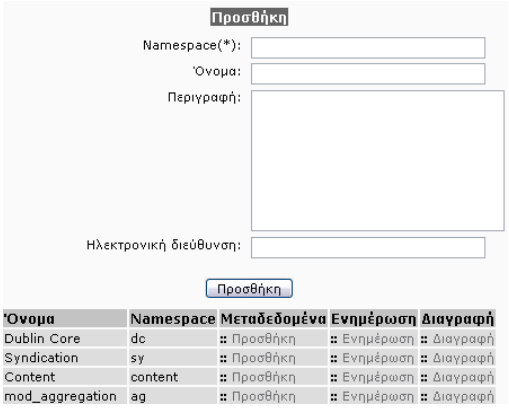
[36] Η λίστα με τα RSS 1.0 modules που έχουν προταθεί, http://web.resource.org/rss/1.0/modules/proposed.html

[37] Dublin Core Metadata Initiative, http://dublincore.org/

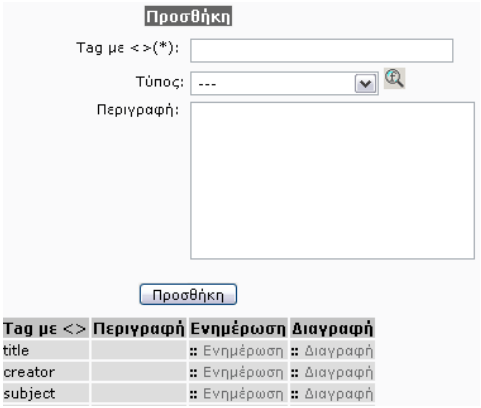
[38] Open Content Syndication Directory Format, http://internetalchemy.org/ocs/directory/0.5/

- Content: Υλοποίηση για την περιγραφή του περιεχομένου των δικτυακών τόπων σε διαφορετικά σχήματα (format). Πρόκειται για περιεχόμενο που μπορεί να περιέχει ενσωματωμένες φωτογραφίες, κώδικα html κτλ.
- mod_aggregation: Παρέχει πληροφορίες για την πρωτότυπη (original) πηγή του RSS νέου στην περίπτωση που γίνεται αναδημοσίευση περιεχομένου από διαφορετική πηγή.
- mod_annotation: Περιέχει στοιχεία για να την υποστήριξη πηγών που σχολιάζουν ή κάνουν αναφορές σε άλλες πηγές. Μερικά παραδείγματα περιλαμβάνουν συστήματα όπως το Usenet ή κάποιες συζητήσεις που γίνονται στο διαδίκτυο. Η υλοποίηση αυτή μπορεί να χρησιμοποιηθεί σε συνδυασμό με άλλα modules για να παρέχουν επιπλέον πληροφορίες για τις πηγές των νέων.
- mod_changedpage: Περιέχει πληροφορίες για την κοινοποίηση της ανανέωσης ενός RSS αρχείου.
- mod_dcterms: Εισάγει τα στοιχεία του προτύπου μεταδεδομένων «Qualified Dublin Core» για να χρησιμοποιηθούν σε συνδυασμό με τα μεταδεδομένα του προτύπου Dublin Core.
- mod_event: Παρέχει πληροφορίες για γεγονότα που θα συμβούν στο μέλλον κατηγοριοποιώντας τα με βάση αυτά που θα πραγματοποιηθούν στο κοντινό μέλλον.
- mod_prism: Το PRISM (Publishing Requirements for Industry Standard Metadata) είναι ένα πρότυπο για την ανταλλαγή και επαναχρησιμοποίηση του περιεχομένου που αφορά ηλεκτρονικές εκδόσεις (π.χ. βιβλία). Προτείνει μάλιστα τη χρήση υπαρχόντων προτύπων όπως τα XML, RDF, Dublin Core και διαφόρων άλλων που χρησιμοποιούνται για την περιγραφή τοποθεσιών γλωσσών και ημερομηνίας / χρόνου. Παράλληλα, εισάγει ένα μικρό αριθμό περιγραφών XML.
- mod_richequiv: Η υλοποίηση αυτή παρέχει τη δυνατότητα χρήσης XML κώδικα μέσα στις περιγραφές των νέων.
- mod_slash: Το Slash είναι ο κώδικας και η βάση δεδομένων που χρησιμοποιήθηκαν για τη δημιουργία του Slashdot. Χρησιμοποιείται για την περιγραφή κειμένου που προέρχεται από δικτυακούς τόπους που ακολουθούν το πρότυπο Slash.
- mod_taxonomy: Χρησιμοποιείται για την ταξινόμηση των νέων και των καναλιών που περιγράφονται σε ένα αρχείο RSS κάτω από όρους μιας ταξονομίας.

Κάθε ένα από τα παραπάνω modules εισάγει μια σειρά από μεταδεδομένα στα υπάρχοντα στοιχεία ενός RSS αρχείου. Το παρόν προσάρτημα παρέχει έναν μηχανισμό για την προσθήκη, ανανέωση και διαγραφή τόσο των modules όσο και των μεταδεδομένων τους (Σχήμα 38, Σχήμα 39).



Σχήμα 38 Προσθήκη και παρουσίαση στοιχείων για ένα module



Σχήμα 39 Προσθήκη και παρουσίαση στοιχείων για τα μεταδεδομένα ενός module

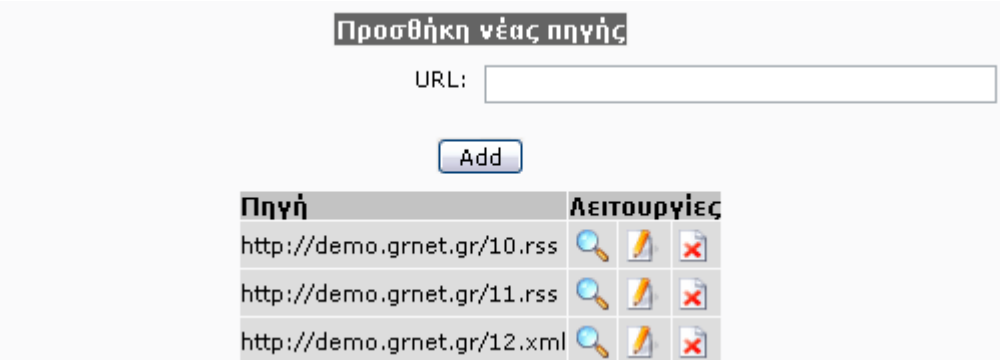
5.5.1.3 Δημιουργία προτύπου για την παρουσίαση των στοιχείων των RSS modules

Κάθε RSS module εισάγει μια σειρά από μεταδεδομένα τα οποία χρησιμοποιούνται στα στοιχεία ενός αρχείου RSS. Τα στοιχεία αυτά προσφέρουν επιπλέον πληροφορία για κάθε νέο ή κανάλι στα οποία εισάγονται. Για το λόγο αυτό έχει δημιουργηθεί ένα πρότυπο (template) που αναλαμβάνει την παρουσίαση των μεταδεδομένων ανάλογα με το στοιχείο στο οποίο χρησιμοποιούνται. Η μετάφραση στα ελληνικά παρέχεται από το χρήστη κατά την προσθήκη των μεταδεδομένων (πεδίο «Τύπος» στο Σχήμα 39).

5.5.1.4 Διαχείριση πηγών: Προσθήκη, ανανέωση, διαγραφή

Η λειτουργικότητα που χρησιμοποιείται συνήθως από τους χρήστες αφορά τη διαχείριση των πηγών συλλογής περιεχομένου. Το στοιχείο που χαρακτηρίζει κάθε πηγή είναι ο σύνδεσμός της (URL). Η προσθήκη γίνεται εύκολα με τη χρήση του πεδίου μιας φόρμας, όπως φαίνεται στο Σχήμα 40, ενώ από κάτω παρουσιάζεται η λίστα των πηγών που έχουν εισαχθεί μέχρι εκείνη τη στιγμή. Το προσάρτημα λογισμικού υποστηρίζει την αυτόματη εύρεση του συνδέσμου του RSS αρχείου για κάποιο δικτυακό τόπο εφόσον ο χρήστης γνωρίζει το σύνδεσμο του δικτυακού τόπου (RSS link autodiscovery). Για κάθε πηγή δίνονται οι ακόλουθες δυνατότητες (κάθε δυνατότητα αντιστοιχεί στα εικονίδια που φαίνονται δίπλα από κάθε πηγή):

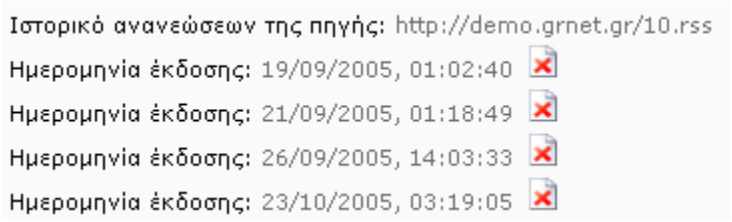
- Παρουσίαση του περιεχομένου της πηγής
- Ανανέωση του περιεχομένου της πηγής τη δεδομένη στιγμή
- Διαγραφή μιας πηγής.



Σχήμα 40 Προσθήκη νέας πηγής

5.5.1.5 Ιστορικό νέων (Διαχείριση εκδόσεων)

Για κάθε πηγή που προστίθεται στο προσάρτημα λογισμικού κρατείται ένα ιστορικό εκδόσεων ανάλογα με τον αριθμό των φορών που έχει γίνει ανανέωση του περιεχομένου της. Για κάθε έκδοση που κρατείται υπάρχει δυνατότητα διαγραφής ενώ εφόσον περάσει ένα εύλογο χρονικό διάστημα, που καθορίζει ο διαχειριστής του συστήματος, γίνεται αυτόματη διαγραφή των παλαιότερων εκδόσεων από το σύστημα.



Σχήμα 41 Παρουσίαση εκδόσεων για μια πηγή

5.5.1.6 Δημιουργία προγράμματος για την αυτόματη ανανέωση των πηγών

Εκτός από τη χειρωνακτική ανανέωση του περιεχομένου μιας πηγής παρέχεται επίσης η δυνατότητα αυτόματης ανανέωσης σε προκαθορισμένα διαστήματα. Η διαδικασία που ακολουθείται περιλαμβάνει την αποθήκευση του συνδέσμου των πηγών την πρώτη φορά που εισάγονται στο προσάρτημα λογισμικού. Στη συνέχεια χρησιμοποιώντας τη βιβλιοθήκη CURL της PHP και τη λειτουργικότητα crontab γίνεται ανανέωση του περιεχομένου όλων των πηγών ανά τακτά χρονικά διαστήματα που ορίζονται από το χρήστη.

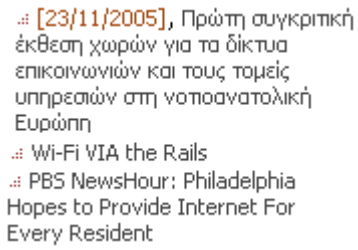
5.5.1.7 Διαχείριση λαθών

Πολλές φορές οι σελίδες που περιέχουν περιεχόμενο RSS επιστρέφουν λάθη (π.χ. HTTP 500 error, HTTP 404 error κτλ.). Στην περίπτωση που στο σύστημα υπάρχουν αποθηκευμένες παλαιότερες εκδόσεις του περιεχομένου παρουσιάζεται μια από αυτές. Αν η πηγή είναι καινούρια το λάθος αποθηκεύεται στο σύστημα και εφόσον τα λάθη επαναληφθούν εμφανίζεται στο χρήστη ένα μήνυμα στη σελίδα του προσαρτήματος λογισμικού.

5.5.1.8 Παρουσίαση περιεχομένου πηγής

Τέλος, παρέχεται η δυνατότητα παρουσίασης των νέων που περιέχει κάθε πηγή σε μια σελίδα. Για κάθε νέο παρουσιάζονται μια σειρά από τα βασικά στοιχεία (τίτλος, σύνδεσμος, περιγραφή) του καθώς και ορισμένα μεταδεδομένα που παρέχουν επιπλέον πληροφορίες (δημιουργός, πνευματικά δικαιώματα κτλ) με βάση το πρότυπο που έχει καταρτιστεί για τη συγκεκριμένη υλοποίηση (Σχήμα 42). Επίσης έχει δημιουργηθεί ένα τμήμα πληροφορίας όπου εισάγονται τα τελευταία νέα μίας ή περισσοτέρων πηγών (Σχήμα 43).

Σχήμα 42 Παρουσίαση των περιεχομένων ενός RSS αρχείου



Σχήμα 43 Τμήμα πληροφορίας με τα τρία τελευταία νέα από τρεις πηγές

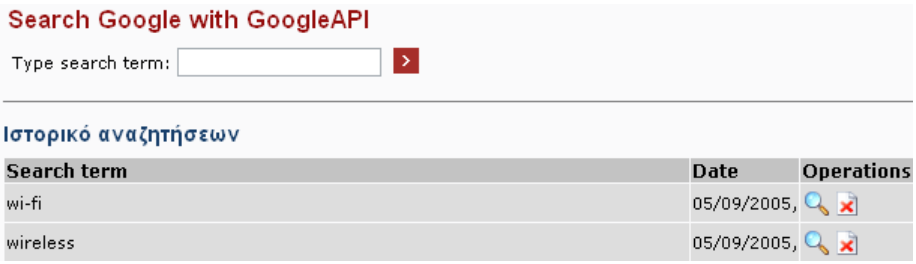
5.5.2 Προσάρτημα λογισμικού GoogleAPI

Το προσάρτημα λογισμικού GoogleAPI υλοποιεί τη λειτουργικότητα που προσφέρει η μηχανή αναζήτησης Google. Η βασική δυνατότητα που παρέχεται σε αυτό είναι η αναζήτηση λέξεων ή φράσεων στη βάση δεδομένων του Google.

Αναζήτηση λέξης – φράσης

Η αναζήτηση μιας λέξης – φράσης πραγματοποιείται με την εισαγωγή της στη φόρμα του ακόλουθου σχήματος (Σχήμα 44). Στην ίδια σελίδα φαίνεται και μια λίστα με τις αναζητήσεις που

έχουν πραγματοποιηθεί στο παρελθόν. Για κάθε αναζήτηση παρέχονται επίσης δυνατότητες παρουσίασης των αποτελεσμάτων για κάθε λέξη – φράση αλλά και διαγραφής αυτών.



Σχήμα 44 Αναζήτηση λέξης – φράσης με το GoogleAPI

Ιστορικό αναζητήσεων

Το προσάρτημα λογισμικού GoogleAPI παρέχει τη δυνατότητα αποθήκευσης των αποτελεσμάτων αναζητήσεων που γίνονται πολλές φορές κρατώντας ένα ιστορικό των αναζητήσεων. Ο χρήστης μπορεί να διαγράψει κάποια από τις εκδόσεις που υπάρχουν αποθηκευμένες.

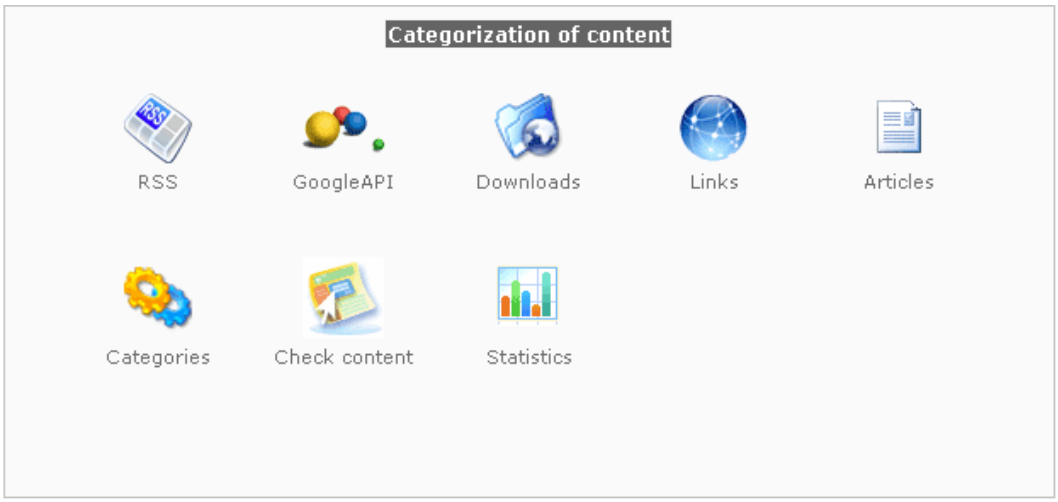
Διαχείριση αποτελεσμάτων αναζήτησης: Αποθήκευση, παρουσίαση και διαγραφή

Το GoogleAPI παρέχει τη δυνατότητα αποθήκευσης μόνο 10 αποτελεσμάτων ανά αναζήτηση. Ωστόσο για την παρούσα εργασία παραμετροποιήθηκε προκειμένου να αποθηκεύει όσα αποτελέσματα επιθυμεί ο χρήστης. Μάλιστα δίνεται η δυνατότητα διαγραφής των αποτελεσμάτων που ο χρήστης δεν επιθυμεί να αποθηκεύσει (Σχήμα 45). Τα αποτελέσματα παρουσιάζονται σε ένα τμήμα πληροφορίας ανάλογα με το θέμα και τον αριθμό που επιλέγει ο χρήστης.

Σχήμα 45 Παρουσίαση αποτελεσμάτων για τον όρο "wi-fi"

5.5.3 Προσάρτημα λογισμικού Metadata

Το προσάρτημα λογισμικού "Metadata" χρησιμοποιείται για την κατηγοριοποίηση του περιεχομένου που συλλέγεται αλλά και για την εξαγωγή στατιστικών στοιχείων για την τελική κατηγοριοποίηση (Σχήμα 46).



Σχήμα 46 Η πρώτη σελίδα του προσαρτήματος λογισμικού Metadata

5.5.3.1 Ανάθεση αδειών για τη διαχείριση του περιεχομένου

Οι άδειες που ανατίθενται για το συγκεκριμένο προσάρτημα λογισμικού αφορούν τη δυνατότητα των χρηστών να κατηγοριοποιούν χειρωνακτικά το περιεχόμενο και την πρόσβαση τους στο αποτέλεσμα της κατηγοριοποίησης. Εκτός από τις άδειες του ίδιου του προσαρτήματος λογισμικού κληρονομούνται οι άδειες των προσαρτημάτων λογισμικού που χρησιμοποιούνται για τη συλλογή του περιεχομένου. Έτσι ένας χρήστης που έχει πρόσβαση μόνο σε ένα συγκεκριμένο φάκελο με αρχεία θα μπορεί να κατηγοριοποιήσει μόνο το περιεχόμενο αυτού.

5.5.3.2 Παρουσίαση περιεχομένου που έχει συλλεχθεί

Ο χρήστης έχει τη δυνατότητα να ελέγξει το περιεχόμενο όλων των προσαρτημάτων που χρησιμοποιούνται για τη συλλογή. Επιλέγοντας για παράδειγμα το σύνδεσμο "Downloads" εμφανίζεται μια λίστα με όλα τα αρχεία στα οποία έχει πρόσβαση. Από εκεί υπάρχουν δύο επιλογές:

- Κατηγοριοποίηση του περιεχομένου χειρωνακτικά. Η επιλογή αυτή χρησιμοποιείται όταν το προσάρτημα λογισμικού της κατηγοριοποίησης δεν έχει ενεργοποιηθεί με τη δημιουργία του δικτυακού τόπου. Επιλέγοντας το εικονίδιο που φαίνεται στο Σχήμα 47

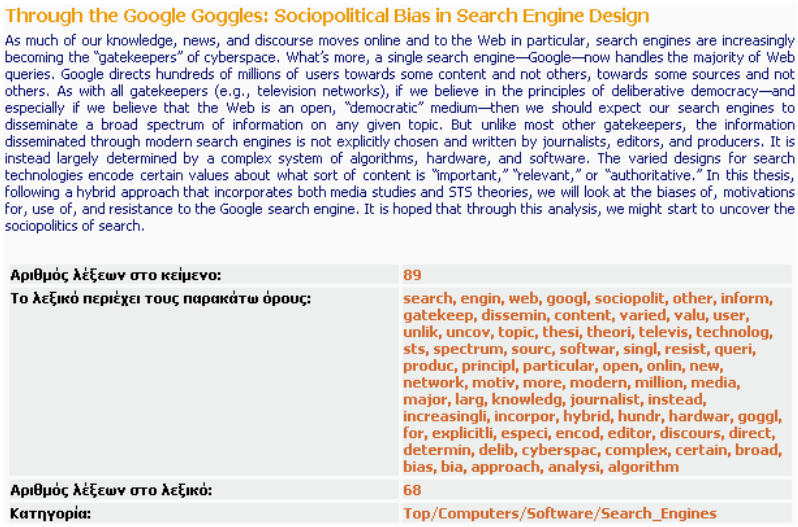
γίνεται κατηγοριοποίηση όλου του περιεχομένου που έχει εισαχθεί μέσω του μηχανισμού που έχει επιλεγεί (στο συγκεκριμένο παράδειγμα κατηγοριοποιούνται όλα τα κείμενα που έχουν εισαχθεί μέσω του μηχανισμού Downloads).

- Παροχή στοιχείων για ένα συγκεκριμένο τμήμα κειμένου (αρχείο). Αναλύεται στην ενότητα 5.5.3.3.

Σχήμα 47 Παρουσίαση περιεχομένου που είναι αποθηκευμένο στο μηχανισμό διαχείρισης αρχείων

5.5.3.3 Παροχή στοιχείων για κάθε τμήμα κειμένου που κατηγοριοποιείται

Για κάθε τμήμα κειμένου που εισάγεται προς κατηγοριοποίηση παρέχονται κάποια στοιχεία που δίνουν μια πλήρη εικόνα γι' αυτό. Τα στοιχεία που δίνονται είναι: α) τίτλος κειμένου, β) περιγραφή, γ) αριθμός λέξεων στο κείμενο, δ) οι ρίζες των λέξεων του κειμένου, ε) αριθμός των λέξεων που περιέχονται στο λεξικό καθώς και στ) η κατηγορία στην οποία κατηγοριοποιείται το κείμενο. Με αυτό τον τρόπο ο χρήστης έχει τη δυνατότητα να ελέγξει το αρχικό κείμενο καθώς και το αποτέλεσμα της κατηγοριοποίησης.


Through the Google Goggles: Sociopolitical Bias in Search Engine Design

As much of our knowledge, news, and discourse moves online and to the Web in particular, search engines are increasingly becoming the "gatekeepers" of cyberspace. What's more, a single search engine—Google—now handles the majority of Web queries. Google directs hundreds of millions of users towards some content and not others, towards some sources and not others. As with all gatekeepers (e.g., television networks), if we believe in the principles of deliberative democracy—and especially if we believe that the Web is an open, "democratic" medium—then we should expect our search engines to disseminate a broad spectrum of information on any given topic. But unlike most other gatekeepers, the information disseminated through modern search engines is not explicitly chosen and written by journalists, editors, and producers. It is instead largely determined by a complex system of algorithms, hardware, and software. The varied designs for search technologies encode certain values about what sort of content is "important," "relevant," or "authoritative." In this thesis, following a hybrid approach that incorporates both media studies and STS theories, we will look at the biases of, motivations for, use of, and resistance to the Google search engine. It is hoped that through this analysis, we might start to uncover the sociopolitics of search.

Αριθμός λέξεων στο κείμενο:	89
Το λεξικό περιέχει τους παρακάτω όρους:	search, engin, web, googl, sociopolit, other, inform, gatekeep, dissemin, content, varied, valu, user, unlik, uncov, topic, thesi, theori, televis, technolog, sts, spectrum, sourc, softwar, singl, resist, queri, produc, principl, particular, open, onlin, new, network, motiv, more, modern, million, media, major, larg, knowledg, journalist, instead, increasingli, incorpor, hybrid, hundr, hardwar, goggl, for, explicitli, especi, encod, editor, discours, direct, determin, delib, cyberspac, complex, certain, broad, bias, bia, approach, analysi, algorithm
Αριθμός λέξεων στο λεξικό:	68
Κατηγορία:	Top/Computers/Software/Search_Engines

Σχήμα 48 Στοιχεία για ένα τμήμα κειμένου

5.5.3.4 Παρουσίαση τελικών κατηγοριών

Με την ολοκλήρωση της διαδικασίας κατηγοριοποίησης τα κείμενα που είναι αποθηκευμένα αποδίδονται σε κάποια κατηγορία. Το υποσύστημα κατηγοριοποίησης παρέχει δύο τρόπους παρουσίασης των κατηγοριών που προέκυψαν. Ο πρώτος τρόπος (όπως φαίνεται στο Σχήμα 49) αφορά τη δημιουργία μιας ειδικής σελίδας όπου εμφανίζονται οι κατηγορίες που προέκυψαν και το περιεχόμενο που κατηγοριοποιήθηκε σε κάθε μια. Για κάθε αντικείμενο παρέχεται ένας σύνδεσμος όπου δίνονται τα πλήρη στοιχεία γι' αυτό, όπως περιγράφηκε στο Σχήμα 48.

Σχήμα 49 Παρουσίαση των τελικών κατηγοριών που προέκυψαν κατά την κατηγοριοποίηση

Ο δεύτερος τρόπος αφορά την παροχή της ιεραρχίας κατηγοριοποίησης μέσω του προτύπου RSS (έκδοση 2.0). Συγκεκριμένα, για κάθε αντικείμενο που κατηγοριοποιήθηκε επιτυχώς δίνονται τα στοιχεία: τίτλος, σύνδεσμος, περιγραφή (εφόσον υπάρχει) και κατηγορία στην οποία ανήκει.

```
<?xml version="1.0" encoding="iso-8859-7" ?>
- <rss version="2.0">
  - <channel>
      <title>ATL CME Platform</title>
      <link>http://hungary.ebusiness.uoc.gr/</link>
      <description>Η περιγραφή του site</description>
      <language>el</language>
      <copyright>ATL CME Platform © Copyright 2004</copyright>
      <webMaster>webmaster@grnet.gr</webMaster>
      <lastBuildDate>Tue, 13 Dec 2005 01:03:27 +0200</lastBuildDate>
    + <image>
    - <item>
        <title>Issues in Mobile E-Commerce</title>
        <link>http://hungary.ebusiness.uoc.gr/content/modules/downloads/acls_danae1.txt</link>
        <description />
        <category>Top/Computers/Software/Operating_Systems/Unix/BSD</category>
      </item>
    - <item>
        <title>Wallet Concept Description</title>
        <link>http://hungary.ebusiness.uoc.gr/content/modules/downloads/acls_danae2.txt</link>
        <description />
        <category>Top/Computers/Programming/Languages</category>
      </item>
```

Σχήμα 50 Εξαγωγή τελικών κατηγοριών με το πρότυπο RSS

5.5.3.5 Χειρισμός τμημάτων κειμένου που δεν κατηγοριοποιήθηκαν

Στην περίπτωση που κάποιο κείμενο δεν έχει κατηγοριοποιηθεί παρέχεται η δυνατότητα εκ των υστέρων κατηγοριοποίησης. Όλα τα κείμενα που ανήκουν σε αυτό το είδος εισάγονται σε μια κατηγορία με την ένδειξη «Μη κατηγοριοποιημένα» οπότε ο χρήστης που έχει την αντίστοιχη άδεια μπορεί να αντιστοιχήσει το κείμενο σε κάποια κατηγορία που θεωρεί ότι είναι σωστή.

5.5.3.6 Παροχή στατιστικών

Μια επιπρόσθετη δυνατότητα του υποσυστήματος κατηγοριοποίησης αφορά την παροχή στατιστικών για το περιεχόμενο. Στατιστικά παρέχονται για κάθε πηγή περιεχομένου (ανάλογα με το μηχανισμό συλλογής) χωριστά αλλά και συνολικά για όλο το περιεχόμενο. Στο Σχήμα 51 παρουσιάζονται τα στατιστικά για το περιεχόμενο που συγκεντρώθηκε μέσω RSS και μέσω του μηχανισμού διαχείρισης αρχείων καθώς και συνολικά για όλο το περιεχόμενο. Τα στοιχεία είναι ποσοτικά και ποιοτικά.

Ποσοτικά στοιχεία

Τα ποσοτικά στοιχεία παρέχονται συνολικά για όλο το περιεχόμενο και για κάθε μηχανισμό χωριστά. Αυτά είναι τα ακόλουθα:

- Αριθμός αποθηκευμένων πηγών (ισχύει μόνο για το RSS): είναι το πλήθος των πηγών RSS feeds που υπάρχουν αποθηκευμένες στο δικτυακό τόπο

- Αριθμός text segments: για το RSS είναι ο αριθμός των νέων που περιέχουν συνολικά οι πηγές ενώ για τους άλλους μηχανισμούς είναι το πλήθος των αντικειμένων που έχουν εισαχθεί (αρχεία, άρθρα, σύνδεσμοι). Ο αριθμός αυτός παρέχεται και συνολικά για όλο το περιεχόμενο και αποτελεί το άθροισμα όλων των κειμένων που έχουν εισαχθεί στο δικτυακό τόπο
- Αριθμός κειμένων που έχει κατηγοριοποιηθεί: Πρόκειται για το σύνολο των κειμένων που έχουν ανατεθεί σε μια κατηγορία. Ο αριθμός αυτός παρέχεται και για το σύνολο του περιεχομένου
- Ποσοστό κατηγοριοποίησης: Είναι το κλάσμα του αριθμού των κειμένων που έχει κατηγοριοποιηθεί προς τον αρχικό αριθμό κειμένων που εισήχθησαν στο δικτυακό τόπο. Το στοιχείο αυτό παρέχεται και για το περιεχόμενο συνολικά.

Ποιοτικά στοιχεία

Όπως έχει αναφερθεί και παραπάνω τα ποιοτικά στοιχεία παρέχονται εφόσον κάποιος χρήστης έχει αξιολογήσει το αποτέλεσμα της κατηγοριοποίησης. Καθώς δεν έχουν νόημα για το σύνολο του περιεχομένου παρέχονται για κάθε μηχανισμό χωριστά. Αυτά είναι:

- Δεδομένα που κατηγοριοποιήθηκαν σωστά: Πρόκειται για τον αριθμό των κειμένων που κατηγοριοποιήθηκαν σε κάποια κατηγορία την οποία θα επέλεγε και ο χρήστης
- Δεδομένα που κατηγοριοποιήθηκαν λανθασμένα: Ο αριθμός των κειμένων για τα οποία επελέγη κάποια κατηγορία με την οποία ο χρήστης δε συμφωνεί
- Ακρίβεια: Περιγράφει την ικανότητα του συστήματος να κατηγοριοποιεί σωστά τα δεδομένα τα οποία έχουν κατηγοριοποιηθεί

Στατιστικά για το περιεχόμενο

Περιεχόμενο που συγκεντρώθηκε μέσω RSS

Αριθμός αποθηκευμένων πηγών	33
Αριθμός text segments	526
Αριθμός κειμένων που έχει κατηγοριοποιηθεί	383
Ποσοστό κατηγοριοποίησης	72.81%

Ποιοτικά στοιχεία	
Δεδομένα που κατηγοριοποιήθηκαν σωστά	170
Δεδομένα που κατηγοριοποιήθηκαν λανθασμένα	174
Ακρίβεια	0.56

Περιεχόμενο που συγκεντρώθηκε από Downloads

Ποσοτικά στοιχεία	
Αριθμός text segments	195
Αριθμός κειμένων που έχει κατηγοριοποιηθεί	193
Ποσοστό κατηγοριοποίησης	98.97%

Ποιοτικά στοιχεία	
Δεδομένα που κατηγοριοποιήθηκαν σωστά	73
Δεδομένα που κατηγοριοποιήθηκαν λανθασμένα	102
Ακρίβεια	0.62

Στατιστικά για όλο το περιεχόμενο	
Αριθμός text segments	724
Αριθμός κειμένων που έχει κατηγοριοποιηθεί	579 ή 79.97%
Δεδομένα που κατηγοριοποιήθηκαν σωστά	243
Δεδομένα που κατηγοριοποιήθηκαν λανθασμένα	277

Σχήμα 51 Στατιστικά για το περιεχόμενο

6 Πειράματα

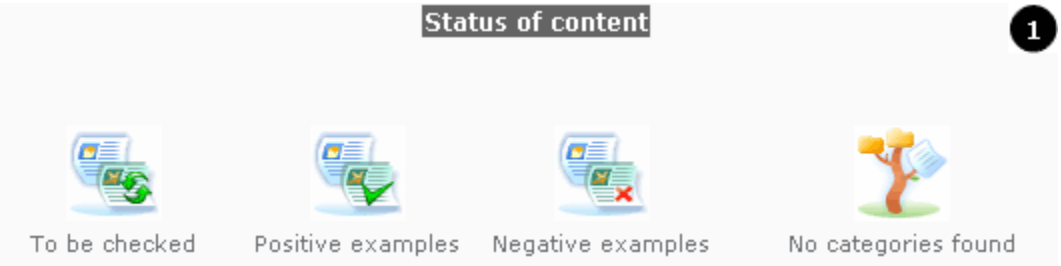
6.1 Εισαγωγή

Στο κεφάλαιο 1 περιγράφονται τα πειράματα που διεξήχθησαν για την αξιολόγηση της απόδοσης του αλγορίθμου κατηγοριοποίησης. Όπως αναφέρθηκε στα κεφάλαια 4 και 5 η κατηγοριοποίηση του περιεχομένου γίνεται με βάση την ιεραρχία που παρέχεται από τον κατάλογο DMOZ. Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκε το στιγμιότυπο που δημιουργήθηκε στις 6 Σεπτεμβρίου 2005. Το στιγμιότυπο αυτό περιλαμβάνει περίπου 800000 κατηγορίες ενώ η θεματική ενότητα που χρησιμοποιήθηκε είναι η "Computers". Στην υλοποίηση αυτή αφαιρέθηκαν κατηγορίες, η περιγραφή των οποίων χρησιμοποιούσε μονοπάτια που ανήκαν σε άλλες θεματικές ενότητες. Η τελική ταξονομία περιλαμβάνει 9000 κατηγορίες περίπου. Η κατηγοριοποίηση κάθε κομματιού περιεχομένου απαιτεί περίπου 3 δευτερόλεπτα περίπου. Στον χρόνο αυτό ενσωματώνεται ο χρόνος ανάλυσης και εξαγωγής του λεξικού για κάθε ένα από τα κείμενα.

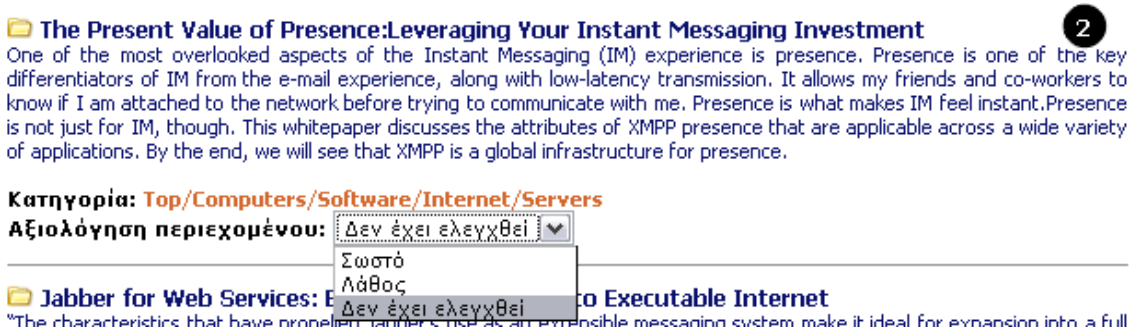
Η διαδικασία για τη διεξαγωγή των πειραμάτων περιλαμβάνει τη συλλογή, αποθήκευση και αξιολόγηση του περιεχομένου από διάφορες πηγές. Η αξιολόγηση του περιεχομένου αυτού πραγματοποιείται μέσω χειρωνακτικά μαρκαρισμένων συλλογών. Η διαδικασία περιλαμβάνει την συλλογή των εγγράφων και την επιλογή ενός ή περισσοτέρων ειδικών οι οποίοι μαρκάρουν κάθε έγγραφο με την ένδειξη σωστό / λάθος. Η ένδειξη «Σωστό» δείχνει ότι η αντιστοίχηση του εγγράφου στην κατηγορία που προέκυψε από τον αλγόριθμο είναι ακριβής. Η ένδειξη «Λάθος» αναπαριστά τη διαφωνία του χρήστη με την κατηγορία στην οποία αποδόθηκε το έγγραφο. Στο περιβάλλον υλοποίησης ο χρήστης έρχεται σε επαφή μόνο με τα έγγραφα στα οποία έχει αποδοθεί κάποια κατηγορία. Στη συνέχεια αφού ολοκληρωθεί η διαδικασία μαρκαρίσματος παρέχονται κάποια ποσοτικά και ποιοτικά στοιχεία σχετικά με την κατηγοριοποίηση. Τα ποσοτικά στοιχεία αφορούν τον αρχικό αριθμό των κειμένων και το ποσοστό από αυτά που κατηγοριοποιήθηκε ενώ τα ποιοτικά αφορά τον υπολογισμό της ακρίβειας. Οι μετρικές αυτές έχουν οριστεί στην ενότητα 4.4.2.6 ενώ η σελίδα εξαγωγής των στατιστικών παρουσιάζεται στο Σχήμα 51.

Τα βήματα της διαδικασίας για την αξιολόγηση του περιεχομένου περιγράφονται στα σχήματα που ακολουθούν. Στο Σχήμα 52 παρουσιάζεται η πρώτη σελίδα της διεπαφής. Υπάρχουν τέσσερις επιλογές. Η πρώτη περιέχει όλο το περιεχόμενο προς αξιολόγηση, η δεύτερη όλα το υλικό που θεωρείται ότι έχει κατηγοριοποιηθεί σωστά, η τρίτη επιλογή παρουσιάζει περιεχόμενο που θεωρείται ότι έχει κατηγοριοποιηθεί σε λάθος κατηγορία και η τέταρτη επιλογή περιγράφει το

περιεχόμενο που δεν έχει αντιστοιχηθεί σε καμία από τις κατηγορίες. Στο Σχήμα 53 ο χρήστης θα πρέπει να αποφασίσει εάν το κείμενο έχει κατηγοριοποιηθεί σωστά ή όχι. Η αρχική επιλογή είναι «Δεν έχει ελεγχθεί». Τέλος, στο Σχήμα 54 παρουσιάζονται τα στατιστικά που παρέχονται εφόσον έχει γίνει η αξιολόγηση.



Σχήμα 52 Η πρώτη σελίδα για το χρήστη που ελέγχει το περιεχόμενο



Σχήμα 53 Έλεγχος ενός αρχείου που έχει εισαχθεί στο υποσύστημα κατηγοριοποίησης

Σχήμα 54 Στατιστικά στοιχεία που παρέχονται συνολικά για τα αρχεία που έχουν εισαχθεί στο υποσύστημα κατηγοριοποίησης

6.2 Σενάρια και αποτελέσματα

Το περιεχόμενο που συγκεντρώθηκε στα πλαίσια των πειραμάτων για την παρούσα εργασία περιλαμβάνει πέντε σενάρια. Αυτά είναι:

- Γενικό περιεχόμενο
- Περιεχόμενο που προτείνεται από το συγγραφέα της δημοσίευσης «Critical themes in electronic commerce research: a meta-analysis» το οποίο είναι κατηγοριοποιημένο ανά τομέα σύμφωνα με το συγγραφέα
- Συλλογή 100 δημοσιεύσεων από το ACM που αφορούν την ασφάλεια
- Συλλογή RSS ενός χρήστη που ενδιαφέρεται για την περιοχή του open-source
- Συλλογή των RSS νέων του Slashdot για 5 ημέρες.

Σενάριο 1

Το πρώτο σενάριο αφορά περιεχόμενο από διαφορετικές θεματικές ενότητες. Περιλαμβάνει 325 νέα που συγκεντρώθηκαν από διαφορετικές RSS πηγές καθώς και 37 δημοσιεύσεις συμβουλευτικών εταιρειών. Στόχος του πειράματος αυτού είναι η επίδειξη της απόδοσης του αλγορίθμου κατηγοριοποίησης όταν το περιεχόμενο είναι γενικό και εφαρμόζεται στο σύνολο της ιεραρχίας.

Στοιχεία	**RSS**	**Δημοσιεύσεις**
Αριθμός text segments	325	37
Αριθμός κειμένων που κατηγοριοποιήθηκαν	216	36
Ποσοστό κατηγοριοποίησης	66.46%	97.3%
Δεδομένα που κατηγοριοποιήθηκαν σωστά	103	15
Δεδομένα που κατηγοριοποιήθηκαν σε λάθος κατηγορία	93	19
Ακρίβεια	0.52	0.58

Πίνακας 15 Αποτελέσματα του περιεχομένου που χρησιμοποιήθηκε στο 1ο σενάριο

Ο Πίνακας 15 παρουσιάζει όλα τα στοιχεία που μπορούν να εξαχθούν από το πρώτο πείραμα. Όπως φαίνεται τα νέα που προέρχονται από RSS κατηγοριοποιούνται σε ποσοστό περίπου 66% σε αντίθεση με τις δημοσιεύσεις στις οποίες το ποσοστό κατηγοριοποίησης φθάνει στο 97%. Αν και τα γενικά ποσοστά κατηγοριοποίησης έχουν μεγάλη απόκλιση φαίνεται παρακάτω ότι τα δεδομένα που κατηγοριοποιήθηκαν σωστά από τα RSS νέα είναι περισσότερα από αυτά των δημοσιεύσεων. Αυτό εξηγείται επειδή τα θέματα των δημοσιεύσεων ανήκουν σε αρκετά κλαδιά του DMOZ οπότε δε μπορεί να εξαχθεί ένα ακριβές αποτέλεσμα όσον αφορά την κατηγοριοποίηση.

Από την άλλη πλευρά παρατηρείται ένα αρκετά μεγάλο ποσοστό νέων το οποίο δεν κατηγοριοποιείται. Αυτό εξηγείται γιατί τα περισσότερα νέα περιέχουν πολύ λίγες λέξεις στην περιγραφή τους άρα υπάρχει και μεγαλύτερη δυσκολία στην εύρεση σημαντικών λέξεων για τη δημιουργία του λεξικού για κάθε νέο.

Σενάριο 2

Το δεύτερο σενάριο αφορά δημοσιεύσεις που προτείνονται από το συγγραφέα μιας επιστημονικής δημοσίευσης ως επιπλέον πηγές για διάβασμα. Οι δημοσιεύσεις αυτές είναι χωρισμένες σε 16 επιμέρους κατηγορίες. Στόχος του πειράματος αυτού είναι να διαπιστώσουμε σε ποιο ποσοστό ο αλγόριθμος κατηγοριοποίησης θα μπορούσε να πλησιάσει την κατηγοριοποίηση που θα είχε στο μυαλό του κάποιος χρήστης (εν προκειμένω ο συγγραφέας μιας επιστημονικής δημοσίευσης).

Ο Πίνακας 16 παρουσιάζει τις κατηγορίες που όρισε ο συγγραφέας και το ποσοστό των κειμένων που κατηγοριοποιήθηκαν σωστά σε αυτές τις κατηγορίες. Κοιτάζοντας κανείς προσεκτικά τον πίνακα αυτό και συσχετίζοντας τα αποτελέσματα που προκύπτουν με την ιεραρχία του DMOZ μπορεί να εξάγει τα παρακάτω συμπεράσματα:

- Θεματικές περιοχές οι οποίες δεν είναι ορισμένες στην κατηγορία που χρησιμοποιείται στα πλαίσια αυτού του πειράματος (“Computers”) έχουν ποσοστό επιτυχίας αρκετά χαμηλό ή και πολύ μικρό. Παράδειγμα αποτελεί η ενότητα M-Commerce η οποία είναι σχετικά καινούρια και δεν υπάρχει ως αυτοτελής κατηγορία. Ωστόσο ορισμοί της βρίσκονται κάτω από διάφορα κλαδιά της ιεραρχίας του DMOZ όπως “Science”, “Business”, “News & Technology”.
- Μικρό ποσοστό επιτυχίας παρουσιάζουν ακόμη κατηγορίες οι οποίες δεν αποτελούν έννοιες που αναπαρίστανται από την ιεραρχία του DMOZ. Για παράδειγμα η κατηγορία “Technology Adoption” δεν υφίστανται ως έννοια άρα δεν περιέχεται και αυτοτελής στην ιεραρχία που χρησιμοποιείται. Μάλιστα επειδή είναι αρκετά γενική και κάποιες δημοσιεύσεις μπορεί να περιέχουν έννοιες ορισμένες από την ιεραρχία παρουσιάζει επιτυχία 20%.
- Αρκετές από τις δημοσιεύσεις περιλαμβάνουν λέξεις – κλειδιά οι οποίες δεν είναι πολύ καλά ορισμένες στην κατηγορία “Computers”. Ένα τέτοιο παράδειγμα αποτελεί η έννοια E-commerce η οποία βρίσκεται ορισμένη σε δύο κλαδιά του DMOZ (Business: E-Commerce και Computers: Software: Business: E-commerce). Αυτό σημαίνει ότι εάν μια δημοσίευση περιγράφει τεχνολογίες E-commerce τότε πιθανότητα θα κατηγοριοποιηθεί στην ενότητα “Computers: Programming” αν και ο χρήστης έχει στο μυαλό του την κατηγορία “E-commerce”.

- Δημοσιεύσεις που ανήκουν σε πολύ καλά ορισμένες κατηγορίες του κλαδιού "Computers" (δηλαδή δεν ανήκουν σε κάποια άλλη θεματική ενότητα) π.χ. Agents παρουσιάζουν μεγάλο ποσοστό κατηγοριοποίησης. Ειδικότερα η έννοια "Agents" περιέχεται στις εξής κατηγορίες: "Computers: Artificial Intelligence: Agents", "Computers: Programming: Agents" και Computers: Artificial Life: Agents.

Κατηγορία	**Ποσοστό κειμένων που κατηγοριοποιήθηκε σωστά**
Agents	100%
Architecture	100%
Auctions	75%
B2B	66%
B2C	25%
CRM	30%
E-Service	75%
M-Commerce	0%
P2P	50%
Regulation	66%
Research	100%
Security	0%
Strategy	66%
Supply Chain	50%
Technology Adoption	20%
Trust	50%

Πίνακας 16 Περιεχόμενο που συλλέχθηκε για το 2ο πείραμα

Ο Πίνακας 17 συνοψίζει τα αποτελέσματα της κατηγοριοποίησης για το δεύτερο πείραμα. Παρατηρούμε ότι το ποσοστό της κατηγοριοποίησης είναι αρκετά υψηλό αν και ο αριθμός των δημοσιεύσεων που κατηγοριοποιήθηκαν σωστά δεν είναι πολύ μεγαλύτερος από τις δημοσιεύσεις που αντιστοιχήθηκαν σε λανθασμένη κατηγορία.

Στοιχεία	Δημοσιεύσεις
Αριθμός text segments	55
Αριθμός κειμένων που κατηγοριοποιήθηκαν	54
Ποσοστό κατηγοριοποίησης	98%
Δεδομένα που κατηγοριοποιήθηκαν σωστά	26
Δεδομένα που κατηγοριοποιήθηκαν σε λάθος κατηγορία	22
Ακρίβεια	0,28

Πίνακας 17 Αποτελέσματα κατηγοριοποίησης για το 2ο πείραμα

Σενάριο 3

Στο πείραμα αυτό χρησιμοποιήθηκε μια συλλογή από 100 δημοσιεύσεις του περιοδικού ACM στο θέμα της ασφάλειας (security). Στόχος του πειράματος είναι να βρεθεί το ποσοστό των κειμένων που θα κατηγοριοποιηθούν σε κάποια συναφή κατηγορία. Ο Πίνακας 18 συνοψίζει τα αποτελέσματα της κατηγοριοποίησης. Όπως φαίνεται το ποσοστό της κατηγοριοποίησης είναι 100% ωστόσο το ποσοστό των κειμένων που κατηγοριοποιείται σωστά είναι αρκετά χαμηλό. Αυτό εξηγείται αν μελετηθεί η δομή της ιεραρχίας του DMOZ για την κατηγορία "Security":

- Computers: Security
- Business: Business Services: Fire and Security
- Society: Issues: Economic: Social Security
- Society: Issues: Warfare and Conflict
- Computers: Security: Consultants: General and Freelance: North America: United States.

Όπως παρατηρείται η περιγραφή της έννοιας είναι κατακερματισμένη σε διάφορα κλαδιά του DMOZ. Αυτό σημαίνει ότι εάν κάποιο κείμενο αναφέρεται για παράδειγμα στην τρίτη κατηγορία τότε η κατηγοριοποίησή του δεν μπορεί να γίνει σωστά αφού χρησιμοποιείται η πρώτη περιοχή γνώσης.

Στοιχεία	Δημοσιεύσεις
Αριθμός text segments	100
Αριθμός κειμένων που κατηγοριοποιήθηκαν	100
Ποσοστό κατηγοριοποίησης	100%
Δεδομένα που κατηγοριοποιήθηκαν σωστά	32
Δεδομένα που κατηγοριοποιήθηκαν σε λάθος κατηγορία	61
Ακρίβεια	0,68

Πίνακας 18 Αποτελέσματα κατηγοριοποίησης για το 3ο πείραμα

Σενάριο 4

Το περιεχόμενο που χρησιμοποιείται για το σενάριο αυτό περιλαμβάνει νέα που προέρχονται από RSS και αφορούν open-source. Πρόκειται για πηγές που έχουν επιλεγεί από ένα χρήστη και αφορούν τις εξελίξεις και τις τεχνολογίες που σχετίζονται με open-source, wireless κτλ. Συνολικά συγκεντρώθηκαν 141 κείμενα και από αυτά κατηγοριοποιήθηκαν 112 (Πίνακας 19). Το ποσοστό της κατηγοριοποίησης είναι 79,4% ενώ το ποσοστό των κειμένων που κατηγοριοποιείται σωστά είναι 42%. Ελέγχοντας τις πηγές παρατηρούμε ότι όσες αναφέρονται σε νέα για linux, στις εκδόσεις από linux κατηγοριοποιούνται σωστά κατά ένα πολύ μεγάλο ποσοστό.

Στοιχεία	Δημοσιεύσεις
Αριθμός text segments	141
Αριθμός κειμένων που κατηγοριοποιήθηκαν	112
Ποσοστό κατηγοριοποίησης	79,4%
Δεδομένα που κατηγοριοποιήθηκαν σωστά	48
Δεδομένα που κατηγοριοποιήθηκαν σε λάθος κατηγορία	64
Ακρίβεια	0,57

Πίνακας 19 Αποτελέσματα κατηγοριοποίησης για το 4ο πείραμα

Σενάριο 5

Το περιεχόμενο που χρησιμοποιείται στα πλαίσια του σεναρίου αυτού αποτελείται από τα RSS feeds που συλλέγονται από το δικτυακό τόπο www.slashdot.org σε χρονικό διάστημα 5 ημερών. Το αντικείμενο που διαπραγματεύονται αφορά τόσο την περιοχή των υπολογιστών όσο και γενικά σχόλια που γίνονται από τους χρήστες. Στόχος του πειράματος είναι η αποτύπωση της ικανότητας εύρεσης του αλγορίθμου να αναγνωρίζει ότι το περιεχόμενο που περιέχεται δεν είναι σωστό.

Ο Πίνακας 20 αποτυπώνει τα αποτελέσματα της κατηγοριοποίησης. Παρατηρούμε ότι το ποσοστό της κατηγοριοποίησης είναι 83% αλλά τα μισά από τα κείμενα κατηγοριοποιούνται σε λανθασμένες κατηγορίες. Το αποτέλεσμα αυτό είναι αναμενόμενο αφού το περιεχόμενο δεν αφορά το αντικείμενο της περιοχής γνώσης της ιεραρχίας που χρησιμοποιείται.

Στοιχεία	**Περιεχόμενο από το Slashdot**
Αριθμός text segments	60
Αριθμός κειμένων που κατηγοριοποιήθηκαν	50
Ποσοστό κατηγοριοποίησης	0,83
Δεδομένα που κατηγοριοποιήθηκαν σωστά	25
Δεδομένα που κατηγοριοποιήθηκαν σε λάθος κατηγορία	25
Ακρίβεια	0,52

Πίνακας 20 Αποτελέσματα της κατηγοριοποίησης του περιεχομένου του 5^{ου} πειράματος

6.3 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάστηκε μια σειρά από πειράματα τα οποία υλοποιούν συγκεκριμένα σενάρια με στόχο την αξιολόγηση της ικανότητας του αλγορίθμου να κατηγοριοποιεί το περιεχόμενο υπό συγκεκριμένες προϋποθέσεις. Συνοψίζοντας, μπορούμε να πούμε ότι:

- Ο αλγόριθμος είναι αποδοτικός όταν το περιεχόμενο προς κατηγοριοποίηση είναι πλούσιο όσον αφορά τη σημασιολογία του
- Το αποτέλεσμα της κατηγοριοποίησης δεν εξαρτάται από το πλήθος των λέξεων του κειμένου όσο από την «ποιότητα» τους (αν χρησιμοποιούνται συχνά)
- Το ποσοστό αλλά και η ακρίβεια της κατηγοριοποίησης συνδέονται με τον ορισμό της κατηγορίας στην ιεραρχία του DMOZ. Παρατηρείται ότι όσο πιο κατακερματισμένη είναι η περιγραφή μιας κατηγορίας στα κλαδιά του δέντρου τόσο πιο διεσπαρμένες είναι οι κατηγορίες που επιστρέφονται ως σχετικές ενώ και η τελική κατηγορία μπορεί να μην είναι ακριβής.

7 Επίλογος

7.1 Συμπεράσματα

Η παρούσα μεταπτυχιακή εργασία ασχολήθηκε με την κατηγοριοποίηση περιεχομένου που συλλέγεται αυτόματα ή ημιαυτόματα από το διαδίκτυο. Η διαδικασία αυτή προτείνεται ως τμήμα ενός συστήματος διαχείρισης περιεχομένου ΕΛ/ΑΚ και στα πλαίσια της εργασίας εντάσσεται λειτουργικά σε ένα τέτοιο σύστημα. Η συλλογή του περιεχομένου από το διαδίκτυο πραγματοποιείται με δύο διαφορετικούς τρόπους: RSS και GoogleAPI. Στην πρώτη περίπτωση το περιεχόμενο προέρχεται από δικτυακούς τόπους που χρησιμοποιούν τα πρότυπα RSS/ATOM ενώ στη δεύτερη το περιεχόμενο εισάγεται στο σύστημά μας χρησιμοποιώντας τη λειτουργικότητα της μηχανής αναζήτησης Google. Η κατηγοριοποίηση του κειμένου στηρίζεται σε έναν ιεραρχικό κατάλογο που υπάρχει διαθέσιμος στο διαδίκτυο, το DMOZ τον οποίο χρησιμοποιούν σήμερα οι περισσότερες μηχανές αναζήτησης. Η διαδικασία της κατηγοριοποίησης περιλαμβάνει την λεξικογραφική ανάλυση του κειμένου που θα υποβάλλουμε στην διαδικασία, τον στατιστικό προσδιορισμό των σημαντικών λέξεων και τέλος την αλγοριθμική απόφαση για την κατηγορία στην οποία ανήκει. Οι χρόνοι λήψης της απόφασης είναι τέτοιοι που επιτρέπουν τη χρήση των αλγορίθμων σε περιβάλλον διαδικτύου. Τέλος, το αποτέλεσμα της κατηγοριοποίησης (εκτός από την λήψη απόφασης για την τοποθέτηση σε κάποια κατηγορία του υπό εξέταση αντικειμένου) εξάγεται σε RSS αλλά και σε μια εσωτερική σελίδα του συστήματος ενώ παρέχεται μια σειρά από στατιστικά στοιχεία που αφορούν το περιεχόμενο.

Στα πλαίσια της εργασίας διενεργήθηκαν μια σειρά από πειράματα σε συλλογές κειμένων που αφορούσαν επιστημονικές δημοσιεύσεις, αποτελέσματα αναζητήσεων σε μια λέξη / φράση, RSS feeds. Στην γενική περίπτωση μπορεί κανείς να συμπεράνει από τα αποτελέσματα της κατηγοριοποίησης ότι ο αλγόριθμος (ή σωστότερα οι αλγόριθμοι που αποτελούν τη διαδικασία κατηγοριοποίησης) που προτάθηκε παραπάνω είναι αποτελεσματικός στις περιπτώσεις που το κείμενο περιέχει σημαντικές λέξεις (ανεξάρτητα από το πλήθος τους) ενώ κείμενα που περιέχουν λέξεις του καθημερινού λεξιλογίου παρουσιάζουν χαμηλό ποσοστό κατηγοριοποίησης. Τέλος, περιεχόμενο που νοηματικά ανήκει σε πολλές κατηγορίες της ιεραρχίας παρουσιάζει σημαντική απώλεια στην ακρίβεια του τελικού αποτελέσματος.

7.2 Προτεινόμενες επεκτάσεις – ανοικτά θέματα

Ο αλγόριθμος κατηγοριοποίησης που προτάθηκε στο πλαίσιο αυτής της εργασίας εφαρμόστηκε αποτελεσματικά σε μια σειρά κειμένων που προέρχονται από διαφορετικές πηγές. Ωστόσο θα πρέπει να πραγματοποιηθούν περαιτέρω πειράματα σε διάφορες συλλογές δεδομένων που παρέχονται στο διαδίκτυο ή που θα δημιουργήσει ο ίδιος προκειμένου αποκτηθεί ακόμα καλύτερη άποψη για τις επιδόσεις της διαδικασίας. Βελτιώσεις και επεκτάσεις μπορούν να πραγματοποιηθούν σε διάφορα μέρη του συστήματος και αφορούν τόσο την υλοποίηση του (π.χ. νέες τεχνολογίες που μπορούν να υιοθετηθούν) όσο και την αλγοριθμική σχεδίαση με την χρήση πιο αποδοτικών τεχνικών.

Στο επίπεδο της τεχνολογίας κανείς μπορεί να συζητήσει για:

- **Χρήση XML βάσης δεδομένων**

Όπως αναφέρθηκε και σε προηγούμενα κεφάλαια η βάση δεδομένων που χρησιμοποιεί το ΣΔΠ ATL CME είναι σχεσιακή. Ωστόσο η πληροφορία που συλλέγεται και κατηγοριοποιείται αφορά ημι-δομημένο περιεχόμενο που περιγράφεται σε XML μορφή. Η χρήση μιας XML βάσης δεδομένων (π.χ. eXist) θα βελτίωνε το χρόνο που απαιτείται στην επεξεργασία (ιεραρχία DMOZ και περιεχόμενο που συλλέγεται) και τη μετατροπή των δεδομένων σε σχεσιακό σχήμα ενώ θα συνέβαλλε στη συνολική μείωση του χρόνου της κατηγοριοποίησης.

- **Διάθεση της διαδικασίας σαν δικτυακή υπηρεσία (web service) ώστε να μπορεί να χρησιμοποιηθεί και από άλλα ΣΔΠ**

Μια μελλοντική επέκταση του συστήματος αφορά τη διάθεσή του ως δικτυακή υπηρεσία όπου θα μπορεί ο κάθε χρήστης να κατηγοριοποιεί το περιεχόμενό του. Η διάθεση της υπηρεσίας αυτής θα συμβάλλει στη μεταφορά γνώσης μεταξύ των χρηστών αλλά και την καλύτερη οργάνωση των συλλογών των εγγράφων τους.

Οι αλγόριθμοι που χρησιμοποιούνται στο σύστημα μπορούν να βελτιωθούν με την:

- **Εκμετάλλευση της γνώσης που παράγεται στο σύστημα από προηγούμενες κατηγοριοποιήσεις**

Στα πλαίσια της παρούσας μεταπτυχιακής εργασίας δόθηκε έμφαση στη δημιουργία ενός αλγορίθμου κατηγοριοποίησης ο οποίος να μπορεί να εφαρμοστεί σε περιεχόμενο που προέρχεται από διάφορες πηγές. Η διαδικασία που ακολουθείται «δεν έχει μνήμη». Αυτό σημαίνει ότι κάθε φορά που ένα κείμενο κατηγοριοποιείται όλες οι ενέργειες γίνονται από την αρχή αγνοώντας γνώση που μπορεί να έχει προέλθει από προηγούμενες κατηγοριοποιήσεις. Στο σημείο αυτό το σύστημα θα μπορούσε να μαθαίνει από το παρελθόν προκειμένου να αποφεύγονται περιττές

αναζητήσεις σε όλο το σύνολο του καταλόγου του DMOZ. Πρακτικά αυτό σημαίνει ότι οι κατηγορίες του DMOZ μπορούν να συσχετιστούν με συχνά χρησιμοποιούμενες λέξεις – κλειδιά ώστε κείμενα που περιέχουν τέτοιες λέξεις να αναζητούνται καταρχήν στο σύνολο αυτών των κατηγοριών.

- **Επέκταση των τεχνικών ανάλυσης του κειμένου: αλλαγή των λιστών των συχνά χρησιμοποιούμενων λέξεων, αλλαγή του τρόπου εξαγωγής του λεξικού**

Η ανάλυση του κειμένου στην προσέγγιση που ακολουθήθηκε πραγματοποιείται εκτελώντας συντακτική και στατιστική ανάλυση. Η συντακτική ανάλυση περιλαμβάνει την αφαίρεση των σημείων στίξης, που μπορεί να υπάρχουν στο κείμενο, καθώς και των συχνά χρησιμοποιούμενων λέξεων που δεν προσφέρουν στην σημασιολογία του. Οι λίστες λέξεων που χρησιμοποιήθηκαν δεν αναφέρονται σε κάποιο συγκεκριμένο θέμα αλλά είναι γενικές και αφορούν το σύνολο της αγγλικής γλώσσας. Ωστόσο, υπάρχουν άλλες λίστες που είναι εξειδικευμένες στην περιγραφή ενός θέματος (π.χ. για μια εξειδικευμένη λίστα που αναφέρεται σε υπολογιστές η λέξη "Computers" είναι κοινή και μπορεί να αφαιρεθεί) και θα μπορούσαν να χρησιμοποιηθούν για την εξαγωγή ενός μικρότερου και περιεκτικότερου λεξικού.

Τέλος, ο τρόπος που επιλέχθηκε για την εξαγωγή των λέξεων – κλειδιών στηρίζεται μόνο στη στατιστική ανάλυση του κειμένου και στον αριθμό εμφάνισης των λέξεων στο κείμενο. Η προσέγγιση που περιγράφεται στο [31] και το [36] προτείνει την χρήση οντολογιών και εννοιών για την περιγραφή ενός κειμένου. Η χρήση οντολογιών μπορεί να περιορίσει το εύρος των κατηγοριών τις οποίες πρέπει κανείς να ελέγξει και να προσφέρει συμπεράσματα για την γνώση που περιγράφουν τα κείμενα.

- **Χρήση των σελίδων που περιγράφονται στο DMOZ ως μεταδεδομένα στις υπάρχουσες κατηγορίες**

Η κατηγοριοποίηση της πληροφορίας στην προσέγγιση που ακολουθείται βασίζεται στην ιεραρχία των κατηγοριών που παρέχεται από το DMOZ. Δε χρησιμοποιείται όμως καθόλου το περιεχόμενο που βρίσκεται κάτω από αυτές τις κατηγορίες. Το περιεχόμενο αυτό αποτελείται από ένα σύνολο δικτυακών τόπων με την περιγραφή τους. Το σύνολο των λέξεων των περιγραφών μπορεί να χρησιμοποιηθεί για να προσδιορίσει καλύτερα μια κατηγορία με τη μορφή μεταδεδομένων [36].

Τέλος σημαντικό τοπικό ενδιαφέρον έχει η **επέκταση της κατηγοριοποίησης σε ελληνικά κείμενα.** Το παρόν σύστημα διαχειρίζεται μόνο αγγλικά κείμενα αφού ακόμη δεν υπάρχουν

αξιόπιστες προσεγγίσεις εξαγωγής των ριζών των ελληνικών λέξεων καθώς και δυνατότητες εξαγωγής μεταδεδομένων από κείμενο. Μια πρόσφατη προσπάθεια συλλογής και κατηγοριοποίησης των σελίδων του ελληνικού διαδικτύου έχει προταθεί στο [6] όπου ακολουθούνται οι κανόνες της γραμματικής του Τριανταφυλλίδη για την εξαγωγή ριζών. Ακόμη θα πρέπει να σημειωθεί ότι ο κατάλογος του DMOZ περιέχει ένα πολύ μικρό κομμάτι του ελληνικού διαδικτύου το οποίο περιορίζεται κυρίως στο χώρο της τέχνης και ως εκ τούτου δε μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση γενικών κειμένων.

8 Βιβλιογραφία

[1]. Steinbach, M., Karypis, G., and Kumar, V. A comparison of document clustering techniques. KDD Workshop on Text Mining, 2000.

[2]. Cheng, D., Vempala, S., Kannan, R., and Wang, G. A divide-and-merge methodology for clustering. In Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2005.

[3]. Qin, H. A Review of Clustering Algorithms as Applied in IR. Master's Thesis, University of Illinois at U-C, 1999.

[4]. Yang, Y. A study on thresholding strategies for text categorization. In Proceedings of the 24th ACM SIGIR, pp. 137-145. ACM Press, 2001.

[5]. Pelleg, D. and Moore, A. Accelerating Exact k-means Algorithms with Geometric Reasoning. In the Proceedings of the Fifth International Conference on Knowledge Discovery in Databases, pp. 277-281. AAAI Press, 1999.

[6]. Eirinaki, M., Labos, H., and Vazirgiannis, M. Archiving the Greek Web, Technical Report 2003, http://www.db-net.aueb.gr.

[7]. Cimiano, P., Hotho, A., and Staab, S. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In Proceedings of the European Conference on Artificial Intelligence, pp. 435–439, 2004.

[8]. Jain, A.K., Murty M.N., and Flynn P.J. Data Clustering: A Review. ACM Computing Surveys, Vol. 31, No. 3. pp. 264-323, 1999.

[9]. Sahoo, N. Incremental Hierarchical Clustering of Text Documents. http://www.andrew.cmu.edu/user/nsahoo/draft6.pdf

[10]. Zeng, H. J., He, O. C., Chen, Z., Ma, W. Y., and Ma, J. Learning to cluster web search results. In the Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 210 – 217, 2004.

[11]. Savaresi, S. M., and Boley, D. L. On the performance of bisecting K-means and PDDP. 1st SIAM Conference on Data Mining, pp.1-14, 2001.

[12]. Zhang, Y., Zincir-Heywood, N., and Milios, E. Term-based Clustering and Summarization of Web Page Collections. Canadian Conference on AI, pp. 60-74, 2004.

[13]. Hoong, D. C., and Buyya, R. Guided Google: A Meta Search Engine and its Implementation using the Google Distributed Web Services. International Journal of Computers and Applications, Vol. 26, 2004. http://www.gridbus.org/papers/guidedgoogle.pdf.

[14]. Labrou, Y., and Finin, T. Yahoo! as Ontology - Using Yahoo! Categories to Describe Documents. In the Proceedings of the eighth international conference on Information and knowledge management, pp. 180-187. ACM Press, 1999.

[15]. Cusumano, M. A. Google: What It Is and What It Is Not. Communications of the ACM Vol. 48, No. 2, pp. 15-17, 2005.

[16]. Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E. A. Automatic Document Metadata Extraction using Support Vector Machines. In the Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, pp. 37 – 48, 2003.

[17]. Thorsten, J. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In the Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 137-142, 1998.

[18]. Huang, C-C., Chuang, S-L., and Chien, L-F. Using a Web-Based Categorization Approach to Generate Thematic Metadata from Texts. ACM Transactions on Asian Language Information Processing (TALIP), Vol. 3, Issue 3, pp. 190 – 212. ACM Press, 2004.

[19]. Yang, Y., and Liu, X. A re - examination of text categorization methods. In the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42-49. ACM Press, 1999.

[20]. Liu, H., Peng, R., Shaozhi, Y., and Xing, Li. An Efficient Centroid Based Chinese Web Page Classifier. In Proceedings of the 1st Asia-Pacific Advanced Network Research Workshop, pp 9-14, 2003.

[21]. Lertnattee, V., and Theeramunkong, T. Effect of term distributions on centroid-based text categorization. Information Sciences—Informatics and Computer Science: An International Journal, Vol. 158, Issue 1, pp. 89-115. Elsevier Science Inc., 2004.

[22]. Attardi, G., Gullì, A., Dato, D., and Tani, C. Towards Automated Categorization and Abstracting of Web Sites. http://www.di.unipi.it/~gulli/papers/submitted/automated_classification.htm.

[23]. Kongovi, M., Guzman, J. C., and Dasigi, V. Text Categorization: An Experiment Using Phrases. Lecture Notes In Computer Science, Vol. 2291, pp. 213-228. Springer-Verlag, 2002.

[24]. Sebastiani, F. Machine Learning in Automated Text Categorization. ACM Computing Surveys (CSUR), Vol. 34 , Issue 1, pp. 1-47. ACM Press, 2002.

[25]. Liu, H., Ou, J., Qi, Q., and Xiao, Y. Text Categorization. http://users.cs.dal.ca/~qiufen/pdfs/6403a.pdf.

[26]. Dumais, S., and Chen, H. Hierarchical Classification of Web Content. Annual ACM Conference on Research and Development in Information Retrieval, pp. 256-263. ACM Press, 2000

[27]. Sebastiani, F. Classification of Text, Automatic. Encyclopedia of Language & Linguistics 2nd Edition, Section: Applications of natural language processing, Vol. 14. Elsevier Science Publishers, 2006. http://www.math.unipd.it/~fabseb60/Publications/Publications.html.

[28]. Sebastiani, F. Text Categorization. Text Mining and its Applications, pp. 109-129. WIT Press, 2005.

[29]. Yang, Y. An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval, Vol. 1, Issue 1-2, pp. 69-90. Kluwer Academic Publishers, 1999.

[30]. Najork, M., and Heydon, A. High-Performance Web Crawling. Handbook of massive data sets, pp. 25-45. Kluwer Academic Publishers, 2002.

[31]. Halkidi, M., Nguyen, B., Varlamis, I., and Vazirgiannis, M. THESUS: Organizing Web document collections based on link semantics. The VLDB Journal - The International Journal on Very Large Data Bases, Vol. 12, Issue 4, pp. 320-332. Springer-Verlag New York, Inc., 2003.

[32]. Kuhlins, S., and Tredwell, R. Toolkits for Generating Wrappers. Lecture Notes In Computer Science, Vol. 2591, pp. 184-198. Springer-Verlag, 2002.

[33]. Bertino, E., and Ferrari, E. XML and Data Integration. IEEE Internet Computing, Vol. 5, Issue 6, pp. 75-76. IEEE Educational Activities Department, 2001.

[34]. Madnick, S. and Siegel, M. Seizing the Opportunity: Exploiting Web Aggregation. MIS Quarterly Executive, Vol. 1, No. 1, 2002. http://web.mit.edu/sloan-msa/Papers/3.12.pdf.

[35]. Dave, P., Logasa Bogen II, P., Karadkar P., U., Francisco-Revilla, L., Furuta, R., and Shipman, F. Dynamically growing hypertext collections. In the Proceedings of the fifteenth ACM conference on Hypertext and hypermedia, pp. 171-180. ACM Press, 2004.

[36]. Kim, S., Alani, H., Hall, W., Lewis H., P., Millard E., D., Shadbolt R., N., and Weal J., M. Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web. IEEE Intelligent Systems, Vol. 18, No. 1, pp. 14-21, 2003.

[37]. Huang, C-C., Chuang, S-L., and Chien, L-F. Categorizing Unknown Text Segments for Information Extraction Using a Search Result Mining Approach. In the Proceedings of the International Joint Conference on Natural Language Processing, pp. 576-586, 2004.

[38]. Chuang, S-L., and Chien, L-F. A Practical Web-based Approach to Generating Topic Hierarchy for Text Segments. In the Proceedings of the thirteenth ACM conference on Information and knowledge management, pp. 127-136. ACM Press, 2004.

[39]. Huang, C-C., Chuang, S-L., and Chien, L-F. Mining the Web for Generating Thematic Metadata from Textual Data. The 20th IEEE International Conference on Data Engineering (ICDE), 2004.

[40]. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., and Milios, E. Semantic similarity methods in wordNet and their application to information retrieval on the web. In the proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 10-16. ACM Press, 2005.

[41]. Porter, M. F. An Algorithm for Suffix Stripping. Program. 14(3): pp. 130–137, 1980.

[42]. Lovins, J. B. Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics (11), pp. 22–31, 1968.

[43]. Amer-Yahia, S., Du, F., and Freire, J. A comprehensive solution to the XML-to-relational mapping problem. In the Proceedings of the 6th annual ACM international workshop on Web information and data management, pp. 31-38. ACM Press, 2004.

[44]. Khan, L., and Rao, Y. A Performance Evaluation of Storing XML Data in Relational Database Management Systems. In the Proceedings of the 3rd international workshop on Web information and data management, pp. 31-38. ACM Press, 2001.

[45]. Süß, C. An Approach to the model-based fragmentation and relational storage of XML-documents. GI-Workshop Grundlagen von Datenbanken, pp. 98-102, 2001.

[46]. Shimura, T., Yoshikawa, M., and Uemura, S. Storage and Retrieval of XML Documents using Object-Relational Databases. Lecture Notes In Computer Science; Vol. 1677, pp. 206-217. Springer-Verlag, 1999.

[47]. Noelle, G. XML data processing and Relational Database Systems. GCA XML Europe' 99, Conference Proceedings, pp. 713-719, 1999.

[48]. Schmidt, A., Kersten, M., Windhouwer, M., and Waas, F. Efficient Relational Storage and Retrieval of XML Documents. Selected papers from the Third International Workshop WebDB 2000 on The World Wide Web and Databases, pp. 137-150. Springer-Verlag, 2000.

[49]. Robertson, J. Open-source content management systems. KM Column, 2004. http://www.steptwo.com.au/papers/kmc_opensource/.

[50]. Robertson, J. So, what is a content management system?. KM Column, 2003. http://www.steptwo.com.au/papers/kmc_what/index.html.

[51]. Zoellick, B. Technology Review: Moving into the Black with On-Demand Content Management. The Gilbane Conference on Content Management Technologies, 2004. http://www.gilbane.com/case_studies/technology_review_case_study.html.

[52]. Treese, W. Open systems for collaboration. netWorker, Vol. 8, Issue 1, pp. 13-16. ACM Press, 2004.

[53]. Denton, W. The Classification & Evaluation of Content Management Systems. Vol. 11, No. 2, pp. 2-13, 2003. http://www.gilbane.com/artpdf/GR11.2.pdf.

[54]. Berk, M. The Content Management Threshold, 2002. http://www.insightexec.com/cgi-bin/kasbrowse.cgi?action=detail&id=4853.

[55]. Grant, A. Content Management Systems, 2000. http://www.ukoln.ac.uk/nof/support/help/papers/cms/.

[56]. Doyle, B. Open Source Content Management Systems Redux. The Gilbane Report, Vol. 11, No. 3, 2003. http://www.gilbane.com/artpdf/GR11.3.pdf.

[57]. Geisler, G., Giersch, S., and McArthur, D. Creating Virtual Collections in Digital Libraries: Benefits and Implementation Issues. In the Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, pp. 210-218. ACM Press, 2002.

[58]. Bergmark, D. Collection synthesis. In the Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, pp. 253-262. ACM Press, 2002.

[59]. Calado, P., Goncalves, M., Fox, E., Ribeiro-Neto, B., Laender, A., S. da Silva, A., Reis, D., Roberto, P., and Vieira, M. The Web-DL Environment for Building Digital Libraries from the Web. In the Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, pp. 346-357. IEEE Computer Society, 2003.

[60]. Plas, L., Pallotta, V., Rajman, M., and Ghorbel, H. Automatic Keyword Extraction from Spoken Text. A Comparison of two Lexical Resources: the EDR and WordNet. In the Proceedings of the LREC 2004 international conference, pp. 2205-2208, 2004.

[61]. Matsuo, Y., and Ishizuka, M. Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. International Journal on Artificial Intelligence Tools, Vol. 13, No. 1, pp. 157-169, 2004.

[62]. Hammersley, B. Content Syndication with RSS. O'Reilly, ISBN 0-596-00383-8, 2003.

[63]. Boiko, B. Content Management Bible. Wiley, ISBN 076454862X, 2001.

[64]. Nong, Y. The Handbook of Data Mining. Lawrence Erlbaum Associates Inc., ISBN 0805840818, 2003.

[65]. Holly, Y. Content and Workflow Management for Library Websites: Case Studies. Idea Group Inc., ISBN 1-59140-533-5, 2005.

[66]. Abbass, H., Newton, C., and Sarker, R. Data Mining: A Heuristic Approach. Idea Group Inc., ISBN 1-930708-25-4, 2002.

[67]. Berry, M., and Linoff, G. Data Mining Techniques: For Marketing, Sales, and Customer Support. Wiley, ISBN 0471179809, 1997.

[68]. Mueller, J. Mining Google Web Services: Building Applications with the Google API. Sybex, ISBN 0782143334, 2004.

[69]. Hand, D., Mannila, H., and Smyth, P. Principles of Data Mining. MIT Press, ISBN 026208290x, 2001.

[70]. Gordon, M., Lindsay, R., and Fan, W. Literature-based discovery on the World Wide Web. ACM Transactions on Internet Technology (TOIT), Vol. 2, Issue 4, pp. 261-275. ACM Press, 2002.

[71]. Bonnet, Ph., and Bressan, S. Extraction and Integration of Data from Semi-structured Documents into Business Applications. In the Proceedings of the International Conference on Applications of Prolog (INAP' 97), 1997.

[72]. Liu, L., Pu, C., and Han W. XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources. In the Proceedings of the 16th International Conference on Data Engineering, pp. 611-621. IEEE Computer Society, 2000.

[73]. Noga, M., and Völkel, M. From Web Pages to Web Services with wal. In the Proceedings of the NCWS 2003, Mathematical Modelling in Physics Engineering and Cognitive Science, 2003.

[74]. Wei, H., Buttler, D., and Pu, C. Wrapping data into XML. Sigmod Record, Vol. 30 (3), pp. 33-38, 2001.

[75]. Liu, Z., Ng, W., and Lim, E. Personalized Web Views for Multilingual Web Sources. IEEE Internet Computing, Vol. 8, Issue 4, pp. 16-22. IEEE Educational Activities Department, 2004.

[76]. Crescenzi, V., Mecca, G., and Merialdo, P. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. In the Proceedings of the 27th International Conference on Very Large Data Bases, pp. 109-118, 2001.

[77]. Huffman, S. and Steier, D. A Navigation Assistant for Data Source Selection and Integration. In the Proceedings of AAAI Fall Symposium: AI Applications in Knowledge Navigation and Retrieval, AAAI Press. http://ai.eecs.umich.edu/people/huffman/hufpubs.html.

[78]. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. and Watkins, C. Text Classification using String Kernels. The Journal of Machine Learning Research, Vol. 2, pp. 419-444. MIT Press, 2002.

[79]. Zahn, C. T. Graph theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers, Vol. 20 Issue 1, pp. 68-86, 1971.

[80]. Firat, A., Madnick, S., Yahaya, N. A., Kuan, C. W., Bressan, S. Information Aggregation using the Cameleon# Web Wrapper. http://ideas.repec.org/p/mit/sloanp/18231.html.

[81]. Aggarwal, C., Gates, S. and Yu, P. On the merits of building categorization systems by supervised clustering. In the Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 352-356. ACM Press, 1999.

[82]. Denton, W. The Trend Towards Distributed Content Management. The Gilbane Report, Vol. 12, No. 2, 2004. http://www.gilbane.com/artpdf/GR12.2.pdf.

[83]. Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. Structure and Evolution of Blogspace. Communications of the ACM, Vol. 47, Issue 12, pp. 35-39. ACM Press, 2004.

[84]. Rosenbloom, A. The Blogosphere. Communications of the ACM, Vol. 47, Issue 12. ACM Press, 2004.

[85]. Hourihan, M. What We're Doing When We Blog. O' Reilly Network, 2002. http://www.oreillynet.com/pub/a/javascript/2002/06/13/megnut.html.

[86]. Zambonini, D. Is Web 2.0 killing the Semantic Web?. O' Reilly Network, 2005. http://www.oreillynet.com/pub/wlg/8013?CMP=OTC-TY3388567169.

[87]. Hof, R. It's A Whole New Web. Business Week, 2005. http://www.businessweek.com/magazine/content/05_39/b3952401.htm.

[88]. Ohmukai, I., Takeda, H. and Numa, K. Personal Knowledge Publishing Suite with Weblog. In the Proceedings of the WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.

[89]. Sauer, I., Bialek, D., Efimova, E., Schwartlander, R., Pless, G., and Neuhaus, P. "Blogs" and "Wikis" Are Valuable Software Tools for Communication Within Research Groups. The Journal of Artificial Organs, Vol. 29, Issue 1, pp. 82-83, 2005.

[90]. Smith, H., and McKeen, J. Enterprise Content Management: What's Your Strategy?. Queen's Centre for Knowledge-Based Enterprises, WP 03-08, 2003. http://business.queensu.ca/knowledge/workingpapers/working/working_03-08.pdf.