ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
(ΠΡΩΗΝ)ΤΜΗΜΑ ΕΦΑΡΜΟΓΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗ ΔΙΟΙΚΗΣΗ ΚΑΙ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ
ΤΜΗΜΑ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

"Αξιολόγηση Λογισμικών Ανοικτού Κώδικα για την Εξόρυξη Δεδομένων στον Παγκόσμιο Ιστό"

ΧΡΥΣΟΒΕΡΓΗΣ Π. ΑΘΑΝΑΣΙΟΣ **ΚΑΝΑΚΗΣ Π. ΣΠΥΡΑΓΓΕΛΟΣ**

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΑΝΤΖΟΥΛΑΤΟΣ Σ. ΓΕΡΑΣΙΜΟΣ

ΠΑΤΡΑ 2015

TECHNOLOGICAL EDUCATIONAL INSTITUTE OF WEST GREECE

COMPUTER SCIENCE IN ADMINISTRATION AND ECONOMY

DIPLOMA THESIS

" Evaluation of Open Source Software for Data Mining on the Web "

CHRYSOVERGIS P. ATHANASIOS **KANAKIS P. SPYRAGGELOS**

SUPERVISOR: ANTZOULATOS S. GERASIMOS

PATRAS 2015

ΕΥΧΑΡΙΣΤΙΕΣ

Θα θέλαμε, αρχικά, να εκφράσουμε τις θερμές μας ευχαριστίες στον καθηγητή μας Κύριο Αντζουλάτο Γεράσιμο, επιβλέποντα της πτυχιακής αυτής εργασίας, πρωταρχικά για την ευκαιρία που μας προσέφερε και την εμπιστοσύνη που μας έδειξε αναθέτοντάς μας αυτή τη εργασία, την πολύτιμη καθοδήγηση του σε κάθε δύσκολο βήμα και στάδιο κατά την εκπόνηση της εργασίας, την αμέριστη συμπαράστασή του για την κάθε δυσκολία που ανέκυπτε στην πορεία καθώς και τις σπουδαίες γνώσεις και εμπειρίες που αποκομίσαμε κατά τη διάρκεια της συνεργασίας μας.

Επίσης, θα θέλαμε να ευχαριστήσουμε τους δικούς μας ανθρώπους που στάθηκαν δίπλα μας σε όλη τη διάρκεια της φοιτητικής ζωής μας, στηρίζοντάς μας ψυχολογικά, ηθικά και οικονομικά, δίνοντάς μας ταυτόχρονα κουράγιο σε κάθε δυσκολία που αντιμετωπίσαμε.

ΠΕΡΙΛΗΨΗ

Ο Παγκόσμιος Ιστός, εξαιτίας της ραγδαίας εξάπλωσής του τις τελευταίες δεκαετίες, έχει μετατραπεί στην μεγαλύτερη ευρέως προσπελάσιμη πηγή δεδομένων και πληροφοριών. Αυτή η διόγκωση του Παγκόσμιου Ιστού καθιστά ολοένα και πιο επιτακτική την ανάγκη ανάπτυξης σύγχρονων εργαλείων και τεχνικών Εξαγωγής Πληροφοριών από τον Ιστό (Web Mining) τα οποία έχουν στόχο την εξόρυξη χρήσιμης γνώσης για τον εκάστοτε χρήστη από τους διάφορους διαδικτυακούς τόπους. Εξαιτίας του τεράστιου όγκου των διαθέσιμων πληροφοριών, η αποδοτική εξόρυξη των τμημάτων αυτών των πληροφοριών που θα αφορούν τις συγκεκριμένες ανάγκες των εκάστοτε χρηστών βασίζεται στην κατασκευή προτύπων των χρηστών που βασίζονται στις δραστηριότητές τους, στη χρήση που κάνουν στον Παγκόσμιο Ιστό και τις λοιπές αλληλεπιδράσεις τους με το Διαδίκτυο.

Η Εξόρυξη στον Ιστό, ανάλογα με τα δεδομένα που χρησιμοποιούνται κατά τη διαδικασία της εξόρυξης, χωρίζεται σε τρεις διαφορετικές κατηγορίες: την Εξόρυξη Περιεχομένου στον Ιστό (Web Content Mining), την Εξόρυξη Δομής στον Ιστό (Web Structure Mining) και την Εξόρυξη Χρήσης στον Ιστό (Web Usage Mining). Η Εξόρυξη Περιεχομένου στον Ιστό αφορά στην εξαγωγή δεδομένων από ήμι-δομημένα και μη-δομημένα κείμενα, συντεταγμένα κυρίως σε HTML, με σκοπό την κατασκευή προτύπων, η Εξόρυξη Δομής στον Ιστό αφορά στην υποβοήθηση των χρηστών μέσω της εξόρυξης URLs και άλλων διαδικτυακών συνδέσεων ενώ η Εξόρυξη Χρήσης στον Ιστό εξάγει πληροφορίες σχετικά με την αλληλεπίδραση των χρηστών με το Διαδίκτυο.

Αυτή η ραγδαία διεύρυνση του Παγκόσμιου Ιστού οδήγησε, βέβαια, και σε υψηλούς ρυθμούς ανάπτυξης εφαρμογών, οι οποίες χρησιμοποιούν σε πολύ μεγάλο βαθμό τις τεχνικές της Εξόρυξης στον Ιστό. Οι εφαρμογές αυτές εντάσσονται, για παράδειγμα, στον ευρύ χώρο του ηλεκτρονικού εμπορίου (e-commerce) όπου ο καταναλωτής μπορεί πλέον να λαμβάνει πλήρως προσωποποιημένες υπηρεσίες, στις προσπάθειες προσωποποιημένων λογισμικών για τους χρήστες του διαδικτύου (προσωποποιημένοι περιηγητές και portals) καθώς και η κατανόηση κοινωνικών ομάδων που δημιουργούνται στα πλαίσια διαδικτυακών ιστότοπων όπως τα μέσα κοινωνικής δικτύωσης. Αυτές οι πολυπληθείς εφαρμογές είναι που έχουν οδηγήσει στην ανάπτυξη λογισμικών που παρέχουν σε ιδιώτες και επιχειρήσεις τη δυνατότητα να επωφεληθούν από τα πλεονεκτήματα της Εξόρυξης στον Ιστό.

Σκοπός της εργασίας αυτής είναι η παρουσίαση της έννοιας της Εξόρυξης στον Ιστό και των διάφορων μορφών που αυτή δύναται να λάβει καθώς και των ήδη υπαρχουσών τεχνικών που την εξυπηρετούν. Επίσης, θα γίνει παρουσίαση των επικρατέστερων λογισμικών που υπάρχουν με τη μορφή εμπορικού προϊόντος καθώς και αξιολόγηση αυτών.

ABSTRACT

The World Wide Web, through its rapid growth in the last few decades, has been transformed into the largest widely accessible source of knowledge and raw information. This growth increases the need of modern tool and Web Mining techniques' development in order users to be able to extract useful knowledge. Due to the enormous volume of information available, the effective extraction of it, concerning the exact needs of every World Wide Web user, is based on the construction of patterns for each and every user based on their actinities and their integrated usage of the Web.

Web Mining according to the data to be mined can be broadly divided into three different categories: Web Contect Mining, Web Structure mining and Web Usage Mining. Web Content Mining is the process of extracting data from semi-structural or non-structural documents, mostly written in HTML, in order to construct patterns. Web Structure Mining aims to assist users through the extraction of URLs and other internet connections. Finally, Web Usage Mining extracts information concerning the interaction of every user with the Web.

The aforementioned growth of the World Wide Web leads, of course, into high application development rates, applications using Web Mining techniques. These applications refer to several sectors of human activity, i.e. e-commerce, where every customer can receive fully personalized services, personalized portals and browsers and efforts to understand social groups' behavior in terms of social media. These numerous applications are the ones who have lead to the development of software that enable individuals and companies to take advantage of Web Mining.

This Thesis aims to present Web Mining and its categories as long as the techniques that serve it. Additionally, the authors will attempt to present the predominant open source software packages existing as commercial products and their evaluation.

ΠΕΡΙΕΧΟΜΕΝΑ

ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ ΚΑΙ ΕΙΚΟΝΩΝ

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

ΕΙΣΑΓΩΓΗ

Κάθε επιστημονική εξέλιξη και κάθε νέα ιδέα εξαρτώνται τόσο από το βαθμό στον οποίο το θεωρητικό της υπόβαθρο αναγκάστηκε να ξεπεράσει τα όριά του και να σχεδιάσει καινούρια αλλά και από τις εφαρμογές αυτών και τον τρόπο που επηρεάζουν την ανθρώπινη δραστηριότητα εντασσόμενες σε αυτή.

Υπό το πλαίσιο αυτής της κοινά αποδεκτής αλήθειας, θα προσπαθήσουμε να αναπτύξουμε τόσο το θεωρητικό υπόβαθρο της Εξόρυξης στον Ιστό όσο και να παρουσιάσουμε την υλοποίησή της υπό τη μορφή πακέτων λογισμικού.

Ωστόσο, οι δύσκολες οικονομικές συνθήκες που ο δυτικός κόσμος βιώνει στην εποχή μας, καθιστούν δύσκολη την προμήθεια τέτοιων πακέτων λογισμικού από τη μεριά πανεπιστημιακών ιδρυμάτων, μικρομεσαίων εταιρειών που προσπαθούν να επιβιώσουν ή ιδιωτών που προσπαθούν να αναπτύξουν δραστηριότητες έχοντας μικρή οικονομική δυνατότητα. Γι' αυτό το λόγο, στην παρούσα εργασία θα παρουσιάσουμε πακέτα λογισμικού ανοικτού κώδικα τα οποία παρέχονται στους χρήστες χωρίς οικονομικό αντάλλαγμα.

Εκτός, όμως, από το σαφές οικονομικό όφελος των πακέτων λογισμικού ανοικτού κώδικα, τα συγκεκριμένα προγράμματα έχουν και ένα ακόμη ενδιαφέρον: ως επί το πλείστων αφορούν χρήστες οι οποίοι μπορούν να τα εξελίξουν και να τα αναπτύξουν περεταίρω.

Η συγγραφή της παρούσας εργασίας έχει γίνει με τρόπο τέτοιο ώστε να επιτρέπεται η ανάγνωση κάθε κεφαλαίου, χωρίς να απαιτούνται γνώσεις από προαναφερθείσες παραγράφους. Επίσης, θα θέλαμε να σημειώσουμε πως η επιλογή των πακέτων λογισμικού έχει γίνει με υποκειμενικά κριτήρια, αφού δεν υπάρχει κάποια καταγεγραμμένη συγκριτική δοκιμή όλων των υπάρχοντών προγραμμάτων αυτού του τύπου.

INTRODUCTION

Every scientific development and every new idea depend on the extent that the corresponding theoretical background was obliged to overcome its boundaries and design new ones but also on their applications and the way they affect human activity when they integrate.

In the context of this commonly accepted truth, we will make the effort to present the theoretical background of Web Mining and its implementation in the form of software packages as well.

However, the unfavourable economic situation governing the developed and developing West make it difficult for Universities, small and medium size corporations who are trying to survive and individuals who are trying to develop new business activities to purchase such software. Hence, in this thesis we will present open source software, which is being distributed for free.

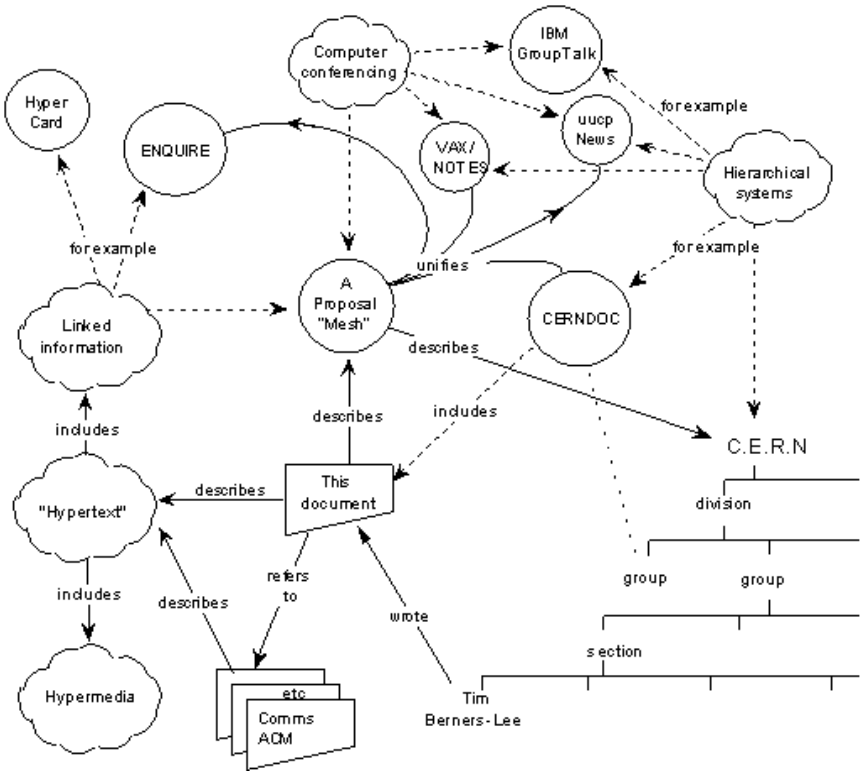
Except for the financial advantage this group of open source software exhibits, they are interesting in another aspect as well: they mostly refer to users who can further develop them.

The writing of this thesis has been done in such a way that every reader is allowed to study every chapter without any prior knowledge demanded. Moreover, we would like to note that the choice of the software packages to be presented has been done on subjective criteria since there is no comparative study on open source Web Mining software.

ΚΕΦΑΛΑΙΟ 1
ΕΙΣΑΓΩΓΗ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ ΚΑΙ ΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

1.1 Εισαγωγή στο Διαδίκτυο και τον Παγκόσμιο Ιστό [1, 26]

Το Μάρτιο του 1989, ο Tim Berners-Lee, τότε εργαζόμενος στο Ευρωπαϊκό Κέντρο Πυρηνικών Ερευνών στη Γενέυη της Ελβετίας (Centre Europeene pour La Recherche Nucleaire – CERN), αναγνωρίζοντας τα πλεονεκτήματα ενός συστήματος διαχείρισης και οργάνωσης της πληροφορίας που βασίζεται σε δομές υπερσυνδέσμων έναντι των ήδη υπαρχόντων ιεραρχικών συστημάτων, κατέθεσε στη διεύθυνση του ερευνητικού κέντρου την πρόταση του με τίτλο «Διαχείριση Πληροφορίας: Μια πρόταση». Έτσι, προτάθηκε για πρώτη φορά η κατασκευή ενός απλού πρωτοκόλλου το οποίο θα μπορεί να αναζητά πληροφορίες αποθηκευμένες σε απομακρυσμένα υπολογιστικά συστήματα μέσω απλών δικτύων. Επίσης, η πρόταση του περιελάβανε την περιγραφή ενός απλού τρόπου ώστε οι πληροφορίες να μπορούν να διακινούνται σε μια ενιαία μορφή και ταυτόχρονα να υπάρχει η δυνατότητα προσπέλασής τους μέσω υπερσυνδέσεων αυτών σε αρχεία άλλων χρηστών. Ουσιαστικά, η πρόταση αυτή οδηγούσε στην κατασκευή ενός συστήματος διανεμημένων υπερκειμένων, όπως αυτό που φαίνεται στην εικόνα που ακολουθεί.



Εικόνα 1.1: Ο Παγκόσμιος Ιστός, όπως αυτός προτάθηκε από τον Tim Berners-Lee. [26]

Η αρχική αντίδραση της επιστημονικής κοινότητας στην πρόταση του Berners-Lee ήταν ιδιαίτερα χλιαρή. Ωστόσο, ο ίδιος δε σταμάτησε να προωθεί την ιδέα του και

επιχείρησε επανακυκλοφορία της πρότασής του το 1990. Τότε ήταν που βρήκε την απαραίτητη υποστήριξη και ξεκίνησε να εργάζεται πάνω στις ιδέες του.

Ο Παγκόσμιος Ιστός δημιουργήθηκε, εν τέλει, το 1990. Η ονομασία που αποδόθηκε στο δημιούργημα αυτό ήταν "World Wide Web", ο πρώτος εξυπηρετητής ήταν ο httpd ενώ ταυτόχρονα κατασκευάστηκε και ο πρώτος περιηγητής, ο "WorldWideWeb". Αυτή θεωρείται και η «χρονική στιγμή μηδέν» της σύγχρονης ιστορίας του Παγκόσμιου Ιστού.

Τρία χρόνια μετά την παρουσίαση του Παγκόσμιου Ιστού από τον Berners-Lee, το 1993, ο Marc Andreesen και η ομάδα του λάνσαραν τον πρώτο περιηγητή για UNIX σε γραφικό περιβάλλον, το "Mosaic for X". Για τους επόμενους μήνες, δημιουργούνταν συνεχώς διάφορες εκδόσεις του λογισμικού αυτού για όλα τα λειτουργικά συστήματα (UNIX, Macintosh, Windows). Έτσι, τα τρία διασημότερα λειτουργικά συστήματα της εποχής είχαν πλέον τη δυνατότητα να προσφέρουν στους χρήστες τους μια εύκολη στη χρήση διεπαφή μέσω της οποίας μπορούσαν να περιηγηθούν στον Παγκόσμιο Ιστό με ένα μόνο κλικ. Απόρροια της προσπάθειας αυτής ήταν η δημιουργία της εταιρείας Mosaic Communications η οποία μετονομάστηκε σε Netscape Communications και κατασκεύασε το εμπορικό προϊόν του περιηγητή Netscape. Η δημιουργία του Παγκόσμιου Ιστού από τον Tim Berners-Lee και του Mosaic από τον Marc Andreesen θεωρούνται οι ακρογωνιαίοι λίθοι του οικοδομήματος του Παγκόσμιου Ιστού όπως το γνωρίζουμε σήμερα.

Ωστόσο, ο Παγκόσμιος Ιστός θα ήταν μια άχρηστη εφεύρεση αν δεν υπήρχε το Διαδίκτυο μέσω του οποίου γίνονται όλες οι επικοινωνίες που εξυπηρετούν τους σκοπούς ύπαρξης του Παγκόσμιου Ιστού. Η γέννηση του Διαδικτύου χρονολογείται περί την περίοδο του Ψυχρού Πολέμου με τη δημιουργία του δικτύου υπολογιστών ARPANET (Advanced Research Projects Agency NETwork), το οποίο δημιουργήθηκες στις Ηνωμένες Πολιτείες της Αμερικής από το αντίστοιχο Υπουργείο Εθνικής Άμυνας με σκοπό τη διατήρηση του ελέγχου των πυραύλων και των βομβαρδιστικών αεροσκαφών στην περίπτωση επίθεσης των αντιπάλων με πυρηνικά όπλα. Με τη βοήθεια του ARPANET, το οποίο χρησιμοποιήθηκε αρχικά το 1969 σε μια πρώιμη μορφή του, οι επιστήμονες συνέδεσαν πολυάριθμους ηλεκτρονικούς υπολογιστές εγκατεστημένους σε 40 διαφορετικές περιοχές. Ακολουθεί η ανάπτυξη του πρωτοκόλλου TCP/IP το 1973 από τους Vinton Cerf και Bob Kahn, ένα πρωτόκολλο το οποίο επιτρέπει τη διασύνδεση διαφορετικών δικτύων μεταξύ τους έτσι ώστε να μπορούν να επικοινωνούν και να ανταλλάσσουν δεδομένα μεταξύ τους. Οι δύο αυτές εφευρέσεις αποτελούν τη βάση γύρω από την οποία δομήθηκε το σύγχρονο Διαδίκτυο.

Ωστόσο, κατά τη συγκεκριμένη χρονική περίοδο, η κατάσταση που επικρατεί στην ανταλλαγή πληροφοριών είναι χαοτική. Τα δίκτυα τα οποία είχαν κατασκευαστεί ήταν πολυάριθμα, ο όγκος της πληροφορίας που είχε αρχίσει να συσσωρεύεται ήταν πολύ μεγάλος και η ανταλλαγή αυτής γινόταν σε παγκόσμιο επίπεδο. Έτσι, γεννήθηκε η επιτακτική ανάγκη από τη μεριά των χρηστών για τεχνικές εύρεσης πληροφοριών μέσα στον Παγκόσμιο Ιστό με τρόπο αποδοτικό και αποτελεσματικό. Η εκπλήρωση της ανάγκης αυτής οδήγησε στην ανάπτυξη μηχανών αναζήτησης. Η πρώτη από αυτές λανσαρίστηκε το 1993 από μια ομάδα φοιτητών και είχε την ονομασία Excite. Ακολούθησαν οι μηχανές αναζήτησης EINet Galaxy, Yahoo!, Lycos, Altavista και άλλες. Το 1998 λανσαρίστηκε η -ίσως- πιο επιτυχημένη μέχρι σήμερα μηχανή

αναζήτησης, το Google, ενώ 5 χρόνια αργότερα, το 2003, η Microsoft έθεσε στη διάθεση του κοινού τη μηχανή αναζήτησης MSN , γνωστή πλέον ως Bing.

Παρά την πολύ μεγάλη πρόοδο στις τεχνολογίες των μηχανών αναζήτησης, για λόγους που θα συζητηθούν εκτενώς στο επόμενο κεφάλαιο, η αναζήτηση δεδομένων μέσω αυτών δε δύναται να καλύψει πλήρως τις σύγχρονες ανάγκες των χρηστών. Έτσι, ξεκίνησε η ανάπτυξη τεχνικών οι οποίες μέσα από ένα σύνολο αρχείων έχουν τη δυνατότητα να εξορύξουν γνώση, τεχνικές οι οποίες ανήκουν σε ένα χώρο γνωστό ως Εξόρυξη Δεδομένων (Dart Mining). Στη συνέχεια, οι τεχνικές αυτές εφαρμόστηκαν σε αρχεία τα οποία βρίσκονται στον Παγκόσμιο Ιστό και οι αντίστοιχες τεχνικές ανήκουν στο χώρο της Εξόρυξης στον Ιστό, τεχνικές οι οποίες είναι μέρος της παρούσας εργασίας. Μια μικρή εισαγωγή στην εξόρυξη Δεδομένων θα γίνει σε ακόλουθη ενότητα ενώ η Εξόρυξη στον Ιστό θα συζητηθεί ενδελεχώς στο Κεφάλαιο 2.

1.2 Εξόρυξη Δεδομένων [1, 13]

1.2.1 Εισαγωγή

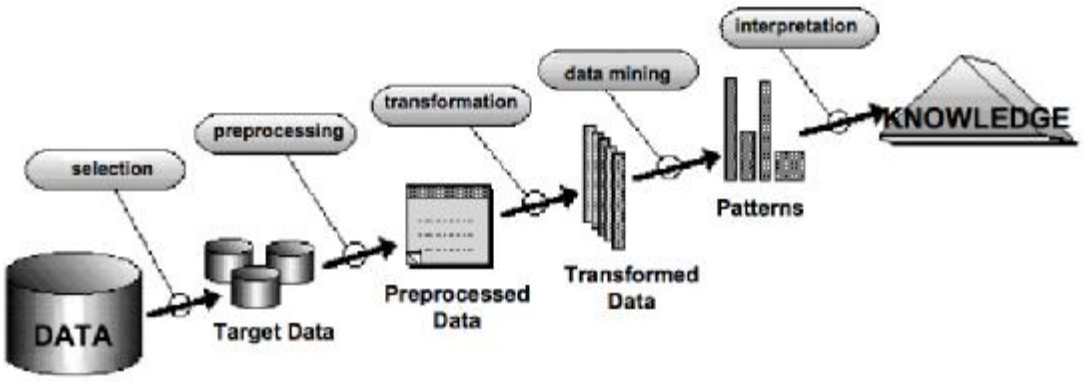
Στη σύγχρονη εποχή έχουμε την καθημερινή δημιουργία ενός τεράστιου όγκου πληροφοριών αφού υπάρχει η δυνατότητα -σχεδόν- αυτόματης καταγραφής κάθε τομέα της ανθρώπινης δραστηριότητας. Καθώς, λοιπόν, ο όγκος της συσσωρευόμενης στον Παγκόσμιο Ιστό πληροφορίας αυξάνεται, η κατανόησή της γίνεται συνεχώς πιο δύσκολη. Μέσω των τεχνολογιών πληροφοριακών συστημάτων είναι δυνατή η ηλεκτρονική διαχείριση μεγάλου όγκου πληροφοριών καθώς και η ανάδειξη πιθανώς χρήσιμης γνώσης καλά κρυμμένης μέσα σε αυτόν. Η Εξόρυξη Δεδομένων αποτελεί μια τεχνική για την ανάδειξης αυτής της καλά κρυμμένης γνώσης. Η δημοτικότητα της Εξόρυξης Δεδομένων αυξάνεται συνεχώς συγκριτικά με τις παραδοσιακές τεχνικές αναζήτησης γνώσης λόγω της δύναμής της να βρίσκει τις γνώσεις που αναζητεί ο εκάστοτε χρήστης σε πολύ μεγάλο όγκο δεδομένων.

Στον επιστημονικό χώρο έχουν επικρατήσει τρεις διαφορετικοί ορισμοί για την έννοια της Εξόρυξης Δεδομένων, οι οποίοι είναι οι παρακάτω:

> Η Εξόρυξη Δεδομένων είναι η διαδικασία της ανακάλυψης ουσιωδών νέων συσχετίσεων, προτύπων και τάσεων κατά τη διέλευση μέσω μεγάλων ποσοτήτων δεδομένων τα οποία είναι αποθηκευμένα σε καταθετήρια, χρησιμοποιώντας τεχνολογίες ανακάλυψης προτύπων και στατιστικών και μαθηματικών τεχνικών.
>
> Η Εξόρυξη Δεδομένων είναι η ανάλυση μεγάλων συνόλων δεδομένων με σκοπό την εύρεση ανύποπτων συσχετίσεων και τη συνόψιση αυτών με νέους τρόπους οι οποίοι είναι τόσο κατανοητοί όσο και χρήσιμοι στους κτήτορες των δεδομένων.
>
> Η Εξόρυξη Δεδομένων είναι ένα διεπιστημονικό πεδίο το οποίο συγκεντρώνει τεχνικές από τη μηχανική μάθηση, την ανακάλυψη προτύπων, τη στατιστική, τις βάσεις δεδομένων και τις μεθόδους

οπτικοποίησης για την αντιμετώπιση του ζητήματος της εξόρυξης πληροφοριών από βάσεις δεδομένων.

Η διαδικασία της Εξόρυξης Δεδομένων χωρίζεται σε 4 σημαντικά και διακριτά στάδια: τη συλλογή δεδομένων, την προετοιμασία των δεδομένων, την ανάδειξη προτύπων και την ανάλυση των προτύπων. Η συλλογή των δεδομένων περιλαμβάνει τη δημιουργία ενός συνόλου δεδομένων από διάφορες πηγές βάσει συγκεκριμένων χαρακτηριστικών που προσδιορίζονται από τις ανάγκες του εκάστοτε χρήστη. Οι συλλεχθείσες πληροφορίες, εξαιτίας του ότι προέρχονται από διάφορες πηγές είναι ετερογενείς και γι' αυτό το λόγο χρειάζεται να ομαλοποιηθεί το συγκεκριμένο σύνολο και τα δεδομένα στο τέλος να αναπαρασταθούν με μορφή που να τα καθιστά πιο διαχειρίσιμα. Η διαδικασία αυτή είναι το δεύτερο από τα προαναφερθέντα στάδια, η προετοιμασία των δεδομένων. Στη συνέχεια, γίνεται προσπάθεια εύρεσης κοινών στοιχείων μεταξύ των συγκεκριμένων αρχείων, τα οποία δημιουργούν ένα κοινό πρότυπο των αρχείων την συλλογής. Τέλος, στο στάδιο της ανάλυσης των προτύπων, το πρότυπο που κατασκευάσαμε το μορφοποιούμε έτσι ώστε να λειτουργεί σαν εργαλείο το οποίο εφαρμοζόμενο βοηθάει στην Εξόρυξη Δεδομένων. Σχηματικά, η Εξόρυξη Δεδομένων φαίνεται στην εικόνα που ακολουθεί.



Εικόνα 1.2: Τα στάδια της Εξόρυξης Δεδομένων. [1]

Ανάλογα με τους σκοπούς της, η Εξόρυξη Δεδομένων χωρίζεται κατά κύριο λόγο σε δύο μεγάλες κατηγορίες. Η πρώτη από αυτές αποσκοπεί στην κατασκευή ενός περιγραφικού μοντέλου (descriptive model). Ένα περιγραφικό μοντέλο αναπαριστά με συνοπτικό τρόπο τα κύρια χαρακτηριστικά που είναι κοινά μέσα σε ένα σύνολο δεδομένων. Ουσιαστικά, είναι το σύνολο των σημείων των δεδομένων τα οποία καθιστούν δυνατή τη μελέτη του. Τυπικά, ένα περιγραφικό μοντέλο κατασκευάζεται με μη διευθυνόμενες μεθόδους (undirected data mining) με αποτέλεσμα τη δημιουργία προτύπων τα οποία, όμως, αναμένουν το χρήστη για να ερμηνευτούν. Η δεύτερη κατηγορία είναι αυτή που αποσκοπεί στην κατασκευή ενός μοντέλου πρόβλεψης (predictive model). Ένα μοντέλο πρόβλεψης επιτρέπει στο χρήστη να προβλέψει την τιμή μιας συγκεκριμένης μεταβλητής σε κάποια συγκεκριμένη χρονική στιγμή στο μέλλον. Η μεταβλητή αυτή αποκαλείται μεταβλητή-στόχος (target variable). Ένα μοντέλο πρόβλεψης δημιουργείται χρησιμοποιώντας γνωστές τιμές των μεταβλητών, πιθανότατα και τιμές αυτών από χρονικές στιγμές στο παρελθόν ενώ καθορίζεται εξαρχής το ανεχόμενο σφάλμα των τιμών που προβλέπονται.

1.2.2 Τεχνικές Εξόρυξης

Οι τεχνικές Εξόρυξης Δεδομένων εφαρμόζονται μετά την προετοιμασία του συνόλου των αρχείων μέσα από το οποίο θα προσπαθήσουμε να εξάγουμε τη γνώση. Σκοπός τους είναι η ανακάλυψη των προτύπων που κρύβονται πίσω από αυτά τα δεδομένα βάσει συγκεκριμένων χαρακτηριστικών. Οι τεχνικές αυτές είναι χωρισμένες σε έξι μεγάλες κλάσεις, οι οποίες και περιγράφονται συνοπτικά ακολούθως:

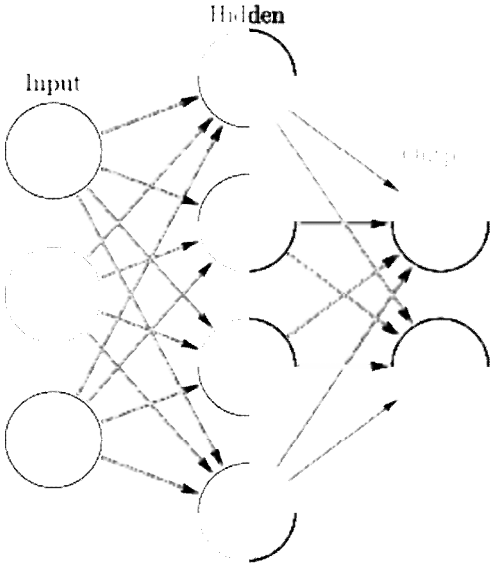
1. Ταξινόμηση: Η ταξινόμηση είναι μια τεχνική όπου καθορίζουμε εξαρχής τον αριθμό των κατηγοριών ή των τάξεων στις οποίες θα εντάξουμε τα αρχεία μας, πριν ξεκινήσει η διαδικασία ανάδειξης των υπάρχοντών προτύπων. Εδώ ορίζεται η μεταβολή τάξης (class variable) η οποία παίρνει τιμές ανάλογα με την υπό συζήτηση τάξη. Αυτή η τεχνική χαρακτηρίζεται ως «μάθησης με επίβλεψη» αφού πρώτα κατασκευάζεται ένα υποτυπώδες πρότυπο βασισμένο σε ένα σύνολο δεδομένων τα οποία ονομάζονται «δεδομένα εκπαίδευσης» (training data). Το πρότυπο αυτό στη συνέχεια εφαρμόζεται σε αυτά τα δεδομένα εκπαίδευσης έτσι ώστε να διαφανούν οι ανακρίβειες και τα πιθανά λάθη κατά την εφαρμογή της ταξινόμησης. Η διαδικασία αυτή είναι γνωστή ως υπερπροσαρμογή (overfitting) και αποσκοπεί στην κατασκευή μιας γνωστής δομής που θα χαρακτηρίζει τη συμπεριφορά και τη δράση της μεθόδου της ταξινόμησης, μια μέθοδος η οποία στη συνέχεια θα γενικευτεί σε νέα-άγνωστα δεδομένα.
2. Ομαδοποίηση: Αυτή είναι η κλάση η οποία περιλαμβάνει όλες τις διαδικασίες οι οποίες αποσκοπούν στην εύρεση στοιχείων που έχουν κοινά χαρακτηριστικά. Αυτές είναι τεχνικές μάθησης χωρίς επίβλεψη αφού δε γνωρίζουμε εκ των προτέρων τον αριθμό των ομάδων στις οποίες θα καταταγούν τα αρχεία μας αλλά ο αριθμός αυτός οριστικοποιείται με το τέλος της εφαρμογής τους. Η ομαδοποίηση τοποθετεί τα αρχεία σε ομάδες ή τάξεις βασιζόμενη στην αρχή της μεγιστοποίησης του βαθμού ομοιότητας εντός μιας ομάδας ή τάξης και ελαχιστοποίησης αυτού μεταξύ διαφορετικών εξ ’αυτών. Οι ομάδες οι οποίες εν τέλει δημιουργούνται μπορεί να είναι πιθανολογικού χαρακτήρα, δίνοντας πάντα τη συγκεκριμένη πιθανότητα ένταξης του αρχείου δεδομένων στη συγκεκριμένη ομάδα.
3. Ανίχνευση ανωμαλιών: Η μέθοδος αφορά στην ανίχνευση δεδομένων τα οποία θεωρούνται ασυνήθιστα εξαιτίας μεταβολών που έχουν υποστεί ή εξαιτίας αποκλίσεων που παρουσιάζουν από τα υπόλοιπα δεδομένα της συλλογής. Αυτές οι ανωμαλίες ελέγχονται για να διαπιστωθεί αν πρόκειται για λάθη ή –στην ενδιαφέρουσα περίπτωση- για εξαιρετικά δεδομένα.
4. Συνόψιση: Χρησιμοποιεί μεθόδους μέσω των οποίων λαμβάνουμε μια πιο συμπαγή αναπαράσταση του συνόλου των δεδομένων, συμπεριλαμβανομένων εικονικών αναπαραστάσεων και γραπτών εκθέσεων.
5. Ανίχνευση κανόνων συσχέτισης: Ένας κανόνας συσχέτισης είναι μια δήλωση πιθανολογικού χαρακτήρα αναφορικά με τη συνύπαρξη συγκεκριμένων χαρακτηριστικών μέσα σε μια βάση δεδομένων. Κατά την ανίχνευση κανόνων συσχέτισης, ουσιαστικά γίνεται μια προσπάθεια εύρεσης πιθανών σχέσεων που να συνδέουν μεταξύ τους τις διάφορες μεταβλητές του προβλήματος. Η κλάση αυτή εφαρμόζεται κατά κύριο λόγο σε σύνολα δεδομένων αραιών συναλλαγών.

6. Παλινδρόμηση: Αποτελεί μια προσπάθεια εύρεσης μιας χαρακτηριστικής συνάρτησης που να μοντελοποιεί τα δεδομένα επιτυγχάνοντας το ελάχιστο δυνατό σφάλμα.

1.2.3 Μεθοδολογίες Εξόρυξης Δεδομένων

Οι μεθοδολογίες της Εξόρυξης Δεδομένων, οι οποίες όπως αναφέρθηκε παραπάνω, κατατάσσονται στις 6 προαναφερθείσες κλάσσεις είναι πολυάριθμες. Ωστόσο, οι πιο σημαντικές είναι τα νευρωνικά δίκτυα, τα δέντρα αποφάσεων, οι γενετικοί αλγόριθμοι και η εξαγωγή κανόνων.

Ένα νευρωνικό δίκτυο ή ένα τεχνητό νευρωνικό δίκτυο είναι ένα σύστημα το οποίο αναγνωρίζει πρότυπα και κάνει προβλέψεις. Οι τεχνικές Εξόρυξης Δεδομένων όπως τα νευρωνικά δίκτυα έχουν τη δυνατότητα να αναγνωρίζουν τα κοινά χαρακτηριστικά που υπάρχουν σε ένα σύνολο δεδομένων και να μοντελοποιούν τις σχέσεις που τα συνδέουν.

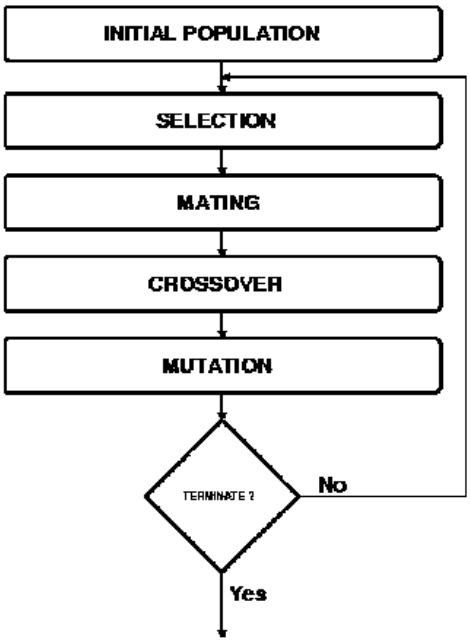


Εικόνα 1.3: Σχηματική αναπαράσταση ενός νευρωνικού δικτύου. Φαίνονται τα τρία κύρια στρώματα από τα οποία κατασκευάζεται, το στρώμα των δεδομένων εισόδου, οι έξοδοι καθώς και ένα ενδιάμεσο, κρυφό στον χρήστη στρώμα όπου επιτελούνται όλες οι διαδικασίες, όπως η φάση εκπαίδευσης. [13]

Αυτές οι τεχνικές, λοιπόν, χρησιμοποιούνται πλέον σε ευρεία κλίμακα στο χώρο των επιχειρήσεων αφού ουσιαστικά καθιστούν τις επιχειρηματικές εφαρμογές, των οποίων η απόδοση ενδιαφέρει ιδιαίτερα αφού είναι άρρηκτα συνδεδεμένη με οικονομικά οφέλη, ιδιαίτερα ευφυείς. Ωστόσο, τα νευρωνικά δίκτυα μειονεκτούν στο ότι κατασκευάζουν ιδιαίτερα πολύπλοκα μοντέλα τα οποία είναι κατά περιπτώσεις δύσκολο να γίνουν κατανοητά και να ερμηνευτούν. Τα τεχνητά νευρωνικά δίκτυα είναι προσαρμοστικά, μη γραμμικά συστήματα τα οποία ουσιαστικά μαθαίνουν να εφαρμόζουν στα δεδομένα μια συνάρτηση, την οποία και κατασκευάζουν κατά τη φάση εκπαίδευσης (training phase) με τη βοήθεια ενός συνόλου δεδομένων εκπαίδευσης. Στο τέλος της φάσης της εκπαίδευσης, με τις τιμές των παραμέτρων του μοντέλου γνωστές, τα τεχνητά νευρωνικά δίκτυα είναι έτοιμα να παρέχουν στο χρήστη το μοντέλο το οποίο χαρακτηρίζει το σύνολο των δεδομένων.

Μια, ακόμη, μεθοδολογία Εξόρυξης Δεδομένων είναι τα δέντρα αποφάσεων. Τα δέντρα αποφάσεων είναι διαγράμματα ροών όπου κάθε κόμβος αντιστοιχεί σε μια δοκιμή πάνω σε κάποια χαρακτηριστική τιμή, κάθε κλάδος αναπαριστά κάποιο από τα αποτελέσματα της δοκιμής αυτής ενώ τα «φύλλα» του δέντρου αντιστοιχούν στις κλάσεις. Είναι σαφές, λοιπόν, ότι το μοντέλο που κατασκευάζεται με τη μέθοδο αυτή εντάσσεται συνήθως στην κλάση της ταξινόμησης. Τα δέντρα αποφάσεων διαμερίζουν το σύνολο των εισόδων σε μικρές ομάδες, όπου κάθε μια από αυτές αντιστοιχεί σε μια κλάση, μέσω μιας διαδοχής δοκιμών. Έτσι, σε κάθε εσωτερικό κόμβο του δέντρου η μέθοδος ελέγχει την τιμή κάποιας μεταβλητής εισόδου και στη συνέχεια κατασκευάζονται οι κλάδοι στους οποίους αναγράφονται τα πιθανά αποτελέσματα των δοκιμών αυτών. Τερματικό σημείο της μεθόδου αυτής είναι τα φύλλα του δέντρου τα οποία προσδιορίζουν την κλάση των δεδομένων στην οποία θα καταλήξει ο χρήστης αν ακολουθήσει συγκεκριμένη σειρά αποφάσεων. Τα δέντρα αποφάσεων είναι επίσης μοντέλα προβλέψεων τα οποία χρησιμοποιούνται για τον καθορισμό της πορείας που θα ακολουθήσει ο χρήστης αναφορικά με κάποια ανάγκη του ανάλογα με τη στατιστική πιθανότητα που αποδίδεται σε κάθε διαφορετικό ενδεχόμενο. Εξαιτίας του εύληπτου σχήματός του το οποίο έχει τη μορφή δέντρου και της δυνατότητάς τους να παράγουν εύκολα κανόνες, τα δέντρα αποφάσεων είναι η προτιμώμενη μέθοδος για την κατασκευή εύκολα κατανοητών και ερμηνεύσιμων μοντέλων.

Συνεχίζοντας, συναντούμε τους γενετικούς αλγόριθμους οι οποίοι αποτελούν μια προσπάθεια ενσωμάτωσης ιδεών της Φύσης. Η κεντρική ιδέα πίσω από τους γενετικούς αλγόριθμους είναι πως η βέλτιστη λύση κατασκευάζεται αν καταφέρουμε να συνδυάσουμε όλα τα αξιόλογα τμήματα κάποιων άλλων λύσεων, όπως ακριβώς κάνει και η ίδια η Φύση κατά τη σύνθεση του DNA των έμβιων οργανισμών.



Εικόνα 1.4: Δομική αναπαράσταση ενός γενετικού αλγόριθμου. [1]

Σκοπός κάθε γενετικού αλγόριθμου είναι αν παρέχει στο αντίστοιχο πρόβλημα τη βέλτιστη δυνατή λύση. Είναι ο καλύτερος τρόπος αντιμετώπισης προβλημάτων για τα οποία έχουμε πολύ λίγες πληροφορίες και δύνανται να λειτουργούν σε γενικευμένα

περιβάλλοντα, με ελάχιστη εξειδίκευση, αφού κατασκευάζουν πολύ γενικούς αλγόριθμους. Η μόνη πληροφορία την οποία χρειάζεται η μέθοδος αναφορικά με τη λύση αυτού του ανεπαρκώς ορισμένου προβλήματος είναι η κατάσταση που θα προκύψει όταν επιλεχθεί η βέλτιστη λύση. Έτσι, γνωρίζοντας την επιθυμητή συμπεριφορά ενός συστήματος όταν αυτό επιλυθεί με το βέλτιστο τρόπο, ο γενετικός αλγόριθμος δύναται να παρέχει με πολύ υψηλά ποσοστά απόδοσης τη βέλτιστη αυτή λύση. Όταν ένας γενετικός αλγόριθμος χρησιμοποιείται κατ' αυτόν τον τρόπο, δηλαδή για την επίλυση ενός προβλήματος, τότε η διαδικασία αυτή χωρίζεται σε τρία διακριτά στάδια. Αρχικά, οι λύσεις του προβλήματος κωδικοποιούνται σε απλές αναπαραστάσεις, τα χρωμοσώματα. Στη συνέχεια, μια λειτουργία καταλληλόλητας (fitness function) αποφασίζει ποια από τις συγκεκριμένες λύσεις είναι οι καταλληλότερες και αυτές προτιμώνται όσον αφορά την επιβίωση και την αναπαραγωγή τους. Τέλος, διάφορες μεταλλαγές παράγουν μια νέα γενιά στοιχείων επανασυνδυάζοντας χαρακτηριστικά των πατρικών. Έτσι, μια νέα γενιά στοιχείων αναφορικά με την επίλυση του αρχικού προβλήματος επανατοποθετείται στην αρχή αυτού του κύκλου. Τερματικό σημείο της διαδικασίας ορίζεται αυτό στο οποίο θα οδηγηθούμε στη βέλτιστη λύση του προβλήματος.

Τέλος, θα αναφερθούμε στην τεχνική της εξαγωγής κανόνων συσχέτισης. Κατά τη μεθοδολογία αυτή, υπάρχουν τρία κριτήρια για την εκτίμηση των αλγορίθμων. Αυτά είναι το εύρος της χρήσης του αλγορίθμου, η μορφή εξάρτησης από το μαύρο κουτί (black-box) και η μορφή της εξαγωγής της περιγραφής. Αντίστοιχα, τα τρία αυτά κριτήρια αφορούν το αν ο αλγόριθμος χρησιμοποιείται από τη σκοπιά της παλινδρόμησης ή της ταξινόμησης, τον αλγόριθμο εξόρυξης που εφαρμόζεται στο μαύρο κουτή (εξαρτημένος ή ανεξάρτητος) και τους παραγόμενους κανόνες που αξίζουν προσοχής (πρόβλεψης έναντι περιγραφικών). Ωστόσο, είναι σημαντικό να τονιστεί πως θα πρέπει κάθε φορά να ικανοποιούνται τρία γενικά κριτήρια: η ποιότητα του εξαγόμενου κανόνα, η επεκτασιμότητα του αλγόριθμου καθώς και η συνέπεια αυτού. Ένας κανόνας, εν γένει, αποτελείται από δύο τιμές, μια προγενέστερη και μια επακόλουθη. Η προγενέστερη μπορεί να έχει είτε μια είτε πολλές προϋποθέσεις οι οποίες να ικανοποιούνται για δεδομένη ακρίβεια του προβλήματος ώστε να αληθεύει η επακόλουθη ενώ αυτή έχει μόνο μια προϋπόθεση. Έτσι, κατά την εξόρυξη ενός κανόνα από μια βάση δεδομένων στοχεύουμε στις τιμές της προγενέστερης και της επακόλουθης μεταβλητής, στην ακρίβεια και στην κάλυψη όλων των πιθανών περιπτώσεων. Βέβαια είναι αναγκαίο να επισημανθεί πως ενώ μπορούμε να κατασκευάσουμε πρότυπα μέσω της τεχνικής της εξαγωγής κανόνων, αυτό δε σημαίνει ότι πάντα το αριστερό μέρος (if part) όταν ικανοποιείται θα προκαλεί το δεξιό μέρος (then part). Στον τομέα της εξόρυξης δεδομένων, τα τεχνητά νευρωνικά δίκτυα καθώς και τα δέντρα αποφάσεων έχουν ένα σημαντικό μειονέκτημα, ότι παράγουν αδιαφανή μοντέλα. Προσπαθώντας να καλύψουμε την ανάγκη για διαφανή μοντέλα που να χαρακτηρίζονται από μεγάλη ακρίβεια, κατευθυνόμαστε στην επιλογή της τεχνικής της εξαγωγής κανόνων. Ακριβέστερα, έχει αποδειχθεί ότι η δυνατότητα επεξήγησης είναι μια λειτουργία ζωτικής σημασίας για τα συστήματα τεχνητής νοημοσύνης, σε σημείο που η ικανότητά τους να γεννούν ακόμη και τις υποτυπώδεις επεξηγήσεις να θεωρείται κρίσιμη για την αποδοχή τους ή μη. Αυτή την ανάγκη προσπαθεί να καλύψει η τεχνική της εξαγωγής κανόνων αφού από το γεγονός ότι σκοπός της πλειονότητας των συστημάτων εξόρυξης δεδομένων είναι η υποστήριξη διαδικασιών λήψης αποφάσεων είναι εμφανής η ανάγκη για παροχή επεξηγήσεων από αυτά.. Ωστόσο, πολλά συστήματα είναι αδιαφανή και έτσι θεωρούνται «μαύρα κουτιά». Όπως προαναφέρθηκε, η βασικότερη προϋπόθεση για

να είναι χρήσιμος ένας κανόνας είναι μαζί με αυτόν να παρέχεται και η ακρίβειά του αλλά και το ποσοστό κάλυψης των πιθανών ενδεχομένων (ουσιαστικά αναφερόμαστε στη συχνότητα χρησιμοποίησης του συγκεκριμένου κανόνα). Βέβαια, οι Craven και Shavlik όρισαν πέντε κριτήρια για την εξαγωγή κανόνων τα οποία έχουν ως ακολούθως:

- Σαφήνεια: Αφορά την έκταση στην οποία οι εξαγόμενες παραστάσεις είναι ανθρωπίνως κατανοητές.
- Πιστότητα: Αφορά το βαθμό στον οποίο οι εξαγόμενες αναπαραστάσεις μοντελοποιούν επακριβώς τα δίκτυα από τα οποία εξήχθησαν.
- Ακρίβεια: Αποτελεί μέτρο τη δυνατότητας της εξαγόμενης αναπαράστασης να κάνει προβλέψεις σε περιπτώσεις που δεν έχουν μελετηθεί στο παρελθόν.
- Επεκτασιμότητα: Αφορά στην ικανότητα της μεθόδου να επεκταθεί σε μεγαλύτερα δίκτυα με μεγαλύτερο αριθμό συνδέσεων.
- Γενικότητα: Αφορά την έκταση στην οποία η χρήση της μεθόδου απαιτεί ειδική εκπαίδευση.

ΚΕΦΑΛΑΙΟ 2
ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΙΣΤΟ

2.1 Εισαγωγή

Ο Παγκόσμιος Ιστός μπορεί να περιγράφει συνοπτικά ως ένα ογκώδες οικοδόμημα κατασκευασμένο από ετερογενείς και διαμοιρασμένες σε διάφορες περιοχές του πληροφορίες. Ωστόσο, το γεγονός ότι αυτές οι πληροφορίες είναι τόσο «άτακτα» διαμοιρασμένες στο χώρο που ο Παγκόσμιος Ιστός ορίζει, καθιστά αδύνατη την πλήρη εκμετάλλευση της δύναμης τους. Αυτή η κατάσταση δεν υπήρξε αντικείμενο έντονης μελέτης κατά το χρονικό διάστημα όπου το Διαδίκτυο και οι υπηρεσίες του αφορούσαν συγκεκριμένες μόνο ομάδες του πληθυσμού. Ωστόσο, στις μέρες μας όπου και η εξάπλωση του σε κάθε είδους ανθρώπινη δραστηριότητα εξελίσσεται μανιωδώς, είναι επιτακτική η παροχή συγκεκριμένων εργαλείων προς τους χρήστες που θα τους προσφέρουν τη δυνατότητα να ανακαλύψουν γνώση μέσα από αυτό με τρόπο αποδοτικό και αποτελεσματικό υπό οποιαδήποτε σκοπιά.

Ειδικότερα, εξαιτίας του όγκου και της ποικιλίας των πληροφοριών που περιέχονται στο διαδίκτυο αλλά και τη δυναμική φύση του ίδιου του Παγκόσμιου Ιστού, ο χρήστης είναι ουσιαστικά υπερφορτωμένος από πληροφορίες. Αυτή η υπερφόρτωση του δημιουργεί σημαντικά προβλήματα όταν επιθυμεί να αλληλοεπιδράσει με το διαδίκτυο. Η αλληλεπίδραση αυτή μπορεί να πάρει αρκετές μορφές. Αρχικά, η πιο διαδεδομένη μορφή αλληλεπίδρασης χρήστη με τον Παγκόσμιο Ιστό είναι αυτή που αφορά στην αναζήτηση πληροφοριών. Εδώ, ο χρήστης είτε περιηγείται στον Ιστό είτε χρησιμοποιεί κάποια υπηρεσία αναζήτησης (search service ή search engine). Στη δεύτερη περίπτωση, ο χρήστης εισάγει λέξεις-κλειδιά που συνδέονται με την αναζήτησή του και λαμβάνει μια λίστα ιστοσελίδων ταξινομημένη βάσει της σχετικότητάς τους με τις λέξεις-κλειδιά. Η υπηρεσία αυτή μειονεκτεί σημαντικά όσον αφορά στην ακρίβειά της αλλά και στην ανάκληση των σχετικών ιστοσελίδων αφού η κατάταξη σε ομάδες του συνόλου της πληροφορίας στον Παγκόσμιο Ιστό είναι πρακτικά αδύνατη. Επίσης, ένας χρηστής είναι πιθανό να επιθυμεί να δημιουργήσει νέα γνώση μέσα από τις φαινομενικά ασύνδετες πληροφορίες που μπορεί να συναντήσει στο Διαδίκτυο. Η διαφορά αυτής της αλληλεπίδρασης με την προηγούμενη είναι πως στη συγκεκριμένη περίπτωση θεωρούμε δεδομένη την ύπαρξη ενός συνόλου δεδομένων στον Ιστό και από αυτά επιθυμούμε να εξάγουμε πιθανά χρήσιμη γνώση. Αντίθετα, κατά την προηγούμενη αλληλεπίδραση δεν έχουμε δεδομένο όγκο πληροφορίας αλλά αυτός μεταβάλλεται ανάλογα με τις λέξεις-κλειδιά που εισάγουμε. Μια τρίτη μορφή που λαμβάνει η αλληλεπίδραση του χρήστη με το Διαδίκτυο είναι αυτή της προσωποποίησης πληροφοριών, με δεδομένο ότι οι χρήστες διαφέρουν μεταξύ τους όσον αφορά στο περιεχόμενο της πληροφορίας που επιθυμούν να τους παρουσιαστεί αλλά και στον τρόπο με τον οποίο επιθυμούν να γίνει η παρουσίαση αυτή. Τέλος, μια από τις εξίσου διαδεδομένες μορφές αλληλεπίδρασης των χρηστών με το Διαδίκτυο είναι όταν επιθυμούν να μάθουν πληροφορίες για τους καταναλωτές ή για άλλους μεμονωμένους χρήστες. Η εξαγωγή αυτής της γνώσης σχετίζεται με την προαναφερθείσα μορφή αλληλεπίδρασης που επιθυμεί τη δημιουργία προσωποποιημένων πληροφοριών αφού η προσωποποίηση της πληροφορίας μας δείχνει ουσιαστικά τις επιθυμίες των χρηστών.

Τα προβλήματα που υπεισέρχονται σε αυτές τις μορφές αλληλεπίδρασης μπορούν να επιλυθούν είτε άμεσα, είτε έμμεσα με χρήση τεχνικών Εξόρυξης στον Ιστό. Ως άμεση επίλυση ονομάζουμε τη χρήση εφαρμογών που βασίζονται σε τεχνικές Εξόρυξης στον Ιστό για την απευθείας επίλυση των προβλημάτων αυτών. Αντιθέτως, έμμεση θεωρείται η προσέγγιση κατά την οποία οι τεχνικές Εξόρυξης στον Ιστό αποτελούν μέρος ενός μεγαλύτερου και πολυπλοκότερου προγράμματος το οποίο πιθανότατα να συναντήσει κατά τη λειτουργία του τα προβλήματα αυτά.

Πριν την ανάπτυξη, ωστόσο, της τεχνικής της Εξόρυξης στον Ιστό είχε ήδη αναπτυχθεί η Εξόρυξη Δεδομένων. Η Εξόρυξη Δεδομένων χρησιμοποιείται για την ανακάλυψη ενός έγκυρου, καινοτομικού, πιθανότατα χρήσιμου και καθόλα κατανοητού προτύπου σε μια συλλογή δεδομένων εντός μιας βάσης δεδομένων. Η τεχνική αυτή, όμως, αδυνατεί να λειτουργήσει εύρυθμα όταν εφαρμόζεται σε μια βάση μη δομημένων και ετερογενών πληροφοριών εντός του Διαδικτύου. Έτσι, η προγραμματιστική κοινότητα οδηγήθηκε στην ανάπτυξη μιας προηγμένης τεχνικής που να ανταποκρίνεται στις προαναφερθείσες προσδοκίες, την Εξόρυξη στον Ιστό που θα συζητηθεί στις επόμενες ενότητες.

2.2 Ορισμός της Εξόρυξης στον Ιστό [1, 2, 3]

Η Εξόρυξη στον Ιστό είναι μια πολυσχιδής τεχνολογία η οποία συνδυάζει και συντονίζει αρκετά ερευνητικά πεδία, όπως η Εξόρυξη Δεδομένων, η Υπολογιστική Γλωσσολογία, η Στατιστική, η Πληροφορική και άλλα. Παρά το γεγονός ότι η επιστημονική κοινότητα δεν έχει καταλήξει σε ένα σαφή ορισμό της τεχνολογίας αυτής, ο πιο γενικός και συνήθης είναι ο ακόλουθος:

> Η Εξόρυξη στον Ιστό είναι η διαδικασία της ανακάλυψης προτύπων p που περιέχονται σε μια μεγάλη συλλογή εγγράφων C και η οποία μπορεί να αναπαρασταθεί ως $\xi: C \rightarrow p$. [2]

Αυτός ο ορισμός μοιάζει αρκετά όμοιος με αυτόν της Εξόρυξης Δεδομένων. Ωστόσο, υπάρχουν συγκεκριμένες ειδοποιοί διαφορές μεταξύ αυτών των δύο τεχνικών που βασίζονται σε συγκεκριμένα μοναδικά χαρακτηριστικά της Εξόρυξης στον Ιστό. Αρχικά, η πηγή δεδομένων της Εξόρυξης στον Ιστό είναι το Διαδίκτυο. Έτσι, ενώ η εξόρυξη σε βάσεις δεδομένων, διαδικτυακές καταγραφές και προφίλ χρηστών βασίζεται απόλυτα στην Εξόρυξη Δεδομένων, η Εξόρυξη στον Ιστό χρησιμοποιεί τον παγκόσμιο Ιστό ως ενδιάμεσο λογισμικό. Επίσης, ο Παγκόσμιος Ιστός αποτελείται από κόμβους αρχείων και υπερσυνδέσμους. Έτσι, κάποιο πρότυπο το οποίο θα αναγνωριστεί μπορεί να οφείλεται είτε στο περιεχόμενο του αρχείου αυτού είτε στη δομή του Ιστού. Τέλος, τα αρχεία που βρίσκονται διαθέσιμα στο Διαδίκτυο είναι είτε ημι-δομημένα είτε μη-δομημένα. Αντίθετα, τα αρχεία που αφορούν την Εξόρυξη Δεδομένων σε μια βάση δεδομένων είναι πάντα δομημένα. Έτσι, κάποιες από τις παραδοσιακές τεχνικές Εξόρυξης Δεδομένων δε μπορούν να εφαρμοστούν στην Εξόρυξη στον Ιστό και άλλες, ενώ δύνανται να εφαρμοστούν απαιτούν προεπεξεργασία των αρχείων.

Από τα παραπάνω είναι εύκολο να αναρωτηθεί κάποιος ποια η διαφορά της Εξόρυξης στον Ιστό με την Αναζήτηση Πληροφοριών στον Ιστό. Ο ορισμός της δεύτερης είναι ο παρακάτω:

> Η Αναζήτηση Πληροφοριών στον Ιστό είναι η διαδικασία εύρεσης ενός υποσυνόλου S από έναν επαρκή αριθμό αρχείων σχετικά με ένα ερώτημα q μέσα από μία μεγάλη συλλογή αρχείων C και η οποία μπορεί να αναπαρασταθεί ως $\xi: (C, q) \rightarrow S$. [2]

Η μεγάλη διαφορά μεταξύ της Εξόρυξης στον Ιστό και της Αναζήτησης Πληροφοριών στον Ιστό είναι πως έχουν ολοκληρωτικά διαφορετικούς στόχους. Παρά το γεγονός ότι η Εξόρυξη στον Ιστό είναι πιο σύγχρονη τεχνολογία, δεν αποσκοπεί στην αντικατάσταση της Αναζήτησης Πληροφοριών στον Ιστό. Ακριβέστερα, δεν την αντικαθιστά αλλά οι δύο αυτές τεχνικές αλληλοσυμπληρώνονται. Κάθε μια τους έχει τα δικά της πλεονεκτήματα και τις δικές της χαρακτηριστικές εφαρμογές. Ωστόσο, αυτό που εξασφαλίζει η Εξόρυξη στον Ιστό ως μεταγενέστερη τεχνολογία είναι μεγαλύτερη ακρίβεια στην αναζήτηση γνώσης καθώς και μια προηγμένη οργάνωση αυτής, γεγονός το οποίο εισάγει τα συστήματα αναζήτησης πληροφοριών στην επόμενη γενιά.

2.3 Ταξινόμηση της Εξόρυξης στον Ιστό [1, 2, 3, 9]

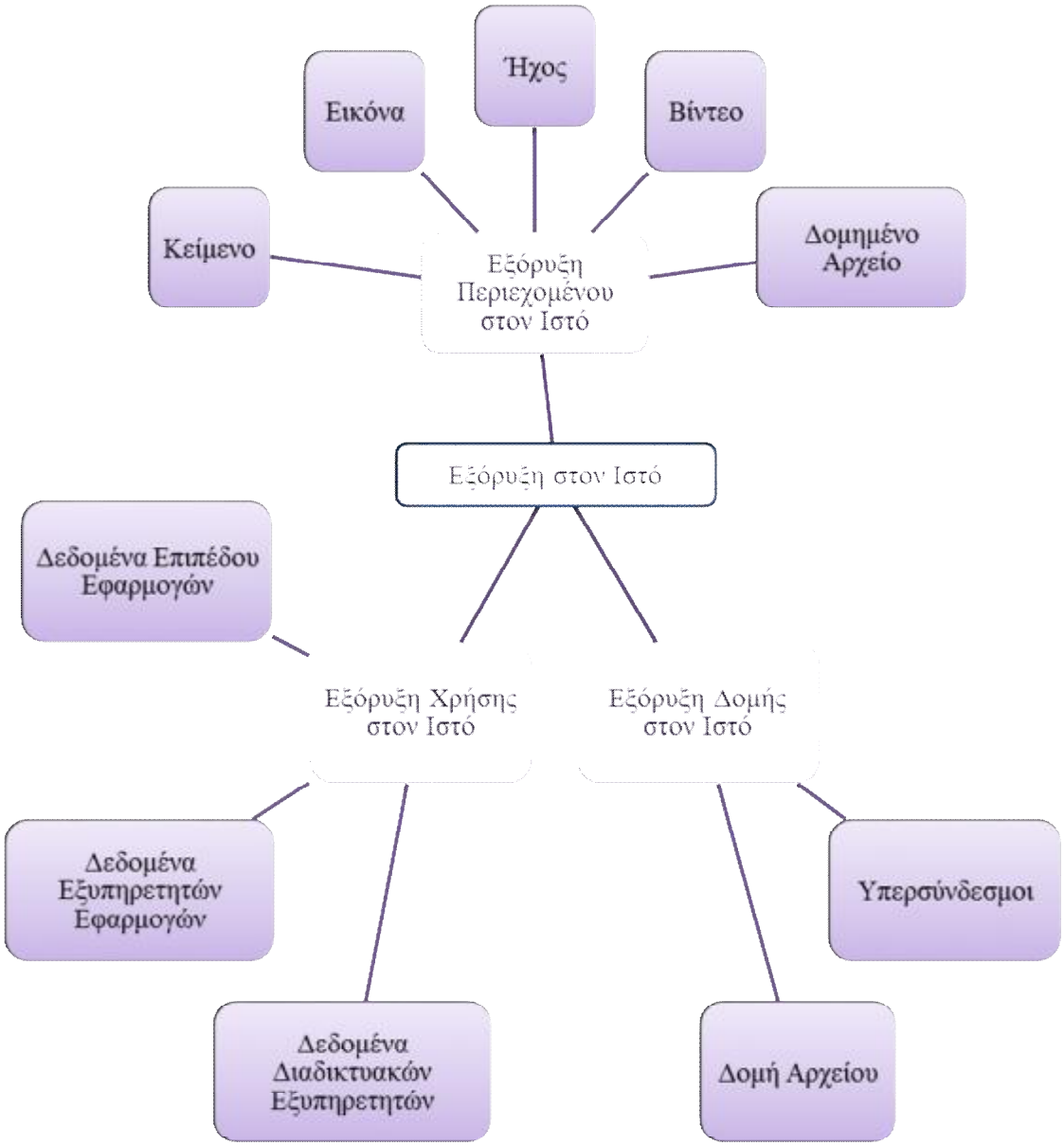
Ανάλογα με το είδος των προς εξόρυξη δεδομένων, η Εξόρυξη στον Ιστό χωρίζεται σε τρεις κατηγορίες, όπως φαίνεται στο Σχήμα 1. Αυτές οι κατηγορίες περιγράφονται ακολούθως.

- **Εξόρυξη Περιεχομένου στον Ιστό**

Η Εξόρυξη Περιεχομένου στον Ιστό είναι η διαδικασία με την οποία εξάγεται χρήσιμη πληροφορία μέσα από τα περιεχόμενα των διαδικτυακών αρχείων. Τα δεδομένα-περιεχόμενο ουσιαστικά είναι μια συλλογή στοιχείων τα οποία μια σελίδα έχει σχεδιαστεί να περιλαμβάνει. Αυτά μπορεί να είναι στοιχεία που περιλαμβάνουν είτε ήχο, είτε εικόνα, είτε βίντεο, είτε κείμενο είτε ακόμη και δομημένα αρχεία όπως είναι για παράδειγμα οι λίστες και οι πίνακες. Η Εξόρυξη Περιεχομένου στον Ιστό είναι ο πιο ευρέως διερευνώμενος τύπος Εξόρυξης στον Ιστό, έρευνα η οποία περιλαμβάνει θέματα όπως η ανακάλυψη και παρακολούθηση ενός θέματος (discovery and tracking), η εξαγωγή προτύπων συσχέτισης (association patterns) και η κατηγοριοποίηση των ιστοσελίδων. Η έρευνα στα πλαίσια της Εξόρυξης Περιεχομένου στον Ιστό υποβοηθείται έντονα από τεχνικές που έχουν αναπτυχθεί για διάφορους άλλους επιστημονικούς κλάδους όπως η Αναζήτηση Πληροφοριών (Information Retrieval – IR), οι Βάσεις Δεδομένων (Database) και η Επεξεργασία σε Φυσική Γλώσσα (Natural Language Processing – NLP). Οι σύγχρονες τάσεις στην Εξόρυξη Περιεχομένου στον Ιστό την έχει χωρίσει σε δύο μεγάλες κατηγορίες, την Εξόρυξη Κειμένου (Text Mining) και την Εξόρυξη Πολυμέσων (Multimedia Mining). Παρά το γεγονός ότι η Εξόρυξη Πολυμέσων στον Ιστό έλκει όλο και μεγαλύτερο μερίδιο του συνολικού ενδιαφέροντος, η Εξόρυξη Κειμένου είναι ο θεμέλιος λίθος της διαδικασίας καθώς και η πιο σημαντική διεργασία αφού παραδοσιακά το κείμενο είναι ο κινητήριος μοχλός της διακίνησης πληροφορίας. Εκτός αυτού, η Εξόρυξη Πολυμέσων, εξαιτίας της ύπαρξης εικόνας, αφορά σε πολύ μεγαλύτερο βαθμό τους κλάδους της Επεξεργασίας Εικόνας (Image Processing) και της Όρασης Υπολογιστών (Computer vision), με αποτέλεσμα τα πολυμέσα να μην αφορούν τόσο έντονα τον κλάδο της Εξόρυξης στον Ιστό.
Η Εξόρυξη Περιεχομένου στον Ιστό μπορεί να θεωρηθεί από δύο διαφορετικές

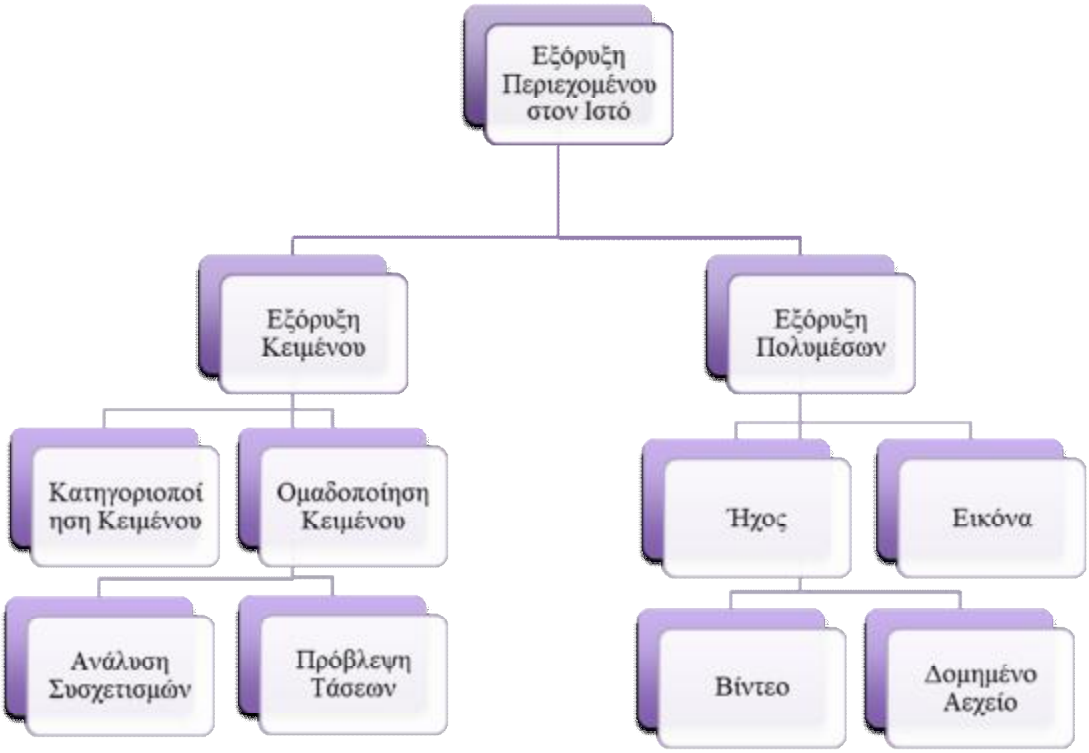
σκοπιές, από αυτή της Αναζήτησης Πληροφοριών και από αυτή των Βάσεων Δεδομένων. Στην πρώτη περίπτωση, σκοπός είναι η υποβοήθηση της διαδικασίας αναζήτησης πληροφοριών με την ταυτόχρονη βελτίωση της ποιότητάς τους μέσω ενός φιλτραρίσματος που έχει ως κριτήρια τις επιθυμίες του εκάστοτε χρήστη. Στη δεύτερη περίπτωση, γίνεται η προσπάθεια μοντελοποίησης των δεδομένων που υπάρχουν στο Διαδίκτυο και η οργάνωσή τους σε σύνολα τα οποία θα μπορούν να αποτελούν την έξοδο σε πιο ραφιναρισμένες αναζητήσεις από αυτές που απλά περιλαμβάνουν λέξεις-κλειδιά.

Προχωρώντας την ταξινόμηση περαιτέρω, μπορούμε να διαχωρίσουμε τα διάφορα είδη Εξόρυξης Κειμένου σε τέσσερεις νέες κατηγορίες με αποτέλεσμα αναφορικά με την Εξόρυξη Περιεχομένου στον Ιστό να παίρνουμε την ταξινόμηση που βλέπουμε στο Σχήμα 2.



Σχήμα 2.1: *Ταξινόμηση της Εξόρυξης στον Ιστό*

Η Κατηγοριοποίηση Κειμένου (Text Categorization) έχει ως σκοπό την κατάταξη κάθε αρχείου εντός μιας συλλογής C σε μία ή περισσότερες κατάλληλες κλάσεις, δεδομένης προ υπάρχουσας ταξινόμησης. Έτσι, οι χρήστες επωφελούνται κατά την αναζήτηση κειμένου από τη δυνατότητα που τους δίνεται να το αναζητούν μέσα σε συγκεκριμένη κλάση. Οι δύο γνωστότεροι αλγόριθμοι για την εκτέλεση αυτής της λειτουργίας είναι ο αλγόριθμος του k-Εγγύτερου Γείτονα και ο αλγόριθμος Naïve Bayes. Το σημείο που διαφοροποιεί την Κατηγοριοποίηση Κειμένου από την αμέσως επόμενη τεχνική, δηλαδή την Ομαδοποίηση Κειμένου, είναι πως στην ομαδοποίηση δεν ορίζεται κάποια ταξινόμηση εκ των προτέρων. Στόχος της ομαδοποίησης κειμένου είναι ο διαχωρισμός των κειμένων που περιλαμβάνονται σε μια συλλογή C σε ομάδες έτσι ώστε η ομοιότητα των χαρακτηριστικών μεταξύ των αρχείων διαφορετικών ομάδων να ελαχιστοποιείται και η ομοιότητα των χαρακτηριστικών μεταξύ των αρχείων της ίδιας ομάδας να μεγιστοποιείται. Προς αυτήν την κατεύθυνση έχουν προταθεί πολυάριθμοι αλγόριθμοι οι οποίοι μπορούν να καταταχθούν σε δύο κατηγορίες, αυτή της Ιεραρχικής Ομαδοποίησης (Hierarchical Clustering) και αυτή της Διαμεριστικής Ομαδοποίησης (Partitional Clustering). Η ανάλυση συσχετισμού αφορά στην εύρεση της σχέσης που υφίσταται μεταξύ φράσεων και μεταξύ λέξεων σε ένα κείμενο ενώ, τέλος, η τεχνική της Πρόβλεψης Τάσεων αφορά στη πρόβλεψη της τιμής που θα έχουν συγκεκριμένα δεδομένα σε συγκεκριμένη στιγμή στο μέλλον.



Σχήμα 2.2: *Το σύνολο των κατηγοριών Εξόρυξης Πληροφοριών στον Ιστό.*

- **Εξόρυξη Δομής στον Ιστό**

Η τυπική μορφή του διαγράμματος κάποιας θέσης στον Παγκόσμιο Ιστό αποτελείται από ιστοσελίδες υπό μορφή κόμβων και υπερσυνδέσμους που παίζουν το ρόλο ακμών

που συνδέουν σχετιζόμενες ιστοσελίδες μεταξύ τους. Η Εξόρυξη Δομής στον Ιστό αποτελεί τη διαδικασία της ανακάλυψης δομικών πληροφοριών εντός του Παγκόσμιου Ιστού, δηλαδή πληροφορίες που να καταδεικνύουν τον τρόπο με τον οποίο οι διαδικτυακές δομές συνδέονται μεταξύ τους. Αποτέλεσμα της Εξόρυξης Δομής είναι η ανακάλυψη του μοντέλου της τοπολογίας του εκάστοτε διαδικτυακού τόπου. Το μοντέλο αυτό, στη συνέχεια, μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση των Ιστοσελίδων και να βοηθήσει στη δημιουργία νέας πληροφορίας, όπως για παράδειγμα πληροφορία που αφορά στις ομοιότητες και στις διαφορές μεταξύ συγκεκριμένων ιστοσελίδων. Ανάλογα με το είδος των δομικών πληροφοριών που συμμετέχουν, η Εξόρυξη Δομής χωρίζεται σε δύο υποκατηγορίες, αυτή των Υπερσυνδέσεων και αυτή της Δομής Αρχείου. Ο Υπερσύνδεσμος είναι μια δομική μονάδα η οποία συνδέει μια συγκεκριμένη τοποθεσία εντός μιας ιστοσελίδας με μια δεύτερη τοποθεσία, είτε εντός της ίδιας ιστοσελίδας, είτε εντός κάποιας δεύτερης. Στην πρώτη περίπτωση, ο υπερσύνδεσμος ονομάζεται intra-document hyperlink ενώ στη δεύτερη inter-document hyperlink. Επίσης, το περιεχόμενο μιας ιστοσελίδας μπορεί να οργανωθεί υπό μορφή δέντρου, βασισμένη τις διάφορες HTML και XML ετικέτες που περιέχονται στη σελίδα.

Η Εξόρυξη Δομής στον Ιστό βρίσκει μεγάλη απήχηση στις προσπάθειες μελέτης των Μέσων Κοινωνικής Δικτύωσης. Μέσω της ανάλυσης αυτών μπορούμε να ανακαλύψουμε συγκεκριμένους τύπους ιστοσελίδων ανάλογα με τους εισερχόμενους και τους εξερχόμενους συνδέσμους. Η Εξόρυξη Δομής κάνει χρήση της δομής των υπερσυνδέσμων στο Διαδίκτυο με σκοπό την εφαρμογή της ανάλυσης των μέσων κοινωνικής δικτύωσης ώστε να μοντελοποιηθεί η υφιστάμενη δομή του ίδιου του Παγκόσμιου Ιστού. Εφαρμογές τέτοιου τύπου αφορούν στον υπολογισμό της σειρά εμφάνισης κάποιας ιστοσελίδας αλλά και της σχετικότητάς της με κάποιο συγκεκριμένο ερώτημα-αναζήτηση, η κατηγοριοποίηση ιστοσελίδων, η ανακάλυψη μικροκοινοτήτων στον Ιστό, η μέτρηση του ποσοστού στο οποίο μια ιστοσελίδα μπορεί να θεωρηθεί ολοκληρωμένη και η ανάδειξη της ιεραρχίας των υπερσυνδέσμων ιστοσελίδων ενός συγκεκριμένου domain ώστε να είναι η δυνατή η μελέτη του τρόπου με τον οποίο η ροή πληροφορίας επηρεάζει το σχεδιασμό των ιστοσελίδων.

- **Εξόρυξη Χρήσης στον Ιστό**

Με τον όρο Εξόρυξη Χρήσης στον Ιστό εννοούμε την εφαρμογή τεχνικών Εξόρυξης Δεδομένων με σκοπό την ανακάλυψη προτύπων χρήσης που προκύπτουν από τα καταγεγραμμένα δεδομένα χρήσης του Παγκόσμιου Ιστού έτσι ώστε να καταλήξουμε στην κατανόηση και στην καλύτερη εξυπηρέτηση των αναγκών των εφαρμογών που βασίζονται στο Διαδίκτυο. Τα καταγεγραμμένα αυτά δεδομένα περιλαμβάνουν την ταυτότητα των χρηστών καθώς και τη συμπεριφορά και τις συνήθειες αυτών κατά την περιήγησή τους στον Παγκόσμιο Ιστό. Ουσιαστικά, λοιπόν, κατά την Εξόρυξη Χρήσης γίνεται μια προσπάθεια κατανόησης των αλληλεπιδράσεων που ο ίδιος ο χρήστης επιθυμεί με το διαδίκτυο. Η μεγάλη διαφορά της κατηγορίας αυτής με τις δύο προηγούμενες (Εξόρυξη Περιεχομένου και Εξόρυξη Δομής) είναι πως ενώ στις άλλες δύο γίνεται χρήση πρωτογενών δεδομένων που υπάρχουν στον Παγκόσμιο Ιστό, στην Εξόρυξη Χρήσης γίνεται εξαγωγή πληροφορίας από τα δευτερογενή δεδομένα που παράγονται κατά τη διάδραση χρήστη και Παγκόσμιου Ιστού. Και αυτό το είδος Εξόρυξης στον Ιστό υπόκειται σε περαιτέρω υποδιαίρεση η οποία βασίζεται

στο είδος των δεδομένων τα οποία λαμβάνονται υπόψιν. Έτσι, υπάρχει η Εξόρυξη Χρήσης που βασίζεται σε δεδομένα που παρέχουν οι διαδικτυακοί εξυπηρετητές. Τα δεδομένα αυτά συλλέγονται από τους εξυπηρετητές και συνήθως περιλαμβάνουν διευθύνσεις IP, αναφορές στις ιστοσελίδες και τις αντίστοιχες χρονικές στιγμές που αυτές προσπελάστηκαν. Επίσης, υπάρχει το είδος Εξόρυξης Χρήσης που βασίζεται στα δεδομένα που παρέχουν οι εξυπηρετητές των εφαρμογών αλλά και αυτό που βασίζεται στα δεδομένα που προέρχονται από το επίπεδο των ίδιων των εφαρμογών. Στην πρώτη περίπτωση, αναφερόμαστε σε δεδομένα που προέρχονται από κυρίως εμπορικές εφαρμογές και τους εξυπηρετητές τους και συνδέονται με τη δυνατότητα αυτή των εφαρμογών να μπορούν να χρησιμοποιηθούν ως το υπόβαθρο για την ανάπτυξη απλών εφαρμογών ηλεκτρονικού εμπορίου. Τέτοια δεδομένα είναι για παράδειγμα καταγραφές εμπορικών δραστηριοτήτων και η αποθήκευση αυτών στα αρχεία των εξυπηρετητών. Αντίθετα, η περίπτωση των δεδομένων που προέρχονται από το επίπεδο των ίδιων των εφαρμογών αφορά στην καταγραφή οποιασδήποτε δραστηριότητας αφορά την εφαρμογή αυτή και τη χρήση της από κάθε μεμονωμένο χρήστη.

Η Εξόρυξη Χρήσης μπορεί, επίσης, να θεωρηθεί από δύο διαφορετικές οπτικές γωνίας. Από τη μία μεριά υπάρχει η προσέγγιση η οποία χαρτογραφεί τα δεδομένα χρήσης σε σχεσιακούς πίνακες πριν εφαρμοστεί η οποιαδήποτε τεχνική εξόρυξης. Αντιθέτως, στη δεύτερη περίπτωση γίνεται άμεση εφαρμογή των τεχνικών εξόρυξης με απαραίτητη προϋπόθεση, όμως, την ύπαρξη σαφώς καθορισμένης προεπεξεργασίας των δεδομένων. Είναι γεγονός ότι αυτή η προεπεξεργασία των δεδομένων είναι καίριας σημασίας και αυτό επειδή ο διαχωρισμός μεταξύ μεμονωμένων χρηστών και διαφορετικών συνεδριών του εξυπηρετητή είναι δύσκολος εξαιτίας της παρουσίας proxy εξυπηρετητών.

Οι εφαρμογές στις οποίες χρησιμοποιούνται οι τεχνικές Εξόρυξης Χρήσης στον Ιστό είναι πολυάριθμες αλλά για καλύτερη διαχείρισή τους μπορούν να κατηγοριοποιηθούν σε δύο ευρείες ομάδες. Η πρώτη ομάδα από αυτές τις ομάδες αφορά στην κατασκευή ενός προφίλ για κάθε χρήστη εντός ενός διαδραστικού περιβάλλοντος, όπως είναι για παράδειγμα τα Μέσα Κοινωνικής Δικτύωσης. Η δεύτερη βασίζεται στη γνώση που παίρνουμε μαθαίνοντας τις συμπεριφορές και τις επιθυμίες των χρηστών και αφορά στην κατασκευή προτύπων που να καθοδηγούν τους χρήστες. Η Εξόρυξη Χρήσης ενδιαφέρει τόσο τους καταναλωτές όσο και τους διαφόρων τύπων παρόχους πληροφοριών, όπως για παράδειγμα οι κατασκευαστές ιστοσελίδων, οι διαφημιστές που εργάζονται στον Παγκόσμιο Ιστό κτλ. Οι καταναλωτές ενδιαφέρονται για τεχνικές που θα παρέχουν πληροφορίες πάνω στα ενδιαφέροντά τους, τις επιθυμίες τους και τις ανάγκες τους. Από την άλλη μεριά, οι εταιρείες ενδιαφέρονται, μεταξύ άλλων, για την αύξηση της αποδοτικότητας των πληροφοριών που διοχετεύουν στον Ιστό. Αυτό γίνεται μέσω μιας προσπάθειας προσαρμογής των ιστοσελίδων στις ανάγκες των χρηστών είτε με την καθοδήγηση των χρηστών προς τους σκοπούς που η ιστοσελίδα εξυπηρετεί. Αυτό είναι που περιγράφουμε με τον όρο «καθοδήγηση χρηστών».

Ωστόσο, είναι σημαντικό να τονιστεί πως οι διαχωριστικές γραμμές μεταξύ των τριών αυτών μεγάλων κατηγοριών Εξόρυξης στον Ιστό αλλά και μεταξύ των επιμέρους υποκατηγοριών τους, δεν είναι σαφείς και απόλυτες. Στην πράξη, ο διαχωρισμός τους είναι πολύ δύσκολος αφού σπάνια χρησιμοποιούνται μεμονωμένα ενώ στις περισσότερες εφαρμογές χρησιμοποιούνται συνδυασμοί αυτών. Για παράδειγμα, κατά

την Εξόρυξη Περιεχομένου περιλαμβάνονται και τεχνικές Εξόρυξης Δομής αφού τα διαδικτυακά αρχεία είναι πιθανόν εκτός από κείμενο να περιέχουν και συνδέσμους.

Στον παρακάτω πίνακα (Πίνακας 1) βλέπουμε συγκεντρωτικά τα κυριότερα χαρακτηριστικά των κατηγοριών Εξόρυξης στον Ιστό.

	Εξόρυξη στον Ιστό			
	Εξόρυξη Περιεχομένου		Εξόρυξη Δομής	Εξόρυξη Χρήσης
	Αναζήτηση Πληροφοριών	Βάσεις Δεδομένων		
Μορφή Δεδομένων	Μη-δομημένα Ήμι-δομημένα	Ήμι-δομημένα Ιστοσελίδες υπό μορφή ΒΔ	Οικοδομήματα Συνδέσμων	Διάδραση
Είδος Δεδομένων	Αρχεία Κειμένου Αρχεία Υπερσυνδέσμων	Αρχεία Υπερσυνδέσμων	Οικοδομήματα Συνδέσμων	Καταγραφές Εξυπηρετητών, Περιηγητών & Εφαρμογών
Αναπαράσταση	Bag-of-Words Όροι, φράσεις Σχεσιακή	Γράφημα Σχεσιακή	Γράφημα	Γράφημα Πίνακες Συσχέτισης
Μέθοδος	TFIDF Μηχανική Μάθηση Στατιστική	Αλγόριθμοι ILP Κανόνες Συσχέτισης	Αλγόριθμοι	Μηχανική Μάθηση Στατιστική Κανόνες Συσχέτισης
Κατηγορίες Εφαρμογών	Κατηγοριοποίηση Ομαδοποίηση Εύρεση Κανόνων Εξόρυξης Εύρεση Προτύπων σε Κείμενα Μοντελοποίηση Χρήστη	Εύρεση Συχνών Υποδομών Ανακάλυψη Δομής Ιστοσελίδων	Κατηγοριοποίηση Ομαδοποίηση	Κατασκευή & Διαχείριση Ιστοσελίδων Διαφήμιση Μοντελοποίηση Χρήστη

Πίνακας 2.1: *Συγκεντρωτικά χαρακτηριστικά των κατηγοριών Εξόρυξης στον Ιστό.*

Στις ακόλουθες ενότητες περιγράφονται οι συνηθέστερα εφαρμοζόμενες τεχνικές για την εξόρυξη δεδομένων στον Ιστό με κάθε μια από τις προαναφερθείσες κατηγορίες.

2.4 Εξόρυξη Περιεχομένου στον Ιστό [1, 2, 3, 8, 10]

2.4.1 Εισαγωγή

Η όρος της Εξόρυξης Περιεχομένου στον Ιστό αφορά στη διαδικασία της εξαγωγής χρήσιμης πληροφορίας από τα περιεχόμενα αρχείων που βρίσκονται διαθέσιμα στον Παγκόσμιο Ιστό. Τα αρχεία αυτά περιλαμβάνουν δεδομένα τα οποία σκόπιμα έχουν σχεδιαστεί ώστε να μπορούν να μεταφερθούν μέσω του διαδικτύου στους χρήστες και μπορούν να περιλαμβάνουν δεδομένα σε μορφή κειμένου, εικόνας, ήχου, βίντεο ή ακόμη και σε δομημένη μορφή, όπως είναι οι λίστες και οι πίνακες. Στη συγκεκριμένη κατηγορία τεχνικών γίνεται πολύ συχνά η χρήση τεχνικών και από άλλους τομείς όπως η Αναζήτηση Πληροφοριών (Information Retrieval – IR) και η Επεξεργασία σε Φυσική Γλώσσα (Natural Language Processing – NLP).

2.4.2 Δεδομένα στον Παγκόσμιο Ιστό

Η Εξόρυξη Περιεχομένου στον Ιστό χρησιμοποιεί διαφορετικές τεχνικές ανάλογα με τη φύση των αρχείων. Στο συγκεκριμένο πλαίσιο, τα αρχεία διαχωρίζονται σε δομημένα, ημιδομημένα και μη-δομημένα.

Τα δομημένα αρχεία έχουν τη μορφή λιστών, δενδροδιαγραμμάτων ή πινάκων. Αντίθετα, τα μη-δομημένα αρχεία έχουν τη μορφή αρχείων κειμένου. Τα μη-δομημένα αρχεία είναι αντικείμενο επεξεργασίας των τεχνικών εξόρυξης κειμένου (text mining), της επεξεργασίας σε φυσική γλώσσα και της μηχανικής μάθησης. Τέλος, τα ημιδομημένα αρχεία δεν αποτελούνται πλήρως από κείμενο και χαρακτηρίζονται από δομή που δεν είναι γνωστή εκ των προτέρων. Και στην περίπτωση των ημιδομημένων αρχείων χρησιμοποιούνται τεχνικές επεξεργασίας σε φυσική γλώσσα, μηχανικής μάθησης, TINTIN και άλλες.

2.4.3 Διαδικασία Εξόρυξης Περιεχομένου στον Ιστό: Μια γενική θεώρηση

Όποια και αν είναι η φύση των αρχείων που χρησιμοποιούνται για την εξόρυξη περιεχομένου στον Ιστό, η ακολουθούμενη διαδικασία έχει ένα γενικό πρότυπο το οποίο αποτελείται από τα εξής διακριτά στάδια:

- Αρχικά, λαμβάνει χώρα η διαδικασία της εύρεσης των πηγών πληροφορίας. Σε αυτό το στάδιο γίνεται ανάκτηση αρχείων και κατ' επέκταση πληροφοριών από πηγές του διαδικτύου με τεχνικές εξόρυξης πληροφοριών, όπως για παράδειγμα με ταξινόμηση, με ομαδοποίηση και άλλες.
- Στη συνέχεια, ακολουθεί η δύσκολη διαδικασία της επιλογής των αρχείων που θα χρησιμοποιηθούν μέσα από τη συλλογή που συγκεντρώθηκε στο προηγούμενο βήμα καθώς και η προεπεξεργασία αυτών. Παράμετρος της διαδικασίας προεπεξεργασίας είναι η αναπαράσταση της εκάστοτε ιστοσελίδας. Έτσι, μπορούμε να έχουμε δυαδική αναπαράσταση (Binary) όπου σε κάθε αντικείμενο αντιστοιχίζεται μια από τις τιμές 0 και 1 ανάλογα με το αν το αντικείμενο αυτό εμφανίζεται ή όχι στη σελίδα. Επίσης, μπορούμε να έχουμε αναπαράσταση σε συχνότητα όρου (Term Frequency – TF), όπου κάθε όρος θεωρείται πως έχει σημαντικότητα ανάλογη με τον αριθμό των εμφανίσεών του στο αρχείο. Η σημαντικότητα αυτή γίνεται αυτόματα το βάρος του όρου t εντός αρχείου d που συμβολίζεται ως W(t)=TF(d,t). Μπορούμε, ακόμα, να έχουμε αναπαράσταση αντίστροφης συχνότητας αρχείου (Inverse Document Frequency – IDF), όπου η σημαντικότητα κάθε όρου είναι αριθμός αντιστρόφως ανάλογος του αριθμού των αρχείων που τον περιλαμβάνουν. Ο παράγοντας αυτός (IDF factor) υπολογίζεται μέσω της σχέσης

  $$W(t) = IDF(t) = \log N \cdot df(t)$$

  όπου N είναι ο αριθμός των αρχείων που περιλαμβάνονται στη συλλογή και df(t) είναι ο αριθμός των αρχείων που περιλαμβάνουν τον όρο t. Οι δύο τελευταίες αναπαραστάσεις συνδυαζόμενες μας δίνουν την αναπαράσταση TF-IDF, όπου πλέον το βάρος ενός όρου t μέσα σε ένα αρχείο d δίνεται από τη σχέση

$$W(d,t) = TF(d,t) \cdot IDF(t)$$

Επιπροσθέτως, μια από τις πιο κοινές αναπαραστάσεις είναι η WIDF, η οποία είναι μια επέκταση της IDF. Το βάρος ενός όρου t υπολογίζεται πλέον σε όλη την έκταση της συλλογής και δίνεται από τη σχέση

$$W(d,t) = \frac{TF(d,t)}{\sum diTF(i,t)}$$

όπου ο μετρητής i τρέχει σε όλη την έκταση της συλλογής.
Τέλος, ένας από τους πιο σημαντικούς τρόπους αναπαράστασης δεδομένων είναι μέσω της κατασκευής διανυσμάτων. Αυτή η αναπαράσταση θα περιγράφει εκτενώς σε επόμενη παράγραφο.
Η επόμενη παράγραφος αφιερώνεται στη διαδικασία της προεπεξεργασίας.

- Το επόμενο στάδιο ονομάζεται γενίκευση και είναι η φάση όπου εκτιμώνται τα διάφορα πρότυπα. Στο στάδιο αυτό χρησιμοποιούνται διαδικασίες από τη μηχανικά μάθηση και την εξόρυξη δεδομένων έτσι ώστε να αναγνωριστούν γενικά πρότυπα είτε σε μεμονωμένες ιστοσελίδες είτε σε ένα εύρος ιστοσελίδων.
- Ακολουθεί η ανάλυση του προτύπου έτσι ώστε με συγκεκριμένες μετρήσεις να εκτιμηθεί η ακρίβειά του.
- Τέλος, ακολουθεί η διαδικασία με την οποία αποφασίζεται ο τρόπος με τον οποίο θα παρουσιαστεί και θα οπτικοποιηθεί η εξαχθείσα γνώση καθώς και η πραγματοποίηση της παρουσίασης αυτής.

Στα παραπάνω αναφερθήκαμε σε τρεις διαφορετικούς τύπους εξόρυξης γνώσης: εξόρυξη δεδομένων, εξόρυξη περιεχομένου και εξόρυξη κειμένου. Θεωρούμε σημαντικό να διευκρινίσουμε τις διαφορές αυτών των τριών τεχνικών μεταξύ τους. Είδαμε πως η εξόρυξη περιεχομένου στον ιστό βασίζεται σε μεθόδους εξόρυξης δεδομένων. η μεγάλη διαφορά μεταξύ των δύο αυτών τεχνικών είναι πως τα αρχεία στην περίπτωση της εξόρυξης περιεχομένου είναι είτε ημιδομημένα, είτε μη-δομημένα. Στην εξόρυξη δεδομένων, αντίθετα, τα αρχεία είναι δομημένα. αυτό είναι και το επίπεδο στο οποίο η εξόρυξη περιεχομένου διαφοροποιείται από την εξόρυξη κειμένου αφού τα κείμενα είναι απόλυτα μη-δομημένα αρχεία ενώ ο Παγκόσμιος Ιστός απαρτίζεται από μη-δομημένα αλλά και από ημιδομημένα αρχεία. Έτσι, συμπεραίνουμε πως η εξόρυξη περιεχομένου δεν είναι απλά άθροισμα προϋπαρχουσών τεχνικών αλλά για τη σωστή λειτουργία της απαιτείται η ανάπτυξη νέων τεχνικών μέσω εφευρετικών προσεγγίσεων.

2.4.4 Προεπεξεργασία Περιεχομένου

Πριν την ένταξη ενός αρχείου εντός μιας συλλογής έτσι ώστε να χρησιμοποιηθεί κατά την προσπάθεια αναζήτησης πληροφοριών, συνήθως εκτελούνται συγκεκριμένες ενέργειες που αποσκοπούν στην προεπεξεργασία του αρχείου. Στα συνηθισμένα αρχεία (αυτά που δεν περιλαμβάνουν ετικέτες HTML), οι ενέργειες αυτές περιλαμβάνουν κατά κύριο λόγο αλγόριθμους περιστολής, αφαίρεση τερματικών λέξεων, επεξεργασία σημείων στίξης κ.α.. Στην περίπτωση των ιστοσελίδων, οι οποίες είναι και το αντικείμενο της εξόρυξης στον Ιστό, περιλαμβάνονται

επιπρόσθετες ενέργειες όπως για παράδειγμα η αφαίρεση των ετικετών HTML και η αναγνώριση των κύριων τμημάτων κειμένου μέσα στο αρχείο.

Αρχικά γίνεται η εξαγωγή των δεδομένων από μια θέση στον Παγκόσμιο Ιστό. Η διαδικασία αυτή γίνεται αυτόματα μέσω μηχανισμών διάσχισης με το σύστημα να γνωρίζει τις θέσεις στις οποίες θα αναζητήσει πληροφορίες δεδομένου περιεχομένου. Οι μηχανισμοί διάσχισης στο διαδίκτυο θα συζητηθούν εκτενώς σε επόμενη παράγραφο.

Στη συνέχεια εφαρμόζονται αλγόριθμοι περιστολής, οι οποίοι αποσκοπούν στην αναγωγή των κλιτών λέξεων του κειμένου στη βασική τους μορφή (τη ρίζα τους) μέσω της αφαίρεσης καταλήξεων και οι οποίοι βασίζονται σε λεξικά και λεξικά καταλήξεων καθώς και σε κανόνες. Η βασική αυτή μορφή θα πρέπει να είναι πανομοιότυπη με τη μορφολογική ρίζα της λέξης έτσι ώστε να καταλήξουμε στη χρησιμοποίηση ενός περιορισμένου λεξιλογίου για την αναπαράσταση των κειμένων. Η περιστολή βασίζεται στην αποδοχή πως σε πολλές γλώσσες μια λέξη έχει διάφορες συντακτικές μορφές και κάθε μια από αυτές εξαρτάται από την εκάστοτε χρήση της συγκεκριμένης λέξης. Οι μορφές αυτές διαφέρουν τόσο ως προς τον αριθμό (ενικός ή πληθυντικός) όσο και ως προς την πτώση ή την κλίση του ουσιαστικού, του ρήματος, του επιθέτου κτλ. Λεξιλογικά, αυτές οι διαφορετικές μορφές θεωρούνται ως τα αποτελέσματα μιας κοινής ρίζας στην οποία έχουν προσαρτηθεί διαφορετικές μεταξύ τους καταλήξεις, αυξάνοντας τη λεξιλογική ποικιλότητα και συνεπώς και τον όγκο των υπό αναζήτηση λέξεων. Η περιστολή έχει μελετηθεί αρκετά με το πέρασμα του χρόνου αναφορικά με το αν βελτιώνει τη διαδικασία εξόρυξης γνώσης ή όχι. Από όσα προαναφέραμε γίνεται σαφές πως η περιστολή αφενός αυξάνει την ανάκληση και μειώνει τον όγκο των δεδομένων στον οποία θα εφαρμοστούν οι διάφορες τεχνικές αναζήτησης. Ωστόσο, , μπορεί να βλάψει την ακρίβεια της εκάστοτε μεθόδου αφού χρησιμοποιώντας σαν κριτήριο αναζήτησης τη ρίζα μιας λέξης και όχι την ίδια τη λέξη μπορεί να λάβουμε ως αποτέλεσμα της αναζήτησης περιεχόμενο με μηδενική συνάφεια με το αίτημά μας.

Στη συνέχεια αφαιρούνται οι τερματικές λέξεις. Οι τερματικές λέξεις είναι συνήθως συνδετικές λέξεις που χρησιμοποιούνται εντός προτάσεων, άρθρα, προθέσεις κτλ. Αυτή είναι μια τεχνική που χρησιμοποιείται κατά κόρον στην επεξεργασία κειμένου σε φυσική γλώσσα και εξαρτάται κάθε φορά από το εκάστοτε σύστημα αφού δεν υπάρχει μια γενικά ορισμένη λίστα με τερματικές λέξεις. Ωστόσο, αυτό το φίλτρο μπορεί να αφαιρεθεί είτε ολικά, είτε μερικά κατά την αναζήτηση ολόκληρων προτάσεων. Για παράδειγμα, έστω ότι μια τερματική λέξη είναι η λέξη «αυτή». Τότε, θα αντιμετωπίσουμε προβλήματα κατά την αναζήτηση φράσεων που την περιέχουν, όπως για παράδειγμα με τη φράση «ο καιρός αυτή την εβδομάδα» και συνεπώς είναι σκόπιμο να αφαιρούνται οι τερματικές λέξεις για την βελτίωση της απόδοσης του συστήματος. Τέλος, υπολογίζονται δύο συχνότητες. Η πρώτη είναι η συχνότητα της υπό αναζήτηση λέξης ή φράσης στο επίπεδο της συλλογής (DF) και η δεύτερη είναι η συχνότητα των υπό αναζήτηση όρων σε κάθε αρχείο (TF_{TD}). Αυτοί οι δύο υπολογισμοί περιγράφονται καλύτερα στην παράγραφο που ακολουθεί.

Επιπρόσθετες τεχνικές επεξεργασίας κειμένου είναι η αφαίρεση ψηφίων και αριθμητικών όρων των οποίων η παρουσία δεν αλλοιώνει το περιεχόμενο του κειμένου (εξαιρούνται, δηλαδή, από τη διαδικασία αυτή αριθμοί όπως οι ημερομηνίες, οι ώρες, αριθμοί ταυτοποίησης κτλ.), η διαχείριση σημείων στίξης

καθώς και της παύλας «-» όταν αυτή χρησιμοποιείται ως μέσο σύνδεσης λέξεων καθώς και της διάκρισης μεταξύ κεφαλαίων και πεζών γραμμάτων.

Οι ιστοσελίδες όπως προ είπαμε, υπόκεινται σε επιπρόσθετη προεπεξεργασία, η οποία περιλαμβάνει τεχνικές όπως, για παράδειγμα, η αφαίρεση των ετικετών HTML και η αναγνώριση των κυριότερων κομματιών κειμένου . Η αφαίρεση των ετικετών HTML γίνεται με τρόπο παρόμοιο όπως η διαχείριση των σημείων στίξης. Προσοχή πρέπει να δοθεί στο γεγονός πως η HTML είναι εγγενώς μια γλώσσα εικονικής αναπαράστασης. Έτσι, πολλές φορές υπάρχουν διαφορετικά τμήματα κειμένου μέσα σε γραφικές αναπαραστάσεις των οποίων οι ετικέτες δεν πρέπει να συνδεθούν μεταξύ τους ώστε η διαδικασία της αναζήτησης να συνεχιστεί ανεμπόδιστα. Μια, επίσης, σημαντική διαδικασία είναι αυτή της αναγνώρισης των σημαντικότερων κομματιών κειμένου μιας ιστοσελίδας. Είναι εμπειρικά γνωστό ακόμη και στον πιο άπειρο χρήστη του διαδικτύου πως μια ιστοσελίδα περιλαμβάνει μεγάλο όγκο πληροφορίας ο οποίος δεν ανήκει στο κυρίως περιεχόμενο της σελίδας. Παράδειγμα τέτοιων τμημάτων πληροφορίας είναι οι διαφημίσεις, οι μπάρες πλοήγησης, οι δηλώσεις των πνευματικών δικαιωμάτων κτλ. Αυτά τα άχρηστα κατά την εξόρυξη πληροφορίας τμήματα κειμένου μειώνουν την αποδοτικότητα της μεθόδου.

Βέβαια, η προαναφερθείσα διαδικασία είναι γενική και μπορεί να τροποποιηθεί κατάλληλα κατά την εφαρμογή κάποιας από τις πολυάριθμες τεχνικές προεπεξεργασίας δεδομένων. Ωστόσο, η περιγραφή αυτών δεν αποτελεί σκοπό της παρούσας εργασίας και συνεπώς η διαδικασία περιεγράφηκε επιφανειακά.

2.4.5 Κατασκευή Διανυσμάτων

Το Μοντέλο του Διανυσματικού Χώρου αποτελεί μια πολύ γνωστή και συνηθισμένη αναπαράσταση ενός συνόλου αρχείων με τη μορφή διανυσμάτων και είναι θεμελιώδης λίθος για τη διενέργεια ενός μεγάλου εύρους εργασιών, από την αναζήτηση αρχείων μέχρι και την ομαδοποίηση ή την ταξινόμησή τους.

Για την επίτευξη αυτού του μοντέλου, σε κάθε όρο εντός του κειμένου αντιστοιχίζεται ένα βάρος το οποίο εξαρτάται από τις εμφανίσεις του όρου αυτού εντός του κειμένου. Αυτό που επιθυμούμε να καταφέρουμε είναι τον υπολογισμό ενός αποτελέσματος μεταξύ του όρου ή της φράσης υπό αναζήτηση και του αρχείου χρησιμοποιώντας το βάρος του T (term) στο D (document). Αυτός ο υπολογισμός ονομάζεται συχνότητα όρου (term frequency – TF_{TD}). Ωστόσο, χρήση του συγκεκριμένου μεγέθους έχει το μεγάλης σημασίας μειονέκτημα ότι όλοι οι όροι θεωρούνται εξίσου σημαντικοί κατά την αξιολόγηση της σχετικότητάς τους με μια δεδομένη αναζήτηση. Έτσι, εισάγεται ένας μηχανισμός για την εξάλειψη (μερική ή ολική) του φαινομένου των όρων που εμφανίζονται πολύ συχνά σε μια συλλογή αρχείων έτσι ώστε να αποκτούν μεγαλύτεροι σημασία όσοι απομένουν. Μια πιθανή υλοποίηση αυτού του μηχανισμού είναι μέσω της εισαγωγής ενός πολλαπλασιαστή ο οποίος θα δρα πάνω στο βάρος κάθε λέξης με τρόπο αντιστρόφως ανάλογο της συχνότητας εμφάνισής του. Αυτό σημαίνει πως κάθε λέξη που εμφανίζεται έντονα μέσα σε μια συλλογή αρχείων θα υπόκειται σε μια μείωση του βάρους της εξαιτίας του πολλαπλασιαστή αυτού, ο οποίος, όμως, θα αυξήσει το βάρος μιας λέξης η οποία συναντάται λιγότερες φορές. Ωστόσο, ο πολλαπλασιαστής ο οποίος συχνότερα χρησιμοποιείται είναι αυτός που λαμβάνει υπόψιν τη συχνότητα αρχείου (document

frequency – DF_T) η οποία ορίζεται ως ο αριθμός των αρχείων στη συλλογή που περιέχουν τον υπό αναζήτηση όρο. Η χρήση του αριθμού αυτού προτιμάται επειδή κατά τη προσπάθεια υπολογισμού του αποτελέσματος μεταξύ του όρου και του αρχείου είναι σαφώς προτιμότερο να χρησιμοποιούμε μια στατιστική που δρα σε επίπεδο αρχείων (όπως είναι αυτή που αφορά, για παράδειγμα, στον αριθμό των αρχείων που περιέχουν ένα συγκεκριμένο όρο) παρά να χρησιμοποιούμε μια στατιστική που να αφορά την παρουσία του όρου σε ολόκληρη συλλογή αρχείων. Το μόνο που απομένει είναι να οριστεί ο τρόπος με τον οποίο το DF_T θα επιδρά πάνω στο βάρος του κάθε όρου. Ορίζουμε

$$IDF_T = (DF_T)^{-1} = logN \cdot df(t) \quad (2.1)$$

όπου N ο συνολικός αριθμός των αρχείων στη συλλογή. Έτσι, όταν ένα αρχείο συναντάται σπάνια στη συλλογή, ο $\mathbf{IDF_T}$ του είναι υψηλός και το αντίστροφο, όταν συναντάται πολύ συχνά ο $\mathbf{IDF_T}$ του είναι χαμηλός. Τώρα, συνδέουμε μεταξύ τους τους δύο όρους TF_{TD} και $\mathbf{IDF_T}$ με τη σχέση

$$TF - IDF_T = TF_{TD} \cdot IDF_T \quad (2.2)$$

έτσι ώστε σε κάθε όρο T να αντιστοιχεί εντός κάθε κειμένου D ένα συγκεκριμένο βάρος το οποίο θα έχει τα παρακάτω χαρακτηριστικά:

1. Θα παίρνει τις πιο υψηλές τιμές του όταν θα συναντάται πολλές φορές εντός μιας μικρής συλλογής κειμένων με επακόλουθη τη μεταφορά υψηλής δυνατότητας διάκρισης στα κείμενα αυτά.
2. Θα παίρνει χαμηλές τιμές όταν ο όρος θα εμφανίζεται λιγότερες φορές σε ένα αρχείο ή θα εμφανίζεται σε πολλά αρχεία με επακόλουθη τη μεταφορά λιγότερο έντονης δυνατότητας διάκρισης του κειμένου.
3. Θα παίρνει τις χαμηλότερες τιμές όλων όταν ο όρος πρακτικά θα εμφανίζεται σε όλα τα αρχεία.

Έτσι, έχουμε καταφέρει να μπορούμε να θεωρήσουμε κάθε αρχείο υπό τη μορφή διανύσματος το οποίο αποτελείται από τις συνιστώσες που αντιστοιχούν σε όλους του τους όρους του χρησιμοποιούμενου λεξικού. Αυτές οι συνιστώσες συνοδεύονται από τα αντίστοιχα βάρη, όπως αυτά υπολογίζονται από τη σχέση (2.1). Είναι -μάλλον- ευνόητο πως όταν μια λέξη δε συναντάται στο κείμενο συνοδεύεται από μηδενικό βάρος.

Όταν θέλουμε, τώρα, να ποσοτικοποιήσουμε το επίπεδο στο οποίο δύο αρχεία D1 και D2 είναι όμοια μεταξύ τους, χρησιμοποιούμε τη σχέση

$$sim(D1, D2) = \frac{V(D1) \cdot V(D2)}{|V(D1)| \cdot |V(D2)|} \quad (2.3)$$

όπου V(D1) και V(D2) οι αναπαραστάσεις των δύο αρχείων στο διανυσματικό χώρο. Ο αριθμητής της σχέσης (2.3) αποτελείται από το εσωτερικό γινόμενο των δύο διανυσμάτων ενώ ο παρονομαστής από το γινόμενο των αντίστοιχων μηκών τελεσμένο στον Ευκλείδειο χώρο.

Η αναπαράσταση των κειμένων στο διανυσματικό χώρο είναι καίριας σημασίας. Όπως γνωρίζουμε, η εκκίνηση μιας διαδικασίας αναζήτησης πληροφοριών λαμβάνει

χώρα με την εισαγωγή ενός αιτήματος του χρήστη στο σύστημα. Τα αιτήματα αυτά είναι τυποποιημένες εκφράσεις των πληροφοριακών αναγκών του κάθε χρήστη. Ωστόσο, παρά το γεγονός ότι οι ανάγκες των χρηστών είναι συγκεκριμένες, κάθε τέτοιο αίτημα δεν αντιστοιχίζεται αμέσως σε συγκεκριμένο αρχείο της συλλογής. Το πιο πιθανό είναι πως η αναζήτηση θα επιστρέψει έναν αριθμό αποτελεσμάτων τα οποία θα διαφέρουν μεταξύ τους ως προς το βαθμό συσχέτισης με το αρχικό αίτημα. Αυτή η σχετικότητα με το αρχικό αίτημα είναι σημαντική και ποσοτικοποιήσιμη. Μπορεί κανείς να σκεφτεί πως αν μπορεί το αρχείο να εκφραστεί με τη μορφή ενός διανύσματος, το ίδιο μπορεί να συμβεί και στην περίπτωση του αιτήματος το οποίο δεν είναι τίποτε παραπάνω από μια συλλογή λέξεων υπό τη μορφή ενός μικρού κειμένου. Έτσι, το αποτέλεσμα της αναζήτησης σε σχέση με κάθε αρχείο που υπάρχει στα αποτελέσματα θα δίνεται από τη σχέση

$$score(Q, D) = \frac{V(Q) \cdot V(D)}{|V(Q)| \cdot |V(D)|} \quad (2.4)$$

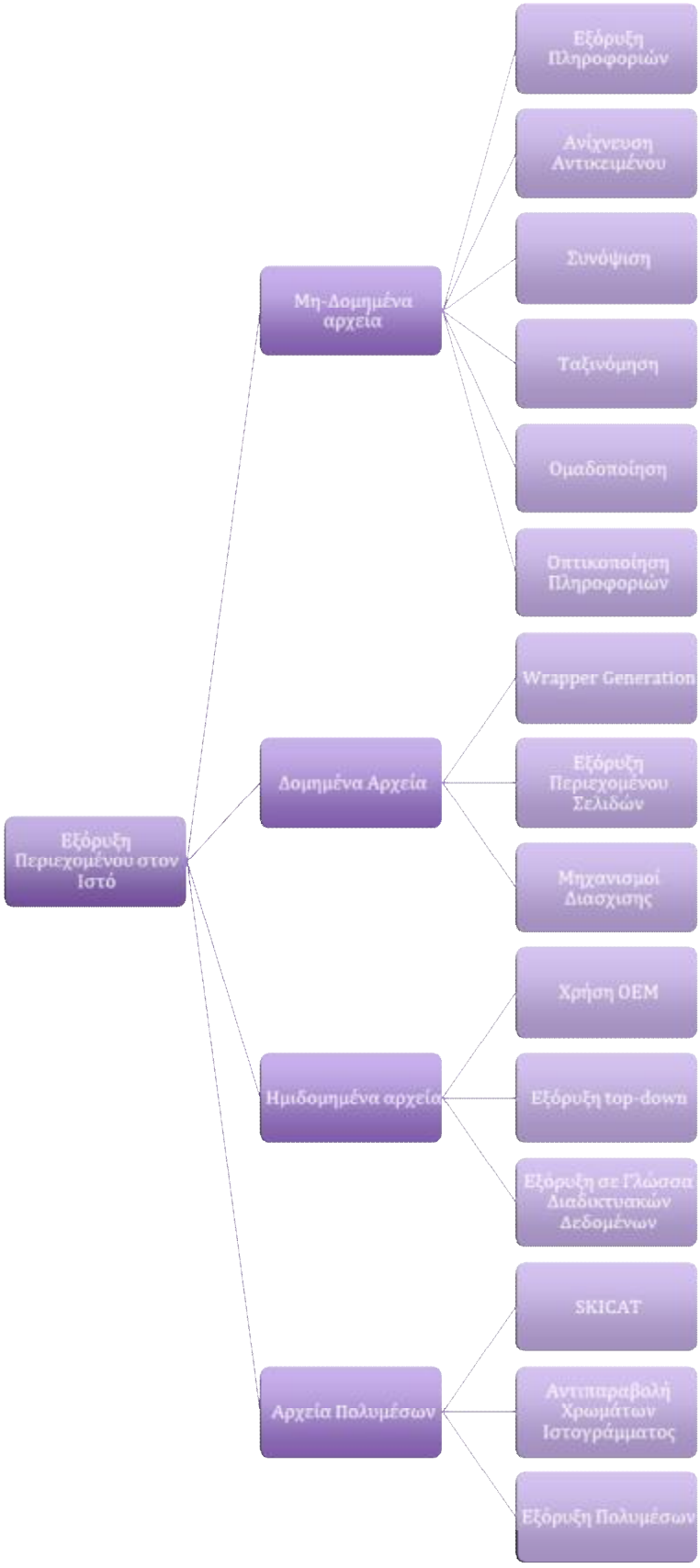
με τις αριθμητικές τιμές που προκύπτουν από αυτή να χρησιμοποιούνται ως κριτήριο για την επιλογή των αρχείων με τη μεγαλύτερη σχετικότητα με το αίτημα. Είναι πιθανό, βέβαια, μέσω της σχέσης (2.4) να προκύψουν υψηλές τιμές συσχέτισης ενός αιτήματος με ένα αρχείο ακόμη και αν δεν ταυτίζονται όλες οι λέξεις του αιτήματος αυτού με το συγκεκριμένο αρχείο. Αυτό μπορεί να συμβεί επειδή, όπως είδαμε παραπάνω, κάποιες πολύ σπάνιες λέξεις μπορεί να συνεισφέρουν στο αποτέλεσμα σε πολύ μεγαλύτερο βαθμό όταν εμφανίζονται σε μικρό αριθμό κειμένων από ότι κάποιες άλλες λέξεις που εμφανίζονται συνεχώς σε όλα τα κείμενα της συλλογής. Από αυτή τη διαδικασία, τα αρχεία με τα υψηλότερα αποτελέσματα θα είναι αυτά που το σύστημα θα δώσει στο χρήστη ως αποτελέσματα της αναζήτησής του ενώ η όλη διαδικασία μπορεί να επιδεχθεί μετατροπών έτσι ώστε τα αποτελέσματα να περιοριστούν περεταίρω.

2.4.6 Τεχνικές Εξόρυξης Περιεχομένου στον Ιστό

Οι τεχνικές εξόρυξης περιεχομένου στον Παγκόσμιο Ιστό είναι πολυάριθμες και αυξάνονται συνεχώς αφού υλοποιούνται συνεχώς είτε νέοι αλγόριθμοι, είτε μεταποιούνται παλαιότεροι με σκοπό να εξυπηρετηθούν οι ανάγκες του εκάστοτε χρήστη. Το ποια από αυτές τις τεχνικές θα χρησιμοποιηθεί κάθε φορά εξαρτάται σε πρώτο βαθμό από το είδος των αρχείων που έχουμε στη διάθεσή μας. Ακολούθως θα δούμε τις κυριότερες τεχνικές για μη-δομημένα, δομημένα, ημιδομημένα αρχεία καθώς και για αρχεία πολυμέσων. Η κατηγοριοποίηση αυτή φαίνεται στο γράφημα που ακολουθεί.

2.4.6.1 Τεχνικές Εξόρυξης Μη-Δομημένων Δεδομένων

Η εξόρυξη περιεχομένου στον Ιστό μπορεί να γίνει σε μη-δομημένα αρχεία όπως είναι για παράδειγμα τα κείμενα. Αποτέλεσμα αυτής της διαδικασίας είναι η εξαγωγή πληροφοριών που ήταν μέχρι πρότινος άγνωστες. Μερικές από τις τεχνικές που χρησιμοποιούνται στα μη-δομημένα δεδομένα είναι η εξόρυξη πληροφοριών, η ανίχνευση αντικειμένου, η συνόψιση, η ταξινόμηση, η ομαδοποίηση και η οπτικοποίηση πληροφοριών. Οι τεχνικές αυτές περιγράφονται ακολούθως.



Σχήμα 2.3: Τεχνικές Εξόρυξης Περιεχομένου στον Ιστό

Εξόρυξη Πληροφοριών

Κατά την εξόρυξη πληροφοριών από μη-δομημένα δεδομένα χρησιμοποιούμε μεθόδους που ελέγχουν την ταύτιση προτύπων. Οι μέθοδοι αυτές ουσιαστικά ανιχνεύουν τις λέξεις/φράσεις κλειδιά και στη συνέχεια βρίσκουν τον τρόπο που αυτές συνδέονται εντός του κειμένου. Η τεχνική αυτή είναι πολύ χρήσιμη όταν ο όγκος του κειμένου είναι πολύ μεγάλος και αποτελεί τη βάση για αρκετές ακόμα τεχνικές που χρησιμοποιούνται στην εξόρυξη σε μη-δομημένα αρχεία. Κατά την εξόρυξη πληροφοριών, τα αρχικά απολύτως μη-δομημένα αρχεία λαμβάνουν μια πιο συγκεκριμένη δομή. Αρχικά, οι πληροφορίες εξάγονται από τη συλλογή των δεδομένων και στη συνέχεια, με χρήση διαφορετικών ειδών κανόνων βρίσκονται οι πληροφορίες που λείπουν. Οι κανόνες που οδηγούν σε λανθασμένες προβλέψεις απορρίπτονται.

Ανίχνευση Αντικειμένου

Η ανίχνευση αντικειμένου είναι μια τεχνική η οποία ελέγχει τα αρχεία που προσπελαύνονται από τους χρήστες και μελετά τα προφίλ των χρηστών. Έτσι, μπορεί να κάνει προβλέψεις για κάθε χρήστη αναφορικά με το ποιο αρχείο θα προσπελάσουν στη συνέχεια. Η ανίχνευση αντικειμένου εφαρμόζεται για την παρακολούθηση ειδήσεων, για παράδειγμα, που αφορούν κάποιο συγκεκριμένο πρόσωπο, μια συγκεκριμένη επιχείρηση κτλ. Έτσι, η τεχνική αυτή μπορεί να χρησιμοποιηθεί σε πολλούς κλάδους με κυριότερους την ιατρική και την εκπαίδευση. Στην πρώτη, για παράδειγμα, ένας γιατρός μπορεί να ενημερώνεται άμεσα για νέες θεραπείες και φάρμακα. Επίσης, με την ανίχνευση αντικειμένου είναι δυνατή η παρακολούθηση του συνόλου της ροής των πληροφοριών που αφορούν το υπό εξέταση θέμα. Μεγάλο μειονέκτημα της τεχνικής αυτής είναι πως είναι πολύ συχνή η εμφάνιση αποτελεσμάτων με ελάχιστη σχετικότητα με το θέμα που μας αφορά.

Συνόψιση

Η διαδικασία με την οποία επιδιώκουμε τη μείωση της έκτασης ενός αρχείου ενώ ταυτόχρονα διατηρούμε ακέραια όλα τα κύρια σημεία του ονομάζεται συνόψιση. Μέσω αυτής, ο χρήστης μπορεί να αποφασίσει αν επιθυμεί να προσπελάσει το αρχείο ή όχι. Η χρονική αποδοτικότητα της τεχνικής αυτής είναι σαφής αν αναλογιστούμε πως ο χρόνος που ο αλγόριθμος χρειάζεται για να κατασκευάσει την περίληψη είναι μικρότερος από αυτόν που ο χρήστης χρειάζεται για να διαβάσει την πρώτη παράγραφο του αρχείου. Το λογισμικό του συστήματος υπολογίζει τα στατιστικά βάρη των προτάσεων και στη συνέχεια εξορύσσει τις πιο σημαντικές προτάσεις του κειμένου. Η τεχνική αυτή δίνει επίσης στο χρήστη την ελευθερία να επιλέξει το ποσοστό του αρχικού κειμένου το οποίο επιθυμεί να εξαχθεί ως συνόψιση αυτού. Τέλος, σημαντικό πλεονέκτημα της μεθόδου είναι ότι μπορεί να λειτουργεί παράλληλα με άλλες τεχνικές, όπως για παράδειγμα η ανίχνευση αντικειμένου.

Ταξινόμηση

Η ταξινόμηση είναι η διαδικασία με την οποία τα αρχεία μιας συλλογής εντάσσονται βάσει κοινών ιδιοτήτων τους σε προκαθορισμένες ομάδες. Η τεχνική αυτή καταμετρά τον αριθμό των λέξεων των κειμένων αλλά δεν επεξεργάζεται το περιεχόμενό τους σε κανένα επίπεδο. Το κυρίως θέμα κάθε ομάδας αποφασίζεται από τον αριθμό των λέξεων. Τελικά, τα αρχεία ταξινομούνται βάσει του θέματός τους. Η τεχνική αυτή είναι χρήσιμη στον τομέα των επιχειρήσεων και των βιομηχανιών για την παροχή υποστήριξης στους πελάτες.

Ομαδοποίηση
Η ομαδοποίηση είναι μια τεχνική που χρησιμοποιείται για την συγκρότηση ομάδων όμοιων αρχείων. Η μεγάλη διαφοροποίηση της ομαδοποίησης με την ταξινόμηση είναι πως οι ομάδες στις οποίες κατατάσσονται τα αρχεία δεν είναι προκαθορισμένες. Επίσης, κάποια αρχεία μπορούν να συμμετέχουν σε περισσότερες από μια ομάδες. Η ομαδοποίηση είναι μια τεχνική ιδιαίτερα χρήσιμη στη διαχείριση πληροφοριακών συστημάτων.

Οπτικοποίηση Πληροφοριών
Σκοπός της τεχνικής αυτής είναι η κατασκευή μιας γραφικής αναπαράστασης των αρχείων. Μέσω αυτής, αναγνωρίζονται τα αρχεία τα οποία έχουν ομοιότητες μεταξύ τους και στη συνέχεια αναπαρίστανται μέσω γραφημάτων και χαρτογραφήσεων. Αποτελεί σημαντικό εργαλείο αφού μέσω αυτής δίνεται η ευκαιρία στους χρήστες να αναλύσουν οπτικά τα περιεχόμενα των αρχείων. Ταυτόχρονα προσφέρονται εργαλεία επεξεργασίας της δημιουργούμενης εικόνας όπως είναι για παράδειγμα η κλίμακα της εικόνας, το ζουμ κτλ. Η οπτικοποίηση πληροφοριών προτιμάται όταν αναζητούμε ένα αντικείμενο μέσα σε έναν μεγάλο όγκο αρχείων.

2.4.6.2 Τεχνικές Εξόρυξης Δομημένων Δεδομένων

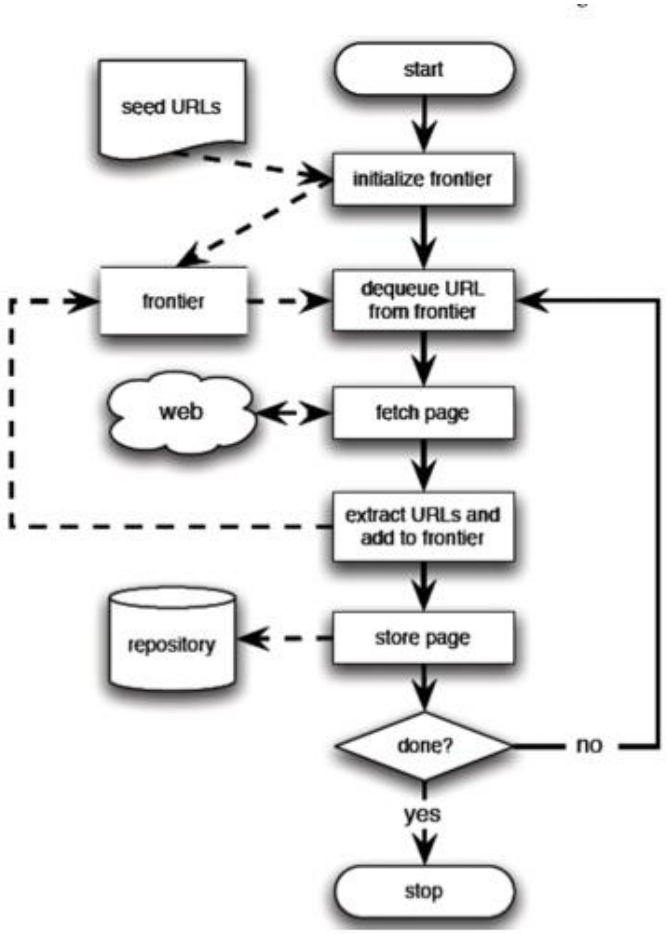
Μηχανισμοί Διάσχισης του Διαδικτύου (Web Crawlers)

Η αναζήτηση στον Παγκόσμιο Ιστό βασίζεται στις τεχνικές αναζήτησης πληροφοριών, ένας κλάδος της επιστήμης της πληροφορικής ο οποίος έχει ως στόχο την εύρεση των επιθυμητών πληροφοριών μέσα από μια μεγάλη συλλογή αρχείων, όπως για παράδειγμα μέσα σε μια βάση δεδομένων. Στον παγκόσμιο ιστό, τα αρχεία αυτά έχουν τη μορφή ιστοσελίδων. Βέβαια, παρά το γεγονός ότι η αναζήτηση στον Παγκόσμιο Ιστό εξαρτάται τόσο έντονα από τις τεχνικές αναζήτησης πληροφοριών, δεν παύει να έχει τα δικά της μοναδικά χαρακτηριστικά και τις δικές της ξεχωριστές τεχνικές.

Ένας σπουδαίος μηχανισμός περιήγησης στο διαδίκτυο είναι ο μηχανισμός διάσχισης (web crawler/web spider/web robot), ο οποίος είναι ουσιαστικά ένα αυτόματο πρόγραμμα το οποίο περιηγείται στο διαδίκτυο με μια καλά ορισμένη και αυτοματοποιημένη μέθοδο. Στη γενική τους μορφή, οι μηχανισμοί διάσχισης εκκινούν έχοντας στη διάθεσή τους προς επίσκεψη μια λίστα ηλεκτρονικών διευθύνσεων (Uniform Resource Locators – URLs). Καθώς ο μηχανισμός επισκέπτεται κάθε μια από αυτές τις διευθύνσεις, εξάγει όλους τους υπερσυνδέσμους που περιλαμβάνονται σε αυτή και τους προσθέτει στην αρχική του λίστα. Ο μηχανισμός επισκέπτεται τις εξαγόμενες από τους υπερσυνδέσμους διευθύνσεις αναδρομικά ακολουθώντας προκαθορισμένους κανόνες. Η διαδικασία αυτή συνεχίζεται μέχρι να αδειάσει πλήρως η λίστα με τις προς επίσκεψη ηλεκτρονικές διευθύνσεις ή μέχρι να ικανοποιηθούν συγκεκριμένες απαιτήσεις από τη μεριά των χρηστών. Γραφικά, η λειτουργία ενός μηχανισμού διάσχισης φαίνεται στην εικόνα που ακολουθεί.

Ο Παγκόσμιος Ιστός, όμως, είναι δυναμικός και αναπτύσσεται με ραγδαίους ρυθμούς με τις περιλαμβανόμενες πληροφορίες να προστίθενται, να διαγράφονται ή και να μετακινούνται συνεχώς. Έτσι, υπάρχει η ανάγκη οι διάφορες διαδικτυακές εφαρμογές

να βοηθιούνται από μηχανισμούς διάσχισης ώστε να είναι μονίμως ενημερωμένες από άποψης πληροφοριών. Οι μηχανισμοί διάσχιση χρησιμοποιούνται σε μεγάλο πλήθος εφαρμογών, ωστόσο η σημαντικότερη συνεισφορά τους βρίσκεται στις μηχανές εύρεσης όπου χρησιμοποιούνται για τη συλλογή των ιστοσελίδων οι οποίες θα αποτελέσουν τη συλλογή των αρχείων όπου θα γίνει η αναζήτηση.



Εικόνα 2.1: Διάγραμμα ροής λειτουργίας ενός τυπικού μηχανισμού διάσχισης. [2]

Υπάρχουν διάφορα είδη μηχανισμών διάσχισης. Για παράδειγμα, υπάρχουν οι μηχανισμοί διάσχισης οι οποίοι επικεντρώνονται στο υπό μελέτη θέμα (focus crawlers) και διασχίζουν σελίδες οι οποίες αφορούν μόνο αυτό ενώ ταυτόχρονα προσπαθούν να ελαχιστοποιήσουν τις εκτός θέματος σελίδες. Υπάρχουν, επίσης, οι επαυξητικοί μηχανισμοί διάσχισης (incremental crawlers) που τείνουν να επανεπισκέπτονται τις σελίδες της λίστας τους για να διασφαλίσουν την εγκυρότητα και το βαθμό στον οποίο τα αποτελέσματά τους είναι ενημερωμένα ενώ υπάρχουν και οι μηχανισμοί διάσμισης με όριο (batch crawlers), οι οποίοι διασχίζουν τον υπό μελέτη χώρο τους μέχρι μια οριακή τιμή αναφορικά με το χρόνο αυτής της διαδικασίας είτε με τον όγκο της συσσωρευόμενης πληροφορίας. Ωστόσο, όλες αυτές οι κατηγορίες έχουν κοινό χαρακτηριστικό το σκοπό ύπαρξής τους που δεν είναι άλλος από τη συλλογή περιεχομένου από ιστοσελίδες.

Wrapper Generators
Οι Wrapper Generators παρέχουν πληροφορίες αναφορικά με τις δυνατότητες των πηγών μας. Αν οι ιστοσελίδες έχουν ήδη καταταχθεί μέσω των παραδοσιακών μηχανών αναζήτησης, τότε η ανάκτηση αυτών γίνεται με χρήση της θέσης τους σε αυτή την κατάταξη μετά από κάποια υποβληθείσα αναζήτηση εκ μέρους του χρήστη. Οι πηγές αφορούν στο είδος των αναζητήσεων στις οποίες αντιστοιχούν καθώς και στο είδος των εξόδων. Ταυτόχρονα, οι wrappers παρέχουν επιπρόσθετα εργαλεία ανάλυσης όπως για παράδειγμα στατιστικά εργαλεία κτλ.

Εξόρυξη Περιεχομένου Σελίδων
Η εξόρυξη περιεχομένου σελίδων είναι μια τεχνική εξόρυξης δομημένων αρχείων η οποία αφορά σε σελίδες που έχουν ήδη καταταχθεί με παραδοσιακές μηχανές αναζήτησης. Αυτός ο τρόπος εξόρυξης οδηγεί στην ταξινόμηση των ιστοσελίδων αυτών.

2.4.6.3 Τεχνικές Εξόρυξης Ημιδομημένων Δεδομένων

Μοντέλο Ανταλλαγής Αντικειμένων (Object Exchange Model -OEM)
Κατά την εξόρυξη ημιδομημένων δεδομένων έχουμε τη δυνατότητα να εξάγουμε πληροφορίες σχετικά με το θέμα της αναζήτησής μας και να τις ενσωματώσουμε όλες μαζί σε μια ομάδα χρήσιμων πληροφοριών που στη συνέχεια θα αποθηκευτεί σε ένα μοντέλο ανταλλαγής αντικειμένων. Το μοντέλο αυτό βοηθάει το χρήστη να κατανοήσει καλύτερα τη δομή των πληροφοριών στον παγκόσμιο ιστό και χρησιμοποιείται προτιμότερα στα πλαίσια ετερογενών και δυναμικών περιβαλλόντων. Το κύριο χαρακτηριστικό των μοντέλων ανταλλαγής αντικειμένων είναι πως είναι αυτοπεριγραφικά, δηλαδή δεν υπάρχει ανάγκη για εκ των προτέρων περιγραφή των αντικειμένων του κάθε μοντέλου.

Εξόρυξη Top-Down (Top-Down Extraction)
Σκοπός της εξόρυξης top-down είναι η εξαγωγή σύνθετων αντικειμένων και η αποδόμησή τους σε βαθμό τέτοιο ώστε το παραγόμενο αντικείμενο να είναι λιγότερο σύνθετο. Αρχικά, για να γίνει η εξαγωγή τέτοιων σύνθετων αντικειμένων θα πρέπει να είμαστε σε θέση να περιγράψουμε το τι θέλουμε να εξορύξουμε. Η τεχνική αυτή χρησιμοποιείται σε πηγές του ιστού οι οποίες είναι πολύ πλούσιες σε δεδομένα και συνεπώς θα πρέπει να είμαστε σε θέση να διακρίνουμε τον τύπο των δεδομένων αυτών. Στη συνέχεια τα δεδομένα αυτά τοποθετούνται σε πίνακες ώστε να μπορεί να γίνει η αναζήτησή μας.

Εξόρυξη σε γλώσσα διαδικτυακών δεδομένων (Web Data Extraction Language)
Σκοπός της εξόρυξης σε γλώσσα διαδικτυακών αντικειμένων είναι η μετατροπή των δεδομένων του Παγκόσμιου Ιστού σε δομημένα δεδομένα με τελικό παραλήπτη το χρήστη. Και σε αυτή την τεχνική, τα δεδομένα αποθηκεύονται σε πίνακες.

2.4.6.4 Τεχνικές Εξόρυξης Δεδομένων Πολυμέσων

SKICAT
Το SKICAT είναι ένα επιτυχημένο σύστημα ανάλυσης αστρονομικών δεδομένων το οποίο έχει τη δυνατότητα να παράγει ψηφιακούς καταλόγους των ουράνιων αντικειμένων. Χρησιμοποιεί τεχνικές μηχανικής μάθησης έτσι ώστε τα αντικείμενα αυτά να μπορούν να πάρουν τη μορφή τάξεων που μπορούν να γίνουν χρήσιμες στους ανθρώπους. Για την κατασκευή του έχει γίνει ενσωμάτωση διάφορων τεχνικών επεξεργασίας εικόνας και ταξινόμησης δεδομένων έτσι ώστε να είναι δυνατή η ταξινόμηση ενός τόσο μεγάλου όγκου δεδομένων.

Αντιπαραβολή Χρωμάτων Ιστογράμματος
Η τεχνική αυτή αφορά σε δύο διαφορετικές εργασίες: την εξίσωση των χρωμάτων του ιστογράμματος και την εξομάλυνσή τους. Η εξίσωση αντιμετωπίζει ένα πολύ σημαντικό πρόβλημα το οποίο σχετίζεται με την παρουσία ανεπιθύμητων αντικειμένων στις επεξεργασμένες εικόνες που ήδη έχει λάβει χώρα η διεργασία τις εξίσωσης. Αυτό το πρόβλημα λύνεται με τη χρήση της εξομάλυνσης.

Εξόρυξη Πολυμέσων
Η εξόρυξη πολυμέσων συνίσταται από τέσσερα κύρια στάδια. Αρχικά, χρησιμοποιείται ένας εκσκαφέας εικόνων ο οποίος εξάγει εικόνες από ανάλογα αρχεία ή από βίντεο. Στη συνέχεια, ένας προεπεξεργαστής δρα πάνω στις εικόνες αυτές και τις αποθηκεύει σε μια βάση δεδομένων. Ένας πυρήνας αναζήτησης χρησιμοποιείται έτσι ώστε κάθε αναζήτηση να μπορεί να αντιστοιχηθεί με τις εικόνες και τα βίντεο που έχουν καταχωρηθεί στη βάση δεδομένων. Τέλος, γίνεται η εξόρυξη πληροφοριών εικόνας για την ανάδειξη προτύπων που πιθανόν υπάρχουν στο σύνολο αυτό.

2.5 Εξόρυξη Δομής στον Ιστό [1, 2, 4, 5, 7, 11]

2.5.1 Εισαγωγή

Η ανάλυση της δομής των διάφορων τοποθεσιών στον Παγκόσμιο Ιστό είναι ένας ερευνητικός τομέας ο οποίος απολαμβάνει ενδιαφέροντος εδώ και δεκαετίες. Ωστόσο, το αυξανόμενο ενδιαφέρον προς την εξόρυξη δεδομένων στον Ιστό έχει δώσει άλλη διάσταση στην έρευνα γύρω από τις δομές αυτές με αποτέλεσμα τη δημιουργία ενός νέου τομέα έρευνας, της Εξόρυξης Συνδέσμων (Link Mining), ο οποίος βρίσκεται στο σταυροδρόμι που συναντώνται η Ανάλυση Συνδέσμων και Υπερσυνδέσμων και η Εξόρυξη στον Ιστό.

Σκοπός της εξόρυξης δομής στον ιστό είναι η ανάδειξη άγνωστων μέχρι πρότινος σχέσεων μεταξύ συγκεκριμένων ιστοσελίδων.

Ο Παγκόσμιος Ιστός έχει την ιδιότητα να αποτελείται από αντικείμενα τα οποία είναι ανομοιόμορφης δομής και περιεχομένου εκτεταμένου όγκου, ιδιότητες που τα διαφοροποιούν ως προς τα παραδοσιακά αρχεία κειμένου. Τα αντικείμενα αυτά έχουν τη μορφή ιστοσελίδων και μπορούν να περιέχουν συνδέσμους άλλων ιστοσελίδων

ή/και να αποτελούν συνδέσμους σε άλλες σελίδες. Τα κύρια χαρακτηριστικά γνωρίσματα των αρχείων αυτών στον Παγκόσμιο Ιστό είναι οι διάφορες HTML επισημάνσεις, η εμφάνιση λέξεων και σημαντικά τμήματα κειμένου. Αυτή η ποικιλία των χαρακτηριστικών, όμως, εκτός από το να δημιουργεί ενδιαφέροντα αρχεία δημιουργεί επιπλέον προβλήματα και νέες προκλήσεις αφού δεν είναι δυνατή η χρήση αλγόριθμων και τεχνικών διαχείρισης βάσεων δεδομένων ή αναζήτησης πληροφοριών.

Όπως προαναφέρθηκε, η εξόρυξη δομής στον ιστό κληρονόμησε σημαντικές τεχνικές από την παραδοσιακή ανάλυση συνδέσμων. Από αυτές τις τεχνικές, οι πιο συχνά εφαρμοζόμενες στο χώρο της εξόρυξης δομής στον ιστό είναι οι παρακάτω:

- Ταξινόμηση βάσει συνδέσμων: Αποτελεί την πιο σύγχρονη αναβάθμιση στην έννοια της εξόρυξης δεδομένων σε διασυνδεδεμένους χώρους (linked domains) και αφορά στην πρόβλεψη της κατηγορίας μιας ιστοσελίδας. Η κατηγορία αυτή προσδιορίζεται από τις λέξεις που συναντώνται στο σύνολο του κειμένου εντός της σελίδας, στους συνδέσμους μεταξύ ιστοσελίδων, στις HTML επισημάνσεις και σε λοιπά χαρακτηριστικά αυτών.
- Ανάλυση ομάδων βάσει συνδέσμων: Αφορά στην εύρεση φυσικά επαγόμενων υποκλάσεων. Τα δεδομένα τμηματοποιούνται σε κλάσεις όμοιων χαρακτηριστικών ενώ τα αντικείμενα με διαφορετική δομή εντάσσονται σε άλλες ομάδες. Η διαφορά της τεχνικής αυτής με την προηγούμενη είναι πως ανήκει στη μη επιβλεπόμενη μάθηση και μπορεί να χρησιμοποιηθεί για την ανάδειξη προτύπων που υπεισέρχονται σε συλλογές δεδομένων.
- Τύπος συνδέσμων: Αφορά στην πρόβλεψη ύπαρξης συνδέσμων, τον τύπο του συνδέσμου μεταξύ δύο ιστοσελίδων ή/και του σκοπού ύπαρξής του.
- Ισχύς συνδέσμων: Αφορά στην αντιστοίχιση των συνδέσμων με τα ανάλογα βάρη.
- Πληθικότητα συνδέσμων: Αφορά στην πρόβλεψη του πλήθους των συνδέσμων μεταξύ συγκεκριμένων αντικειμένων

2.5.2 Αλγόριθμοι Εξόρυξης Δομής στον Ιστό

Μέχρι τη δεκαετία του 1990, τα παραδοσιακά συστήματα αναζήτησης πληροφοριών λειτουργούσαν μέσω μεθόδων βασιζόμενων αποκλειστικά στο περιεχόμενο των ιστοσελίδων. Ωστόσο, ο ραγδαία αυξανόμενος όγκος πληροφοριών κατέστησε αυτό το μηχανισμό αναζήτησης πληροφοριών αναποτελεσματικό αφού τα αποτελέσματα κάθε αναζήτησης ήταν υπεράριθμα σε σχέση με τις ανάγκες του εκάστοτε χρήστη. Συνέπεια αυτής της αναποτελεσματικότητας υπήρξε μια συνεχής προσπάθεια για βελτίωση της λειτουργίας των μηχανών αναζήτησης. Η βελτίωση αυτή εντοπίστηκε στους αλγόριθμους κατάταξης των ιστοσελίδων μέσω ανάλυσης της δομής των συνδέσμων και των υπερσυνδέσμων που περιλαμβάνουν. Προς αυτή την κατεύθυνση έχουν προταθεί αρκετοί αλγόριθμοι, ωστόσο οι επικρατέστεροι είναι τρεις: ο PageRank, ο σταθμισμένος PageRank (Weighted PageRank) και ο HITS (Hyperlink Induced Topic Research). Αυτοί οι αλγόριθμοι περιγράφονται στις παραγράφους που ακολουθούν.

2.5.2.1 Ο αλγόριθμος PageRank

Ο αλγόριθμος PageRank, ο οποίος δημιουργήθηκε το 1998 από τους ιδρυτές της Google Sergey Brin και Lawrence Page, είναι ένας από τους κυρίαρχους αλγόριθμους κατάταξης των ιστοσελίδων και είναι μάλιστα ο αλγόριθμος που χρησιμοποιείται στη μηχανή αναζήτησης της εταιρείας Google.

Ο αλγόριθμος, σε μια γενική περιγραφή του, χρησιμοποιεί ως παραμέτρους του μηχανισμού λειτουργίας του τις υπάρχουσες πληροφορίες αναφορικά με τη διασύνδεση ανάμεσα στις ιστοσελίδες. Έτσι, κατατάσσει τις ιστοσελίδες ανάλογα με τον τρόπο που οι σελίδες συνδέονται μεταξύ τους έτσι ώστε να κατατάσσονται σε υψηλές θέσεις οι σελίδες οι οποίες συνδέονται με σελίδες που επίσης έχουν υψηλές θέσεις στην κατάταξη.

Η μετρική του αλγόριθμου PageRank PR(p) είναι το μέγεθος το οποίο αντιστοιχεί στη σημαντικότητα μιας σελίδας p_i μέσω της ακόλουθης σχέσης

$$PR(p_i) = d + (1 - d)\left[\frac{PR(p_1)}{c_1} + \cdots + \frac{PR(p_n)}{c_n}\right]$$

δηλαδή ως το άθροισμα του βαθμού σημαντικότητας όλων των ιστοσελίδων που διασυνδέονται με την p_i. Η βασική ιδέα πίσω από αυτή τη σχέση είναι η χρησιμοποίηση της πιθανότητας να προσπελασθεί η συγκεκριμένη σελίδα από ένα χρήστη ο οποίος περιηγείται με τυχαίο τρόπο στον Παγκόσμιο Ιστό. Έτσι, στην παραπάνω σχέση, η ιστοσελίδα p_i διασυνδέεται με τις ιστοσελίδες p_j ,j=1,…n όπου c_j ο αριθμός του συνολικού αριθμού συνδέσμων που εξέρχονται από τη σελίδα p_j . Επίσης, d είναι ο συντελεστής απόσβεσης, του οποίου η έννοια συνδέεται με τη συμπεριφορά των χρηστών κατά την περιήγηση στο διαδίκτυο. Ακριβέστερα, έχει παρατηρηθεί ότι η σχετική με συγκεκριμένο θέμα περιήγηση στο διαδίκτυο έχει ανώτατο όριο χρονικής διάρκειας αφού οι χρήστες μπορούν να μείνουν συγκεντρωμένοι στο συγκεκριμένο θέμα μόνο για περιορισμένο χρόνο ενώ με το πέρας αυτού αρχίζει η απόσπαση της προσοχής τους από αντικείμενα άσχετα με το αρχικό θέμα. Κατ' επέκταση, η παράμετρος d απαγορεύει σε άσχετες με το θέμα αντικείμενες να λάβουν υψηλή θέση εξαιτίας της τυχαίας προσπέλασής τους. Έτσι, με την παράμετρο d να έχει την προαναφερθείσα έννοια, ο παράγοντας 1-d συνδέεται με την υπολειπόμενη πιθανότητα ο χρήστης να περιηγηθεί σε κάποιον από τους τυχαίους c_j συνδέσμους της σελίδας p_j. Μια ικανοποιητική τιμή για το δείκτη απόσβεσης είναι το d=0.85 που συνεπάγεται πως κάθε σελίδα διαμοιράζει ισομερώς το 85% του αρχικού μέτρου της σε κάθε μια από τις επόμενες.

Ενώ ο αλγόριθμος PageRank έχει υψηλή απόδοση, συνδέεται με συγκεκριμένα μειονεκτήματα. Αρχικά, μπορεί να χαρακτηριστεί ως στατικός αλγόριθμος του οποίου ο συσσωρευτικός χαρακτήρας συνεπάγεται πως μια δημοφιλής ιστοσελίδα θα παραμένει δημοφιλής γενικά και για μεγάλα χρονικά διαστήματα, ανεξάρτητα από τις προτιμήσεις των χρηστών. Ένα δεύτερο μειονέκτημα είναι πως εξαρτάται από το βαθμό στον οποίο μια ιστοσελίδα είναι δημοφιλής. Ωστόσο, η δημοφιλία μιας ιστοσελίδας δε συνεπάγεται και την απόλυτη ταύτιση του περιεχομένου της με τις αναζητήσεις των χρηστών. Ο αλγόριθμος είναι επίσης πολύ αργός αναφορικά με τον όγκο της πληροφορίας που περιέχεται στον Παγκόσμιο Ιστό ενώ, ταυτόχρονα, δεν

υποστηρίζει κάποια μέθοδο προσωποποίησης των αναζητήσεων των χρηστών με αποτέλεσμα να μην ικανοποιούνται οι συγκεκριμένες ανάγκες κάθε χρήστη.

2.5.2.2 Ο σταθμισμένος αλγόριθμος PageRank

Ο σταθμισμένος αλγόριθμος PageRank είναι μια αναβάθμιση του κλασικού αλγόριθμου PageRank και έχει υλοποιηθεί από τους Wenpu Xing και Ali Ghorbani κατά την προσπάθειά τους να τροποποιήσουν τον αλγόριθμο PageRank με τρόπο τέτοιο που να λαμβάνει υπόψιν του το γεγονός πως κάποιοι σύνδεσμοι είναι πιο σημαντικοί από άλλους και συνεπώς συνεισφέρουν με διαφορετικό τρόπο στην τελική κατάταξη μιας δεδομένης ιστοσελίδας.. Ο σταθμισμένος αλγόριθμος PageRank διαφοροποιείται από τον κλασικό ως προς τον υπολογισμό του μέτρου της θέσης κάθε σελίδας με το να αντιστοιχεί στις πιο σημαντικές σελίδες μεγαλύτερο μέτρο το οποίο τις κατατάσσει αυτόματα σε υψηλότερες θέσεις. Έτσι, αντί να διαιρείται το μέτρο κάθε σελίδας δια τον αριθμό των εξερχόμενων από αυτή συνδέσμων, κάθε σύνδεσμος λαμβάνει ένα δικό του μέτρο ανάλογα με τη σημαντικότητα του ίδιου και όχι της αρχικής σελίδας.

Στο σταθμισμένο PageRank, σε κάθε σύνδεσμο αντιστοιχίζεται ένα βάρος, είτε αυτός ο σύνδεσμος είναι εισερχόμενος (δηλαδή μας οδηγεί στη σελίδα υπό μελέτη) είτε είναι εξερχόμενος (δηλαδή μας οδηγεί από την υπό μελέτη σελίδα σε κάποια άλλη). Αυτό είναι και το μεγάλο πλεονέκτημά του έναντι του κλασικού αλγόριθμου PageRank, ότι δηλαδή εξαρτάται από δύο παραμέτρους, τους αριθμούς δηλαδή των εισερχόμενων και των εξερχόμενων συνδέσμων. Στη βιβλιογραφία, η δημοφιλία των εισερχόμενων και των εξερχόμενων συνδέσμων, το αντίστοιχο βάρος τους δηλαδή, συμβολίζεται με τις συναρτήσεις $W_{in}(v,u)$ και $W_{out}(v,u)$ αντίστοιχα. Η συνάρτηση $W_{in}(v,u)$ αντιστοιχεί στο βάρος ενός εισερχόμενου συνδέσμου (v,u) όπου v ο αριθμός των εισερχόμενων συνδέσμων σε όλες τις σελίδες στις οποίες παραπέμπει η υπό συζήτηση σελίδα και u ο αριθμός των εισερχόμενων ή εξερχόμενων συνδέσμων της σελίδας αυτής. Το βάρος αυτό δίνεται από τη σχέση

$$W_{in}(v,u) = \frac{I_u}{\sum_{p \in R(v)} I_p}$$

με I_u και I_p τους αριθμούς των εισερχόμενων συνδέσμων των σελίδων u και p ενώ το σύνολο R(v) είναι το σύνολο από το οποίο παίρνει τιμές η μεταβλητή v. Αντίστοιχα, η συνάρτηση $W_{out}(v,u)$ αντιστοιχεί στο βάρος ενός εξερχόμενου συνδέσμου και οριζόμενη αντίστοιχα με την $W_{in}(v,u)$ δίνεται από τη σχέση

$$W_{out}(v,u) = \frac{O_u}{\sum_{p \in R(v)} O_p}$$

με O_u και O_p τους αριθμούς των εξερχόμενων συνδέσμων των σελίδων u και p ενώ το σύνολο R(v) είναι το σύνολο από το οποίο παίρνει τιμές η μεταβλητή v.

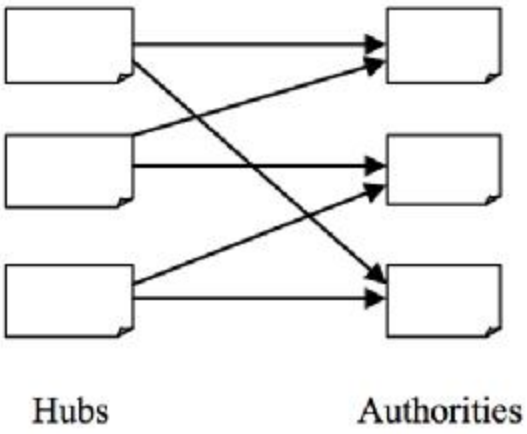
Σύμφωνα με τη φιλοσοφία του σταθμισμένου PageRank, όπως αυτή περιεγράφηκε παραπάνω, η μετρική του αλγόριθμου δίνεται τελικά από τη σχέση

$$PR(u) = (1 - d) + d \sum_{v \in B(u)} PR(v) W_{in}(v,u) W_{out}(v,u)$$

2.5.2.3 Ο αλγόριθμος HITS

Ο αλγόριθμος Hypertext Induced Topic Search (HITS) επινοήθηκε το 1997 από τον Jon Klienberg και βασίζεται σε έναν θεμελιώδη διαχωρισμό των ιστοσελίδων σε hubs και authorities. Ως hub ορίζεται μια σελίδα όταν έχει πολλούς εξερχόμενους υπερσυνδέσμους ενώ ως authority η σελίδα που έχει πολλούς εισερχόμενους υπερσυνδέσμους. Είναι σαφές πως αυτοί οι δύο τύποι ιστοσελίδων συνδέονται μεταξύ τους με τρόπο τέτοιο ώστε οι καλοί hubs να δείχνουν σε καλές authorities και το αντίστροφο, οι καλές authorities να δεικνύονται από καλά hubs. Βέβαια, μία ιστοσελίδα μπορεί να είναι ταυτόχρονα καλός hub και καλή authority.

Για να αντιληφθούμε τον τρόπο λειτουργίας του αλγόριθμου HITS βλέπουμε την παρακάτω εικόνα.



Εικόνα 2.2: Η σύνδεση των hubs και των authorities στον Παγκόσμιο Ιστό. [11]

Ο αλγόριθμος HITS αντιμετωπίζει τον Παγκόσμιο Ιστό ως ένα ορισμένης κατεύθυνσης γράφημα G(V,E), όπου V είναι το σύνολο των κορυφών που αναπαριστούν τις ιστοσελίδες και E το σύνολο των ακμών που αντιστοιχούν στους συνδέσμους. Η φιλοσοφία πίσω από αυτόν συνδυάζει μεθόδους αναζήτησης πληροφοριών βάσει περιεχομένου και ανάλυσης συνδέσμων.

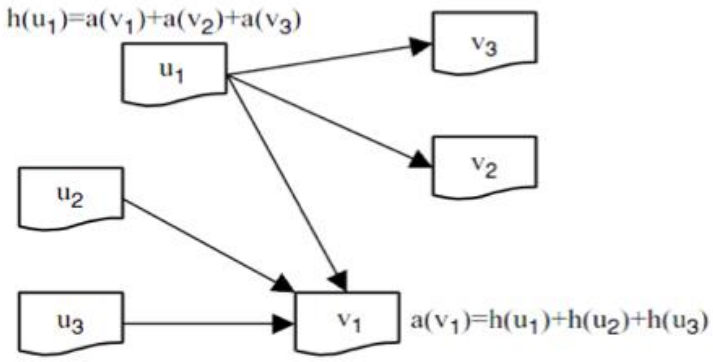
Το πρώτο βήμα του αλγόριθμου HITS είναι η συλλογή ενός αρχικού συνόλου (root set) μέσω μιας μηχανής αναζήτησης. Στη συνέχεια κατασκευάζεται το σύνολο βάσης (base set), μεγέθους μεταξύ 1000 και 1500 σελίδων, το οποίο περιλαμβάνει όλες τις σελίδες που δείχνουν στο root set. Το τρίτο βήμα είναι η κατασκευή ενός γραφήματος που χαρακτηρίζεται από τη δομή του base set. Αυτά τα τρία πρώτα βήματα αποτελούν το στάδιο της δειγματοληψίας του αλγόριθμου. Στη συνέχεια, σε αντίθεση με τον αλγόριθμο PageRank, για κάθε σελίδα δεν υπολογίζεται μόνο ένα μέτρο για την κατάταξή της αλλά δύο, ένα authority score A(x) και ένα hub score H(x) που υπολογίζονται επαναληπτικά μέσω των σχέσεων

$$H_{i+1}(x) = \sum_{(x,s)} A_i(s)$$

$$A_{i+1}(x) = \sum_{(p,s)} H_i(s)$$

όπου κάθε ζεύγος (x,y) αντιστοιχεί σε κάθε υπαρκτό υπερσύνδεσμο μεταξύ των ιστοσελίδων x και y. Ο αλγόριθμος HITS είναι συγκλίνων στις πρώτες περίπου 5 επαναλήψεις.

Μια βασική διαφορά μεταξύ του PageRank και του HITS είναι ο τρόπος με τον οποίο η μετρική του καθενός διαδίδεται εντός του συνόλου που δρα. Η διάδοση αυτή για τον HITS φαίνεται στην παρακάτω εικόνα



Εικόνα 2.3: Διάδοση των hub scores και authority scores στον αλγόριθμο HITS. [11]

Βλέπουμε, λοιπόν, πως σε κάθε βήμα η σελίδα u1 κληρονομεί το hub score της h(u1) από το άθροισμα των authority scores των σελίδων στις οποίες δείχνει (v1, v2 και v3). Στο αμέσως επόμενο βήμα, η σελίδα v1 κληρονομεί το authority score της a(v1) από το άθροισμα των hub scores των σελίδων που δείχνουν σε αυτές. Η διαδικασία αυτή τερματίζεται όταν τα hub score και authority score φτάσουν κάποια ορισμένη τιμή.

Ο αλγόριθμος HITS αντιμετωπίζει σημαντικά μειονεκτήματα. Ένα από αυτά είναι πως αντιμετωπίζει σημαντική δυσκολία να χαρακτηρίσει μια ιστοσελίδα ως hub ή authority επειδή κάποιες ιστοσελίδες μπορούν να παίζουν και τους δύο αυτούς ρόλους ταυτόχρονα. Επίσης, πολλές φορές παράγουν αποτελέσματα τα οποία δεν είναι ιδιαίτερα σχετικά με την αναζήτηση. Αυτό οφείλεται στο γεγονός ότι τα βάρη που παράγει ο αλγόριθμος σε κάθε σελίδα μπορούν να εξισωθούν με αποτέλεσμα να οδηγηθούμε σε λάθος αποτελέσματα. Την ίδια επίδραση στον αλγόριθμο έχει το γεγονός ότι ο αλγόριθμος δίνει την ίδια αξία στους αυτόματα δημιουργούμενους συνδέσμους οι οποίοι, όμως, μπορεί να μη σχετίζονται με αντικείμενα σημαντικά για την εκάστοτε αναζήτηση. Τέλος, ο αλγόριθμος HITS δεν παρουσιάζει υψηλή αποδοτικότητα σε πραγματικό χρόνο με αποτέλεσμα να μη μπορεί να χρησιμοποιηθεί σε ανάλογες μηχανές αναζήτησης.

2.5.2.4 Σύγκριση Αλγόριθμων Εξόρυξης Δομής στον Ιστό

Επιχειρώντας μια μικρή προσπάθεια σύγκρισης των αλγόριθμων που περιγράψαμε παραπάνω, καταλήγουμε στον πίνακα που ακολουθεί. Η σύγκριση αυτή, όπως βλέπουμε, γίνεται σε επίπεδο της τεχνικής που χρησιμοποιείται, της μεθοδολογίας που ο αλγόριθμος ακολουθεί, τις παραμέτρους που χρησιμοποιεί, την ποιότητα των αποτελεσμάτων και τους περιορισμούς που ο κάθε αλγόριθμος αντιμετωπίζει.

Όνομα Αλγόριθμου	PageRank	Σταθμισμένος PageRank	HITS
Τεχνική	Εξόρυξη Δομής στον Ιστό	Εξόρυξη Δομής στον Ιστό	Εξόρυξη Δομής στον Ιστό και Εξόρυξη Περιεχομένου στον Ιστό
Μεθοδολογία	Υπολογίζει ένα μέτρο για κάθε σελίδα σε πραγματικό χρόνο.	Το βάρος κάθε σελίδας υπολογίζεται μέσω των εισερχόμενων και των εξερχόμενων συνδέσμων και στη συνέχεια υπολογίζεται το μέτρο κάθε σελίδας.	Υπολογίζει τα hub score και authority score κάθε σελίδας
Παράμετροι	Εισερχόμενοι σύνδεσμοι	Εισερχόμενοι και εξερχόμενοι σύνδεσμοι	Εισερχόμενοι σύνδεσμοι, εξερχόμενοι σύνδεσμοι και περιεχόμενο σελίδας
Ποιότητα Αποτελεσμάτων	Μέτρια	Υψηλότερη από τον PageRank	Χαμηλότερη από τον PageRank
Περιορισμοί	Τα αποτελέσματα δε δίνονται στο χρόνο της αναζήτησης.	Η σχετικότητα των αποτελεσμάτων αγνοείται.	Προβλήματα αποδοτικότητας και σχετικότητας με την αρχική αναζήτηση.

Πίνακας 2.2: Σύγκριση των αλγόριθμων Εξόρυξης Δομής στον Ιστό.

2.6 Εξόρυξη Χρήσης στον Ιστό [3, 6, 12]

2.6.1 Εισαγωγή

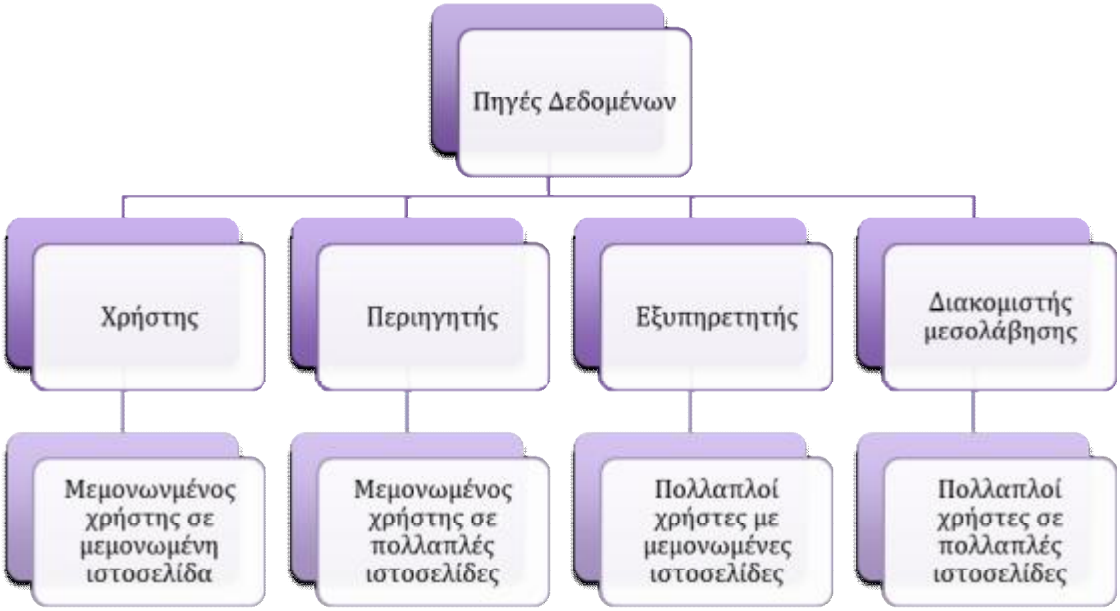
Η εξόρυξη χρήσης στον Ιστό προσπαθεί μέσω συγκεκριμένων τεχνικών να αναγνωρίσει ένα πρότυπο περιήγησης του χρήστη στον Παγκόσμιο Ιστό κατά την αλληλεπίδραση του με αυτόν. Με την σύνδεση κάθε χρήστη με ένα πρότυπο συμπεριφοράς κατά την αλληλεπίδραση με το διαδίκτυο μπορούμε να συνδέσουμε το χρήστη αυτό με όλους τους υπόλοιπους που επιδεικνύουν παρόμοια συμπεριφορά. Τότε, αυτή η ομάδα χρηστών με την παρόμοια διαδικτυακή συμπεριφορά μπορεί να αποτελεί στόχο παροχής προσωποποιημένων υπηρεσιών όπως είναι οι διαφημίσεις προϊόντων τα οποία θεωρούνται κατάλληλα για τον εκάστοτε χρήστη. Επίσης, μέσω της εξόρυξης χρήσης στον ιστό, οι προγραμματιστές έχουν στα χέρια τους ένα σημαντικό εργαλείο ελέγχου της πραγματικότητας και της λειτουργίας των λογισμικών τους αφού μπορούν να ελέγχουν αν η αναμενόμενη κατά το σχεδιασμό του λογισμικού συμπεριφορά ταιριάζει με αυτή που οι χρήστες πράγματι επιδεικνύουν.

2.6.2 Δεδομένα Χρήσης του Παγκόσμιου Ιστού

Υπάρχουν τέσσερεις κύριοι τύποι πηγών δεδομένων που περιέχουν πληροφορίες διαφόρων επιπέδων αναφορικά με τη χρήση του Παγκόσμιου Ιστού. Αυτά τα διάφορα επίπεδα είναι τα παρακάτω:

- Συλλογή δεδομένων στο επίπεδο του χρήστη (client level collection)
- Συλλογή δεδομένων στο επίπεδο του περιηγητή (browser level collection)
- Συλλογή δεδομένων στο επίπεδο του εξυπηρετητή (server level collection)
- Συλλογή δεδομένων στο επίπεδο του διακομιστή μεσολάβησης (proxy level collection)

Στο παρακάτω γράφημα βλέπουμε την κατάταξη των δεδομένων που χρησιμοποιούνται στην εξόρυξη χρήσης στον Ιστό.



Σχήμα 2.4: Οι διάφορες πηγές δεδομένων για επεξεργασία κατά την εξόρυξη χρήσης στον Ιστό.

Η συλλογή δεδομένων στο επίπεδο του χρήστη περιλαμβάνει δεδομένα που συλλέγονται μέσω java μικροεφαρμογών και περιλαμβάνουν στοιχεία της συμπεριφοράς μεμονωμένων χρηστών σε μεμονωμένες ιστοσελίδες. Η συλλογή των δεδομένων αυτών απαιτεί τη συμμετοχή του χρήστη ο οποίος θα πρέπει να επιτρέψει την ενεργοποίηση αυτών των java μικροεφαρμογών. Το επίπεδο αυτό συλλογής δεδομένων χαρακτηρίζεται από το μεγάλο πλεονέκτημα του να μπορεί να καταγράψει κάθε ενέργεια του χρήστη, ακόμα και την πιο στοιχειώδη όπως είναι για παράδειγμα η ανανέωση μιας ιστοσελίδας.

Μια εναλλακτική μέθοδος συλλογής δεδομένων είναι αυτή που αφορά στο επίπεδο του περιηγητή. Η συγκεκριμένη, αυτή, μέθοδος απαιτεί την τροποποίηση του περιηγητή ώστε να καταγράφει τη συμπεριφορά ενός μόνο χρήστη σε περισσότερες του ενός ιστοσελίδες. Οι τροποποιήσεις που μπορούν να γίνουν στον πηγαίο κώδικα των περιηγητών αυξάνουν τις δυνατότητές τους με αποτέλεσμα να μπορούν να παρέχουν πολυσχιδείς πληροφορίες.

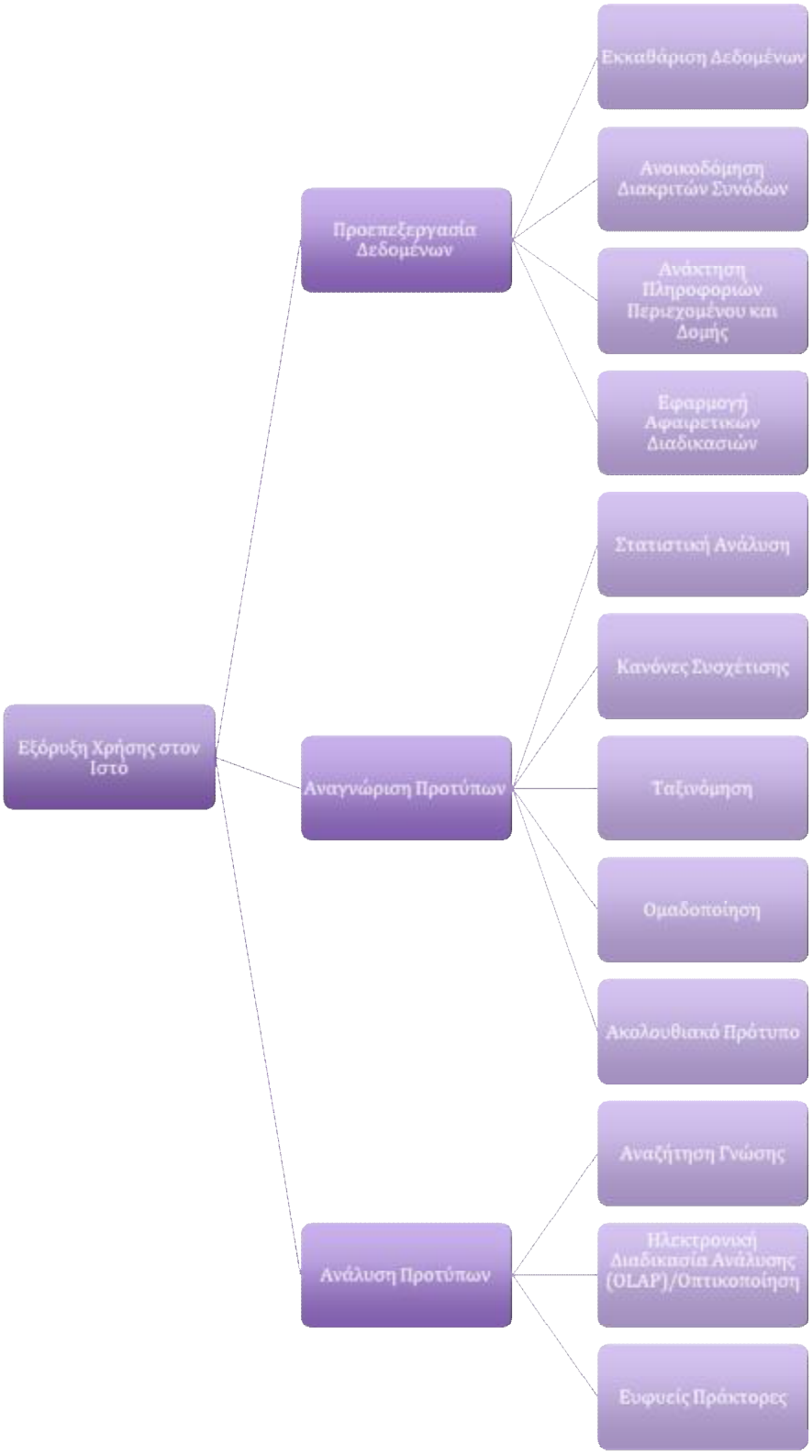
Η συλλογή δεδομένων στο επίπεδο του εξυπηρετητή αφορά στη συλλογή δεδομένων πολλαπλών χρηστών σε μια μόνο ιστοσελίδα. Τα δεδομένα αυτά αποθηκεύονται είτε σε μορφή common log είτε σε μορφή extended log. Ένας εναλλακτικός τρόπος καταγραφής των συγκεκριμένων δεδομένων είναι μέσω αναλυτών δικτύου TCP/IP (TCP/IP packet sniffers). Αυτοί οι αναλυτές δικτύων λειτουργούν παρακολουθώντας την κίνηση του διαδικτύου και ανακτούν άμεσα τα δεδομένα χρήσης.

Τέλος, η συλλογή δεδομένων στο επίπεδο των διακομιστών μεσολάβησης (proxies) βασίζεται στη λειτουργία των proxies που χρησιμοποιούν οι διαδικτυακοί πάροχοι στους πελάτες τους. Οι proxies αυτοί καταγράφουν και αποθηκεύουν τη συμπεριφορά πολλαπλών χρηστών σε πολλαπλές ιστοσελίδες.

2.6.3 Διαδικασία Εξόρυξης Χρήσης στον Ιστό: Μια Γενική Θεώρηση

Στο ακόλουθο σχήμα βλέπουμε τη γραφική αναπαράσταση της διαδικασίας της εξόρυξης χρήσης στον παγκόσμιο Ιστό. Για τα διάφορα στάδια αυτής μπορούμε να πούμε περιληπτικά τα παρακάτω:

- Το πρώτο στάδιο της εξόρυξης χρήσης στον ιστό είναι η προεπεξεργασία των δεδομένων τα οποία από τη φύση τους εμπεριέχουν πολύ θόρυβο. Θεωρείται το πιο δύσκολο στάδιο της διαδικασίας εξόρυξης χρήσης εξαιτίας της παρουσίας ατελών και ασυνεπών δεδομένων στις καταγραφές των εξυπηρετητών. Η προεπεξεργασία των δεδομένων λαμβάνει χώρα σε τέσσερεις διακριτές φάσεις. Στην αρχή γίνεται η εκκαθάριση δεδομένων όπου τα πρωτογενή δεδομένα απαλλάσσονται από το σύνολο του θορύβου καθώς και κάθε άχρηστης πληροφορίας. Επίσης, αφαιρείται οποιαδήποτε πληροφορία αφορά σε αναφορές για το στυλ του αρχείου, τα γραφήματα και τα αρχεία ήχου και εικόνας. Στη δεύτερη φάση λαμβάνει χώρα η αναγνώριση των διακριτών συνόδων των χρηστών στα εκκαθαρισμένα αρχεία και η ανοικοδόμηση αυτών. Στη φάση αυτή αντιμετωπίζουμε το πρόβλημα ότι ο proxy δεν επιτρέπει να χρησιμοποιηθεί η διεύθυνση IP του χρήστη για την ταυτοποίησή του. Ωστόσο, αυτό το εμπόδιο ξεπερνιέται με τη χρήση των cookies. Η τρίτη φάση αποτελείται από την ανάκτηση πληροφοριών αναφορικά με το περιεχόμενο και τη δομή των δεδομένων και η τέταρτη και τελική φάση αφορά στην εφαρμογή τεχνικών αναγνώρισης προτύπων για την ανάπτυξη διαφόρων αφαιρετικών διαδικασιών.
- Το δεύτερο στάδιο της εξόρυξης χρήσης στον ιστό είναι η ανακάλυψη των υπαρχόντων προτύπων. Είσοδος της διαδικασίας αυτής είναι τα αρχεία όπως αυτά δημιουργήθηκαν από τη διαδικασία της προεπεξεργασίας και σκοπός της

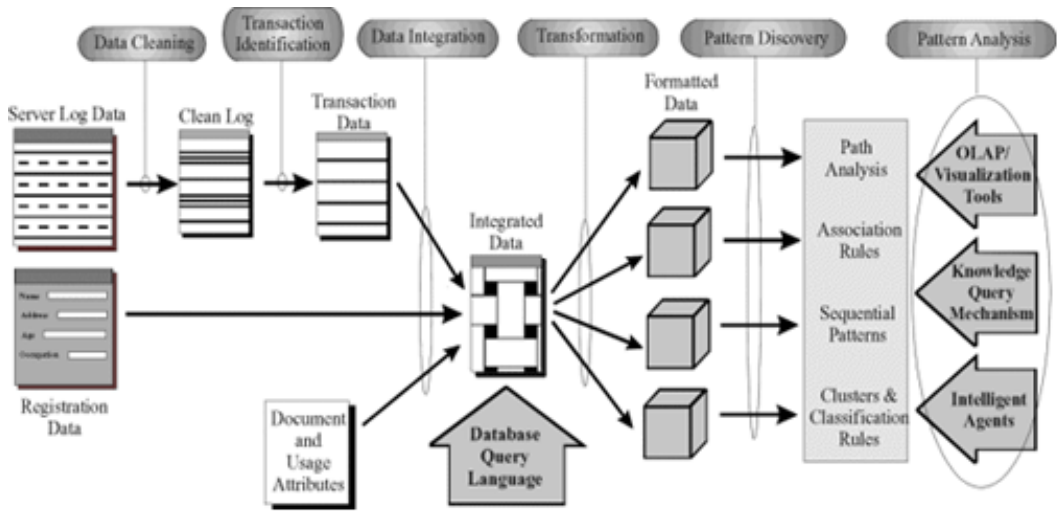


Σχήμα 2.5: Στάδια Εξόρυξης Χρήσης στον Ιστό

είναι η εφαρμογή διάφορων τεχνικών από διάφορα επιστημονικά πεδία (π.χ. στατιστική, αναγνώριση προτύπων μηχανική μάθηση κτλ.). Οι αλγόριθμοι και οι τεχνικές του σταδίου αυτού θα συζητηθούν στις επόμενες παραγράφους.

- Το τελευταίο στάδιο της διαδικασίας εξόρυξης χρήσης στον ιστό είναι η ανάλυση των προτύπων. Στο στάδιο αυτό, όλα τα πρότυπα που αναδύθηκαν κατά την ανακάλυψή τους διαχωρίζονται σε ενδιαφέροντα και μη ενδιαφέροντα, ανάλογα με τις εκάστοτε ανάγκες του κάθε χρήστη. Τα αποτελέσματα αυτής της φάσης χρησιμοποιούνται σε πολυάριθμες εφαρμογές όπως το ηλεκτρονικό εμπόριο, οι βελτιστοποιήσεις των ιστοσελίδων καθώς και η προσωποποίηση αυτών. Υπάρχουν τρεις κυρίαρχες τεχνικές που χρησιμοποιούνται κατά την ανάλυση των προτύπων. Η πρώτη από αυτές αφορά στους μηχανισμούς αναζήτησης γνώσης (Knowledge Query Mechanism) οι οποίοι εφαρμόζουν τη γλώσσα SQL (Structured Query Language) ώστε να εξάγουν τα χρήσιμα πρότυπα από ένα σύνολο αναγνωρισμένων προτύπων. Μια δεύτερη τεχνική ανάλυσης προτύπων είναι η χρήση εργαλείων OLAP (OnLine Analytical Processing – Ηλεκτρονική Διαδικασία Ανάλυσης). Οι έξοδοι των OLAP εργαλείων χρησιμοποιούνται ως είσοδοι στα συστήματα εξόρυξης γνώσης ή σε συστήματα οπτικοποίησης. Επίσης, είναι πιθανή η εφαρμογή τεχνικών οπτικοποίησης όπου η αντιστοίχιση χρωμάτων στις εκάστοτε τιμές των μεταβλητών οδηγεί στη δημιουργία γραφικών προτύπων. Τέλος, υπάρχει η τεχνική της χρήσης ευφυών πρακτόρων για την ανάλυση των αναδεδειγμένων προτύπων.

Τ παραπάνω βήματα φαίνονται εικονικά ακολούθως:



Εικόνα 2.4: Διαδικασία εξόρυξης χρήσης στον Ιστό. [27]

Στον ακόλουθο πίνακα φαίνονται ενδεικτικά κάποια από τα εργαλεία που μπορούν να χρησιμοποιηθούν κατά τη διαδικασία της εξόρυξης χρήσης.

Εργαλείο	Λειτουργία
	Προεπεξεργασία Δεδομένων
Data Preparator	Εκτελεί την εκκαθάριση των αρχείων, την εξόρυξη και την μετατροπή των δεδομένων ώστε να είναι έτοιμα για να χρησιμοποιηθούν στην ανακάλυψη προτύπων.
Sumatra TT	Εργαλείο μετατροπής δεδομένων ανεξάρτητο πλατφόρμας.
Lisp Miner	Εκτελεί την επεξεργασία των δεδομένων μέσω ανάλυσης του click stream και των δεδομένων που συλλέχθηκαν.
SpeedTracer	Εξορύσσει δεδομένα από καταγραφές διαδικτυακών εξυπηρετητών και ανακατασκευάζει τις συνόδους ώστε να μπορούν να αναγνωριστούν
	Ανακάλυψη Προτύπων
SEWEBAR-CMS	Χρησιμοποιείται για την ανακάλυψη προτύπων με ανάδειξη των κανόνων συσχέτισης.
iMiner	Ανακαλύπτει ομάδες δεδομένων χρησιμοποιώντας ασαφείς αλγόριθμους (fuzzy algorithms).
Argonaut	Αναπτύσσει τα πρότυπα χρήσιμων δεδομένων με χρήση ακολουθιακών προτύπων.
MiDas	Ανακαλύπτει πρότυπα περιήγησης μέσα από αρχεία καταγραφών χρήσης στο διαδίκτυο.
	Ανάλυση Προτύπων
Webalizer	Δημιουργεί ιστοσελίδες μετά από την ανάλυση των προτύπων.
Naviz	Εργαλείο οπτικοποίησης το οποίο συνδυάζει γραφικά σε 2 διαστάσεις και ομαδοποίηση την συσχετιζόμενων ιστοσελίδων. Έχει τη δυνατότητα να περιγράφει τα πρότυπα της περιήγησης στο διαδίκτυο.
Webviz	Αναλύει τα πρότυπα και τα εξάγει σε γραφική μορφή.
WebMiner	Εξορύσσει τα χρήσιμα πρότυπα και παρέχει τις συγκεκριμένες πληροφορίες που αναζητά ο εκάστοτε χρήστης.
Stratdyn	Παρέχει οπτικοποίηση των προτύπων.

Πίνακας 2.3: Εργαλεία που χρησιμοποιούνται κατά την εξόρυξη χρήσης στον ιστό.

2.6.4 Τεχνικές Εξόρυξης Χρήσης στον Ιστό

Η ανακάλυψη προτύπων βασίζεται σε αλγόριθμους οι οποίοι έχουν αναπτυχθεί στα πλαίσια διαφορετικών επιστημονικών κλάδων, όπως είναι η στατιστική, η μηχανική μάθηση και η αναγνώριση προτύπων. Στις ακόλουθες παραγράφους δίνονται οι κυριότερες τεχνικές που έχουν εφαρμοστεί για την εξόρυξη σε αρχεία χρήσης κατά τις διάφορες συνόδους των χρηστών στον Παγκόσμιο Ιστό. Κατά τη συγκεκριμένη διαδικασία, με τον όρο «σύνοδος» ορίζουμε κάθε διατεταγμένη ακολουθία ιστοσελίδων που ζητήθηκε από το χρήστη. Βέβαια, όπως αναφέραμε και σε προηγούμενη παράγραφο, η αναγνώριση των διάφορων συνόδων είναι ιδιαίτερα δύσκολη διαδικασία. Οι τεχνικές που θα συζητηθούν περιλαμβάνουν τη στατιστική ανάλυση, τα ακολουθιακά πρότυπα, την ομαδοποίηση, την ταξινόμηση, την εξαγωγή κανόνων συσχέτισης καθώς και τη μοντελοποίηση εξάρτησης.

2.6.4.1 Στατιστική Ανάλυση

Η στατιστική ανάλυση είναι η πιο συνηθισμένη μέθοδος εξόρυξης γνώσης σε μια ιστοσελίδα αναφορικά με τους επισκέπτες της. Έχουν υλοποιηθεί αρκετά εργαλεία ανάλυσης της κίνησης στο διαδίκτυο τα οποία έχουν τη δυνατότητα να παράγουν περιοδικές αναφορές των στατιστικών στοιχείων των κινήσεων αυτών. Έτσι, μπορούμε να έχουμε πρόσβαση σε πληροφορίες όπως το ποιες είναι οι συχνότερα επισκεπτόμενες σελίδες, πόσος είναι ο μέσος χρόνος παραμονής σε κάθε μια από αυτές ή ακόμα πόσο είναι το μέσο μήκος διαδρομής της περιήγησης ενός χρήστη μέσα σε κάθε μία από αυτές. Εκτός από τετριμμένα στατιστικά στοιχεία αναφορικά με τη χρήση των ιστοσελίδων, υπάρχει η δυνατότητα παροχής προηγμένων στατιστικών πληροφοριών, όπως είναι για παράδειγμα η ανάλυση σφαλμάτων. Έτσι, έχουμε επίσης πρόσβαση σε πληροφορίες όπως είναι η ανίχνευση προσπαθειών για μη εξουσιοδοτημένη πρόσβαση σε ιστοσελίδες καθώς και τον τόπο από τον οποίο προέρχονται αυτές οι προσπάθειες. Επίσης, μπορεί να γίνει ανίχνευση των άκυρων URLs που τυπώνονται συχνότερα και, γενικότερα, πληροφορίες οι οποίες αθροιστικά μας παρέχουν γνώση που πιθανότατα θα φανεί χρήσιμη κατά τις προσπάθειες βελτιστοποίησης των συστημάτων, ενίσχυσης του επιπέδου ασφάλειάς τους καθώς και επιλογής των σωστών εμπορικών διαδικασιών για την προώθηση των αγαθών που προσφέρουν.

2.6.4.2 Ακολουθιακά Πρότυπα

Η συγκεκριμένη τεχνική αφορά στην εύρεση προτύπων εντός των συνόδων. Τα πρότυπα αυτά θα αποτελούνται από μια διαδοχή αντικειμένων σε χρονική διάταξη. Τα ακολουθιακά πρότυπα είναι, επίσης, σημαντικά για τα εμπορικά τμήματα εταιρειών που δραστηριοποιούνται διαδικτυακά αφού μπορούν να προβλέψουν πρότυπα μελλοντικών επισκέψεων σε συγκεκριμένους ιστότοπους. Με αυτόν τον τρόπο γίνεται δυνατή η αποτελεσματική τοποθέτηση διαφημίσεων που βασίζονται σε συγκεκριμένα κριτήρια αναφορικά με τους χρήστες του διαδικτύου. Επίσης, τα ακολουθιακά πρότυπα μπορούν να προβλέψουν μελλοντικές τάσεις της συμπεριφοράς των χρηστών και να γίνουν εργαλεία για την ανάλυση ομοιοτήτων.

2.6.4.3 Ταξινόμηση

Η ταξινόμηση είναι η διαδικασία με την οποία τα δεδομένα κατηγοριοποιούνται σε μια από τις προκαθορισμένες τάξεις. Η ταξινόμηση γίνεται με τη βοήθεια αλγόριθμων μάθησης με επίβλεψη όπως είναι για παράδειγμα τα δέντρα αποφάσεων, οι αλγόριθμοι Naïve-Bayes, ο αλγόριθμος k-κοντινότερων γειτόνων κτλ. Τα αποτελέσματα αυτής της διαδικασίας μπορούν να συνδυαστούν με τα αντίστοιχα αποτελέσματα της Εξόρυξης Περιεχομένου και της Εξόρυξης Δομής για καλύτερη κατάταξη των ιστοσελίδων σε μηχανές αναζήτησης, την κατάταξη αρχείων του διαδικτύου καθώς και για την κατασκευή πολυεπίπεδων βάσεων πληροφοριών στον Παγκόσμιο Ιστό. Επιπροσθέτως, σε κάθε μεμονωμένο επιστημονικό κλάδο υπάρχει πάντα η ανάγκη αυτόματης κατάταξης των αντίστοιχων αρχείων ανάλογα με το ακολουθούμενο πρότυπο κατάταξης (δηλαδή ανάλογα με το αν ταξινομούνται βάσει θέματος ή με χρονολογική σειρά κτλ.). Έτσι, βάσει της εξόρυξης χρήσης μπορεί να βελτιωθεί η ποιότητα της ταξινόμησης αυτού του είδους.

2.6.4.4 Εξαγωγή Κανόνων Συσχέτισης

Είναι συχνό φαινόμενο στα δεδομένα που αναλύονται μέσω της Εξόρυξης Χρήσης να συναντώνται ιστοσελίδες οι οποίες πολύ συχνά εμφανίζονται μαζί στην ίδια σύνοδο. Η τεχνική της εξαγωγής κανόνων συσχέτισης χρησιμοποιείται για βρει τον τρόπο με τον οποίο αυτές οι συχνά εμφανιζόμενες μαζί σελίδες συνδέονται μεταξύ τους. Κυριότερη εφαρμογή των αλγόριθμων εξαγωγής κανόνων συσχέτισης είναι αυτή που αφορά στην ανάλυση του «καλαθιού» του καταναλωτή όπου επιθυμείτε η ανακάλυψη των συσχετισμών των προϊόντων που αγοράζουν οι καταναλωτές μαζί. Αυτή η γνώση είναι σημαντική για τους εμπορικούς αντιπροσώπους αφού μπορούν να προωθήσουν διαφορετικά τα προϊόντα τους ανάλογα με τις συνήθειες των καταναλωτών.

2.6.4.5 Ομαδοποίηση

Ομαδοποίηση ονομάζεται η διαδικασία με την οποία τα αντικείμενα ενός συνόλου τοποθετούνται σε ομάδες όμοιων αντικειμένων. Κάθε μια από αυτές τις ομάδες αποτελεί μια νέα συλλογή αντικειμένων μέσα στην οποία η ομοιότητα των περιεχόμενων αντικειμένων είναι μέγιστη και η ομοιότητα με τα αντικείμενα των άλλων ομάδων είναι ελάχιστη. Στα πλαίσια της εξόρυξης χρήσης στον Ιστό, δύο είναι τα διαφορετικά είδη ομάδων τα οποία ενδιαφέρουν, αυτές που αφορούν στους χρήστες και αυτές που αφορούν στις ιστοσελίδες. Η πρώτη κατηγορία ομάδων αφορά στην ομαδοποίηση των χρηστών ανάλογα με τις συμπεριφορές τους που καταδεικνύουν τα όμοια πρότυπα περιήγησης. Αυτή η γνώση είναι χρήσιμη για την εξαγωγή συμπερασμάτων σχετικά με το δημογραφικό χαρακτήρα των χρηστών ώστε το σύνολο των χρηστών να μπορεί να τμηματοποιηθεί βάσει των αναγκών των εφαρμογών του ηλεκτρονικού εμπορίου ή των αναγκών των παρόχων διαδικτυακών υπηρεσιών. Η δεύτερη κατηγορία ομάδων αφορά στην ομαδοποίηση των σελίδων ανάλογα με το περιεχόμενό τους. Είναι ευνόητο πως αυτή η τεχνική είναι σημαντική για τη σωστή λειτουργία των μηχανών αναζήτησης.

ΚΕΦΑΛΑΙΟ 3
ΠΑΚΕΤΑ ΛΟΓΙΣΜΙΚΟΥ ΕΛΕΥΘΕΡΟΥ ΚΩΔΙΚΑ ΓΙΑ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟΝ ΙΣΤΟ

3.1 Εισαγωγή

Στα προηγούμενα Κεφάλαια έγινε εκτενής αναφορά στις έννοιες της Εξόρυξης Δεδομένων και της Εξόρυξης Δεδομένων στον Ιστό καθώς και τις διάφορες υποκατηγορίες αυτών. Τα οφέλη των τεχνικών αυτών στην ποιότητα της παραγόμενης γνώσης καθιστούν επιτακτική την ανάγκη σχεδιασμού και ανάπτυξης πακέτων λογισμικού που να εφαρμόζουν τις τεχνικές αυτές στην πράξη. Ωστόσο, το κόστος των πακέτων αυτών, όπως συμβαίνει άλλωστε και με οποιοδήποτε αντικείμενο τεχνολογικής εξέλιξης, πολλές φορές ξεπερνά την αγοραστική δύναμη μεμονωμένων ιδιωτών, ερευνητικών ομάδων, μικρών επιχειρήσεων και άλλων φορέων. Γι' αυτό κρίνεται σημαντική η ανάπτυξη πακέτων αντίστοιχου λογισμικού ελεύθερου κώδικα. Βέβαια, τα οφέλη του λογισμικού ελεύθερου κώδικα δεν περιορίζονται πάντα στον οικονομικό άξονα αλλά σχετίζονται και με άλλες δυνατότητες που προσφέρονται στους χρήστες, όπως για παράδειγμα η δυνατότητα επέκτασης ενός ήδη υπάρχοντος κώδικα. Γι' αυτούς τους λόγους η παρούσα εργασία στοχεύει στην παρουσίαση τέτοιου είδους πακέτων λογισμικού.

Τα πακέτα λογισμικού για Εξόρυξη στον Ιστό είναι σχεδιασμένα ώστε να δύνανται να περιηγούνται στο διαδίκτυο, ανάμεσα σε πολυάριθμες HTML σελίδες που περιέχουν τόσο κείμενο, δομημένο ή μη, όσο και εικόνες, ήχο ή βίντεο. Στις ιστοσελίδες αυτές αναζητούν πληροφορίες που εξαρτώνται κάθε φορά από τις επιθυμίες του χρήστη με τα αποτελέσματα της αναζήτησης αυτής να παρουσιάζονται στο χρήστη ανάλογα με τη σχετικότητά τους με το αίτημα και ανάλογα με την ικανοποίηση ή μη συγκεκριμένων παραμέτρων που ο χρήστης θέτει, όπως για παράδειγμα χρονικές παράμετροι.

Στο Κεφάλαιο αυτό θα παρουσιάσουμε τα πακέτα λογισμικού ανοιχτού κώδικα για Εξόρυξη στον Ιστό. Όπως θα παρατηρήσει ο οποιοσδήποτε αναγνώστης έχει μια στοιχειώδη εμπειρία από λογισμικό Εξόρυξης στον Ιστό, οι αναφορές που θα γίνουν δεν εξαντλούν το εύρος των υπαρχόντων πακέτων λογισμικού. Ωστόσο, η περιγραφή μας θα περιοριστεί στους κυριότερους εκπρόσωπους αυτής της κατηγορίας λογισμικού. Βέβαια, ένας στείρος διαχωρισμός των πακέτων αυτών ανά κατηγορία είναι δύσκολος αφού τα σύγχρονα πακέτα έχουν τη δυνατότητα να εκτελούν περισσότερες της μιας εργασίες. Έτσι, στην παρούσα εργασία θα περιγράφουν χωρίς να εντάσσονται σε κάποια υποκατηγορία (Εξόρυξη Χρήσης, Δομής ή Περιεχομένου).

3.2 DeiXTo [16]

Εικόνα 3.1: Το λογότυπο του πακέτου λογισμικού DEiXTo. [16]

Το DEiXTo είναι ένα εργαλείο εξόρυξης διαδικτυακών δεδομένων το οποίο έχει αναπτυχθεί -και συνεχίζει να αναπτύσσεται- από τους Έλληνες Φώτη Κόκκορη και Κώστα Ντόνα.

Αποτελεί εφαρμογή για περιβάλλον MS Windows και βασίζεται στο μοντέλο W3C.
Με το DEiXTo, οι χρήστες έχουν τη δυνατότητα να κατασκευάζουν υψηλής ακρίβειας wrappers οι οποίοι να περιγράφουν τα δεδομένα τα οποία θα πρέπει να περισυλλεχθούν από τις διάφορες ιστοσελίδες.

Τα επιμέρους τμήματα του DEiXTo είναι το GUI DEiXTo, το DEiXToBOT και το DEiXToCLE. Το GUI DEiXTo είναι μια απλή εφαρμογή για την εκτέλεση εργασιών εξόρυξης μικρής εμβέλειας και χαρακτηρίζεται από την ιδιαίτερα φιλική προς το χρήστη διεπαφή, όπως αυτή φαίνεται στην εικόνα που ακολουθεί.

Εικόνα 3.2: Η διεπαφή του GUI DEiXTo. [16]

Οι wrappers που κατασκευάζονται με το GUI DEiXTo μπορούν στη συνέχεια να χρησιμοποιηθούν από το DEiXToBOT για την εξόρυξη δεδομένων συγκεκριμένου ενδιαφέροντος. Το DEiXToBOT είναι στην ουσία ένας εξομοιωτής περιηγητή στο διαδίκτυο ο οποίος επειδή έχει ενσωματώσει τις σύγχρονες τεχνολογίες Pearl παρέχει τη δυνατότητα παροχής πλήρως εξατομικευμένων δυνατοτήτων. Στην τεχνολογία του DEiXToBOT βασίζεται και το DEiXToCLE, το οποίο αποτελεί αυτόνομο πρόγραμμα με τη δυνατότητα να εφαρμόσει τους wrappers που έχουν ήδη δημιουργηθεί σε πολλαπλές σελίδες και να παρουσιάσει τα αποτελέσματά του σε διάφορες μορφές.

Τα πλεονεκτήματα του DEiXTo είναι σημαντικά. Αρχικά, μπορεί να εφαρμοστεί σε μεγάλο πλήθος ιστοσελίδων και να παράγει αποτελέσματα υψηλής ακρίβειας. Επίσης, οι παρεχόμενες δυνατότητες οδηγούν στη δημιουργία άρτιων κανόνων εξόρυξης. Τέλος, οι wrappers που κατασκευάζονται με το GUI DEiXTo μπορούν να προγραμματιστούν ώστε να λειτουργούν αυτόματα με αποτέλεσμα την οικονομία χρόνου και προσπάθειας εκ μέρους του χρήστη.

Σχετικά με τις παρεχόμενες δυνατότητες που αναφέρθηκαν προηγουμένως, αυτές είναι πράγματι πολυάριθμες. Το GUI DEiXTo χαρακτηρίζεται, όπως προείπαμε, από μια ιδιαίτερα φιλική για το χρήστη διεπαφή. Το πρόγραμμα λειτουργεί χωρίς να χρειάζεται περεταίρω προγραμματισμός και κατ' επέκταση μπορεί να χρησιμοποιηθεί και από μη εξειδικευμένους χρήστες. Κατασκευάζει wrappers, εκκαθαρίζει ιστοσελίδες από HTML επισημάνσεις, είναι αλγόριθμος που χαρακτηρίζεται από μεγάλη ταχύτητα, υψηλή απόδοση και σημαντική ευελιξία, η ακρίβεια των αποτελεσμάτων του αγγίζει κατά περιπτώσεις το 100%, δύναται να εξορύσσει δεδομένα σε πολλαπλές σελίδες καθώς και να παράγει αποτελέσματα εξάγοντάς τα σε διάφορες μορφές καθώς και σε μορφή XML. Εκτός από κείμενο, έχει τη δυνατότητα να εξορύσσει URLs καθώς και πηγαίο κώδικα HTML, μπορεί να προγραμματιστεί μέσω MS Scheduler ώστε να πραγματοποιεί εξορύξεις αυτόματα και τα παραγόμενα αποτελέσματά του είναι συμβατά με τους εκτελεστικούς αλγόριθμους DEiXTo, δηλαδή το DEiXToBOT και το DEiXTOCLE, λογισμικά τα οποία είναι γνωστά με τον ενιαίο όρο DEiXTo Executor.

Επιπροσθέτως, το DEiXTo Executor παρέχει μερικές δυνατότητες ακόμα. Εφαρμόζει τους αλγόριθμούς του χρησιμοποιώντας τους wrappers του GUI DEiXTo μέσω γραμμής εντολών ενώ υποστηρίζει ακόμη περισσότερες μορφές για την εξαγωγή των αποτελεσμάτων όπως η CVS, η .xls και η .ods. Παρέχει υποστήριξη σε βάσεις δεδομένων μέσω του DBI της Pearl (Database Independent Interface) καθώς και μέσω αρχείων dbconfig. Υποστηρίζει, επίσης, proxies και μπορεί να προγραμματιστεί ώστε τα χρησιμοποιούμενα wrappers να δημιουργούνται αυτόματα χωρίς παρέμβαση του χρήστη. Τέλος, όπως και όλα τα προϊόντα DEiXTo, διατίθεται ελεύθερα, χωρίς οικονομική επιβάρυνση.

3.3 Bixo [17]

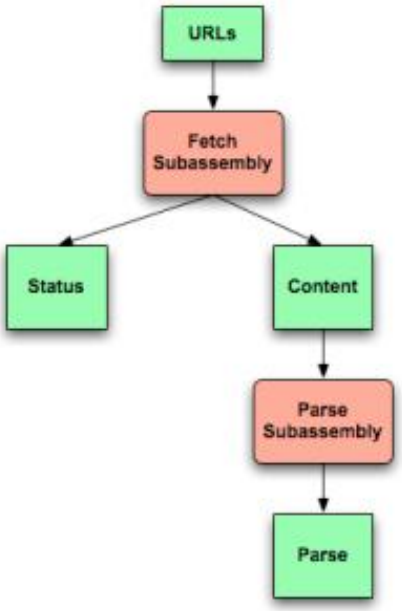
Το Bixo αποτελεί πακέτο λογισμικού το οποίο αναπτύχθηκε με αφορμή την ανάγκη που εξέφρασαν επιχειρήσεις για εργαλεία Εξόρυξης στον Ιστό που να μπορούν να προσαρμοστούν εύκολα σε επικαλύπτουσες ροές εργασίας (Cascading workflows).

Εικόνα 3.3: Το λογότυπο του πακέτου λογισμικού Bixo. [17]

Μέχρι εκείνη τη στιγμή, η μόνη διαθέσιμη τεχνολογία ήταν ο μηχανισμός διάσχισης Apache Nutch. Ωστόσο, η μετατροπή του Nutch με τρόπο τέτοιο ώστε να είναι δυνατή η χρήση του στο συγκεκριμένο περιβάλλον εργασίας κρίθηκε ως αναποτελεσματική αφού θα αποτελούσε επίπονη διαδικασία η οποία, όμως, δε θα παρήγαγε προϊόν μεγάλης αποτελεσματικότητας και ακρίβειας. Έτσι, η δημιουργία ενός νέου λογισμικού αποτελούσε μονόδρομο.

Έτσι αναπτύχθηκε το πακέτο λογισμικού Bixo το οποίο λειτουργεί σε 1.6 ή μεταγενέστερη έκδοση Java Runtime Environment (JRE). Ενώ δεν αποτελεί κάποια μετεξέλιξη του Apache Nutch, κάνει μεγάλη χρήση αυτού όπως και άλλων λογισμικών ανοιχτού κώδικα, όπως το HTTPClient4, το Tika, το Cascading και το Hadoop.

Η αρχιτεκτονική του Bixo είναι αλά δομημένη και αποτελείται από έναν αριθμό επικαλυπτόμενων εντολών και δομών οι οποίες κατά το συνδυασμό τους δημιουργούν ένα περιβάλλον επεξεργασίας δεδομένων το οποίο ξεκινά με ένα σύνολο URLs προς προσπέλαση και τερματίζεται με την εξαγωγή πληροφοριών ως αποτέλεσμα της ανάλυσης ιστοσελίδων HTML.



Εικόνα 3.4: Η αρχιτεκτονική του πακέτου λογισμικού Bixo. [17]

Στη δόμηση της αρχιτεκτονικής του Bixo, όπως αυτή φαίνεται και στην εικόνα που προηγήθηκε, διακρίνουμε δύο σημαντικά τμήματα, αυτό όπου συγκεντρώνεται οι επιθυμητές ιστοσελίδες μέσω των UrlDatum wrappers (Fetch Subassembly) και αυτό όπου γίνεται η επεξεργασία των ιστοσελίδων που συγκεντρώθηκαν (Parse Subassembly). Το Parse Subassembly χρησιμοποιεί το λογισμικό Tika για να διαχειριστεί τη διαδικασία εξόρυξης κειμένου από δεδομένα πολλών μορφών, όπως για παράδειγμα οι HTML σελίδες.

Από τα δύο προαναφερθέντα στάδια, αυτό που είναι το πιο δύσκολο είναι της περισυλλογής των ιστοσελίδων. Κατά το στάδιο αυτό, οι εργασίες γίνονται σε διακριτά στάδια τα οποία σε μια γενική θεώρηση είναι τα εξής: Αρχικά, ομαδοποιούνται οι URLs βάσει ονόματος και στη συνέχεια, για κάθε μια από αυτές τις ιστοσελίδες, εξάγονται οι εκάστοτε διευθύνσεις IP και τα εκάστοτε αρχεία robots.txt ενώ τα URLs φιλτράρονται με εφαρμογή Robot Exclusion Πρωτοκόλλων. Ακολούθως, ομαδοποιούνται οι διευθύνσεις IP συνήθως σε ομάδες των 10 ανά κοινή IP ενώ ο αριθμός αυτών μπορεί να υποστεί περιορισμό ανάλογα με τις επιθυμίες του χρήστη. Επίσης, ορίζονται οι απαιτούμενοι χρόνοι για την εφαρμογή του αλγόριθμου ανάλογα με τον αριθμό των URLs και τη χρονική καθυστέρηση που εισάγει ο μηχανισμός διάσχισης. Στο επόμενο βήμα, τα URLs διαμοιράζονται σε n κατηγορίες ανάλογα με την τιμή n μεταβλητών έτσι ώστε σε κάθε ομάδα να περιλαμβάνονται πάντα URLs με την ίδια IP διεύθυνση. Τέλος, ακολουθεί η περισυλλογή της κάθε ομάδας URLs.

Ένα μεγάλο μειονέκτημα της λειτουργίας του Bixo είναι πως, όπως και στην περίπτωση κάθε κάθετου μηχανισμού διάσχισής, δεν είναι δυνατή η εκ των προτέρων γνώση του χρονικού διαστήματος που θα χρειαστεί για την διάσχιση όλων των URLs αφού το χρονικό διάστημα για τη διάσχιση n URLs εξαρτάται από τον ιστότοπο που απαιτεί το μεγαλύτερο χρονικό διάστημα για τη διάσχισή του, το πλήθος των URLs, την πολιτική που ακολουθούν οι robots.txt, την απόδοση του εξυπηρετητή κτλ. Αυτό το μειονέκτημα έχει γίνει μια σημαντική προσπάθεια να περιοριστεί με την εισαγωγή μιας πολιτικής του μηχανισμού διάσχισης η οποία θέτει χρονικό περιορισμό στο

στάδιο της συλλογής των URLs αναγκάζοντας το σύστημα να συλλέξει όσο περισσότερα από τα URLs σε αυτό το χρονικό διάστημα. Στην περίπτωση που η διαδικασία, πράγματι, χρειαστεί να περιοριστεί χρονικά, το σύστημα δε θα καταφέρει να συλλέξει το σύνολο των URLs αλλά θα έχει καταφέρει να συλλέξει τα πιο σημαντικά από αυτά κάνοντας οικονομία χρόνου και συνεχίζοντας την υπόλοιπη διαδικασία χωρίς περιττές αναμονές.

3.4 JWanalytics [18]

Το jwanalytics είναι ένα πακέτο λογισμικού σε περιβάλλον Java το οποίο έχει ως στόχο την υποστήριξη αποθήκευσης πληροφοριών σε τρισδιάστατα μοντέλα. Εξαιτίας του χαρακτήρα του αυτού χρησιμοποιείται για την αποθήκευση αποτελεσμάτων ανάλυσης υψηλής ποιότητας διαδικτυακών δεδομένων Java εφαρμογών.

Με το jwanalytics, εκτός από την αποθήκευση αυτών των δεδομένων δίνεται και η δυνατότητα χρησιμοποίησης των δεδομένων αυτών σε πραγματικό χρόνο. Έτσι, είναι εφικτή η λήψη αποφάσεων σε πραγματικό χρόνο που θα αφορούν τη λειτουργία της ιστοσελίδας. Για παράδειγμα, ο διαχειριστής κάθε ιστότοπου θα μπορεί να έχει στη διάθεσή του πληροφορίες αναφορικά με το περιεχόμενο της σελίδας, τα προϊόντα που πιθανόν διακινούνται μέσα από την ιστοσελίδα, τις ερωτήσεις που απευθύνουν οι επισκέπτες και λοιπές παρόμοιες πληροφορίες. Έτσι, ανάλογα με την κρίση του θα μπορεί να διαχειριστεί τις πληροφορίες αυτές και να αναμορφώσει την ιστοσελίδα ώστε να καλύπτει καλύτερα τις ανάγκες των χρηστών αλλά και τους εμπορικούς σκοπούς του ιστότοπου.

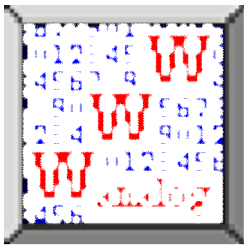
Οι δυνατότητες που παρέχονται στους χρήστες από ένα τέτοιο λογισμικό είναι πολύ ενδιαφέρουσες. Το γεγονός ότι υπάρχει ανταλλαγή πληροφορίας σε πραγματικό χρόνο είναι μια ουσιαστική μορφή επικοινωνίας μεταξύ χρήστη και διαχειριστή του ιστότοπου με το δεύτερο να μπορεί να λαμβάνει πληροφορίες για τον τρόπο με τον οποίο περιηγείται ο πρώτος στο συγκεκριμένο ιστότοπο και έτσι να μπορεί να πάρει αποφάσεις όπως για παράδειγμα την εισαγωγή εκπτώσεων σε πορισμένα προϊόντα, την προώθηση συγκεκριμένου προϊόντος κτλ. Οι αποφάσεις αυτές μπορούν να αφορούν όλους τους χρήστες του ιστού γενικά ή να περιορίζονται σε ορισμένες κατηγορίες, όπως είναι για παράδειγμα κατηγορίες χρηστών ανάλογα με τη γεωγραφική τους θέση. Αυτό είναι και το μεγάλο πλεονέκτημα του jwanalytics, το οποίο δεν ακολουθεί τον τρόπο λειτουργίας των συνηθισμένων εργαλείων για τις ιστοσελίδες τα οποία προσφέρουν τη δυνατότητα διαχείρισης του ιστότοπου μόνο όσον αφορά το συνολικό πληθυσμό των χρηστών της σελίδας και όχι συγκεκριμένες κατηγορίες αυτού. Επίσης, η δυνατότητα λήψης αποφάσεων σε πραγματικό χρόνο είναι σπουδαίο πλεονέκτημα. Κατά τη χρήση των συνηθισμένων πακέτων λογισμικού ανάλυσης ιστοσελίδων, οι διαχειριστές λαμβάνουν αποφάσεις βασιζόμενοι σε πληροφορίες που αντικατοπτρίζουν τη συμπεριφορά των χρηστών σε κάποια στιγμή στο παρελθόν. Αποτέλεσμα αυτού του χαρακτηριστικού είναι πως οι αποφάσεις που λαμβάνονται ουσιαστικά έρχονται να καλύψουν ανάγκες των χρηστών σε κάποια στιγμή στο παρελθόν, ανάγκες οι οποίες μπορεί να έχουν αλλάξει με το πέρασμα ακόμη και ενός μικρού χρονικού διαστήματος. Με το jwanalytics, λοιπόν, διασφαλίζεται το γεγονός ότι οι αποφάσεις που λαμβάνονται αφορούν το παρόν και όχι το παρελθόν.

Η δυνατότητα διαμόρφωσης του ιστότοπου ανάλογα με το προφίλ που παρουσιάζει ο χρήστης κατά την ανάλυση της αλληλεπίδρασής του με την ιστοσελίδα είναι πράγματι σημαντικό εργαλείο και θα πρέπει να συζητηθεί εκτενέστερα. Ιστορικά, τα λογισμικά διαχείρισης εργασιών ιστοτόπων παρέχουν τη δυνατότητα μορφοποίησης του περιβάλλοντος αυτού ως προς όλους τους χρήστες μέσω διαδικασιών όπως είναι τα A/B τεστ. Τα A/B τεστ είναι εργαλεία της επιστήμης του μάρκετινγκ για τον έλεγχο στατιστικών υποθέσεων. Ιδιαίτερα στα πλαίσια του σχεδιασμού ιστοσελίδων, τα A/B τεστ είναι τα εργαλεία τα οποία συνηθέστερα χρησιμοποιούνται για την αναγνώριση των αλλαγών επί της ιστοσελίδας που θα επιφέρουν το μέγιστο επιθυμητό αποτέλεσμα. Το jwanalytics σαφώς και παρέχει τη δυνατότητα αυτή. Προχωράει, όμως, ένα βήμα μπροστά δίνοντας τη δυνατότητα διαμόρφωσης του ιστότοπου για τον κάθε χρήστη χωριστά με εφαρμογή πολυπαραμετρικών μοντέλων που λαμβάνουν ως τιμές εισόδου μεταβλητές όπως η ταχύτητα σύνδεσης κάθε χρήστη, η ανάλυση της οθόνης του, το ιστορικό του στο συγκεκριμένο ιστότοπο και άλλες παρόμοιες.

Ωστόσο, το jwanalytics παρουσιάζει ένα σημαντικό μειονέκτημα. Το jwanalytics είναι μια εφαρμογή η οποία είναι σχεδιασμένη για να δίνεται στο χρήστη ο πλήρης έλεγχος αυτής και όχι για να μπορεί η εφαρμογή να λειτουργεί αυτόνομα. Αυτό συνεπάγεται την ανάγκη για συγγραφή ενός μέρους κώδικα από τη μεριά των χρηστών, κάτι το οποίο μπορεί να συμβεί μόνο από εξειδικευμένους χρήστες και όχι από το γενικό κοινό. Επίσης, το γεγονός ότι τα δεδομένα που συλλέγονται δύνανται να αποθηκεύονται σε τρισδιάστατο πίνακα καθώς και να συνδεθούν άμεσα με τα υπόλοιπα πακέτα λογισμικού που χρησιμοποιούνται από τη μεριά των διαχειριστών της ιστοσελίδας επίσης συνεπάγεται την ανάγκη για ύπαρξη χειριστών με κατάλληλες προγραμματιστικές γνώσεις.

Η ομάδα ανάπτυξης του συγκεκριμένου πακέτου λογισμικού έχει εκφράσει την επιθυμία για το λανσάρισμα μιας νέας έκδοσης του λογισμικού το οποίο θα φέρει επίσης τη δυνατότητα εφαρμογής ενός μοντέλου μελλοντικών προβλέψεων. Το νέο αυτό εργαλείο το οποίο επιθυμείτε να ενσωματωθεί θα επιτρέπει την ακόμη πιο έγκαιρη λήψη αποφάσεων αφού οι διαχειριστές του ιστότοπου θα μπορούν να συμβαδίζουν με τις ανάγκες των χρηστών πριν -ίσως- αυτές εκφραστούν με απόλυτους αριθμούς μετά από ανάλυση ήδη υπαρχόντων δεδομένων. Επίσης, εκτός από την πρόβλεψη της συμπεριφοράς των χρηστών στο σύνολό της, οι εξατομικευμένες προβλέψεις είναι ένας ακόμη στόχος της ομάδας των σχεδιαστών του λογισμικού. Έτσι, θέλουν να πετύχουν ένα εργαλείο το οποίο θα μπορεί να προβλέψει την τελική κατάληξη του πελάτη αναλύοντας τα δεδομένα της διαδρομής του μέσα από τον ιστότοπο, όπως για παράδειγμα αν θα καταλήξει να παραγγείλει κάποιο προϊόν ή όχι.

3.5 Analog [19]



Εικόνα 3.5: Το λογότυπο του πακέτου λογισμικού Analog. [19]

Το analog αποτελεί πρόγραμμα ανάλυσης της χρήσης που γίνεται σε έναν διαδικτυακό εξυπηρετητή. Οι πληροφορίες που παρέχονται στους χρήστες είναι ποικίλλες και περιλαμβάνουν τις πιο δημοφιλείς ιστοσελίδες, τις γεωγραφικές θέσεις από τις οποίες συνδέονται οι χρήστες, τους ιστότοπους που επισκέφθηκαν ακολουθώντας συνδέσμους που βρίσκονται σε άλλες σελίδες και πληροφορίες συναφούς περιεχομένου.

Στο χώρο των εμπορικών προϊόντων Εξόρυξης στον Ιστό, το analog θεωρείται ένας από τους ηγέτες της κατηγορίας του. Ανάμεσα στα πλεονεκτήματά του μπορεί κανείς να συναντήσει τις υψηλές του ταχύτητες. Ακριβέστερα, τελευταίες μετρήσεις έδειξαν τη δυνατότητα του λογισμικού να επεξεργαστεί 56 εκατομμύρια γραμμές δεδομένων από καταγραφές εξυπηρετητών με χρήση chip 266 MHz σε 35 λεπτά, μέγεθος που για το συγκεκριμένο σύστημα αντιστοιχεί σε επεξεργασία 1 GB δεδομένων κάθε 5 λεπτά. Βέβαια, αυτός ο χρόνος εξαρτάται έντονα από το σύστημα στο οποίο εργάζεται ο αλγόριθμος και συνεπώς σε νεότερα μηχανήματα είναι ακόμη μικρότερος. Ένα εξίσου σημαντικό πλεονέκτημα είναι πως μπορεί να επεξεργαστεί πολύ μεγάλα αρχεία, επιπέδου πάνω από 1 δισεκατομμύριο γραμμών. Μπορεί να παρουσιάσει τα αποτελέσματα σε 33 διαφορετικές γλώσσες ενώ εξαιτίας του γεγονότος ότι έχει γραφεί σε τυπική C γλώσσα προγραμματισμού μπορεί να εγκατασταθεί σε όλα τα διάσημα λειτουργικά συστήματα (MacOS, UNIX, MS Windows) με την εγκατάσταση να γίνεται με υπεραπλουστευμένες διαδικασίες μέσω ενός απλού C compiler. Ταυτόχρονα, χαρακτηρίζεται από υψηλή ικανότητα παραμετροποίησης ώστε να ικανοποιεί τις ανάγκες οποιουδήποτε χρήστη. Μπορεί να δημιουργεί περί τα 32 διαφορετικά είδη εκθέσεων των αποτελεσμάτων, ένα από τα οποία φαίνεται στην εικόνα που ακολουθεί.

General Summary		
1.	Host name	analog documentation
2.	Host URL	/~sret1/analog/olddocs/
3.	Time of first request	Mar 1, 2009 07:48
4.	Time of last request	May 16, 2009 23:47
5.	Time last 7 days lasts until	May 16, 2009 23:47
6.	Successful server requests	289,810 Requests
7.	Successful requests in last 7 days	25,117 Requests
8.	Successful requests for pages	27,267 Requests for pages
9.	Successful requests for pages in last 7 days	2,257 Requests for pages
10.	Failed requests	4,718 Requests
11.	Failed requests in last 7 days	351 Requests
12.	Redirected requests	104 Requests
13.	Redirected requests in last 7 days	4 Requests
14.	Distinct files requested	4,828 Files
15.	Distinct files requested in last 7 days	2,969 Files
16.	Distinct hosts served	9,395 Hosts
17.	Distinct hosts served in last 7 days	1,017 Hosts
18.	Corrupt lines in the logfile	662 Lines
19.	Total data transferred	3.120 GB
20.	Total data transferred in last 7 days	264.199 MB

Εικόνα 3.6: Παράδειγμα έκθεσης αποτελεσμάτων της ανάλυσης του analog. Στη συγκεκριμένη εικόνα μπορούμε να δούμε την επισκόπηση των γενικών στατιστικών της συγκεκριμένης ιστοσελίδας στο χρονικό πλαίσιο που έχει ορίσει ο χρήστης. [19]

Τέλος, είναι συμβατό με όλων των ειδών τους εξυπηρετητές αφού το μόνο που απαιτείται είναι το logfile του εξυπηρετητή να είναι γραμμένο σε γλώσσα που να μπορεί να αναγνωρίζει το analog. Το analog, όμως, μπορεί να διαβάσει όλα τα τυπικά format των logfiles οπότε θεωρείται απόλυτα συμβατό με όλους τους εξυπηρετητές.

3.6 GNU Wget [20]

Εικόνα 3.7: Το λογότυπο του πακέτου λογισμικού GNU Wget. [20]

Το GNU Wget είναι ένα πακέτο λογισμικού κατάλληλο για την ανάκτηση αρχείων που ακολουθούν τα πρωτόκολλα HTTP, HTTPS και FTP. Έχει μη-διαδραστική μορφή και λειτουργεί μέσω γραμμής εντολών με αποτέλεσμα να μπορεί να υποστηριχθεί από οποιοδήποτε σύστημα διαθέτει παράθυρο εντολών (terminal) ακόμη και αν δε διαθέτει υποστήριξη από X-Windows κτλ.

Το GNU Wget χαρακτηρίζεται από πλήθος δυνατοτήτων οι οποίες επιτρέπουν την ανάκτηση μεγάλων αρχείων και το mirroring ιστότοπων χωρίς μεγάλη δυσκολία. Αρχικά, το λογισμικό αυτό έχει τη δυνατότητα να επανεκκινεί ματαιωμένες λήψεις, να κάνει αναδρομικό mirroring σε ιστότοπους, να χρησιμοποιεί αρχεία μηνυμάτων τύπου NLS για πολλές διαφορετικές γλώσσες ενώ λειτουργεί σε όλα τα UNIX συστήματα καθώς και σε MS Windows. Υποστηρίζει HTTP proxies, HTTP cookies και HTTP συνδέσεις ενώ μπορεί να λειτουργεί παράλληλα με άλλα προγράμματα του συστήματος. Επίσης, εξαιτίας του μη διαδραστικού του χαρακτήρα μπορεί να λειτουργεί ακόμη και μετά την αποσύνδεση του χρήστη από το σύστημα.

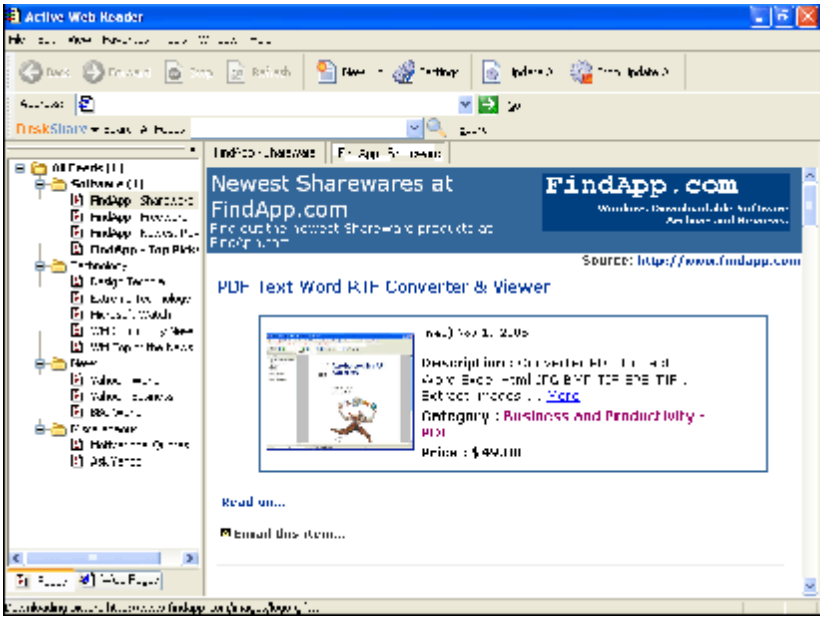
Οι προσπάθειες της ομάδας ανάπτυξης του GNU Wget εστιάζονται επί του παρόντος στη δημιουργία ενός λογισμικού το οποίο θα έχει τη δυνατότητα να υποστηρίξει πολλαπλές ταυτόχρονες συνδέσεις. Είναι αναμενόμενο πως μια τέτοια ανάπτυξη στο λογισμικό θα χρειαστεί αρκετό χρόνο και σημαντικές επεμβάσεις στον ήδη υπάρχοντα κώδικα. Ωστόσο, αυτή η προσπάθεια θεωρείται σημαντική αφού είναι απόλυτα κατανοητό πως δεν επιτρέπεται ένας αργός εξυπηρετητής να καθυστερεί το σύνολο της επεξεργασίας. Βέβαια, αυτή η μετατροπή στον κώδικα δε θα μετατρέψει το GNU Wget σε «επιταχυντή λήψεων» (download accelerator) αφού βάσει των οδηγιών του πρωτοκόλλου HTTP, RFC 2616, οι παράλληλες συνδέσεις σε έναν εξυπηρετητή είναι απαγορευμένες. Εκτός, όμως, της απαγόρευσης αυτής από πλευράς οδηγιών, υπάρχει καλύτερος τρόπος επιτάχυνσης των λήψεων και αυτός περιλαμβάνει τη χρήση του Metalink το οποίο μπορεί να παρέχει στο GNU Wget μια λίστα με εναλλακτικές θέσεις για την ανάκτηση της ίδιας πληροφορίας. Έτσι, το GNU Wget μπορεί να κάνει λήψη διαφορετικών επιμέρους τμημάτων του ίδιου όμως αρχείου από διαφορετικές θέσεις με αποτέλεσμα την επιτάχυνση της λήψης χωρίς την παραβίαση των υπάρχοντών κανόνων. Βέβαια, η τελευταία έκδοση του GNU Wget δεν υποστηρίζει το Metalink αλλά μια τέτοια μετατροπή είναι προγραμματισμένη να υπάρχει στην επόμενη έκδοση.

3.7 Active Web Reader [21]

Ο Active Web Reader αποτελεί λογισμικό το οποίο επιτρέπει την ανάγνωση RSS αρχείων με τρόπο απλό και εύχρηστο. Τα αρχεία τύπου RSS (Really Simple Syndication) είναι αρχεία τύπου XML τα οποία χρησιμοποιούνται στη διανομή πληροφορίας που πηγάζει από τα περιεχόμενα του Παγκόσμιου Ιστού. Παράδειγμα τέτοιων αρχείων αποτελούν οι τίτλοι των ειδήσεων. Τα RSS χρησιμοποιούνται ευρέως από τους διαφόρων ειδών παρόχους διαδικτυακού περιεχομένου για την εύκολη δημιουργία και διάδοση πληροφοριών που περιλαμβάνουν μεταξύ άλλων συνδέσμους σε ειδήσεις, τίτλους γεγονότων, μικρές περιλήψεις αυτών και προωθητικές ενέργειες.

Σε μια προσπάθεια γενικής θεώρησής του, το πακέτο λογισμικού Active Web Reader αποτελεί ένα μηχανισμό συλλογής RSS καθώς και ανάγνωσης αυτών. Ο χρήστης, λοιπόν, μπορεί να εκφράσει τα αντικείμενα του ενδιαφέροντος του και το λογισμικό θα συλλέξει τα αντίστοιχα RSS. Επίσης, θα τα κατηγοριοποιήσει και θα τα αποθηκεύσει σε ξεχωριστούς φάκελους. Εκτός, όμως, από τη δυνατότητα ορισμού των θεμάτων που ενδιαφέρουν το χρήστη, το λογισμικό μπορεί να ανακαλύψει τα θέματα αυτά μέσω επικοινωνίας με τη λίστα των σελιδοδεικτών στον MS Internet Explorer. Τα διαφορετικά αυτά αντικείμενα που ενδιαφέρουν το χρήστη, το λογισμικό μπορεί να τα αναζητά σε διαφορετικούς χρόνους και με διαφορετικές συχνότητες, σύμφωνα πάντα με τις επιθυμίες του χρήστη. Ωστόσο, η αρχική ρύθμιση

του λογισμικού και αυτή που ισχύει σε περίπτωση που δεν του δοθεί αντίθετη εντολή είναι η αναζήτηση σε ημερήσια βάση. Παρέχονται πολλά διαφορετικά στυλ παρουσίασης των αποτελεσμάτων των RSS έτσι ώστε να μπορούν να καλυφθούν οι επιθυμίες διαφορετικών χρηστών και να είναι δυνατή η εξατομίκευση του περιβάλλοντος ανάγνωσης των ειδήσεων. Η χρήση του, όμως, ξεπερνά αυτή του προγράμματος ανάγνωσης ειδήσεων αφού -σε ένα βαθμό- μπορεί να λειτουργήσει ως διαδικτυακός περιηγητής επιτρέποντας την προσπέλαση ιστοσελίδων στο περιβάλλον εργασίας του. Υπάρχει, επίσης, η δυνατότητα αυτόματης αναζήτησης RSS εκ μέρους του λογισμικού κατά την περιήγηση του χρήστη στο διαδίκτυο με χρήση του MS Internet Explorer ενώ οι πληροφορίες και οι ειδήσεις που συλλέγονται μπορούν πολύ εύκολα να σταλθούν στις επαφές του χρήστη μέσω ηλεκτρονικού ταχυδρομείου. Η διεπαφή με το χρήστη είναι αρκετά οικεία και εύχρηστη όπως φαίνεται και στην εικόνα που ακολουθεί ενώ το λογισμικό, τέλος, υποστηρίζεται -δυστυχώς- μόνο από λειτουργικά MS Windows.



Εικόνα 3.8: Το περιβάλλον εργασίας του Active Web Reader. [21]

Προηγουμένως έγινε αναφορά σε δύο σημαντικά χαρακτηριστικά του λογισμικού αυτού στα οποία θέλουμε να αναφερθούμε περεταίρω, δηλαδή την αυτόματη αναζήτηση πληροφοριών αλλά και τη συχνότητα με την οποία είναι δυνατό να γίνεται αυτόματη ανανέωση των RSS. Αυτή η αυτόματη αναζήτηση πληροφοριών (το λογισμικό την αναφέρει ως “Auto Discovery”) είναι ένα από τα χαρακτηριστικά που καθιστούν το Active Web Reader μοναδικό στο είδος του αφού μπορεί να αναζητά αυτόματα τις πληροφορίες που αφορούν στις προτιμήσεις του χρήστη καθώς αυτός απλά περιηγείται στον Παγκόσμιο Ιστό χρησιμοποιώντας τον περιηγητή MS Internet Explorer. Κατά τη διαδικασία αυτή, όταν το λογισμικό ανακαλύψει RSS τα οποία σχετίζονται με τη σελίδα στην οποία βρίσκεται ο χρήστης, εμφανίζει ένα μικρό μήνυμα το οποίο δίνει στο χρήστη τη δυνατότητα να διαλέξει ποια από αυτά τα RSS επιθυμεί να προσθέσει στη λίστα ανάγνωσης του λογισμικού και ακόμα, αν θέλει να παρακολουθεί στο μέλλον τις τοποθεσίες αυτές για νεότερες πληροφορίες. Το δεύτερο, εξίσου σημαντικό, χαρακτηριστικό του λογισμικού αυτού είναι η δυνατότητα χειροκίνητου ορισμού της συχνότητας με την οποία ο χρήστης επιθυμεί να γίνεται προσπάθεια ανανέωσης του περιεχομένου των RSS. Αυτό το χρονικό

διάστημα θα πρέπει να έχει τη μορφή κάποιων ωρών ή κάποιων ημερών με ελάχιστη τιμή τη μία ώρα.

Προσπαθώντας μια αναγνώριση των πλεονεκτημάτων και των μειονεκτημάτων του Active Web Reader καταλήγουμε πως, σε γενικές γραμμές, είναι ένα αρκετά ευφυές και εύχρηστο πακέτο λογισμικού το οποίο, όμως, είναι περιορισμένης χρήσης αφού δεν υποστηρίζει σημαντικά χαρακτηριστικά. Έτσι, έχουμε στα χέρια μας ένα λογισμικό το οποίο αναζητεί RSS με τρόπο απλό, γρήγορο και εύκολο και μπορεί αυτά τα RSS να τα κατηγοριοποιεί σε ομάδες. Τα RSS που θα αναζητηθούν μπορούν να οριστούν από το χρήστη και μόλις συλλεχθούν, ο χρήστης μπορεί να εκτελέσει σε αυτά αλγόριθμους αναζήτησης χρησιμοποιώντας απλές λέξεις ή φράσεις. Σημαντικό είναι, επίσης, το γεγονός ότι υποστηρίζονται όλοι οι σημαντικοί τύποι RSS, δηλαδή οι τύποι 0.9x, 1.x και 2.x. Ταυτόχρονα, είναι ένα πολύ ελαφρύ λογισμικό το οποίο καταλαμβάνει μόλις 1 MB στο σκληρό δίσκο του ηλεκτρονικού υπολογιστή. Τέλος, ο εκάστοτε αναγνώστης μπορεί να διαλέξει το στυλ του περιβάλλοντος που είτε ταιριάζει καλύτερα στην αισθητική του είτε τον βοηθά στην πράξη κατά την ανάγνωση των ειδήσεων. Από την άλλο μεριά, όμως, έχουμε ένα λογισμικό το οποίο αρχικά μπορεί να χρησιμοποιηθεί μόνο από συστήματα τα οποία λειτουργούν με MS Windows. Αυτό δεν ήταν σημαντικό μειονέκτημα στο παρελθόν αφού τα MS Windows κυριαρχούσαν στο χώρο των λειτουργικών συστημάτων για χρήση σε προσωπικούς υπολογιστές. Ωστόσο, οι σύγχρονες τάσεις που θέλουν τους καταναλωτές να έχουν στραφεί και σε εναλλακτικές λειτουργικές πλατφόρμες (MacOS και UNIX) αφήνουν εκτός κοινού χρήσης του Active Web Reader μεγάλο ποσοστό του πληθυσμού. Επίσης, το πακέτο αυτό δεν έχει δυνατότητα συγχρονισμού με άλλα χρησιμοποιούμενα λογισμικά ενώ παρά το γεγονός ότι είναι δυνατή η ομαδοποίηση των RSS, δεν είναι δυνατές διάφορες ενέργειες επεξεργασίας των φακέλων ενώ τα μεμονωμένα αντικείμενα δε μπορούν να λάβουν ετικέτες ώστε να ξεχωρίζουν μεταξύ τους βάσει ονόματος.

3.8 Scrapy [22]

Εικόνα 3.9: Το λογότυπο του πακέτου λογισμικού Scrapy. [22]

Το Scrapy είναι ένα πακέτο λογισμικού το οποίο έχει αναπτυχθεί γύρω από έναν ευφυή μηχανισμό διάσχισης ιστοσελίδων για την εξόρυξη δομικών πληροφοριών αυτών. Το Scrapy μπορεί να χρησιμοποιηθεί στα πλαίσια ενός μεγάλου εύρους εργασιών όπως είναι η εξόρυξη δεδομένων, η επεξεργασία πληροφοριών και η αρχειοποίηση του ιστορικού χρήσης του διαδικτύου. Την ονομασία του την παίρνει από έναν από τους εναλλακτικούς όρους που έχουν αποδοθεί στις διαδικασίες εξόρυξης πληροφορίας στον Παγκόσμιο Ιστό, το web scraping. Ο όρος αυτός αναφέρεται σε διαδικασίες οι οποίες περιστρέφονται γύρω από μη δομημένα αρχεία στο διαδίκτυο, κυρίως HTML αρχεία, και τη μετατροπή αυτών σε εύχρηστες δομημένες μορφές. Οι μορφές αυτές μπορούν, στη συνέχεια, να επεξεργαστούν, να αναλυθούν και να αποθηκευτούν.

Το Scrapy έχει συνταχθεί σε γλώσσα προγραμματισμού Python και ο χειρισμός του γίνεται μέσω γραμμής εντολών και εντολών της συγκεκριμένης γλώσσας. Η πρώτη

επίσημη έκδοση του λογισμικού δόθηκε στο κοινό το 2008 και από τότε το πακέτο υφίσταται συνεχώς εξελικτικές μετατροπές. Ο αρχικός σκοπός της ανάπτυξής του ήταν η Εξόρυξη στον Ιστό. Χρησιμοποιείται, ωστόσο, τόσο ως ένας γενικός μηχανισμός διάσχισης του διαδικτύου, όσο και για την εξόρυξη δεδομένων με χρήση διεπαφών προγραμματισμού εφαρμογών (API – Application Programming Interface).

Το Scrapy παρέχει ένα σημαντικό σύνολο δυνατοτήτων οι οποίες στο σύνολό τους δημιουργούν ένα αποτελεσματικό, εύχρηστο και γρήγορο σε χρόνο απόκρισης λογισμικό. Μεταξύ αυτών, κάποιος μπορεί να συναντήσει τη δυνατότητα επιλογής και εξόρυξης δεδομένων από πηγές τύπου HTML και XML, εκκαθάρισης των δεδομένων με επαναχρησιμοποιούμενα φίλτρα και την εξαγωγή των αποτελεσμάτων σε αρχεία διάφορων τύπων, όπως για παράδειγμα σε μορφές JSON, CSV και XML. Μπορεί, επίσης, να κάνει αυτόματη λήψη εικόνων και λοιπών αρχείων πολυμέσων που σχετίζονται με τα υπό επεξεργασία αρχεία ενώ υπάρχει η δυνατότητα μορφοποίησης του λογισμικού μέσω της εγκατάστασης υπάρχοντών plug-ins ώστε να καλύπτονται οι επιθυμίες και οι ανάγκες κάθε μεμονωμένου χρήστη. Τα plug-ins αυτά αφορούν στη διαχείριση των cookies, τη συμπίεση και την ταυτοποίηση αρχείων HTTP, το χειρισμό των robots.txt, τον ορισμό των παραμέτρων της διαδικασίας της διάσχισης και άλλες παρόμοιες εργασίες. Επιπροσθέτως, υπάρχουν προεγκατεστημένα πρότυπα των spiders έτσι ώστε να γίνεται γρηγορότερα η δημιουργία τους. Με τον όρο “spider” ορίζουμε ένα μικρό μπλοκ εντολών το οποίο ορίζει από ποιο URL θα εξ ορυχθούν οι πληροφορίες, με ποιο τρόπο θα γίνει η επεξεργασία και η εξαγωγή τους, το αν θα αποθηκευτούν κτλ. Για τους χρήστες οι οποίοι ενδιαφέρονται για στατιστική ανάλυση των δεδομένων, το Scrapy είναι εφοδιασμένο με πλήρη συλλογή στατιστικών στοιχείων και μετρικών που είναι χρήσιμα εργαλεία κυρίως για την παρακολούθηση της αποδοτικότητας των spiders και για την αναγνώριση των αιτιών που πιθανόν αυτά να μην λειτουργούν σωστά. Η συγγραφή και η μεταγλώττιση των spiders γίνεται σε κατάλληλη προ εγκατεστημένη κονσόλα ενώ παρέχεται και κατάλληλη βοήθεια για τη σωστή συγγραφή και τη διόρθωση λαθών στον κώδικα. Τέλος, παρέχεται υποστήριξη για τη σωστή διάσχιση των URL που έχουν οριστεί εξαρχής ή έχουν προκύψει ως αποτέλεσμα αυτόματης αναζήτησης ενώ η ταχύτητα διάσχισης αγγίζει τις 500 σελίδες ημερησίως.

Η διαδικασία λειτουργίας του Scrapy είναι πολύ απλή. Αρχικά, γίνεται η επιλογή της ιστοσελίδας από την οποία επιθυμούμε να εξορύξουμε τις απαραίτητες για την εργασία μας πληροφορίες. Στη συνέχεια, θα πρέπει να είμαστε σε θέση να ορίσουμε επακριβώς το είδος των πληροφοριών τις οποίες επιθυμούμε να εξορύξουμε. Αφού το κάνουμε αυτό, είναι η στιγμή να γίνει η συγγραφή του spider το οποίο θα ορίζει το URL από το οποίο θα γίνει η εκκίνηση της διάσχισης, οι κανόνες που θα πρέπει να ακολουθηθούν κατά τη μετάβαση από τον ένα σύνδεσμο στον άλλο καθώς και οι κανόνες που θα ακολουθηθούν κατά την εξόρυξη των δεδομένων. Αφού γίνει η συγγραφή και η μεταγλώττιση και διορθωθούν πιθανά λάθη, το μόνο που απομένει είναι η εφαρμογή του κώδικα αυτού. Με την εφαρμογή του spider ξεκινά η διάσχιση της ιστοσελίδας και τα αποτελέσματα εξάγονται σε προκαθορισμένη μορφή. Επίσης, υπάρχει η δυνατότητα αποθήκευσης των αποτελεσμάτων σε μορφή βάσης δεδομένων. Αφού εξαχθούν τα αποτελέσματα, αυτά μπορούν να προσπελαστούν από το χρήστη οποιαδήποτε στιγμή.

3.9 Trapit [23]

Εικόνα 3.10: Το λογότυπο του πακέτου λογισμικού Trapit. [23]

Το Trapit είναι -ίσως- ο ηγέτης στην Εξόρυξη εξατομικευμένου περιεχομένου στον Ιστό. Δεν αφορά χρήσεις που γίνονται μόνο στα πλαίσια εταιρειών αλλά και ιδιωτικού τύπου. Η ανάπτυξή του έχει χτιστεί γύρω από την κεντρική ιδέα της «ενίσχυσης του σήματος με ταυτόχρονη εξάλειψη του θορύβου» και σκοπός του είναι η παροχή δυνατοτήτων στους χρήστες που σχετίζονται με την ανακάλυψη, την επεξεργασία, την ανταλλαγή και τη δημοσίευση πληροφοριών με τρόπο εύκολο, σαφή και αποτελεσματικό.

Η δημιουργία του Trapit έχει βασιστεί σε τεχνολογία Τεχνητής Νοημοσύνης η οποία αρχικά αναπτύχθηκε από το Υπουργείο Εθνικής Άμυνας των Ηνωμένων Πολιτειών Αμερικής για προγράμματα που αφορούν την εθνική ασφάλεια (Σχέδια Επιτροπής DARPA – Defence Advance Research Projects Agency). Βασιζόμενη σε αυτό το τόσο στέρεο θεωρητικό υπόβαθρο, το Trapit έχει τη δυνατότητα να εκτελεί ανάλυση περιεχομένου υψηλού επιπέδου με αποτέλεσμα να εξάγεται πληροφορία υψηλής συνάφειας με την αναζήτηση του χρήστη. Εκτός, όμως, από την υψηλή αποδοτικότητα και την υψηλή σχετικότητα με των αποτελεσμάτων του, ταυτόχρονα το συγκεκριμένο λογισμικό μπορεί να καυχιέται για τους χαμηλούς χρόνους λειτουργίας του.

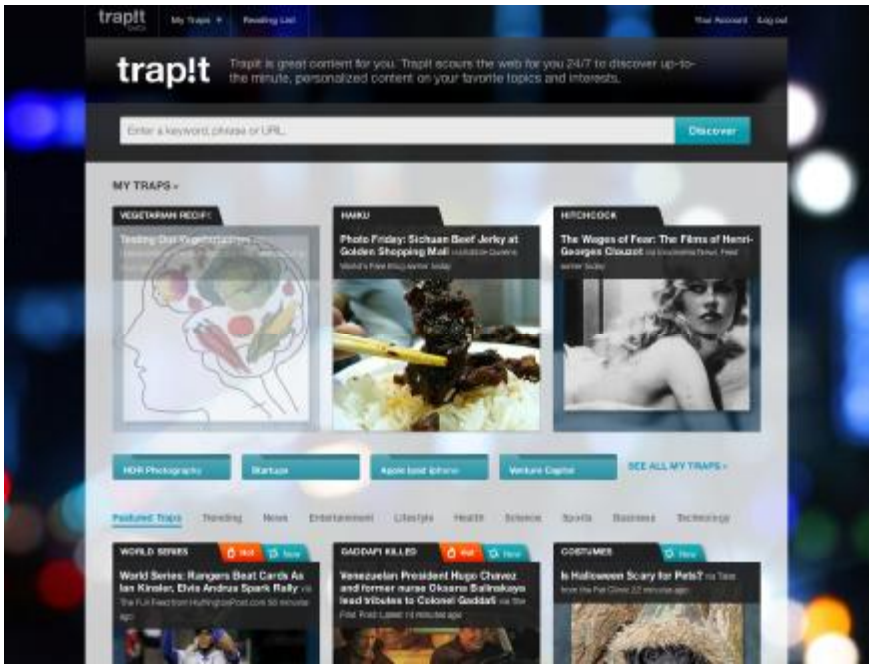
Πρόκειται, πράγματι, για ένα πολύ ευφυές λογισμικό το οποίο κάνει χρήση τεχνικών Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης έτσι ώστε να εκπαιδευτεί στις επιθυμίες του χρήστη και να αρχίσει να λειτουργεί αυτόνομα. Η λειτουργία του βασίζεται στο «στήσιμο παγίδων» (trap = παγίδα), κάθε μια από τις οποίες αφορά και σε ένα διαφορετικό αντικείμενο με σκοπό να πέφτουν σε αυτή την παγίδα όλα τα δεδομένα που σχετίζονται με την πληροφορία προς εξόρυξη. Όπως δηλώνουν και οι κατασκευαστές, παγίδα μπορεί να στηθεί για οποιαδήποτε πληροφορία (“If you can type it, you can trap it”) ενώ πολυάριθμες, διαφορετικές μεταξύ τους παγίδες μπορούν να λειτουργούν ταυτόχρονα. Τα εξορυσσόμενα δεδομένα στη συνέχεια μπορούν να αναλυθούν με κατάλληλα εργαλεία και να εξαχθούν ως δομημένη γνώση. Επιπροσθέτως, όπως συμβαίνει και με κάθε λογισμικό νέας γενιάς, τα αποτελέσματα αυτά μπορούν να κοινοποιηθούν αυτόματα στα μέσα κοινωνικής δικτύωσης.

Το πρώτο στάδιο της λειτουργίας του Trapit είναι η αναζήτηση και η ανακάλυψη της γνώσης. Αυτό γίνεται με την επικοινωνία του χρήστη με το σύστημα με βασικές λέξεις-κλειδιά ή με την υπόδειξη κάποιων URL για να δημιουργηθεί η παγίδα. Από εκεί και πέρα, η διαδικασία εκτυλίσσεται αυτόματα και το λογισμικό δύναται να παρέχει το σχετικό με την αναζήτηση περιεχόμενο σε μικρό χρονικό διάστημα. Υπάρχει, ακόμα, η δυνατότητα συνεχούς ανανέωσης των παγίδων έτσι ώστε ο χρήστης να λαμβάνει συνεχώς τις νεότερες των σχετικών πληροφοριών. Επίσης, το εργαλείο αναζήτησης παρέχει στο χρήστη τη δυνατότητα να ορίζει τον επιθυμητό βαθμό συνάφειας των αποτελεσμάτων με το αντικείμενο που αναζητά.

Ένα πολύ σημαντικό χαρακτηριστικό του Trapit είναι το σύνολο των πηγών στις οποίες γίνεται η αναζήτηση των πληροφοριών. Η διαφορά του συνόλου αυτού με τις πηγές που χρησιμοποιούνται από άλλα παρόμοια λογισμικά είναι πως έχουν εκ των προτέρων κριθεί ως πηγές έγκυρης πληροφορίας και υψηλού επιπέδου περιεχομένου. Σε αυτές τις πηγές συγκαταλέγονται εφημερίδες, περιοδικά, ειδησεογραφικά πρακτορεία, πηγές βίντεο και εικόνας, ιστολόγια σημαντικών προσωπικοτήτων και οργανισμών και podcasts. Στο σύνολό τους, οι πηγές αυτές ξεπερνούν τις 100.000 και έχουν ανακαλυφθεί μέσω των προαναφερθουσών τεχνικών Τεχνητής Νοημοσύνης με αυστηρά κριτήρια ως προς την αυθεντικότητα, την αξιοπιστία και την ποιότητα των πληροφοριών τους. Επιπροσθέτως, εκτός από τις τεχνικές Τεχνητής Νοημοσύνης, χρησιμοποιούνται και τεχνικές Μηχανικής Μάθησης για έρευνα σε μεγάλο βάθος του Παγκόσμιου Ιστού, αναδεικνύοντας πληροφορίες μεγάλης σχετικότητας με την αναζήτηση οι οποίες, όμως, δεν υπήρξαν ιδιαίτερα δημοφιλείς και εξαιτίας αυτού δεν προκύπτουν άμεσα στα αποτελέσματα μιας αναζήτησης πληροφοριών με υψηλή απήχηση. Έτσι, το Trapit έχει τη δυνατότητα να παρέχει στους χρήστες ένα μοναδικό σύνολο πληροφοριών.

Ταυτόχρονα με τις αναζητήσεις του χρήστη, το λογισμικό έχει τη δυνατότητα να μαθαίνει από αυτές το είδος του περιεχομένου που αρέσει στο χρήστη με αποτέλεσμα να παρέχει σε αυτόν όσο το δυνατόν πιο προσωποποιημένες πληροφορίες. Αυτό γίνεται με την ύπαρξη ενός εργαλείου με το οποίο ο χρήστης χαρακτηρίζει τις ιστοσελίδες ανάλογα με το αν του αρέσουν ή όχι και το σύστημα χρησιμοποιεί αυτή την πληροφορία ώστε να μαθαίνει για μελλοντικές αναζητήσεις.

Ένα από τα πιο εντυπωσιακά χαρακτηριστικά του λογισμικού είναι η αισθητική της διεπαφής του με το χρήστη, όπως αυτή φαίνεται στην εικόνα που ακολουθεί.



Εικόνα 3.11: Η διεπαφή του Trapit με το χρήστη. [23]

Σε αυτή τη διεπαφή, το περιεχόμενο που εξορύχθηκε παρουσιάζεται με πολύ κομψό τρόπο, άρτια γραφικά, οργανωμένο με έξυπνο τρόπο έτσι ώστε να μπορεί ο χρήστης να το προσπελάσει γρήγορα και να επιλέξει το τμήμα της πληροφορίας με το οποίο επιθυμεί να ασχοληθεί. Οι τίτλοι αυτών των τμημάτων μπορούν να μορφοποιηθούν σύμφωνα με τις επιθυμίες των χρηστών και επιπροσθέτως να γίνει η επιλογή αν είναι επιθυμητή η εμφάνιση μικρής περίληψης του αρχείου ή όχι.

Αφού γίνει η αναζήτηση της πληροφορίας, υπάρχει η δυνατότητα για κοινοποίηση των πληροφοριών σε άλλους χρήστες, στην προσωπική ιστοσελίδα του χρήστη αν πρόκειται για ιδιώτη ή στην ιστοσελίδα της επιχείρησης αν πρόκειται για εταιρικό χρήστη καθώς και στα μέσα κοινωνικής δικτύωσης. Η κοινοποίηση αυτή γίνεται με μόνο ένα πάτημα πλήκτρου ενώ μπορεί να γίνει και βάσει προγράμματος. Μοναδική απαίτηση για την πραγματοποίηση αυτού είναι να ρυθμιστεί το λογισμικό ώστε να γνωρίζει πότε να κοινοποιεί πληροφορίες κατά τη διάρκεια της ημέρας καθώς και ποιες πληροφορίες να κοινοποιεί. Υπάρχει και το εργαλείο με το οποίο γίνεται αυτόματα κοινοποίηση κάποιου είδους πληροφορίας, τη στιγμή που αυτή ανακαλύπτεται, χωρίς χρονικούς περιορισμούς αλλά αυτόματα έτσι ώστε το κοινό μιας ιστοσελίδας, ενός ιστολογίου ή ενός διαδικτυακού τόπου να τροφοδοτείται αυτόματα με τις νεότερες πληροφορίες. Για την επίτευξη αυτών των αυτόματων διαδικασιών έχουν υλοποιηθεί αντίστοιχες εφαρμογές για χρήση σε smartphones και tablets, οι οποίες μάλιστα έχουν βραβευτεί.

Εκτός από την κοινοποίηση των πληροφοριών, το λογισμικό παρέχει εργαλεία για την στατιστική ανάλυση των εξαγόμενων πληροφοριών. Έτσι, ο διαχειριστής ενός ιστότοπου μπορεί να γνωρίζει πως χρησιμοποιούν οι επισκέπτες του ιστότοπου το συγκεκριμένο περιεχόμενο. Προς αυτή την κατεύθυνση, το λογισμικό συνεργάζεται στενά με το Google Analytics με αποτέλεσμα την παροχή λεπτομερών αναφορών σχετικά με το περιεχόμενο των ιστοσελίδων, τους τρόπους που αυτό χρησιμοποιείται και την επίδραση που έχει στο κοινό το οποίο περιηγείται στην ιστοσελίδα. Με αυτόν τον τρόπο μπορεί ο διαχειριστής να αναμορφώνει συνεχώς τη σελίδα και το περιεχόμενό της και να διαμορφώνει τις κατάλληλες στρατηγικές για την προώθηση του περιεχομένου.

3.10 Pattern [24,25]

Το Pattern είναι μια από τις πιο ολοκληρωμένες σουίτες Εξόρυξης στον Ιστό, η οποία περιλαμβάνει πλήθος εργαλείων για Εξόρυξη Δεδομένων, επεξεργασία δεδομένων σε Φυσική Γλώσσα, Μηχανική Μάθηση, Ανάλυση Δικτύων και Οπτικοποίηση Δεδομένων. Έτσι, χρησιμοποιεί γνωστά εργαλεία όπως το Google, το Twitter, μηχανισμούς διάσχισης, ανάλυση συναισθημάτων, το WordNet, διανυσματικά μοντέλα τεχνικές ομαδοποίησης και λοιπές, παρόμοιες τεχνικές για τη βέλτιστη δυνατή αναζήτηση, επεξεργασία και παρουσίαση διαδικτυακών δεδομένων.

Το Pattern είναι γραμμένο σε γλώσσα προγραμματισμού Python και η μόνη του εξωτερική εξάρτηση είναι από το LSA (Latent Semantic Analytics – Λανθάνουσα Σημασιολογική Ανάλυση) στο εργαλείο pattern.vector, το οποίο απαιτεί το NumPy (το βασικό πακέτο λογισμικού της Python για επιστημονικού τύπου υπολογιστική ανάλυση). Η λειτουργία του γίνεται μέσα από παράθυρο εντολών και η εγκατάστασή του είναι εύκολη στην τέλεση, όπως και τα εργαλεία του στο σύνολό τους αφού

σκοπός της ομάδας που ανέπτυξε το Pattern ήταν η κατασκευή ενός εύκολου στη χρήση λογισμικού ώστε να έχει ανταπόκριση από μεγάλο ποσοστό του πληθυσμού.

Σκοπός του πακέτου λογισμικού αυτού είναι η εξυπηρέτηση των αναγκών εξειδικευμένων χρηστών, οι οποίες είναι συνήθως επιστημονικού χαρακτήρα και ενδιαφέροντος όσο και ανεξειδίκευτων χρηστών. Η σύνταξη των εντολών γίνεται με απλό και άμεσο τρόπο ενώ τα ονόματα των εντολών και των παραμέτρων έχουν επιλεγεί με τρόπο τέτοιο, ώστε να εξηγούν από μόνα τους τη χρήση τους. Το λογισμικό συνοδεύεται από πλήρη οδηγό χρήσης, ο οποίος δε θεωρεί δεδομένη την ύπαρξη προαπαιτούμενων γνώσεων με αποτέλεσμα να είναι εύκολη η χρήση των εργαλείων από χρήστες που δεν έχουν προηγούμενη επαφή με την Python. Ακολούθως, γίνεται μια συνοπτική περιγραφή των εργαλείων που περιλαμβάνει.

Pattern.web

Το pattern.web είναι ένα από τα εργαλεία της σουίτας pattern και περιέχει γνωστές διεπαφές προγραμματισμού εφαρμογών, όπως το Google, το Twitter, το Facebook και άλλες, ένα πρόγραμμα ανάλυσης HTML αρχείων και ένα μηχανισμό διάσχισης. Ουσιαστικά, λοιπόν, περιλαμβάνονται εργαλεία τα οποία χρησιμοποιούν ένα μηχανισμό λήψεως δεδομένων που υποστηρίζει το caching, τα proxies, τα ασύγχρονα αιτήματα και τις ανακατευθύνσεις με σκοπό την εξόρυξη δεδομένων στον Ιστό.

Pattern.en|es|de|fr|it|nl

Το συγκεκριμένο εργαλείο είναι ένα πακέτο εφαρμογών σε Φυσική Γλώσσα για κάθε μια από τις υποστηριζόμενες γλώσσες. Βέβαια, όπως προαναφέραμε στο θεωρητικό μέρος της εργασίας μας, η γλώσσα χαρακτηρίζεται από έναν ασαφή τρόπο δόμησης και λειτουργίας με αποτέλεσμα να πρέπει να χρησιμοποιηθούν στατιστικοί τρόποι προσέγγισης αυτής καθώς και πλήθος κανονικών εκφράσεων. Ο ασαφής, αυτός, χαρακτήρας της γλώσσας καθιστά το εργαλείο αυτό σχεδόν ακριβές και κατά περιπτώσεις λανθασμένο. Επίσης, είναι εφοδιασμένο με αλγόριθμους που αναγνωρίζουν το είδος των λέξεων, την κλίση τους και τη γραμματική τους λειτουργία καθώς και με τη διεπαφή προγραμματισμού εφαρμογών API. Είναι γρήγορο στη λειτουργία του και σύμφωνα με αντίστοιχες μετρήσεις, η ακρίβειά του αγγίζει το 95%.

Pattern.search

Το pattern.search αποτελείται από έναν αλγόριθμο αναζήτησης ο οποίος ανακτά ακολουθίες λέξεων σε επισημασμένο κείμενο. Ακολουθεί τη μέθοδο N-gram και οι αναζητήσεις μπορούν να περιλαμβάνουν τόσο μεμονωμένες λέξεις όσο και φράσεις, μέρη του λόγου, όρους ταξινόμησης και τελεστές για την ανάδειξη σχετικών πληροφοριών.

Pattern.vector

Το pattern.vector αποτελεί εργαλείο μηχανικής μάθησης και βασίζεται στο Μοντέλο

Διανυσματικού Χώρου, όπως αυτό περιεγράφηκε στην παράγραφο 2.4.5. Με βάση το μοντέλο αυτό μπορούμε να επιτελέσουμε ομαδοποίηση, ταξινόμηση καθώς και λανθάνουσα σημασιολογική ανάλυση. Το αρχείο χωρίζεται σε ομάδες λημμάτων ανάλογα με τη ρίζα της κάθε λέξης και στη συνέχεια υπολογίζονται οι αριθμοί

$$TF - IDF_T = TF_{TD} \cdot IDF_T$$

για κάθε μια από αυτές. Το εργαλείο περιλαμβάνει, επίσης, έναν ιεραρχικό αλγόριθμο και έναν αλγόριθμο k-μέσων, έναν αλγόριθμο Naïve Bayes, έναν αλγόριθμο k-εγγύτερων γειτόνων και ένα SVM ταξινομητή.

Pattern.graph

Το pattern.graph παρέχει μια δομή γραφικής αναπαράστασης δεδομένων για την οπτικοποίηση των σχέσεων μεταξύ εννοιών και όρων που ονομάζονται κόμβοι. Στο εργαλείο περιλαμβάνονται αλγόριθμοι εύρεσης του συντομότερου δρόμου που συνδέει δύο διαφορετικούς κόμβους, αλγόριθμους διαμερισματοποίησης των γραφημάτων και κατασκευής ιδιοδιανυσμάτων. Τα γραφήματα αυτά μπορούν να εξαχθούν σε μορφή HTML ενώ η διαχείρισή τους γίνεται μέσω ενός συνηθισμένου προγράμματος περιήγησης στο διαδίκτυο.

Όλα τα προαναφερθέντα εργαλεία μαζί “χτίζουν” ένα πακέτο λογισμικού το οποίο είναι συμβατό με όλα τα λειτουργικά συστήματα. Ο πηγαίος κώδικας διανέμεται υπό την άδεια BSD με αποτέλεσμα να είναι δυνατή η συνδυασμένη χρήση του με παρόμοια λογισμικά όπως είναι το Scrapy που συζητήθηκε παραπάνω, το NLTK, το Pybrain και άλλα.

ΚΕΦΑΛΑΙΟ 4
ΣΥΜΠΕΡΑΣΜΑΤΑ

Δύο διαφορετικού τύπου αλλά συνεργαζόμενες, όπως συζητήθηκε στα προηγούμενα κεφάλαια, τάσεις διαμορφώνουν το πλαίσιο στο οποίο κινείται, αναπτύσσεται και επιχειρεί η σύγχρονη Εξόρυξη Δεδομένων. Αρχικά, ο συνεχώς αυξανόμενος όγκος δεδομένων που καθημερινά αποθηκεύονται στις διάφορες διαδικτυακές θέσεις και διακινούνται μέσω του Παγκόσμιου Ιστού καθιστά επιτακτική την ανάγκη εξεύρεσης ενός τρόπου διαχείρισης αυτών. Από την άλλη μεριά, το γεγονός ότι ο Παγκόσμιος Ιστός αποτελεί -πλέον- το κύριο καταθετήριο πληροφορίας, τον καθιστά το πρώτο μέρος όπου θα αναζητηθούν πληροφορίες για να μετουσιωθούν σε γνώση σε πρώτο στάδιο και στη συνέχεια να χρησιμοποιηθούν ποικιλλοτρόπως.

Αυτές οι δύο τάσεις δρουν συνεργατικά καθιστώντας πλέον την Εξόρυξη στον Ιστό τεχνολογία αιχμής. Η Εξόρυξη στον Ιστό μπορεί να λάβει διαφορετικά πρόσωπα ανάλογα με τις ανάγκες του εκάστοτε χρήστη. Έτσι, μιλάμε για Εξόρυξη Χρήσης όταν ενδιαφερόμαστε για πληροφορίες που αφορούν στην αλληλεπίδραση των χρηστών με τον Παγκόσμιο Ιστό, για Εξόρυξη Περιεχομένου όταν προσπαθούμε να εξάγουμε πολύτιμες πληροφορίες από δεδομένα που είναι αποθηκευμένα σε ιστοσελίδες και για Εξόρυξη Δομής όταν αναζητούμε δομικές πληροφορίες συγκεκριμένων διαδικτυακών θέσεων.

Για την εξυπηρέτηση αυτών των μορφών Εξόρυξης στον Ιστό, έχουν χρησιμοποιηθεί τεχνικές και αλγόριθμοι από πολλούς κλάδους της επιστήμης των υπολογιστών, όπως είναι η Τεχνητή Νοημοσύνη, η Μηχανική Μάθηση, τα Μοντέλα Διανυσματικών Χώρων και άλλα. Ωστόσο, δημιουργήθηκαν και νέα εργαλεία τα οποία και έχουν λάβει τη μορφή πακέτων λογισμικού με εξειδίκευση σε διάφορους τύπους Εξόρυξης στον Ιστό.

Στην εργασία αυτή παρουσιάσαμε πακέτα λογισμικού ελεύθερου κώδικα. Τα κριτήρια για τα πακέτα που επιλέχθηκαν να παρουσιαστούν ήταν -μάλλον- υποκειμενικά, δεδομένης της μη ύπαρξης κάποιας ανάλογης συγκριτικής μελέτης. Παρουσιάσαμε, λοιπόν, λογισμικά τα οποία είτε είχαν κάποιο ιδιαίτερο χαρακτήρα και κάποια ξεχωριστή δυνατότητα είτε ήταν καλά εργαλεία στα χέρια των χρηστών, αποδοτικά και αξιόπιστα.

Θεωρούμε πως η Εξόρυξη στον Ιστό είναι το εργαλείο του μέλλοντος. Ο αντίκτυπός της στις διάφορες εμπορικές εφαρμογές, στο μάρκετινγκ, στην εξόρυξη γνώσης ακαδημαϊκού περιεχομένου και στην πληροφόρηση του κοινού θεωρούμε πως είναι πολύ σπουδαίες και θα τραβήξουν το ενδιαφέρον των προγραμματιστών για μεγάλο χρονικό διάστημα ακόμα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Bing, L., Web Data Mining – Exploring Hyperlinks, Contents and Usage Data, Springer 2011

2. Chakrabarti, S., Mining The Web – Discovering Knowledge from Hypertext Data, Morgan Kaufmann Publishers 2003

3. Han, J., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers 2006

4. Xing, W. & Ghorbani, Al., Weighted PageRank Algorithm, Communication Networks and Services Research, Proceedings Second Annual Conference 2004

5. Gomes da Costa Junior, M. & Gong, Z., Web Structure Mining: An Introduction, Proceedings on the 2005 IEEE International Conference on Information Acquisition 2005

6. Chaudhary, K. & Gupta, K., Web Usage Mining Tools & Tachniques: A Survey, International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June 2013

7. Nithya, T., Link Analysis Algorithm for Web Structure Mining, International Journal of Advanced Research in Computer and Communication Engineering, Volume 2, Issue 8, August 2013

8. Bharanipriya, V. & Kamakshi Prasad, V., Web Content Mining Tools: A Comparative Study, International Journal of Information Technology and knowledge management, Volume 4, Issue 1, 2011

9. Kosala, R. & Blockeel, H., Web Mining Research: A Survey, SIGKDD, Volume 2, Issue 1, 2000

10. Pol, K. & Patil, N., A Survey on Web Content Mining and Extraction of Structured and Semistructured Data, First International Conference on Emerging Trends in Engineering and Technology, 2008

11. Dinuca, C. E., Web Structure Mining, Annals of the University of Petrosani, Economics, 11(4), 2011

12. Soni, R. & Kaur, G., Web Usage Mining: Personalization of Web Usage Data, International Journal of Advanced Research in Computer and Communication Engineering, Volume 3, Issue 2, 2014

13. Jain, N. & Srivastava, V., Data Mining Techniques: A Survey Paper, International Journal of Research in Engineering and Technology, Volume 2, Issue 11, 2013

14. Rajdeepa, B. & Sumathi, Dr. p., Web mining and its Methods, International Journal of Scientific & Engineering Research, Volume 4, Issue 6, 2013

15. Zubi, Z. S., Ranking WebPages Using Web Structure Mining Concepts, Recent Advances in Telecommunications, Signals and Systems

16. DEiXTo Basic Guide - http://deixto.com/wp-content/uploads/DEiXTo_basic_usage_guide_Greek.pdf

17. Bixo Documentation - http://bixo.101tec.com/documentation/

18. Jwanalytics Official Site - https://code.google.com/p/jwanalytics/

19. Analog 6.0 Documentation - http://www.analog.cx/docs/Readme.html

20. GNU Wget Documentation - https://www.gnu.org/doc/doc.html

21.Active Web Reader Official Site - http://www.deskshare.com/awr.aspx

22. Scrapy 0.22 Documentation - http://doc.scrapy.org/en/latest/

23. Trapit Official Site - https://trap.it

24. Pattern Official Site - http://www.clips.ua.ac.be/pages/pattern

25. De Smedt, T. & Daelemans, W., Pattern for Python, Journal of Machine Learning Research 13 (2012) 2063-2067

26. Berners-Lee, T., Information Management: A Proposal, 1989 (Extracted from info.cern.ch/Proposal.html)

27. Zubi, Z. S., Saleh El Riani, M., Recent Advances in Image, Audio and Signal Processing,