Πανεπιστήμιο Πελοποννήσου
Σχολή Θετικών Επιστημών και Τεχνολογίας
Τμήμα Επιστήμης και Τεχνολογίας Υπολογιστών

Μεταπτυχιακή Εργασία

Κρυμμένος Ιστός (Deep Web)

Ονοματεπώνυμο Φοιτητή : Μαρούδας Ιωάννης

ΑΜ : ΠΜΣ2010017

Επιβλέπων Καθηγητής: Βασιλάκης Κωνσταντίνος

Τρίπολη, Μάιος 2013

Πίνακας Περιεχομένων

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

Περίληψη

Η εργασία αυτή έχει ως αντικείμενο τον κρυμμένο ιστό, τα χαρακτηριστικά του γνωρίσματα, τις μεθόδους ιστοσυλλογής του κρυμμένου ιστού καθώς και τον συνδυασμό πολλαπλών πηγών δεδομένων. Οι μηχανές αναζήτησης δεν μπορούν συχνά να εντοπίσουν και να επιστρέψουν στατικές σελίδες που αντιστοιχούν στο περιεχόμενο του κρυμμένου ιστού, με αποτέλεσμα οι ερευνητές να στραφούν σε εναλλακτικές μεθόδους αναζήτησης στον κρυμμένο ιστό. Αυτό συμβαίνει επειδή η πρόσβαση στις πληροφορίες του κρυμμένου ιστού είναι εφικτή μέσω διεπαφών αναζήτησης, στις οποίες ο χρήστης εισάγει ένα σύνολο λέξεων κλειδιών.

Στην εργασία παρουσιάζονται αρχικά οι δύο κύριες κατηγορίες που υπάρχουν στον παγκόσμιο ιστό, ο επιφανειακός και ο κρυμμένος. Ο κρυμμένος ιστός περιέχει τις ακόλουθες κατηγορίες δεδομένων: μη συνδεδεμένες σελίδες, περιεχόμενο βάσεων δεδομένων, σελίδες που δεν έχουν μορφή HTML, σελίδες με εκτελέσιμα ή συμπιεσμένα αρχεία, περιεχόμενο περιορισμένης πρόσβασης αλλά και δυναμικό περιεχόμενο. Επίσης γίνεται παρουσίαση ενός ιστοσυλλέκτη που μπορεί να παράγει αυτόματα ερωτήματα, με σκοπό να εντοπίσει και να λάβει τις σελίδες του κρυμμένου ιστού. Ακολουθεί ένα σύστημα περιήγησης στο περιεχόμενο του κρυμμένου ιστού και περιγράφεται ένας αλγόριθμος με αποτελεσματική λειτουργία στον προσδιορισμό των πιθανών συνδυασμών εισόδου που παράγουν κατάλληλα URLs, τα οποία μπορούν να προστεθούν στον πίνακα μηχανών αναζήτησης. Τέλος, μελετάται ο τρόπος με τον οποίο μπορεί να υπολογιστεί ο σχεδιασμός ερωτήματος στα πλαίσια ενός συστήματος ολοκλήρωσης κρυμμένου ιστού. Επιπλέον παρουσιάζεται ένα αυτόματο σύστημα εξερεύνησης και ερώτησης των πηγών δεδομένων του κρυμμένου ιστού.

Abstract

The topic of this thesis is Deep Web and its characteristics, the crawling methods for Deep Web and the combination of multiple data sources. Search engines often cannot find and return static pages corresponding to the contents of the Deep Web, leading researchers to seek new methods for searching the hidden web. This is because access to hidden web information is possible through query interfaces, where the user enters a set of keywords.

This thesis initially presents the two main categories on the Web, the Surface Web and the Deep Web. Deep Web contains the following data categories: unconnected pages, database contents, pages not following the HTML format, pages with executable or compressed files, restricted content and dynamic content. A Deep Web crawler, which can automatically generate queries in order to locate and retrieve pages within the Deep Web, is also presented. A system which can browse the content of deep web is subsequently presented, followed by an algorithm which effectively determines possible input combinations that produce appropriate URLs that can be added to the search engine table. Finally, the thesis considers the way in which query designs can be calculated within a Deep Web integration system, and additionally presents an automated system for exploring and querying data sources of the Deep Web.

1 Εισαγωγή

Στο κεφάλαιο αυτό γίνεται αναφορά στις δύο κύριες κατηγορίες που υπάρχουν στον παγκόσμιο ιστό, τον επιφανειακό ιστό (surface web) και τον κρυμμένο ιστό (deep web). Επίσης επισημαίνονται οι βασικές διαφορές των δύο κατηγοριών καθώς και τα κύρια χαρακτηριστικά τους.

1.1 Επιφανειακός ιστός (Surface web)

Ο επιφανειακός ιστός (surface web ή visible web ή indexable web) είναι εκείνο το μέρος του παγκόσμιου ιστού που μπορεί να ευρετηριοποιηθεί από τις παραδοσιακές μηχανές αναζήτησης. Στην αντίπερα όχθη, το κομμάτι εκείνο του παγκόσμιου ιστού που δεν είναι προσβάσιμο από τις παραδοσιακές μηχανές αναζήτησης, καλείται κρυμμένος ιστός (deep web ή invisible web)[1].

Πρακτικά, οι μηχανές αναζήτησης κατασκευάζουν μια βάση δεδομένων του ιστού χρησιμοποιώντας κάποια ειδικά προγράμματα (ιστοσυλλέκτες - spiders ή web crawlers) τα οποία ξεκινούν τη διαδικασία σάρωσης από μια λίστα γνωστών σελίδων. Ο ιστοσυλλέκτης φυλάσσει ένα αντίγραφο της κάθε σελίδας και το καταχωρεί σε ένα ευρετήριο, αποθηκεύοντας χρήσιμες πληροφορίες οι οποίες θα επιτρέψουν να ανακληθεί η σελίδα γρήγορα σε μεταγενέστερο χρόνο. Οι υπερσυνδέσεις σε νέες σελίδες, προστίθενται στη λίστα των σελίδων που θα πρέπει να σαρωθούν. Το αποτέλεσμα είναι όλες οι προσπελάσιμες σελίδες που έχουν σχέση με την αναζήτηση, να καταχωρηθούν στο ευρετήριο. Η συλλογή αυτών των αποτελεσμάτων, δηλαδή των προσπελάσιμων σελίδων, καθορίζει τον επιφανειακό ιστό.

1.2 Μηχανές Αναζήτησης

Το κύριο εργαλείο που χρησιμοποιείται για την αναζήτηση πληροφοριών στον παγκόσμιο ιστό είναι οι μηχανές αναζήτησης (web search engines). Τα αποτελέσματα της αναζήτησης ταξινομούνται σε μία λίστα και παρουσιάζονται στις σελίδες αποτελεσμάτων της μηχανής αναζήτησης (SERPs - Search engine results page).

Στον Πίνακας 1 παρατίθεται μια λίστα του χρονικού ιστορικού λειτουργίας των διαφόρων μηχανών αναζήτησης καθώς και της κατάστασης λειτουργίας τους σήμερα.

Έτος	Μηχανή Αναζήτησης	Παρούσα Κατάσταση
1993	W3Catalog	Ανενεργή
	Aliweb	Ανενεργή
1994	WebCrawler	Ενεργή, Aggregator
	Go.com	Ενεργή, Αναζήτηση Yahoo!
	Lycos	Ενεργή
1995	AltaVista	Ανενεργή (Το URL ανακατευθύνεται στο Yahoo!)
	Daum	Ενεργή
	Magellan	Ανενεργή
	Excite	Ενεργή
	SAPO	Ενεργή
	Yahoo!	Ενεργή, δρομολογήθηκε ως κατάλογος

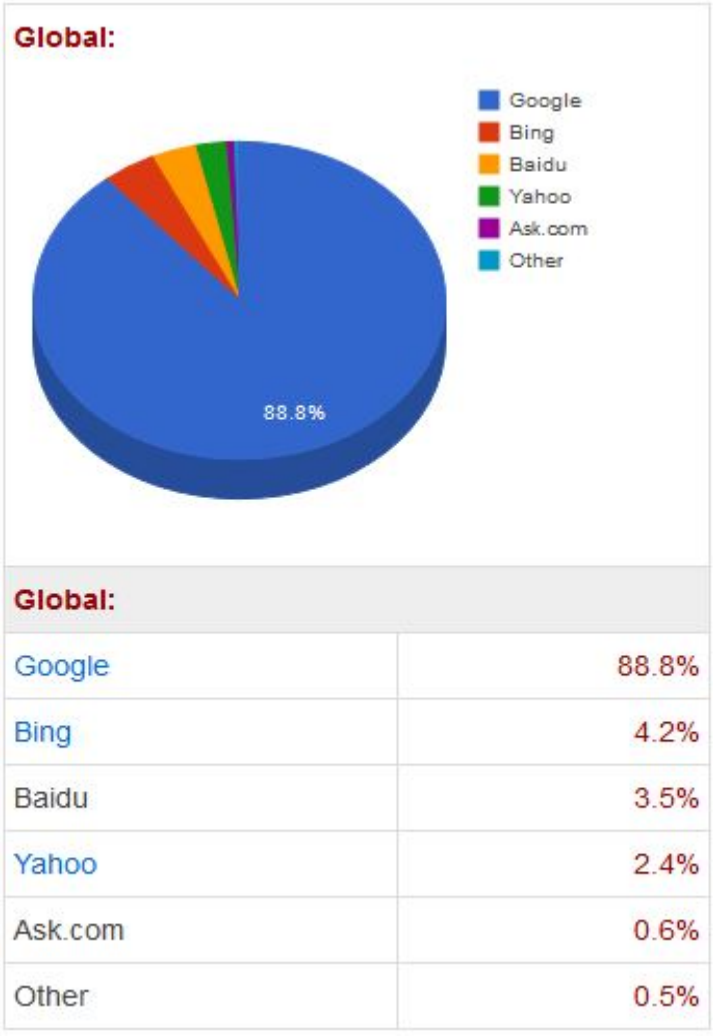
[1] Πηγή: http://en.wikipedia.org/wiki/Deep_Web

Έτος	Μηχανή Αναζήτησης	Παρούσα Κατάσταση
1996	Dogpile	Ενεργή, Aggregator
	Inktomi	Αποκτήθηκε από τη Yahoo!
	HotBot	Ενεργή (lycos.com)
	Ask Jeeves	Ενεργή (ask.com, η επέκταση "Jeeves" καταργήθηκε)
1997	Nothern Light	Ανενεργή
	Yandex	Ενεργή
1998	Google	Ενεργή
	MSN Search	Ενεργή ως Bing
1999	AlltheWeb	Ανενεργή (Το URL ανακατευθύνεται στο Yahoo!)
	GenieKnows	Ενεργή, ξαναπροβλήθηκε ως Yellowee.com
	Naver	Ενεργή
	Taoma	Ενεργή
	Vivisimo	Ανενεργή
2000	Baidou	Ενεργή
	Exalead	Αποκτήθηκε από τη Dassault Systemes
2002	Inktomi	Αποκτήθηκε από τη Yahoo!
2003	Info.com	Ενεργή
2004	Yahoo! Search	Ενεργή
	A9.com	Ανενεργή
	Sogou	Ενεργή
2005	AOL Search	Ενεργή
	Ask.com	Ενεργή
	GoodSearch	Ενεργή
	SearchMe	Έκλεισε
2006	Wikiseek	Ανενεργή
	Quaero	Ενεργή
	Ask.com	Ενεργή
	Live Search	Ενεργή ως Bing
	ChaCha	Ενεργή
	Guruji.com	Ενεργή
2007	Wikiseek	Ανενεργή
	Sproose	Ανενεργή
	Wikia Search	Ανενεργή
	Blackle.com	Ενεργή
2008	Powerset	Ανενεργή (ανακατευθύνει στο Bing)
	Picollator	Ανενεργή
	Viewzi	Ανενεργή
	Boogami	Ανενεργή
	LeapFish	Ανενεργή
	Forestle	Ανενεργή (ανακατευθύνει στοEcosia)
	VADLO	Ενεργή
	DuckDuckGo	Ενεργή, Aggregator
2009	Bing	Ενεργή
	Yebol	Ενεργή
	Mugurdy	Ανενεργή
	Goby	Ενεργή
2010	Black Google Mobile	Ενεργή
	Blekko	Ενεργή
	Cuil	Ανενεργή

Έτος	Μηχανή Αναζήτησης	Παρούσα Κατάσταση
	Yandex	Ενεργή, Αναζήτηση μόνο στα Αγγλικά
	Yammly	Ενεργή
2011	Interred	Ενεργή
2012	Volunia	Ενεργή, μόνο δυναμικοί χρήστες

Πίνακας 1: Αναπαράσταση ενεργών μηχανών αναζήτησης κατά τη περίοδο 1993 - 2012[2]

Αναφορικά με τα ποσοστά χρήσης των μηχανών αναζήτησης, όπως απεικονίζεται στο ακόλουθο σχεδιάγραμμα, η πιο δημοφιλής μηχανή αναζήτησης για το Φεβρουάριου του 2013, παραμένει με διαφορά η μηχανή αναζήτησης του Google.



Global:	
Google	88.8%
Bing	4.2%
Baidu	3.5%
Yahoo	2.4%
Ask.com	0.6%
Other	0.5%

Εικόνα 1: Ποσοστό χρήσης των μηχανών αναζήτησης[3]

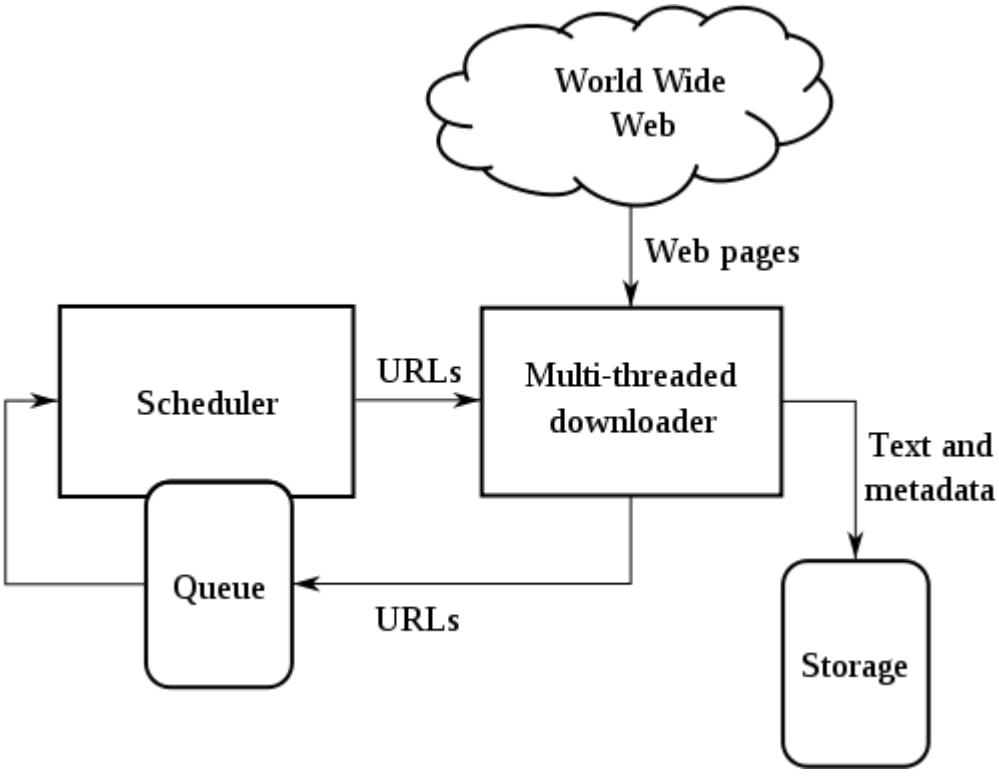
Όπως αναφέρθηκε, οι μηχανές αναζήτησης χρησιμοποιούν ειδικά προγράμματα (ιστοσυλλέκτες - spiders ή web crawlers) για την ανίχνευση των πληροφοριών στον

[2] Πηγή: http://en.wikipedia.org/wiki/Search_engine
[3] Πηγή: http://www.karmasnack.com/about/search-engine-market-share/

παγκόσμιο ιστό. Ο ιστοσυλλέκτης αποτελείται από τέσσερα κύρια μέρη τα οποία αναλύονται ακολούθως[4]:

- **Scheduler**: Αποτελείται από δύο μονάδες. Η μία μονάδα αναλαμβάνει τον εντοπισμό διπλότυπων URLs και η άλλη μονάδα αναλαμβάνει την εισαγωγή των URLs από την ουρά (queue).
- **Downloader**: Ξεκινώντας από τα URLs εκκίνησης, ανακτά τις αντίστοιχες σελίδες, εξάγει τα URLs που περιέχονται στις σελίδες υπό τη μορφή ιστοσελίδων κι εν συνεχεία τα στέλνει στον Scheduler, ο οποίος με τη σειρά του τα προσθέτει στην ουρά (εφόσον δεν αποτελούν διπλό-εγγραφή). Αποτελείται από τις ακόλουθες τρεις μονάδες:
    - DNS Resolving
    - Ανάκτηση σελίδων μέσω του HTTP
    - Συντακτική ανάλυση σελίδων HTML με σκοπό την εξαγωγή υπερσυνδέσμων και λοιπών στατιστικών δεδομένων.
- **Queue**: Δέχεται από τον Scheduler, τα URLs.
- **Μονάδα αποθήκευσης**: Το μέρος στο οποίο αποθηκεύονται οι σελίδες από τον Downloader.

Στην Εικόνα 2 εμφανίζεται η λειτουργία που εκτελείται.



Εικόνα 2: Αρχιτεκτονική του ιστοσυλλέκτη[5]

[4] Πηγή: http://en.wikipedia.org/wiki/Web_crawler
[5] Πηγή: http://en.wikipedia.org/wiki/File:WebCrawlerArchitecture.svg

1.3 Κρυμμένος ιστός εναντίον Επιφανειακού ιστού

Στην υποενότητα 1.3 παρουσιάζονται ορισμένα στατιστικά δεδομένα που αφορούν τις ποιοτικές και ποσοτικές διαφορές του κρυμμένου ιστού (deep web) από τον επιφανειακό (surface web). Τα δεδομένα αντλήθηκαν από τις εργασίες [1-3] και παρατίθενται ακολούθως:

- Ο όγκος των πληροφοριών στον κρυμμένο ιστό υπολογίζεται περίπου 500 φορές μεγαλύτερος από αυτόν του επιφανειακού ιστού.
- Ο κρυμμένος ιστός περιλαμβάνει 7.500 terabytes όγκο δεδομένων, ενώ ο επιφανειακός ιστό μόλις 19.
- Ο ρυθμός ανάπτυξης του κρυμμένου ιστού είναι πολύ υψηλότερος σε σχέση με τον ρυθμό ανάπτυξης του επιφανειακού ιστού.
- Υπολογίζεται ότι αυτή τη στιγμή υπάρχουν περισσότερα από 2.000.000 ιστοσελίδες στον κρυμμένο ιστό.
- Τα ανεξάρτητα έγγραφα στον κρυμμένο ιστό εκτιμώνται περίπου στα 550.000.000, τη στιγμή που ο επιφανειακός ιστός περιέχει περίπου 1.000.000.
- Οι περισσότερες πληροφορίες στον κρυμμένο ιστό, διατηρούνται από ακαδημαϊκά ιδρύματα και ερευνητικούς οργανισμούς, γι' αυτό το λόγο πολλοί ερευνητές του αντικειμένου υποστηρίζουν ότι η ποιότητα των πληροφοριών που βρίσκεται στον κρυμμένο ιστό είναι πολύ υψηλότερη από αυτή που βρίσκεται στον επιφανειακό.
- Το 95% των πληροφοριών που βρίσκονται στον κρυμμένο ιστό είναι προσβάσιμες, χωρίς την απαίτηση εγγραφής ή χρηματικού αντιτίμου.
- Το περιεχόμενο του κρυμμένου ιστού έχει μεγαλύτερη συνάφεια με την απαιτούμενη πληροφορία, σε σχέση με τον επιφανειακό ιστό.
- Περίπου το 55% του περιεχομένου του κρυμμένου ιστού είναι αποθηκευμένο σε βάσεις δεδομένων με συγκεκριμένη θεματική ενότητα.

Με βάση τα παραπάνω και σύμφωνα με μία μελέτη του ερευνητικού ινστιτούτου NEC που δημοσιεύτηκε στο επιστημονικό περιοδικό Nature εκτιμάται ότι οι μηχανές αναζήτησης με το μεγαλύτερο αριθμό ιστοσελίδων στο ευρετήριο τους (όπως το Google ή το Northern Light) δεν υπερβαίνουν το 16% του επιφανειακού ιστού στις καταχωρήσεις τους. Δεδομένου ότι απουσιάζουν τα αποτελέσματα του κρυμμένου ιστού, όταν χρησιμοποιούνται αυτές οι μηχανές αναζήτησης, οι ερευνητές που ψάχνουν στο διαδίκτυο αναζητούν δεδομένα μόνο στο 0,03% των σελίδων που έχουν στη διάθεσή τους σήμερα.

Στις ενότητες που θα ακολουθήσουν θα αναλυθούν περεταίρω οι διαφοροποιήσεις μεταξύ του κρυμμένου και του επιφανειακού ιστού, και θα δοθεί ιδιαίτερη έμφαση στα χαρακτηριστικά του κρυμμένου ιστού.

2 Κρυμμένος ιστός

2.1 Εισαγωγή στον κρυμμένο ιστό (Deep Web)

Στη διεθνή βιβλιογραφία, ο κρυμμένος ιστός (deep web) αναφέρεται και ως "invisible web" ή "hidden web". Η κατανόηση της πρακτικής του έννοιας δεν είναι εύκολα κατανοητή, ωστόσο η έννοια αυτή δεν μπορεί να προσδιοριστεί αυστηρώς από κάποιο ορισμό. Με απλά λόγια αυτό που στην ουσία εννοείται με τον όρο κρυμμένος ιστός είναι το περιεχόμενο εκείνο το όποιο υπάρχει στο διαδίκτυο αλλά δεν μπορεί να προσπελαστεί από τις μηχανές αναζήτησης γενικού σκοπού. Το περιεχόμενο αυτό μπορεί να αφορά αρχεία, σελίδες κειμένου και γενικότερα οποιαδήποτε άλλη πληροφορία η οποία δεν μπορεί να ανακτηθεί από τις μηχανές αναζήτησης γενικού σκοπού.

2.2 Περιεχόμενο και μέγεθος του κρυμμένου ιστού

Για την καλύτερη κατανόηση του είδους και του όγκου πληροφοριών που περιλαμβάνει ο κρυμμένος ιστός, αρχικά αναφέρεται ότι ο τρόπος μέσω του οποίου οι μηχανές αναζήτησης ανακτούν το περιεχόμενο του κρυμμένου ιστού, είναι κάποια ειδικά προγράμματα, οι ιστοσυλλέκτες (web crawlers), οι οποίοι στην ουσία ακολουθούν συνδέσμους (links). Η χρήση της τεχνικής αυτής είναι πολύ αποτελεσματική στην εύρεση πληροφοριών από τον επιφανειακό ιστό (Surface web), όχι όμως κι από τον κρυμμένο ιστό. Αυτό είναι λογικό αν σκεφτεί κανείς ότι ο ιστοσυλλέκτης (web crawler) πλοηγείται στον ιστό μέσω των συνδέσμων που είναι διαθέσιμοι σε κάθε σελίδα. Αν λοιπόν υπάρχει μια σελίδα για την οποία δεν υπάρχει σύνδεσμος σε καμία άλλη, τότε δεν μπορεί να εντοπιστεί από τον ιστοσυλλέκτη. Τέτοιου είδους σελίδες αποτελούν μέρος του κρυμμένου ιστού.

2.2.1 Περιεχόμενο του κρυμμένου ιστού

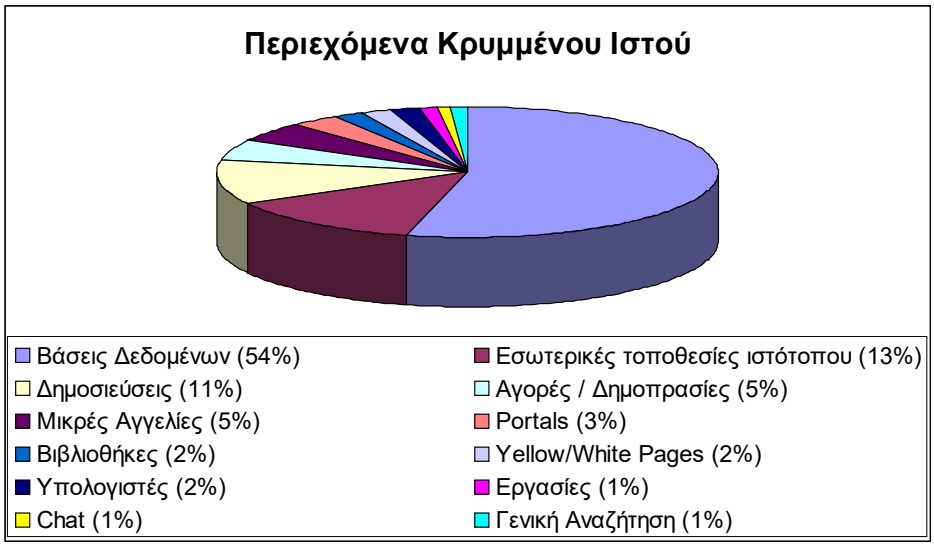
Όπως αναφέρεται στις εργασίες [1,2], ο κρυμμένος ιστός μπορεί να περιέχει τις ακόλουθες κατηγορίες:

- *Μη συνδεδεμένες σελίδες*: Δεν υπάρχει σύνδεσμος που να κατευθύνει σε αυτή τη σελίδα. Πρακτικά με αυτό τον τρόπο εμποδίζονται οι ιστοσυλλέκτες να αποκτήσουν πρόσβαση στο περιεχόμενό τους.
- *Περιεχόμενο βάσεων δεδομένων*: Οι ιστοσυλλέκτες δεν μπορούν να αλληλεπιδράσουν με τις φόρμες αναζήτησης που παρέχονται στους πραγματικούς χρήστες. Όπως θα αναλυθεί και στη συνέχεια το μεγαλύτερο μέρος του κρυμμένου ιστού εμπεριέχεται σε βάσεις δεδομένων. Ο λόγος για την ύπαρξη αυτού του φαινομένου είναι το γεγονός, ότι οι μηχανές αναζήτησης γενικού σκοπού δεν μπορούν να εντοπίσουν τις δυναμικά παραγόμενες πληροφορίες που υπάρχουν στις βάσεις δεδομένων.
- *Σελίδες που δεν έχουν μορφή HTML και περιέχουν κυρίως οπτικοακουστικό υλικό*: Λόγω της μη ύπαρξης "αρκετού" κειμένου, οι μηχανές αναζήτησης δεν μπορούν να δημιουργήσουν λέξεις κλειδιά για να εντοπίσουν αυτό το περιεχόμενο. Για παράδειγμα, μια σελίδα που περιέχει μόνο γραφικά δεν θα

μπορούσε να προσπελαθεί καθώς δεν υπάρχει κάποια λέξη κλειδί που θα μπορούσε να καταχωρηθεί στη μηχανή αναζήτησης.

- *Σελίδες με εκτελέσιμα ή συμπιεσμένα αρχεία*: Συνήθως είναι προσπελάσιμα αλλά πολλές φορές οι μηχανές αναζήτησης τα απορρίπτουν εσκεμμένα (συνήθως για λόγους προστασίας πχ. εκτέλεση αρχείων με κακόβουλο λογισμικό).
- *Περιεχόμενο περιορισμένης πρόσβασης*: Αφορά σελίδες που δεν επιτρέπουν την περιήγηση στο περιεχόμενο τους (ή σε μέρος του περιεχομένου τους). Συνήθως, τέτοιου τύπου ιστοσελίδες χωρίζονται σε δύο μεγάλες κατηγορίες:
    - εκείνες οι οποίες απαιτούν εγγραφή του χρήστη (registration), δηλαδή οι σελίδες εκείνες στις οποίες απαιτείται η εισαγωγή κάποιου "ονόματος χρήστη (username)" και "κωδικού (password)" και ως εκ τούτου δεν μπορούν να προσπελαστούν από τους ιστοσυλλέκτες.
    - εκείνες οι οποίες χρησιμοποιούν ειδικά προγράμματα, τα οποία αποτρέπουν την πρόσβαση των ιστοσυλλεκτών στο περιεχόμενό τους.
- *Δυναμικό Περιεχόμενο*: Περιεχόμενο το οποίο δημιουργείται δυναμικά ανάλογα με τις απαιτήσεις του χρήστη, δηλαδή δυναμικές σελίδες οι οποίες προκύπτουν ως απάντηση (response) σε ένα ερώτημα (query) ή δυναμικές σελίδες που μπορούν να προσπελαστούν μέσω κάποιας φόρμας αναζήτησης. Αυτό είναι σύνηθες στις αναζητήσεις που πραγματοποιούνται σε βάσεις δεδομένων, γι' αυτό το λόγο άλλωστε το μεγαλύτερο μέρος των εγγράφων που βρίσκονται στον κρυμμένο ιστό ενυπάρχει σε τέτοιου τύπου βάσεις δεδομένων.

Σύμφωνα με τον Michael K. Bergman [1], ο κρυμμένος ιστός περιλαμβάνει σε ποσοστά τις ακόλουθες ενότητες περιεχομένων.



Εικόνα 3: Κατανομή του κρυμμένου ιστού με βάση το περιεχόμενο

Όπως παρουσιάζεται και στο γράφημα της Εικόνα 3, το 54% αφορά βάσεις δεδομένων με συγκεκριμένη θεματική ενότητα. Αν σε αυτό το ποσοστό προσθέσουμε

το 13% των εσωτερικών εγγράφων που βρίσκονται σε μεγάλους ιστότοπους και το 11% που αφορά τις δημοσιεύσεις, τότε διαπιστώνουμε ότι η συντριπτική πλειοψηφία του κρυμμένου ιστού (78%) καλύπτεται από αυτές τρεις μεγάλες κατηγορίες.

2.2.2 Μέγεθος του κρυμμένου ιστού

Το ακριβές μέγεθος του κρυμμένου ιστού, είναι ουσιαστικά αδύνατο να υπολογιστεί. Στην ουσία μπορεί μόνο μια εκτίμηση να γίνει για το μέγεθός του. Πιο συγκεκριμένα, η εκτίμηση του όγκου των δεδομένων που υπάρχει στο κρυμμένο ιστό αποτελεί ανοικτό πρόβλημα από το 1998. Στη διάρκεια αυτών των χρόνων αναπτύχθηκαν διάφορες τεχνικές για την εκτίμηση του μεγέθους του. Στην εργασία [3] γίνεται μια σύνοψη όλων των τεχνικών που αναπτύχθηκαν τα τελευταία χρόνια σχετικά με τον υπολογισμό του όγκου των δεδομένων που υπάρχουν στον κρυμμένο ιστό.

Η βασική τεχνική για τον υπολογισμό των σχετικών μεγεθών του όγκου δεδομένων προτάθηκε από τους Bharat και Broder [5] και αφορά το ακόλουθο πιθανοκρατικό μοντέλο, όπου αν θεωρηθούν τα σύνολο A (το οποίο μπορεί να αντιπροσωπεύει το σύνολο των αποτελεσμάτων URL που επιστρέφονται από ένα ερώτημα) και το σύνολο B (που αφορά το δείγμα των αποτελεσμάτων URL που επιλέγονται από τα πρώτα πχ. 100 αποτελέσματα) τότε:

- P(A), η πιθανότητα ένα στοιχείο να ανήκει στο A.
- P(A∩B | A), η πιθανότητα ένα στοιχείο να ανήκει στη τομή των A και B και ταυτόχρονα να ανήκει και στο A.

τότε,

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

ισοδύναμα προκύπτει ότι:

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

Από τις δύο προηγούμενες εξισώσεις εξάγεται η εξίσωση:

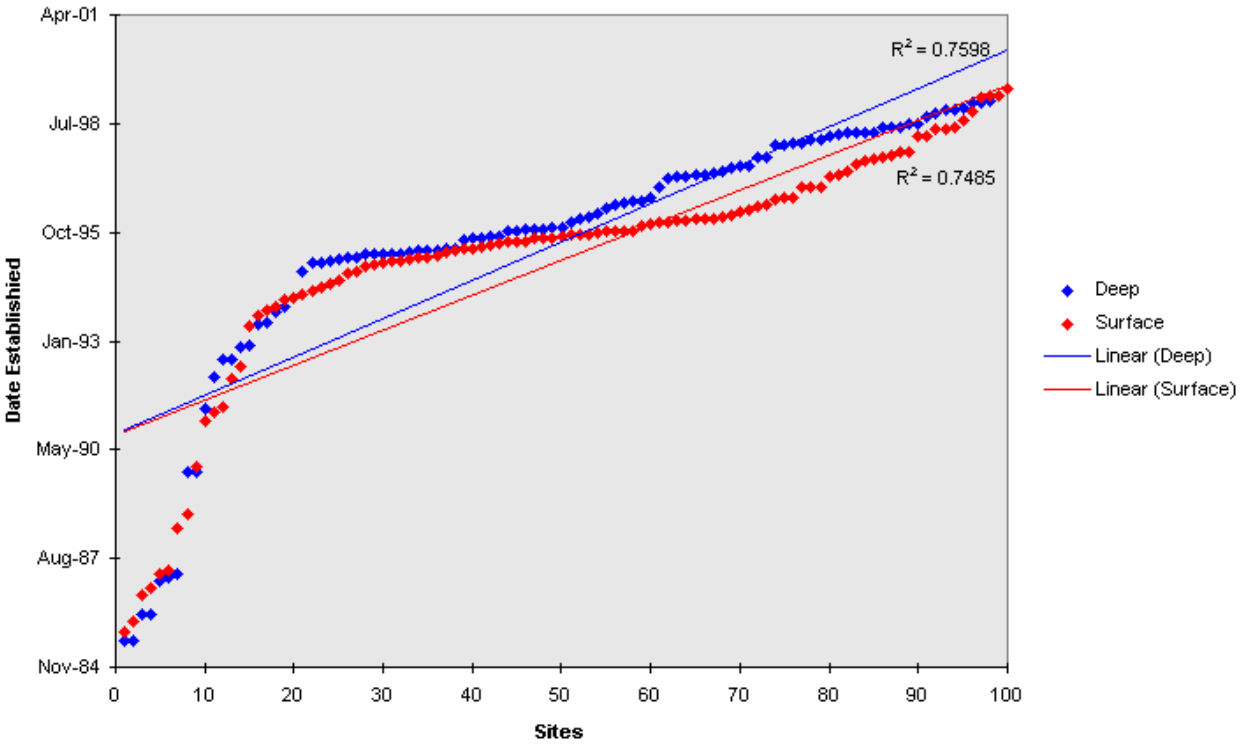
$$\frac{|A|}{|B|} = \frac{P(A \cap B|B)}{P(A \cap B|A)}$$

Σύμφωνα με έρευνα που διεξήχθη στο Πανεπιστήμιο Berkeley της Καλιφόρνια το 2001, ο κρυμμένος ιστός αποτελείται περίπου από 91.000 terabytes, εν αντιθέσει με τον επιφανειακό ιστό, όπου είναι περίπου 167 terabytes. Περίπου 550 εκατομμύρια ανεξάρτητων εγγράφων είναι διαθέσιμα στον κρυμμένο ιστό, τη στιγμή που ο επιφανειακός ιστός διαθέτει περίπου ένα εκατομμύριο [4].

Ο κρυμμένος ιστός είναι η μεγαλύτερη σε ανάπτυξη αποθήκη πληροφοριών στο διαδίκτυο. Οι πληροφορίες που υπάρχουν στον κρυμμένο ιστό υπολογίζονται ότι

είναι περίπου 500 φορές μεγαλύτερες σε όγκο από αυτές που είναι διαθέσιμες στον παγκόσμιο ιστό. Επίσης η ποιότητα (αξιοπιστία πηγών) του περιεχομένου του είναι 1.000 με 2.000 φορές καλύτερη σε σχέση με τον επιφανειακό ιστό [2].

Στην Εικόνα 4, όπως αυτή παρουσιάζεται στην εργασία [4] απεικονίζεται ο ρυθμός ανάπτυξης του επιφανειακού και του κρυμμένου ιστού σε σχέση με το χρόνο.



Εικόνα 4: Σύγκριση του ρυθμού ανάπτυξης επιφανειακού και κρυμμένου ιστού

Συμπερασματικά, μπορεί να ειπωθεί ότι όσο αυξάνεται ο όγκος πληροφοριών στον επιφανειακό ιστό, τόσο θα υπάρχει μια αύξηση στον όγκο των πληροφοριών που υπάρχουν στον κρυμμένο ιστό.

2.3 Λόγοι χρήσης του κρυμμένου ιστού

Υπάρχουν διάφορες αιτίες για τις οποίες ο χρήστης θα επέλεγε τη χρήση του κρυμμένου ιστού για την εύρεση των πληροφοριών που επιθυμεί, ωστόσο όπως συμβαίνει και στα περισσότερα πράγματα, έτσι και η χρήση του κρυμμένου ιστού έχει κάποια πλεονεκτήματα και μειονεκτήματα.

2.3.1 Περιορισμοί των μηχανών αναζήτησης

Οι μηχανές αναζήτησης συνέβαλαν τα μέγιστα στην δυνατότητα των ανθρώπων να έχουν πρόσβαση στο περιεχόμενο του διαδικτύου και είναι ένα σημαντικό συστατικό μιας επιτυχούς έρευνας. Ωστόσο υπάρχουν αρκετοί λόγοι για τους οποίους μπορεί να μην είναι η βέλτιστη επιλογή. Ακολούθως παρατίθενται κάποιοι από τους λόγους [2,4]:

- *Ποιότητα αποτελεσμάτων*: Πολλές μηχανές αναζήτησης επιστρέφουν αποτελέσματα βάση της συνάφειας ταξινόμησης σύμφωνα με τις λέξεις-κλειδιά που πληκτρολογεί ο χρήστης. Η ποιότητα των αποτελεσμάτων δεν λαμβάνεται υπόψη με αποτέλεσμα να εμφανίζονται πολλές σελίδες που το περιεχόμενό τους είναι αμφιβόλου ποιότητας. Η ποιότητα αφορά κυρίως δύο παραμέτρους, η πρώτη σχετίζεται με την αξιοπιστία των δεδομένων και η άλλη με τη συσχέτιση που υπάρχει μεταξύ των δεδομένων που αντλούνται και του ερωτήματος που έχει τεθεί.
- *Ποσότητα αποτελεσμάτων*: Οι μηχανές αναζήτησης επιστρέφουν πληθώρα αποτελεσμάτων, χωρίς ωστόσο να υπάρχει διακριτοποίηση ανάμεσα στις ποιοτικές σελίδες και σε εκείνες που παρέχουν επισφαλείς πληροφορίες. Επίσης, ένα άλλο σημαντικό θέμα που προκύπτει, είναι το γεγονός ότι ένα υψηλό ποσοστό των αποτελεσμάτων που προκύπτουν δεν είναι προσπελάσιμα από τον χρήστη, δηλαδή ενώ οι μηχανές αναζήτησης εμφανίζουν στον αριθμό των αποτελεσμάτων τους κάποιες σελίδες, αυτές στην πράξη δεν είναι προσβάσιμες. Ο λόγος που συμβαίνει αυτό είναι επειδή σύμφωνα με τον ορισμό του κρυμμένου ιστού, κάποιες σελίδες αντλούν δεδομένα από "κρυφές" βάσεις δεδομένων. Για τη λειτουργία των βάσεων δεδομένων στον κρυμμένο ιστό, θα υπάρξει περεταίρω ανάλυση στα κεφάλαια που ακολουθούν.
- *Επιφανειακή αναζήτηση*: Είναι σύνηθες το φαινόμενο στα αποτελέσματα που προκύπτουν από τις μηχανές αναζήτησης να εμφανίζονται μόνο 2-3 σελίδες από έναν ιστότοπο κι όχι σελίδες που βρίσκονται σε "βαθύτερο επίπεδο". Για παράδειγμα σ' ένα μεγάλο ιστότοπο με πολλές σελίδες, μπορεί να επιστραφούν μόνο η αρχική σελίδα και 1-2 ακόμα. Αυτό συμβαίνει λόγω της επιθυμίας των σχεδιαστών των μηχανών αναζήτησης να μειώσουν το υπολογιστικό κόστος, καθώς είναι αρκετά χρονοβόρο για τους ιστοσυλλέκτες να ανιχνεύσουν τον ιστό. Αυτό όμως έχει σαν αποτέλεσμα να αποκλείονται από τα αποτελέσματα τα περιεχόμενα ενός ποιοτικού ιστότοπου με πολλά επίπεδα σελίδων.
- *Προτιμήσεις των μηχανών αναζήτησης*: Ανάλογα με το πώς ο χρήστης έχει ρυθμίσει τις προτιμήσεις του, μια μηχανή αναζήτησης ενώ μπορεί να βρει και να εμφανίσει κάποια διαθέσιμα έγγραφα, τελικά τα αποκλείει λόγω των ρυθμίσεων που έχουν καθοριστεί από τον χρήστη.
- *Όριο εμφάνισης μηχανών αναζήτησης*: Κάποιες μηχανές αναζήτησης μπορεί να έχουν κάποιο όριο στον αριθμό των αποτελεσμάτων που μπορούν να εμφανίσουν από έναν ιστότοπο. Για παράδειγμα, η μηχανή αναζήτησης του Google επιτρέπει 1-2 αποτελέσματα στο ευρετήριο του από έναν ιστότοπο για ένα θέμα αναζήτησης. Πρακτικά αυτό σημαίνει ότι για να προβάλει κάποιος τις υπόλοιπες σχετικές σελίδες από τον ιστότοπο, πρέπει να επιλέξει την αντίστοιχη επιλογή.
- *Επιχειρήσεις και δημοτικότητα*: Οι περισσότερες μηχανές αναζήτησης ταξινομούν τα αποτελέσματά τους σύμφωνα με το πόσο δημοφιλής είναι η

κάθε σελίδα. Για παράδειγμα, η μηχανή αναζήτησης του Google επιστρέφει αποτελέσματα βάση του ποιες είναι οι πιο δημοφιλής και ευρέως γνωστές ιστοσελίδες για το θέμα της αναζήτησης. Ένα σημαντικό κριτήριο που λαμβάνεται υπόψη είναι το πόσες σελίδες συμπεριλαμβάνουν στα περιεχόμενα τους ως σύνδεσμο την πιο "δημοφιλή ιστοσελίδα". Αυτό ωστόσο δεν σημαίνει απαραίτητα ότι αυτές είναι καλύτερες σελίδες για την κάλυψη των αναγκών. Επίσης, πρέπει να τονισθεί ότι πίσω από τις περισσότερες μηχανές αναζήτησης είναι επιχειρήσεις που ως βασική επιδίωξη έχουν το κέρδος. Ένα μεγάλο λοιπόν κέρδος προσφέρεται σε αυτές, μέσω των εσόδων από διαφημίσεις και ως γνωστό πολλές επιχειρήσεις επενδύουν πολλά χρήματα ώστε να καταφέρουν να εμφανίζεται ο ιστοχώρος τους στην κορυφή των αποτελεσμάτων.

2.3.2 Οφέλη από τη χρήση του κρυμμένου ιστού

Τα οφέλη που έχει ο χρήστης από την αναζήτηση πληροφοριών στο κρυμμένο ιστό, είναι σε άμεση συσχέτιση με εκείνους τους λόγους, για τους οποίους οι μηχανές αναζήτησης πολλές φορές δεν είναι σε θέση να μας επιστρέψουν τις πληροφορίες που χρειαζόμαστε τόσο ποσοτικά, όσο και ποιοτικά [1,2,4].

Οι κυριότεροι λόγοι για τους οποίους ο χρήστης θα επέλεγε τον κρυμμένο ιστό παρουσιάζονται ακολούθως:

- *Μη διαθεσιμότητα στον ιστό*: Ο κυριότερος λόγος για τη χρήση του κρυμμένου ιστού είναι η αδυναμία εύρεσης πολλών πληροφοριών στον επιφανειακό ιστό. Όπως αναφέρθηκε, οι μηχανές αναζήτησης γενικού σκοπού εκτελούν ειδικά προγράμματα (ιστοσυλλέκτες, spiders ή crawlers) τα οποία ανιχνεύουν τα περιεχόμενα των σελίδων, ακολουθώντας τους συνδέσμους που βρίσκονται σε κάθε σελίδα. Τα περισσότερα όμως περιεχόμενα που βρίσκονται στον κρυμμένο ιστό, υπάρχουν μέσα σε βάσεις δεδομένων άρα παραμένουν κρυμμένα από τις μηχανές αναζήτησης. Έτσι, τα προγράμματα που χρησιμοποιούνται για την αναζήτηση πληροφοριών στον κρυμμένο ιστό, παρέχουν πληθώρα αποτελεσμάτων που οι παραδοσιακές μηχανές αναζήτησης δεν δύναται να τις εμφανίσουν.
- *Εξειδίκευση*: Συνήθως οι πόροι-δεδομένα του κρυμμένου ιστού εστιάζουν σε συγκεκριμένα θέματα, έτσι δύνεται η δυνατότητα στους χρήστες να ανακτήσουν πολύ πιο συγκεκριμένα και περιεκτικά αποτελέσματα.
- *Εστίαση*: Οι διεπαφές αναζήτησης στον κρυμμένο ιστό σχεδιάζονται με βάση τον τύπο αναζήτησης που γίνεται.
- *Ποιότητα - αξιοπιστία*: Τα περιεχόμενα του κρυμμένου ιστού πολλές φορές δημιουργούνται από οργανισμούς και ιδρύματα που έχουν την εποπτεία των θεμάτων που καλύπτουν και γι’ αυτό το λόγο οι αντίστοιχες πηγές είναι και πιο αξιόπιστες.

3 Μέθοδοι ιστοσυλλογής για τον κρυμμένο ιστό

Στα κεφάλαια που προηγήθηκαν, αναλύθηκε η έννοια του κρυμμένου ιστού καθώς και τα ιδιαίτερα χαρακτηριστικά του γνωρίσματα. Η πρόσβαση στον μεγάλο όγκο πληροφοριών του παγκόσμιου ιστού είναι συνήθως εφικτή μέσω διεπαφών αναζήτησης, στις οποίες ο χρήστης εισάγει ένα σύνολο λέξεων-κλειδιών. Όπως έχει ήδη αναφερθεί, δεν υπάρχουν στατικοί σύνδεσμοι για τις σελίδες στον κρυμμένο ιστό, γεγονός που καθιστά τις μηχανές αναζήτησης ανίσχυρες στην προσπάθεια τους να εντοπίσουν και να επιστρέψουν ως αποτέλεσμα τις σελίδες αυτές. Το πρόβλημα αυτό, έχει απασχολήσει πλήθος ερευνητών, καθώς οι σελίδες αυτές περιέχουν πολλές φορές υψηλής ποιότητας περιεχόμενο.
Το κεφάλαιο αυτό πραγματεύεται τη μεθοδολογία συλλογής πληροφοριών για τις σελίδες του κρυμμένου ιστού. Στις ενότητες που ακολουθούν θα γίνει η ανάλυση τριών μεθοδολογιών οι οποίες ασχολούνται με το θέμα.

Πιο συγκεκριμένα στην ενότητα 3.1, παρουσιάζεται μια προσέγγιση σύμφωνα με την οποία ένας ιστοσυλλέκτης μπορεί να παράγει αυτόματα ερωτήματα, με σκοπό να εντοπίσει και να λάβει τις σελίδες του κρυμμένου ιστού. Στην ενότητα αυτή, όπως θα φανεί και στη συνέχεια, δίνεται έμφαση στις βάσεις δεδομένων κειμένου (textual databases) οι οποίες υποστηρίζουν ερωτήματα λέξεων-κλειδιών.

Εν συνεχεία, στην ενότητα 3.2, περιγράφεται ένα σύστημα περιήγησης στο περιεχόμενο του κρυμμένου ιστού. Το σύστημα αυτό, σχετίζεται με τον προ-υπολογισμό υποβολών ερωτημάτων για κάθε φόρμα HTML και τη μετέπειτα προσθήκη των προκυπτουσών σελίδων HTML στα ευρετήρια μιας μηχανής αναζήτησης. Η τεχνική αυτή χρησιμοποιείτε από την Google και είναι σε θέση να προσφέρει σε περισσότερα από χίλια ερωτήματα ανά δευτερόλεπτο συνδέσμους προς σελίδες του κρυμμένου ιστού. Επίσης, γίνεται η παρουσίαση ενός νέου αλγορίθμου που χρησιμεύει στην επιλογή έγκυρων τιμών εισόδου σε πλαίσια κειμένου, τα οποία δέχονται λέξεις-κλειδιά προς διεξαγωγή αναζήτησης, καθώς επίσης και ενός αλγορίθμου για την αναγνώριση πλαισίων κειμένου που δέχονται τιμές ενός συγκεκριμένου τύπου. Τέλος, προτείνεται ένας αλγόριθμος με αποτελεσματική λειτουργία στον προσδιορισμό των πιθανών συνδυασμών τιμών για πλαίσια κειμένου μιας φόρμας που παράγουν κατάλληλα URLs, τα οποία μπορούν να προστεθούν στο πίνακα μηχανών αναζήτησης.

Τέλος, στην ενότητα 3.3, θα παρουσιαστεί το σύστημα DeLa, το οποίο επανακατασκευάζει ένα μέρος μιας βάσης δεδομένων του κρυμμένου ιστού. Αυτό επιτυγχάνεται μέσω της αποστολής ερωτημάτων με τη χρήση φορμών HTML, παράγοντας αυτόματα περιτυλίγματα (wrappers) κανονικών εκφράσεων, με σκοπό την εξαγωγή δεδομένων από τις ιστοσελίδες και την αποθήκευση αυτών σ' ένα πίνακα με ετικέτες.

3.1 Λήψη Περιεχομένου Κειμένων από τον Κρυμμένο Ιστό μέσω ερωτημάτων λέξεων-κλειδιών

Στην ενότητα αυτή, παρουσιάζεται μια μεθοδολογία κατασκευής ενός αποδοτικού και λειτουργικού ιστοσυλλέκτη κρυμμένου ιστού, ο οποίος είναι σε θέση να εντοπίζει και να "κατεβάζει" αυτόνομα σελίδες από τον κρυμμένο ιστό. Η μεθοδολογία αυτή έχει προταθεί από τους A. Ntoulas, P. Zerfos και J. Cho [6].

3.1.1 Πλαίσιο

3.1.1.1 Μοντέλο βάσης δεδομένων του κρυμμένου ιστού

Ένας ιστοχώρος του κρυμμένου ιστού μπορεί να κατηγοριοποιηθεί ανάλογα με τον τύπο των πληροφοριών που διαθέτει είτε ως *βάση δεδομένων κειμένου* (textual database), είτε ως *δομημένη βάση δεδομένων* (structured database). Βάση δεδομένων κειμένου είναι ένας δικτυακός τόπος που περιέχει έγγραφα απλού κειμένου (plain-text). Πολλές φορές τα έγγραφα απλού κειμένου δεν έχουν καλά ορισμένη δομή και γι' αυτό οι περισσότερες βάσεις δεδομένων κειμένου επιτρέπουν στους χρήστες να πληκτρολογούν λέξεις-κλειδιά σε ένα μοναδικό πλαίσιο κειμένου (Εικόνα 5).

Εικόνα 5: Απλή διεπαφή αναζήτησης με μοναδικό πλαίσιο κειμένου για εισαγωγή λέξεων-κλειδιών

3.1.1.2 Ένας γενικός αλγόριθμος ιστοσυλλογής (crawling) του κρυμμένου ιστού

Ένας γενικός αλγόριθμος ιστοσυλλογής του κρυμμένου ιστού για να προσπελάσει σελίδες του δικτυακού τόπου μέσω μιας φόρμας αναζήτησης χρειάζεται να ακολουθήσει τα εξής τρία βήματα.

1. Να δημιουργήσει ένα ερώτημα που θα τεθεί στο δικτυακό τόπο.
2. Να λάβει τη σελίδα με τον πίνακα αποτελεσμάτων.
3. Να επιλέξει τους συνδέσμους προς το περιεχόμενο που θεωρείται ενδιαφέρον και να τους ζητήσει από τον ιστοχώρο ώστε να λάβει τις αντίστοιχες σελίδες.

Τα βήματα αποτυπώνονται στην Εικόνα 6.

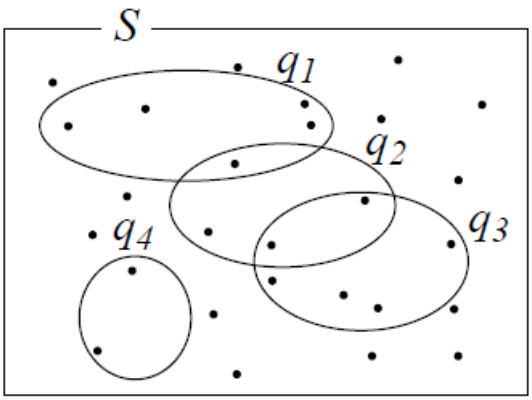
```
(1) while ( there are available resources ) do
        // select a term to send to the site
(2)     qi = SelectTerm()
        // send query and acquire result index page
(3)     R(qi) = QueryWebSite( qi )
        // download the pages of interest
(4)     Download( R(qi) )
(5) done
```

Εικόνα 6: Γενικός αλγόριθμος ιστοσυλλογής (crawling) του κρυμμένου ιστού

Το πιο σημαντικό βήμα είναι η επιλογή του ερωτήματος που θα υποβάλει ο αλγόριθμος. Ως πετυχημένη αναζήτηση θεωρείται αυτή που θα φέρει τα περισσότερα αποτελέσματα που πληρούν τα κριτήρια που έχει θέσει ο χρήστης.

3.1.1.3 Ο φορμαλισμός του προβλήματος

Για να διατυπωθεί τυπικά η έννοια του προβλήματος της επιλογής ερωτημάτων, θα βασιστούμε στη θεωρεία συνόλων. Ας θεωρήσουμε ότι ο ιστοσυλλέκτης λαμβάνει σελίδες από ένα δικτυακό τόπο μέσα από ένα σύνολο σελίδων S. Στην Εικόνα 7 με τις κουκίδες αναπαρίσταται κάθε σελίδα του δικτυακού τόπου, με q_i κάθε ερώτημα που μπορεί να τεθεί σε ένα δικτυακό τόπο, με τη συσχετιζόμενη έλλειψη οι σελίδες που το q_i επιστρέφει (υποσύνολο του S). Επίσης, κάθε q_i χαρακτηρίζεται από ένα "βάρος" το οποίο υποδηλώνει το κόστος από τη χρησιμοποίηση του ερωτήματος.



Εικόνα 7: Ο φορμαλισμός του προβλήματος της βέλτιστης επιλογής ερωτήματος με βάση τη θεωρία συνόλων

Σύμφωνα με τα παραπάνω ως βέλτιστη λύση είναι η εύρεση εκείνων των ερωτημάτων q_i τα οποία επιστρέφουν όσο το δυνατόν περισσότερες σελίδες που αφορούν το χρήστη, ενώ ταυτόχρονα επιτυγχάνεται το μικρότερο δυνατό κόστος, όπου το κόστος μπορεί να μετριέται:

- σε χρόνο,
- σε εύρος ζώνης του δικτύου,
- σε αριθμό των αλληλεπιδράσεων με το δικτυακό τόπο,
- ή σε συνδυασμό όλων των παραπάνω.

Η παρακάτω σχέση υποδηλώνει το συνολικό κόστος από ένα ερώτημα q_i

$$Cost(q_i) = c_q + c_r P(q_i) + c_a P(q_i)$$

όπου:

- $P(q_i)$: το σύνολο των σελίδων που θα επιστραφούν εάν τεθεί το ερώτημα q_i,
- c_q : το κόστος υποβολής ερωτήματος (θεωρείται σταθερό και ανεξάρτητο της ερώτησης),

- c_r : το κόστος για τη λήψη της σελίδας που περιέχει τα αποτελέσματα της αναζήτησης (θεωρείται σταθερό και ανεξάρτητο της ερώτησης),
- c_d : το καθορισμένο κόστος από τη λήψη ενός εγγράφου που πληροί τα κριτήρια αναζήτησης (θεωρείται σταθερό και ανεξάρτητο της σελίδας).

Υπάρχουν περιπτώσεις όπου ορισμένα έγγραφα που επιστρέφονται από το ερώτημα q_i να έχουν ήδη ληφθεί από προηγούμενα ερωτήματα. Σε τέτοιες περιπτώσεις, ο ιστοσυλλέκτης μπορεί να μην προβεί στη λήψη αυτών των εγγράφων. Αυτό συνεπάγεται τη μεταβολή της σχέσης του κόστους:

$$Cost(q_i) = c_q + c_r P(q_i) + c_d P_{new}(q_i)$$

Όπου:

- $P_{new}(q_i)$ είναι ο αριθμός των νέων σελίδων που επιστρέφονται από την υποβολή του ερωτήματος και οι οποίες δεν είχαν συμπεριληφθεί σε προηγούμενες αναζητήσεις.

Από τα παραπάνω προκύπτει ότι ο φορμαλισμός του στόχου διατυπώνεται ως εξής:

Να βρεθεί το σύνολο των ερωτημάτων $q_1, \ldots, q_n$ που μεγιστοποιούν τη συνάρτηση $P(q_1 \vee \ldots \vee q_{i-1} \vee q_{i-1})$, σύμφωνα με τον περιορισμό:

$$\sum_{i=1}^{n} \mathrm{Cost}(q_i) \leq t$$

όπου t είναι το μέγιστο ποσό πόρων που μπορεί να χρησιμοποιήσει ο ιστοσυλλέκτης.

3.1.2 Επιλογή λέξεων-κλειδιών

Δοθέντος ότι ο στόχος είναι να ληφθούν όσο το δυνατόν περισσότερα μοναδικά έγγραφα από μία βάση δεδομένων κειμένου, αυτό επιτυγχάνεται μέσω κάποιας απ' τις τρεις ακόλουθες επιλογές.

- Τυχαία (Random): Επιλέγονται τυχαίες λέξεις-κλειδιά για να χρησιμοποιηθούν στη βάση δεδομένων πιστεύοντας ότι ένα τυχαίο ερώτημα θα επιστρέψει ένα λογικό αριθμό εγγράφων που πληρούν τα κριτήρια αναζήτησης.
- Γενική Συχνότητα (Generic-frequency): Αναλύεται μια γενική συλλογή εγγράφων που συλλέγεται από άλλες πηγές (για παράδειγμα από τον ιστό) και βρίσκεται η γενική συχνότητα κατανομής της κάθε λέξης-κλειδί. Με βάση τη γενική κατανομή, ξεκινάμε με τη λέξη-κλειδί με τη μεγαλύτερη συχνότητα εμφάνισης και την υποβάλλουμε ως ερώτημα στη βάση δεδομένων του κρυμμένου ιστού για να εξαχθεί το αποτέλεσμα. Η διαδικασία συνεχίζεται με όλες τις λέξεις-κλειδιά σύμφωνα με τη μεγαλύτερη συχνότητα εμφάνισης. Ακολουθείται η ίδια διαδικασία μέχρι να εξαντληθούν οι πόροι που έχει στη διάθεσή του ο ιστοσυλλέκτης.

- Προσαρμοστική (Adaptive): Αναλύονται τα έγγραφα που επεστράφησαν από προηγούμενα ερωτήματα στη βάση δεδομένων του κρυμμένου ιστού και εκτιμώνται οι λέξεις-κλειδιά που είναι πιθανότερο να επιστρέψουν το μεγαλύτερο πλήθος εγγράφων. Σύμφωνα με την ανάλυση αυτή βρίσκεται το καλύτερο ερώτημα και η διαδικασία επαναλαμβάνεται.

3.1.2.1 Υπολογισμός του αριθμού των σελίδων που θα επιστραφούν

Για να βρεθεί το καλύτερο ερώτημα είναι αναγκαίο να γίνει μια εκτίμηση για το πόσα νέα έγγραφα θα ληφθούν εάν τεθεί το ερώτημα q_i ως επόμενο ερώτημα. Θεωρείται ότι τέθηκαν τα ερωτήματα q_i, ... q_{i-1} και τώρα πρέπει να γίνει εκτίμηση του $P(q_1 \vee \ldots \vee q_{i-1} \vee q_{i-1})$ για κάθε επόμενο ερώτημα q_i και να επιλεχθεί το q_i που δίνει τη μέγιστη τιμή. Για την εκτίμηση του $P(q_i)$ μπορούν υπάρξουν διάφοροι τρόποι όπως για παράδειγμα αυτοί που ακολουθούν:

i. Εκτιμητής ανεξαρτησίας (*Independence estimator*): Χρησιμοποιεί τον τύπο $P(q_i) = P(q_i \mid q_1 \vee \ldots \vee q_{i-1} \vee q_{i-1})$, θεωρώντας ότι η εμφάνιση του όρου q_i, είναι ανεξάρτητη από τους όρους q_1, …, q_{i-1}.
ii. Εκτιμητής Zipf (*Zipf estimator*): Υπολογίζει τη συχνότητα με την οποία εμφανίζεται ένας όρος μέσα σε μια συλλογή κειμένων η οποία προκύπτει από την εξίσωση:
$f = \alpha(r + \beta)^{-\gamma}$
όπου:

- r: είναι η κατάταξη του όρου βάση της συχνότητας, δηλαδή τη πρώτη θέση θα την έχει ο όρος με την μεγαλύτερη συχνότητα.
- α, β και γ: είναι σταθερές εξαρτώμενες από τη συλλογή κειμένων.

3.1.2.2 Αλγόριθμος επιλογής ερωτήματος

Στόχος του ιστοσυλλέκτη κρυμμένου ιστού είναι να λαμβάνει το μέγιστο αριθμό μοναδικών εγγράφων συμμορφούμενος με τον περιορισμό στους διαθέσιμους πόρους. Σύμφωνα με αυτό ο ιστοσυλλέκτης πρέπει να λάβει υπόψη του δύο συντελεστές:

1. Τον αριθμό των νέων εγγράφων που θα ανακτηθούν από το ερώτημα q_i,
2. Το κόστος από τη χρησιμοποίηση του ερωτήματος q_i.

Έτσι αν δύο ερωτήματα επιστρέφουν το ίδιο πλήθος νέων εγγράφων και το ένα έχει μεγαλύτερο κόστος απ' το άλλο, τότε επιλέγεται το ερώτημα με το μικρότερο κόστος.

Σύμφωνα με αυτό ο ιστοσυλλέκτης του κρυμμένου ιστού ποσοτικοποιεί την απόδοση του ερωτήματος και στη συνέχεια επιλέγει το κατάλληλο ερώτημα q_i με βάση την ακόλουθη μετρική αποδοτικότητας:

$$Efficiency(q_i) = \frac{P_{new}(q_i)}{Cost(q_i)}$$

όπου:

- $P_{new}(q_i)$: το πλήθος των νέων εγγράφων που επιστρέφονται χρησιμοποιώντας το ερώτημα q_i,
- $Cost(q_i)$: το κόστος που προκύπτει από τη χρήση του ερωτήματος q_i.

Ο αλγόριθμος επιλογής επόμενου ερωτήματος (Εικόνα 8) επιστρέφει το πλήθος των νέων εγγράφων που επιστρέφονται για κάθε μονάδα κόστους και μπορεί να χρησιμοποιηθεί για να υποδείξει το πόσο καλά δαπανώνται οι πόροι όταν τίθεται το ερώτημα q_i. Σύμφωνα με τον αλγόριθμο, ο ιστοσυλλέκτης υπολογίζει το βαθμό αποδοτικότητας για κάθε ερώτημα q_i και επιλέγει αυτό με τη μεγαλύτερη τιμή με αποτέλεσμα ο ιστοσυλλέκτης να λάβει τον μέγιστο αριθμό νέων εγγράφων και παράλληλα να προσπαθεί να μεγιστοποιήσει το όφελος του σε κάθε βήμα.

Parameters:
T: The list of potential query keywords
Procedure
(1) Foreach t_k in T do
(2) Estimate Efficiency$(t_k) = \frac{P_{new}(t_k)}{Cost(t_k)}$
(3) done
(4) return t_k with maximum Efficiency(t_k)

Εικόνα 8: Αλγόριθμος επιλογής επόμενου ερωτήματος

3.1.2.3 Βελτιστοποιημένη μέθοδος μέτρησης απόδοσης των ερωτημάτων

Ο αλγόριθμος επιλογής επόμενου ερωτήματος, πρέπει να υπολογίζει την αποδοτικότητα των ερωτημάτων για κάθε πιθανό ερώτημα q_i. Αυτό είναι ιδιαίτερα δαπανηρό αν γίνεται εξ αρχής για κάθε q_i σε κάθε επανάληψη του αλγόριθμου. Η διαδικασία εκτίμησης του $P(q_i|q_1 \vee \ldots \vee q_{i-1})$ μπορεί να επιταχυνθεί σημαντικά, χρησιμοποιώντας ένα πίνακα στατικών ερωτημάτων (*query statistics table).* Στον πίνακα αυτόν καταχωρείται το πόσες φορές εμφανίζεται η λέξη-κλειδί q_i στα έγγραφα από τα αντίστοιχα ερωτήματα $q_1, \ldots, q_{i-1}$.

Ένα παράδειγμα για την κατανόηση της βελτιστοποιημένης μεθόδου μέτρησης απόδοσης των ερωτημάτων είναι αν λάβει 50 έγγραφα στα οποία ο όρος "model" εμφανίζεται 10 φορές. Σύμφωνα μ' αυτά τα δεδομένα υπολογίζεται $P(\text{model} \mid q_1 \vee \ldots \vee q_{i-1}) = 10/50 = 0.2$ (Πίνακας 2).

Term t_k	$N(t_k)$
model	10
computer	38
digital	50
Total pages: 50	

Πίνακας 2: Πίνακας στατιστικών ερωτημάτων έπειτα από τα ερωτήματα $q_1, \ldots, q_{i-1}$

Ας υποθέσουμε ότι υποβάλλουμε ως επόμενο ερώτημα τον όρο «computer» και λαμβάνουμε 20 νέες σελίδες, όπου οι 12 περιέχουν τον όρο «model» και οι 18 τον όρο «disk» (Πίνακας 3).

Term t_k	$N(t_k)$
model	12
computer	20
disk	18
New pages: 20	

Πίνακας 3: Πίνακας στατιστικών ερωτημάτων για το νέο ερώτημα q_i = computer

Ακολούθως, ο Πίνακας 2 ενημερώνεται προσθέτοντας απλά σε αυτόν τα δεδομένα που απεικονίζει ο Πίνακας 3, οπότε προκύπτει ο Πίνακας 4.

Term t_k	$N(t_k)$
model	10+12 = 22
computer	38+20 = 58
disk	0+18 = 18
digital	50+0 = 50
Total pages: 50 + 20 = 70	

Πίνακας 4: Αποτελέσματα έπειτα από q1, …, qi ερωτήματα

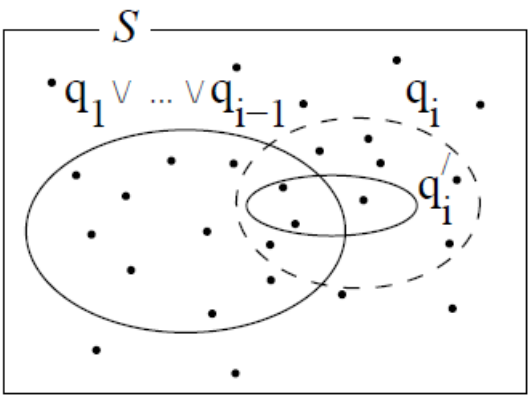
Σύμφωνα με τα δεδομένα που απεικονίζει ο Πίνακας 4, ο όρος «model» εμφανίζεται σε 22 έγγραφα από τις 70 συνολικά σελίδες άρα $P(\text{model} \mid q_1 \vee \ldots \vee q_{i-1}) = 22/70 = 0.3$. Με αυτό το τρόπο μπορούν να υπολογιστούν όλοι οι όροι.

3.1.2.4 Δικτυακοί τόποι που περιορίζουν τον αριθμό των αποτελεσμάτων

Όταν τίθεται ένα ερώτημα που επιστρέφει πολλά αποτελέσματα (μεγάλο αριθμό σελίδων) ορισμένοι δικτυακοί τόποι του κρυμμένου ιστού επιτρέπουν την ανάκτηση μόνο μέρους των σελίδων αυτών. Ένα χαρακτηριστικό παράδειγμα του περιορισμού αυτού είναι η ιστοσελίδα Open Directory Project[6] η οποία επιστρέφει έως 10.000 αποτελέσματα για κάθε αναζήτηση.

Όπως διαφαίνεται, αυτού του είδους οι περιορισμοί επιδρούν άμεσα στον ιστοσυλλέκτη του κρυμμένου ιστού. Ο τρόπος επιλογής ερωτημάτων που παρουσιάστηκε παραπάνω υποθέτει ότι για κάθε πιθανό ερώτημα q_i μπορεί να βρει το $P(q_i|q_1 \vee \ldots \vee q_{i-1})$. Για να βρεθεί το ποσοστό των εγγράφων σε ολόκληρη τη βάση δεδομένων κειμένου χρησιμοποιείται η μέθοδος που προαναφέρθηκε, δηλαδή ότι για κάθε ερώτημα q_i βρίσκεται το $P(q_i|q_1 \vee \ldots \vee q_{i-1})$. Σε κάθε περίπτωση όμως όπου επιστρέφεται μόνο ένα μέρος αποτελεσμάτων για κάθε ερώτημα, τότε η τιμή του P δεν μπορεί να είναι ακριβής με αποτέλεσμα να επηρεάζεται η αποδοτικότητα του ιστοσυλλέκτη.

[6] http://www.dmoz.org/



Εικόνα 9: Δικτυακός τόπος με περιορισμούς στα αποτελέσματα που επιστρέφει

Ο περιορισμός όμως που παρουσιάζουν ορισμένοι δικτυακοί τόποι κρυμμένου ιστού μπορεί να παρακαμφθεί και να εκτιμηθεί ορθός η τιμή του P. Στην Εικόνα 9, ο προς εξέταση δικτυακός τόπος αναπαρίσταται από ένα ορθογώνιο και οι σελίδες από κουκίδες. Θεωρείται ότι τέθηκαν ερωτήματα $q_1, \dots, q_{i-1}$ τα οποία όμως επέστρεψαν μικρότερο αριθμό αποτελεσμάτων από το μέγιστο επιτρεπτό του δικτυακού τόπου (αριστερός κύκλος). Επίσης τίθεται το ερώτημα q_i αλλά λόγω περιορισμού αποτελεσμάτων επιστρέφεται το σύνολο q_i' (δεξιά έλλειψη με συνεχή γραμμή) αντί του συνόλου q_i (έλλειψη με διάστικτη γραμμή).

Εδώ λοιπόν, ενημερώνεται ο στατιστικός πίνακας ερωτήματος με ακριβείς πληροφορίες για το επόμενο βήμα. Έτσι για κάθε πιθανό ερώτημα q_{i+1} βρίσκεται $P(q_{i+1}|q_1 \vee \dots \vee q_i)$ από την παρακάτω εξίσωση:

$$P(q_{i+1}|q_1 \vee \dots \vee q_i) = \frac{1}{P(q_1 \vee \dots \vee q_i)} \cdot [P(q_{i+1} \wedge (q_1 \vee \dots \vee q_{i-1})) + P(q_{i+1} \wedge q_i) - P(q_{i+1} \wedge q_i \wedge (q_1 \vee \dots \vee q_{i-1}))]$$

Στην ανωτέρω εξίσωση, το $P(q_1 \vee \dots \vee q_i)$ υπολογίζεται βάση της εξίσωσης για το $P(q_i)$, η οποία παρουσιάστηκε προηγουμένως. Αντίστοιχα, τα $P(q_{i+1} \wedge (q_1 \vee \dots \vee q_{i-1}))$ και $P(q_{i+1} \wedge q_i \wedge (q_1 \vee \dots \vee q_{i-1}))$, υπολογίζονται με τον άμεσο έλεγχο των εγγράφων που έχουν ληφθεί από τα ερωτήματα $q_1 \dots q_{i-1}$. Για τον υπολογισμό του $P(q_{i+1} \wedge q_i)$, θεωρείται το q_i' ως ένα τυχαίο δείγμα του q_i και δίνεται η εξίσωση:

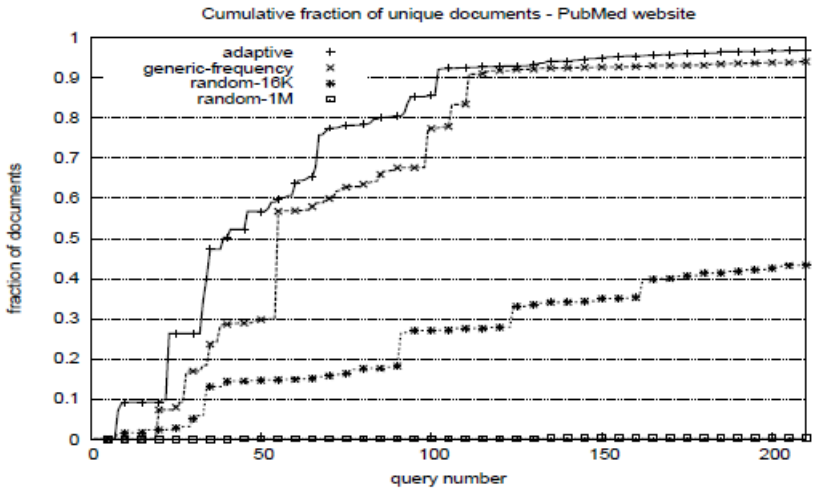
$$\frac{P(q_{i+1} \wedge q_i)}{P(q_{i+1} \wedge q_i')} = \frac{P(q_i)}{P(q_i')}$$

Από την ανωτέρω εξίσωση, προκύπτει $P(q_{i+1} \wedge q_i)$, αφού κάνουμε αντικατάσταση της τιμής της αρχικής εξίσωσης βρίσκουμε $P(q_{i+1}|q_1 \vee \dots \vee q_i)$.

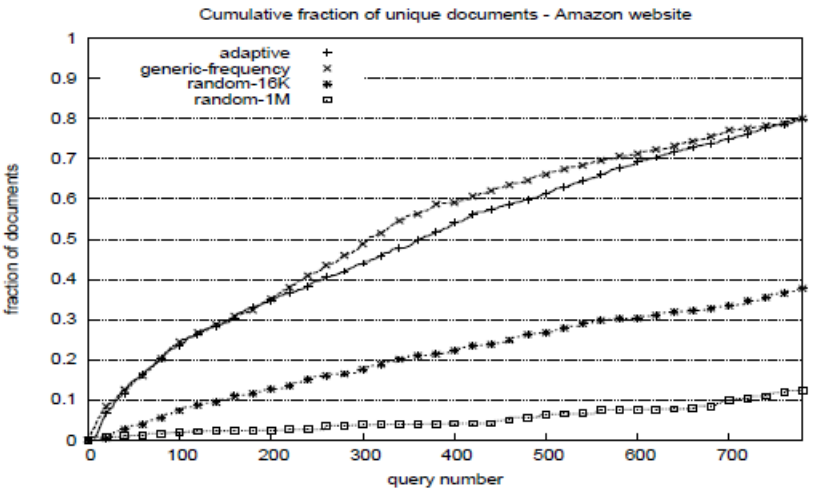
3.1.3 Πειραματική αξιολόγηση μεθοδολογίας

Ένα από τα κύρια θέματα είναι η εξέλιξη του μέτρου κάλυψης κατά την υποβολή ερωτημάτων σε έναν δικτυακό τόπο. Το κύριο ενδιαφέρον είναι η αναλογία των

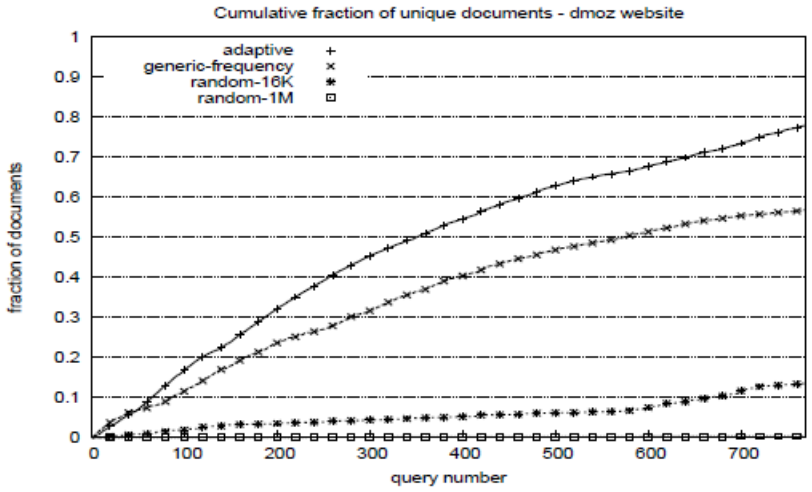
εγγράφων που είναι αποθηκευμένα στον κρυμμένο ιστό και τα οποία μπορούμε να λάβουμε, κάνοντας παράλληλα αναζήτηση για νέες λέξεις-κλειδιά, με βάση τις τρεις μεθοδολογίες που εξετάστηκαν στην Ενότητα 3.1.2. Σύμφωνα με τα παραπάνω εξετάζεται το $P(q_1 \vee \ldots \vee q_i)$, καθώς αυξάνεται το i. Στα γραφήματα των ακόλουθων Εικόνων Εικόνα 10, Εικόνα 11 και Εικόνα 12 απεικονίζεται η μετρική κάλυψης για κάθε μια από τις τρεις μεθοδολογίες εύρεσης λέξεων-κλειδιών, ως προς τον αριθμό των ερωτημάτων που τίθενται στους ιστοχώρους. Συγκεκριμένα για τους ιστοχώρους της PubMed (Εικόνα 10), της Amazon (Εικόνα 11) και της dmoz (Εικόνα 12):



Εικόνα 10: Κάλυψη μεθοδολογιών για τη PubMed



Εικόνα 11: Κάλυψη μεθοδολογιών για την Amazon

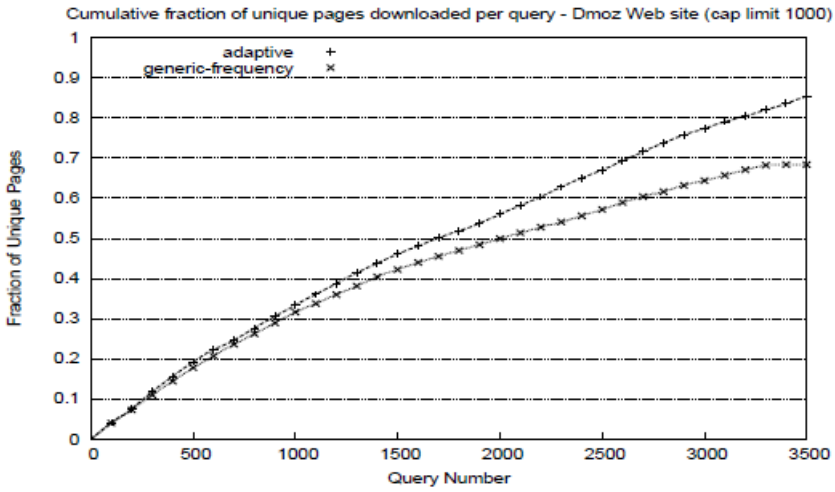


Εικόνα 12: Κάλυψη μεθοδολογιών για την dmoz

Με βάση τα ανωτέρω γραφήματα η μεθοδολογία *γενικής συχνότητας* (generic-frequency) και *προσαρμοστική* (Adaptive) λειτουργούν αποδοτικότερα από τον αλγόριθμο *τυχαίας επιλογής* (Random).

Η αξιολόγηση των αλγορίθμων *γενικής συχνότητας* και *προσαρμοστικής*, δίνουν ποικίλα αποτελέσματα., ανάλογα με τους ιστοχώρους που εφαρμόζονται. Παρατηρώντας τους ιστοχώρους της Pubmed και της dmoz ο αλγόριθμος *γενικής συχνότητας* είναι αποδοτικότερος, ενώ για τον ιστοχώρο της Amazon ο προσαρμοστικός αλγόριθμος εμφανίζεται αποδοτικότερος. Συμπερασματικά ο προσαρμοστικός αλγόριθμος αποδίδει καλύτερα όταν ο ιστοχώρος πραγματεύεται συγκεκριμένο θέμα. Παράδειγμα και πειραματική απόδειξη της τελευταίας πρότασης αποτελεί το γεγονός ότι ο προσαρμοστικός αλγόριθμος απαιτεί 83 ερωτήματα για να λάβει περίπου το 80% των εγγράφων που είναι αποθηκευμένα στον δικτυακό τόπο της Pubmed, ενώ ο αλγόριθμος γενικής συχνότητας απαιτεί 106 ερωτήματα για να καλύψει το ίδιο ποσοστό.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η αποδοτικότητα των δύο επικρατέστερων αλγορίθμων, σε συνάρτηση με τον περιορισμό που έχουν οι δικτυακοί τόποι στον αριθμό των αποτελεσμάτων που επιστρέφουν. Για το λόγο αυτό διερευνήθηκε η κάλυψη που επιφέρουν οι δύο προαναφερθέντες αλγόριθμοι κατά την ιστοσυλλογή από τον ιστοχώρο της dmoz, θέτοντας ως περιορισμό 1.000 αποτελέσματα για κάθε ερώτημα (Εικόνα 13).



Εικόνα 13: Κάλυψη μεθοδολογιών για την dmoz με περιορισμό αποτελεσμάτων

Με βάση την Εικόνα 12 όπου ο περιορισμός των αποτελεσμάτων ανά ερώτημα ήταν στις 10.000, στην Εικόνα 13 ο περιορισμός αυτός κατέρχεται στα 1.000. Το αποτέλεσμα όσον αφορά το ποιος είναι πιο αποδοτικός αλγόριθμος μπορεί να μην αλλάζει, ωστόσο μια χρήσιμη παρατήρηση που εντοπίζεται, είναι ότι τόσο ο προσαρμοστικός αλγόριθμος όσο και ο αλγόριθμος γενικής συχνότητας, πρέπει να υποβάλουν μεγαλύτερο αριθμό ερωτημάτων για να επιτύχουν το ίδιο ποσοστό κάλυψης με την περίπτωση άνευ περιορισμού.

3.1.4 Σύνοψη μεθοδολογίας

Συμπερασματικά, στην παρούσα υποενότητα παρουσιάστηκε ο τρόπος κατασκευής ενός ιστοσυλλέκτη κρυμμένου ιστού, ο οποίος μπορεί αυτόματα να θέτει ερωτήματα σε έναν δικτυακό τόπο του κρυμμένου ιστού και να λαμβάνει σελίδες από αυτόν. Σύμφωνα με αποτελέσματα πειραμάτων που παρατίθενται στην εργασία [1], σε πραγματικούς δικτυακούς τόπους του κρυμμένου ιστού διαφάνηκε η αποδοτικότητα της παραπάνω μεθοδολογίας. Στη μεθοδολογία αυτή υπήρξαν πειράματα που επέτρεψαν τη λήψη περίπου του 90% των εγγράφων από δικτυακό τόπο του κρυμμένου ιστού ο οποίος περιέχει 14.000.000 έγγραφα, έπειτα από περίπου 100 ερωτήματα.

3.2 Ο αλγόριθμος της Google για ιστοσυλλογή του κρυμμένου ιστού

Στην ενότητα αυτή παρουσιάζεται ένα σύστημα περιήγησης στο περιεχόμενο του κρυμμένου ιστού, το οποίο σχετίζεται με τον προ-υπολογισμό υποβολών ερωτημάτων για κάθε φόρμα HTML και τη μετέπειτα προσθήκη των προκυπτουσών σελίδων HTML στα ευρετήρια μιας μηχανής αναζήτησης. Στη συνέχεια παρουσιάζεται ένας αλγόριθμος για την επιλογή έγκυρων τιμών εισόδου σε πλαίσια κειμένου τα οποία παίρνουν ως είσοδο λέξεις-κλειδιά. Ακόμη παρουσιάζεται ένας αλγόριθμος για την αναγνώριση τιμών εισόδου όπου σε πλαίσια κειμένου όπου οι αποδεκτές τιμές είναι συγκεκριμένου τύπου και τέλος ένας αλγόριθμος όπου προσδιορίζει τους

συνδυασμούς εισόδων σε διαφορετικά στοιχεία που είναι πιθανό να παράγουν περισσότερα URLs για πρόσθεση στα ευρετήρια της μηχανής αναζήτησης.

Στη συνέχεια παρουσιάζεται μία μεθοδολογία για την ανάδειξη του περιεχομένου στον κρυμμένο ιστό με τα εξής καινοτόμα χαρακτηριστικά:

1. Δοκιμές εκτίμησης επιπέδου πληροφόρησης (informativeness tests) όπου αξιολογούνται *τα πρότυπα ερωτήσεων*, δηλ. οι συνδυασμοί που δίνονται ως είσοδοι στις φόρμες αναζήτησης. Σε κάθε πρότυπο ερώτησης εισάγονται διαφορετικά σύνολα ως είσοδοι στη φόρμα και στη συνέχεια εξετάζεται το κατά πόσο οι σελίδες HTML που παράγονται είναι αρκετά διαφορετικές μεταξύ τους. Τα πρότυπα που παράγουν διαφορετικές σελίδες θεωρούνται καλοί υποψήφιοι ανάδειξης του κρυμμένου ιστού.

2. Περιλαμβάνεται ένας αλγόριθμος που διατρέχει τα πρότυπα ερωτήσεων και αναγνωρίζει τα κατάλληλα για ανάδειξη/εμφάνιση του κρυμμένου ιστού. Ο αλγόριθμος προσπαθεί να αντισταθμίσει από τη μία πλευρά την παραγωγή λίγων URLs (και υποβολή αντίστοιχων ερωτημάτων), και από την άλλη πλευρά την καλύτερη κάλυψη του περιεχομένου.

3.2.1 Επεξεργασία φορμών HTML

Σε ένα έγγραφο HTML μπορεί να δημιουργηθεί μία φόρμα που είναι χρήσιμη για την εισαγωγή δεδομένων και την αποστολή αυτών των δεδομένων στον εξυπηρέτη (server) προς επεξεργασία. Μια φόρμα προσδιορίζεται από την ετικέτα "Form" η οποία μπορεί να περιέχει στοιχεία εισόδου όπως είναι τα πεδία κειμένου, στοιχεία επιλογής (check boxes και radio buttons), κουμπιά υποβολής κ.ά.

```
<form action="http://jobs.com/find" method="get">
  <input type="hidden" name="src" value="hp">
  Keywords:  <input type="text" name="kw">
  State:  <select name="st"> <option value="Any"/>
    <option value="AK"/> ...  </select>
  Sort By:  <select name="sort"> <option value="salary"/>
    <option value="startdate"/> ...  </select>
  <input type="submit" name="s" value="go">
</form>
```

Εικόνα 14: Φόρμα HTML για την αναζήτηση εργασίας με βάση λέξεις-κλειδιά

Στην Εικόνα 14 φαίνεται μία φόρμα HTML για την αναζήτηση εργασίας με βάση λέξεις-κλειδιά. Με τη μέθοδο "get" που χρησιμοποιείται, τα δεδομένα προστίθενται στο τέλος του URL που ορίζει η ιδιότητα "action" και χωρίζονται από αυτό με τον χαρακτήρα "&". Με την ετικέτα "input" ορίζονται τα περισσότερα στοιχεία της φόρμας. Οι κυριότερες ιδιότητες της ετικέτας είναι η "type" η οποία καθορίζει το τύπο του στοιχείου της φόρμας (πεδίο κειμένου ή περιοχή κειμένου ή κουμπί

επιλογών ή κουτί πολλαπλών επιλογών ή κουμπί). Η ιδιότητα "name" ορίζει το όνομα του στοιχείου της φόρμας, το οποίο πρέπει να είναι μοναδικό. Τέλος, με την ιδιότητα "value" δίνεται μια αρχική τιμή στο στοιχείο της φόρμας.

Όταν υποβάλλεται μία φόρμα η εφαρμογή πλοήγησης αποστέλλει στον εξυπηρέτη μια αίτηση HTTP, η οποία περιλαμβάνει τις εισαχθείσες τιμές. Στην περίπτωση χρήσης της μεθόδου "get" οι παράμετροι ενσωματώνονται στο URL της αίτησης HTTP όπως για παράδειγμα http://jobs.com/find?src=hp&kw=chef&st=Any&sort=salary&s=go. Αντίθετα, με τη μέθοδο "post", οι παράμετροι αποστέλλονται στο σώμα της αίτησης HTTP και το URL ορίζει μόνο τη σελίδα που θέλουμε να προσπελάσουμε (εν προκειμένω http://jobs.com/find). Έτσι διαπιστώνεται πως τα URLs που διαμορφώνονται με τη μέθοδο "get" ορίζουν μονοσήμαντα τα δεδομένα, ενώ αυτά που διαμορφώνονται με τη μέθοδο "post" δεν έχουν αυτή την ιδιότητα.

Με δεδομένο ότι οι μηχανές αναζήτησης προσδιορίζουν τις σελίδες με βάση το URL τους, η μεθοδολογία που αναλύεται δίνει έμφαση στη μέθοδο "get" προκειμένου να λαμβάνεται περιεχόμενο που μπορεί να καταχωρηθεί στα σχετικά ευρετήρια των μηχανών αναζήτησης. Ταυτόχρονα, παραλείπονται τα στοιχεία που παραπέμπουν σε προσωπικά δεδομένα (είσοδοι τύπου password, πεδία με όνομα username, login κ.ο.κ.).

3.2.2 Το πρόβλημα της ανάδειξης

Το πρόβλημα ανάδειξης του κρυμμένου ιστού ορίζεται ως η επιλογή ενός συνόλου ερωτήσεων για υποβολή στη φόρμα, για την εισαγωγή ερωτήσεων στη φόρμα υπάρχουν δύο τύπου εισόδων:

- Πρώτον οι *είσοδοι επιλογής* (selection inputs), οι οποίες ορίζουν το περιεχόμενο που επιθυμεί ο χρήστης. Οι είσοδοι αυτές λαμβάνουν τις τιμές τους είτε από μια λίστα επιλογών είτε από μία περιοχή κειμένου.

- Δεύτερον οι *είσοδοι παρουσίασης* (presentation inputs), οι οποίες ελέγχουν μόνο τις τιμές αποτελεσμάτων, όπως είναι η σειρά ταξινόμησης.

Τα πρότυπα ερωτήσεων χρησιμοποιούνται σε συλλογές υποβολών, και πιο συγκεκριμένα για να αντιμετωπιστεί το πρόβλημα της επιλογής ενός καλού συνόλου υποβολών από φόρμες. Ένα πρότυπο ερωτήσεων ορίζει ένα υποσύνολο από εισόδους μιας φόρμας του διαδικτύου ως "*δεσμευτικές εισόδους*" και τις υπόλοιπες ως "*ελεύθερες εισόδους*". Οι πολλαπλές υποβολές φορμών μπορούν να παραχθούν με την ανάθεση διαφορετικών τιμών στις δεσμευτικές εισόδους. Σε αναλογία με τις ερωτήσεις SQL, το πρότυπο ερωτήσεων αντιπροσωπεύει όλα τα ερωτήματα της μορφής

select * from D where P_B

όπου το P_B περιλαμβάνει μόνο τις επιλογές που επιβάλλονται από τις δεσμευτικές εισόδους της φόρμας. Ο αριθμός των δεσμευτικών εισόδων ισούται με τη *διάσταση* ενός προτύπου.

Το πρόβλημα διασπάται σε δύο μικρότερα προβλήματα:

1. Επιλογή κατάλληλου συνόλου προτύπων ερωτήσεων,
2. Επιλογή κατάλληλων τιμών για τις παραμέτρους του προτύπου ερωτήσεων. Για παράδειγμα, αν ένα πεδίο είναι λίστα επιλογών τότε χρησιμοποιούνται όλες οι επιλογές, ενώ για ένα πεδίο κειμένου πρέπει να εκτιμηθούν οι τιμές εισόδου του.

Ο τελικός στόχος είναι η επιλογή κατάλληλων ερωτήσεων από πολλές διαφορετικές φόρμες ώστε να επιτευχθεί καλή κάλυψη από ένα μικρό αριθμό υποβολών ανά ιστοσελίδα με αποτέλεσμα οι σελίδες που θα εμφανίζονται να είναι καλοί υποψήφιοι για την εισαγωγή τους στο πίνακα της μηχανής αναζήτησης.

3.2.3 Επιλογή προτύπων ερωτήσεων

Η ακριβής επιλογή των διαστάσεων των προτύπων ερωτήσεων, εξαρτάται από το είδος της κάθε βάσης δεδομένων. Οι πολύ μεγάλες βάσεις δεδομένων είναι πιθανό να έχουν βέλτιστα πρότυπα με περισσότερες εισόδους, ενώ οι μικρότερες βάσεις δεδομένων πιθανό να έχουν πρότυπα ερωτήσεων με λιγότερες εισόδους.

Τα πρότυπα ερωτήσεων αξιολογούνται με βάση τη διακριτότητα (distinctness) των σελίδων που παράγονται από τις αντίστοιχες υποβολές φορμών. Τα σύνολα σελίδων που παράγονται από το πρότυπο ερωτήσεων ομαδοποιούνται με βάση την ομοιότητα του περιεχομένου τους. Ένα πρότυπο ερωτήσεων καλείται *πληροφοριακό* (informative) όταν τα σύνολα σελίδων που παράγονται περιέχουν επαρκώς διακριτά μέλη ενώ καλείται *μη πληροφοριακό* όταν τα σύνολα σελίδων που παράγονται δεν περιέχουν επαρκώς διακριτά μέλη.

Το σύστημα που περιγράφεται στην ενότητα 3.2 έχει *n* εισόδους, όπου κάθε μία μπορεί να χρησιμοποιηθεί ή όχι, κατά συνέπεια τα διαφορετικά πρότυπα ερωτήσεων που μπορούν να δημιουργηθούν είναι 2^n-1 (δεν λογίζεται το πρότυπο που δεν χρησιμοποιεί καμία είσοδο). Προκειμένου να μην δοκιμασθούν και τα 2^n-1 πρότυπα ερωτήσεων, πρέπει να ακολουθηθεί μια στρατηγική που θα εξετάζει τον χώρο των υποψήφιων προτύπων και θα εξετάζει μόνο όσα είναι πληροφοριακά. Η στρατηγική που ακολουθούμε είναι να ξεκινάμε από τα πρότυπα ερωτήσεων που έχουν μία δεσμευτική είσοδο και στη συνέχεια να συνεχίζουν με τα πρότυπα ερωτήσεων που έχουν περισσότερες εισόδους. Η διαισθητική εξήγηση αυτής της στρατηγικής είναι ότι η «πληροφοριακότητα» ενός προτύπου σχετίζεται με την «πληροφοριακότητα» του προτύπου που επεκτείνει (δηλαδή που έχει μια επιπλέον δεσμευτική είσοδο). Αν ένα πρότυπο ερωτήσεων έχει διάσταση x και από τα x πρότυπα ερωτήσεων που επεκτείνει κανένα δεν είναι πληροφοριακό, τότε είναι απίθανο το πρότυπο ερωτήσεων να είναι πληροφοριακό.

Αφού τελειώσει η αναζήτηση για τα κατάλληλα πρότυπα ερωτήσεων, προστίθενται τα URLs που παράχθηκαν από τα πληροφοριακά πρότυπα ερωτήσεων στα ευρετήρια μιας μηχανής αναζήτησης. Για την μη ύπαρξη διπλότυπων περιεχομένων προστίθενται μόνο εκείνα που έχουν διαφορετικές υπογραφές.

Η *υπογραφή* αποτελεί το εργαλείο για την επίλυση του προβλήματος της ταυτοποίησης ενός μηνύματος ή του περιεχομένου μιας ιστοσελίδας. Είναι ένα μαθηματικό σύστημα που χρησιμοποιείται για την απόδειξη της γνησιότητας ενός ψηφιακού περιεχομένου. Η *υπογραφή* είναι ουσιαστικά ένα σύνολο από στοιχεία που περιγράφουν και ταυτοποιούν το περιεχόμενο ενός μηνύματος ή μιας ιστοσελίδας, και είναι ανεκτική σε μικρές αλλαγές που γίνονται στο περιεχόμενο.

Οι ιδιότητες μιας υπογραφής είναι οι ακόλουθες: πρώτον η υπογραφή δεν πρέπει να γνωρίζει την μορφοποίηση HTML δεδομένου ότι οι είσοδοι παρουσίασης συχνά αλλάζουν τη διάταξη της ιστοσελίδας. Δεύτερον, η υπογραφή δεν πρέπει να γνωρίζει τη διάταξη των όρων, επειδή η αναδιάταξη τους είναι μια κοινή λειτουργία. Τρίτον, πρέπει να είναι ανεκτική σε μικρές αλλαγές και τροποποιήσεις του περιεχόμενου της ιστοσελίδας και τέλος η υπογραφή δεν πρέπει να περιλαμβάνει τις ίδιες τις τιμές των εισόδων.

3.2.4 Παραγωγή τιμών εισόδου

Πολλές από τις φόρμες HTML έχουν πλαίσια κειμένου, κάποιες από αυτές έχουν λίστα επιλογών (μενού) και απαιτούν να έχουν εισαχθεί έγκυρες τιμές στα πλαίσια κειμένου πριν την ανάκτηση οποιουδήποτε αποτελέσματος.

Υπάρχουν δύο τρόποι χρήσης των πλαισίων κειμένου:

- Ο πρώτος τρόπος είναι οι λέξεις που εισάγονται στα πλαίσια κειμένου να χρησιμοποιούνται για την ανάκτηση όλων των δεδομένων μιας βάσης δεδομένων κειμένου που περιέχουν συγκεκριμένες λέξεις (βιβλία, συγγραφείς).

- Δεύτερος τρόπος είναι το πλαίσιο κειμένου να είναι χρήσιμο ως βοήθημα επιλογής σε ένα συγκεκριμένο γνώρισμα σε μια πρόταση "where" μιας ερώτησης SQL. Το γνώρισμα μπορεί να λαμβάνει διακριτές τιμές, όπως ο ταχυδρομικός κώδικας (Τ.Κ.), ή να είναι ένα στιγμιότυπο ενός συνεχούς τύπου δεδομένων, π.χ. τριπλέτες ακέραιων που αναπαριστούν ημερομηνίες.

Οι δυο παραπάνω τρόποι χρήσης έχουν ως αποτέλεσμα τη διάκριση των πλαισίων κειμένου σε δυο τύπους:

- τα πλαίσια κειμένου ελεύθερης εισόδου,

- τα πλαίσια κειμένου που απαιτούν συγκεκριμένο τύπο δεδομένων εισόδου.

Τα πλαίσια κειμένου με συγκεκριμένο τύπο δεδομένων εισόδου όταν έχουν άκυρες καταχωρήσεις επιστρέφουν ως αποτέλεσμα σελίδες σφάλματος και γι' αυτό είναι σημαντικό να οριστεί σωστά ο τύπος δεδομένων.

3.2.4.1 Πλαίσια κειμένου ελεύθερης εισόδου

Για τα πλαίσια κειμένου ελεύθερης εισόδου υιοθετείται μια επαναληπτική προσέγγιση ώστε να βρεθούν υποψήφιες λέξεις-κλειδιά, από ένα αρχικό σύνολο λέξεων-κλειδιών. Οι λέξεις-κλειδιά τίθενται ως τιμές για το πλαίσιο κειμένου και κατασκευάζεται ένα πρότυπο ερωτήσεων με μόνη είσοδο το πλαίσιο κειμένου. Οι αντίστοιχες υποβολές φορμών επιστρέφουν σελίδες, από τον οποίων τα κείμενα εξάγονται επιπλέον λέξεις-κλειδιά. Οι λέξεις-κλειδιά που εξάγονται χρησιμεύουν στην ενημέρωση του συνόλου των υποψηφίων τιμών του πλαισίου κειμένου. Η διαδικασία αυτή επαναλαμβάνεται έως ότου να μην εξάγονται άλλες λέξεις-κλειδιά ή να επαληθευθεί κάποια άλλη συνθήκη τερματισμού. Κατά τον τερματισμό, ένα υποσύνολο των υποψηφίων λέξεων-κλειδιών επιλέγεται ως το σύνολο τιμών για το πλαίσιο κειμένου.

Στην επαναληπτική προσέγγιση οι τεχνικές που υιοθετούνται είναι:

1. *Επιλογή υποψηφίων λέξεων-κλειδιών*, στην αρχή εφαρμόζεται η δοκιμή για να διαπιστωθεί εάν ένα πρότυπο ερωτήσεων είναι πληροφοριακό στο πρότυπο που αφορά το πλαίσιο κειμένου, λαμβάνοντας υπόψη ένα αρχικό σύνολο τιμών εισόδου. Επιπρόσθετες λέξεις-κλειδιά επιλέγονται από τις λέξεις των ιστοσελίδων που παράγονται από τις υποβολές φορμών. Ακόμη επιλέγονται λέξεις από μία σελίδα με την προϋπόθεση ότι σχετίζονται περισσότερο με το περιεχόμενο της συγκεκριμένης σελίδας. Σε αυτή τη περίπτωση, χρησιμοποιείται η δημοφιλής μετρική ανάκτησης πληροφοριών TF-IDF. Η *συχνότητα όρου* (term frequency, TF) μετράει τη σημασία της λέξης σε μια συγκεκριμένη ιστοσελίδα. Η *αντίστροφη συχνότητα κειμένου* (inverse document frequency, IDF) μετρά τη σημασία της λέξης σε όλες τις πιθανές ιστοσελίδες. Οπότε, η μετρική TF-IDF εξισορροπεί τη σημασία της λέξης σε μια σελίδα με τη γενική σημασία της, με αποτέλεσμα να επιλέγονται οι κορυφαίες *N* λέξεις στη σελίδα φόρμας, με βάση την ταξινόμηση των λέξεων κατά τις μετρικές TF-IDF τους.

2. Ε*πιλογή τιμών για τα πλαίσια κειμένου*. Για να περιοριστεί ο αριθμός των URLs που παράγονται από τη φόρμα χρησιμοποιούνται περιορισμοί που αφορούν το μέγιστο αριθμό λέξεων-κλειδιών για ένα πλαίσιο κειμένου. Θεωρούνται αρχικά οι N επικρατέστερες λέξεις-κλειδιά της σελίδας που αντιστοιχούν σε κάθε υποψήφια λέξη-κλειδί. Ακολούθως οι λέξεις-κλειδιά ομαδοποιούνται και λαμβάνεται τυχαία μια λέξη-κλειδί από κάθε ομάδα. Τελικά ταξινομούνται οι λέξεις-κλειδιά που επιλέχθηκαν με βάση το μήκος της σελίδας που προκύπτει από την αντίστοιχη υποβολή φόρμας και η ταξινομημένη λίστα διατρέχεται μέχρι να βρεθεί ο επιθυμητός αριθμός λέξεων-κλειδιών.

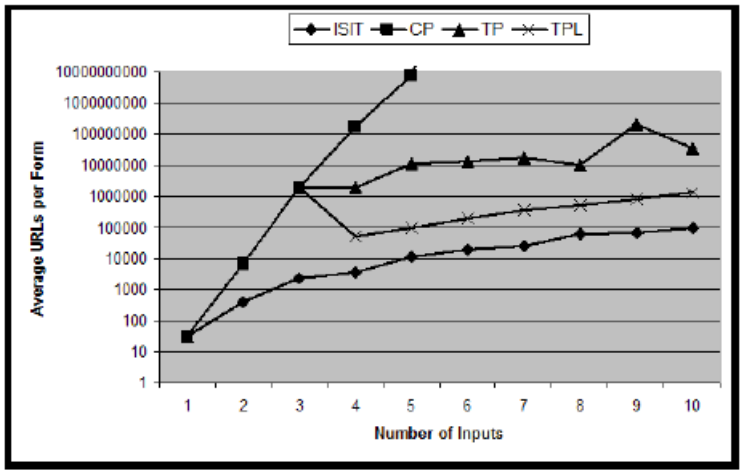
3.2.4.2 *Πλαίσια κειμένου με συγκεκριμένο τύπο δεδομένων εισόδου*

Στην εργασία [7] παρατηρείται ότι υπάρχουν λίγοι τύποι δεδομένων που εάν αναγνωριστούν μπορούν να χρησιμοποιηθούν ως ευρετήρια σε πολλά πεδία ορισμού και συνακόλουθα να εμφανιστούν σε πολλές φόρμες. Για παράδειγμα ο ταχυδρομικός κώδικας χρησιμοποιείται ως είσοδος σε πολλά πεδία ορισμού και η ημερομηνία χρησιμοποιείται ως δείκτης πολλών πεδίων ορισμού.

Στην εργασία [7] και υπό τη θεώρηση ότι μία υποβολή σελίδας που περιέχει πλαίσιο κειμένου με συγκεκριμένο τύπο δεδομένων εισόδου παράγει ως αποτέλεσμα σελίδες που έχουν πληροφοριακή αξία όταν στο πλαίσιο έχει εισαχθεί κατάλληλη ως προς τον τύπο τιμή, πραγματοποιήθηκαν δοκιμές εκτίμησης επιπέδου πληροφόρησης χρησιμοποιώντας γνωστές τιμές για δημοφιλής τύπους δεδομένων εισόδου. Θεωρήθηκαν πεπερασμένοι (π.χ. ταχυδρομικοί κώδικες), όπου οι τιμές λαμβάνονταν με δειγματοληψία και συνεχείς τύποι δεδομένων εισόδου όπου χρησιμοποιήθηκαν σύνολα από ομοιόμορφα κατανεμημένες τιμές που αντιστοιχούσαν σε διαφορετικά δεδομένα.

3.2.5 Πειραματικά αξιολόγηση

Στην πειραματική αξιολόγηση, συγκρίνεται ο αλγόριθμος *ISIT* με άλλους αλγόριθμους που έχουν δημιουργηθεί για τον ίδιο σκοπό. Στην Εικόνα 15, παρουσιάζεται η σύγκριση μεταξύ του αλγόριθμου *ISIT* και τεσσάρων άλλων αλγορίθμων ως προς την αποδοτικότητα του καθενός με διαφορετικές τιμές εισόδου στις φόρμες. Σημειώνεται ότι οι αλγόριθμοι θεωρούν το ίδιο σύνολο υποψήφιων δεσμευτικών τιμών εισόδου για κάθε φόρμα (για παράδειγμα, τα φίλτρα που επιλέχθηκαν από τον αλγόριθμο *ISIT* και χρησιμοποιούνται για να περιορίσουν τις δεσμευτικές τιμές εισόδου, εφαρμόζονται εξίσου και στους άλλους αλγόριθμους).



Εικόνα 15: Σύγκριση του μέσου αριθμού URLs ανά φόρμα για τους τέσσερεις αλγόριθμους

Με βάση το παραπάνω γράφημα της Εικόνα 15, όλοι οι αλγόριθμοι παρουσιάζουν την ίδια αποδοτικότητα για μια μόνο είσοδο (*n=1*) ενώ για n=3 όλοι οι αλγόριθμοι (εκτός του ISIT) παράγουν τον ίδιο αριθμό URLs. Αυξάνοντας τον αριθμό των τιμών

εισόδου για n>3, διαφαίνεται ότι η διαφορά της απόδοσης μεταξύ των αλγορίθμων παραμένει σχετικώς σταθερή.

Ο πλέον μη πρακτικός αλγόριθμος είναι ο CP, ο οποίος παράγει περισσότερα από 100.000.000 URLs για n=4. Για τον ίδιο αριθμό τιμών εισόδου (n=4), οι αλγόριθμοι TP και TPL παράγουν λιγότερα URLs εν συγκρίσει με τον CP ωστόσο αυτά εξακολουθούν να θεωρούνται πολλά, ενώ ο αλγόριθμος ISIT παράγει λιγότερα URLs από όλους τους αλγόριθμους ακόμα για φόρμες με περισσότερες τιμές εισόδου. Ο αλγόριθμος ISIT παράγει για n=10 τον ίδιο αριθμό URLs με τον αλγόριθμο TPL για n=5 (όπου ο TPL είναι ο δεύτερος καλύτερος σε απόδοση αλγόριθμος).

3.3 Εξαγωγή δεδομένων και ανάθεση ετικετών για βάσεις δεδομένων

Στην εργασία [8], παρουσιάζεται το σύστημα DeLa, το οποίο επανακατασκευάζει ένα μέρος μιας βάσης δεδομένων του κρυμμένου ιστού. Για να επιτευχθεί αυτό στέλνονται ερωτήματα μέσω φορμών HTML που αυτόματα παράγουν περιτυλίγματα (wrappers) κανονικών εκφράσεων για την εξαγωγή δεδομένων από ιστοσελίδες. Τα δεδομένα αυτά αποθηκεύονται σε πίνακα με ετικέτες.

Το πρόβλημα που πρέπει να λυθεί είναι η αυτόματη εξαγωγή δεδομένων από μια ιστοσελίδα και στη συνέχεια να ανατεθούν ετικέτες με νόημα στα δεδομένα αυτά. Το πρόβλημα εντοπίζεται στις πολύπλοκες φόρμες HTML για την επερώτηση βάσεων δεδομένων από τους χρήστες και όχι μέσω αναζητήσεων με λέξεις-κλειδιά. Για τη λύση του προβλήματος επιτρέπεται η εξαγωγή δεδομένων από τις ιστοσελίδες αυτές, με στόχο τον εύκολο χειρισμό των δεδομένων και την ανάλυση αυτών περεταίρω. Στην επίλυση του προβλήματος βασική προϋπόθεση είναι η λύση να είναι αυτόματη και γρήγορη.

Τρείς λόγοι κάνουν το πρόβλημα δυσεπίλυτο:

1. Οι φόρμες HTML σχεδιάστηκαν ώστε να τις χειρίζονται άνθρωποι, με αποτέλεσμα να μην αναγνωρίζονται όλα τα στοιχεία τους από τα συστήματα.

2. Το περιτύλιγμα (wrapper) που παράχθηκε για την κάθε ιστοσελίδα πρέπει να είναι πολύπλοκο για να εξάγει όχι μόνο τα εμφανή δεδομένα αλλά και αυτά που βρίσκονται σε εμφωλευμένες δομές.

3. Το περιτύλιγμα (wrapper) στηρίζεται στη δομή ετικετών της HTML, οι οποίες μπορεί να μην αντικατοπτρίζουν την αληθινή δομή της βάσης δεδομένων.

Το σύστημα DeLa βασίζεται σε 2 γενικότερες παρατηρήσεις:

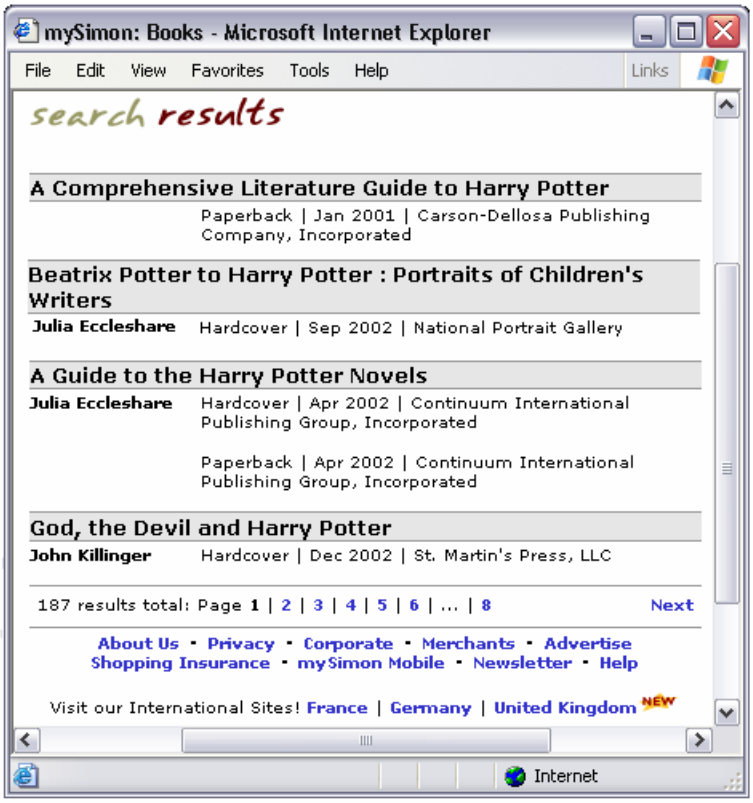
- *Πρώτη παρατήρηση*: τα δεδομένα που περιέχονται στις δυναμικά παραγόμενες ιστοσελίδες διαμοιράζονται μια κοινή δομή από ετικέτες HTML και καταγράφονται συνεχόμενα στις ιστοσελίδες αυτές. Με αυτή τη μεθοδολογία παράχθηκαν αυτόματα περιτυλίγματα (wrappers) κανονικών εκφράσεων για την εξαγωγή των δεδομένων και την ενσωμάτωσή τους σ' έναν πίνακα.

- *Δεύτερη παρατήρηση*: η φόρμα που περιέχεται σε μια ιστοσελίδα, μέσω της οποίας υποβάλλουν οι χρήστες τα ερωτήματά τους, της σκιαγραφεί τη δομή της σχεσιακής βάσης δεδομένων του ιστοχώρου. Με βάση αυτό εξάγονται οι ετικέτες των φορμών HTML και στη συνέχεια γίνεται αντιστοίχιση με τις στήλες του πίνακα δεδομένων, ώσπου να ανατεθούν ετικέτες σε όλα τα χαρακτηριστικά των εξαγόμενων δεδομένων.

3.3.1 Το Σύστημα DeLa

3.3.1.1 Μοντέλο Δεδομένων

Ως παράδειγμα έχουμε μια ιστοσελίδα από ένα ηλεκτρονικό κατάστημα βιβλίων με φόρμα αναζήτησης. Για να αναζητήσουν οι χρήστες στη βάση δεδομένων του βιβλιοπωλείου ένα συγκεκριμένο βιβλίο πληκτρολογούν στη θέση "Τίτλος" (Title) ένα τίτλο βιβλίου που αναζητούν, ώστε να λάβουν σε μία νέα σελίδα τις αντίστοιχες ονομασίες βιβλίων.

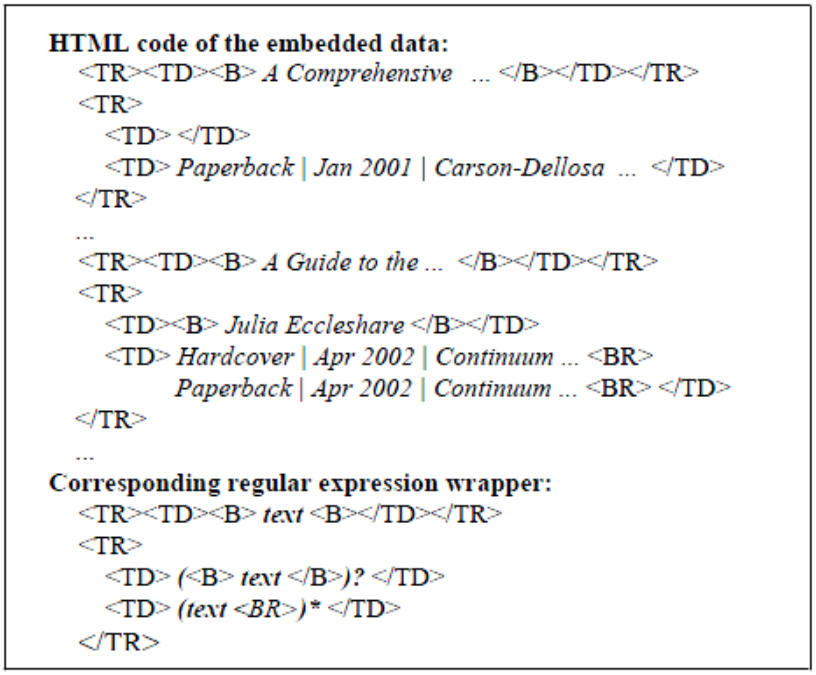


Εικόνα 16: Παράδειγμα σελίδας αποτελεσμάτων

Στην Εικόνα 16, τέσσερα αντικείμενα είναι αυτά που σχετίζονται με το ερώτημα "Harry Potter" και έχουν ως ιδιότητες το συγγραφέα, τον τίτλο και τις εκδόσεις.

Θεωρώντας ένα αλφάβητο συμβόλων Σ και μία λέξη "text" που δεν ανήκει στο Σ, μια κανονική έκφραση στο Σ είναι μια συμβολοσειρά στο Σ ∪ {text, *, ?, |, (,)}, που ορίζεται ως εξής:

- Η κενή συμβολοσειρά ϵ κι όλα τα στοιχεία στο Σ ∪ {text} είναι κανονικές εκφράσεις.
- Αν οι Α και Β είναι κανονικές εκφράσεις, τότε και οι ΑΒ, (Α|Β) και $(A)^?$ είναι κανονικές εκφράσεις, όπου η (Α|Β) συνεπάγεται ότι ισχύει το ένα από τα 2 και η $(A)^?$ ισοδυναμεί με την έκφραση (Α|ϵ).
- Αν η Α είναι κανονική έκφραση, τότε και η $(A)^*$ είναι κανονική έκφραση, όπου η $(A)^*$ συνεπάγεται το ένα από τα ϵ ή Α ή ΑΑ ή κ.ο.κ.



Εικόνα 17: Κώδικας HTML για δεδομένα και το αντίστοιχο περιτύλιγμα (wrapper)

Στην Εικόνα 17 φαίνεται ο κώδικας HTML για το πρώτο και το τρίτο βιβλίο της Εικόνα 16 όπως επίσης και το αντίστοιχο περιτύλιγμα (wrapper) κανονικών εκφράσεων για την εξαγωγή στιγμιότυπων βιβλίων.

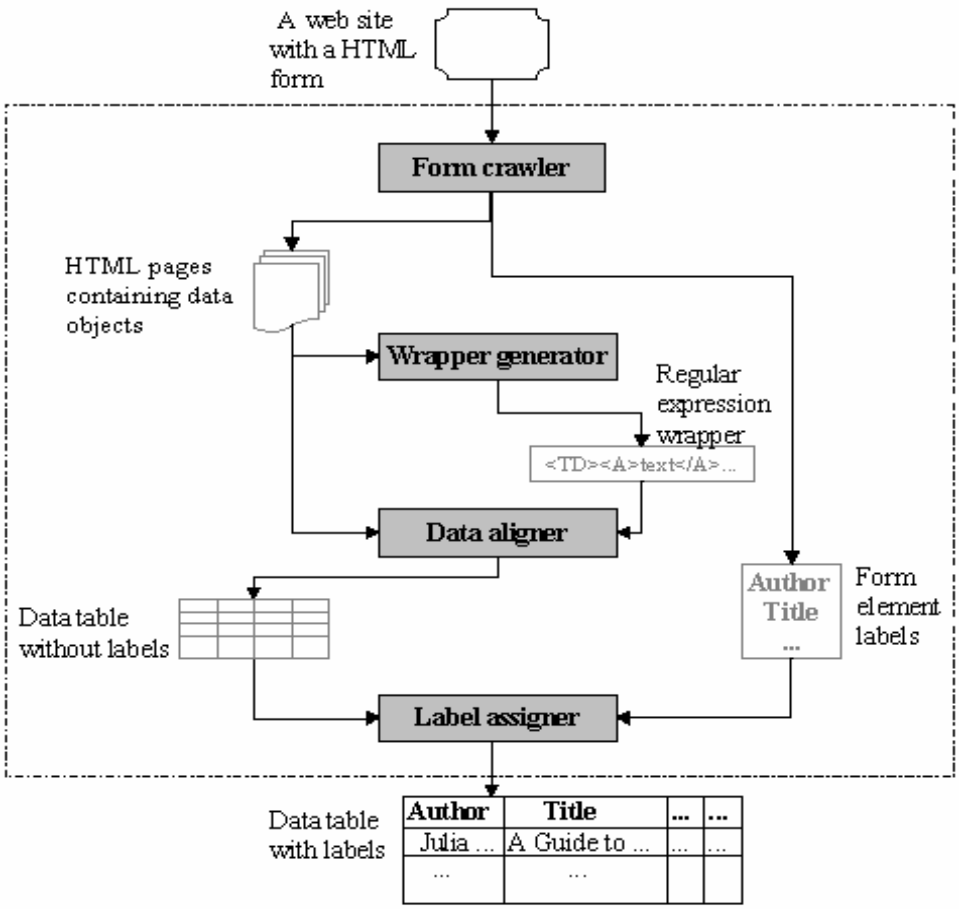
3.3.1.2 Η Αρχιτεκτονική του συστήματος DeLa

Το σύστημα DeLa αποτελείται από τέσσερα στοιχεία:

- *Τον ιστοσυλλέκτη φορμών* (form crawler): Σε μία ιστοσελίδα με μια φόρμα αναζήτησης HTML, ο ιστοσυλλέκτης φορμών συλλέγει τις ετικέτες κάθε στοιχείου που περιέχεται στη φόρμα και αποστέλλει τα ερωτήματα από τα στοιχεία της φόρμας για να αποκτήσει τις σελίδες-αποτελέσματα που περιέχουν τα δεδομένα. Χρησιμοποιήθηκε ο ιστοσυλλέκτης του κρυμμένου ιστού, HiWe (Hidden Web Exposer), ο οποίος κατασκευάστηκε με την εξής λογική: "οι περισσότερες φόρμες σχετίζονται με κάποιο περιγραφικό κείμενο και σκοπός τους είναι να βοηθήσουν το χρήστη να κατανοήσει τη σημασία του στοιχείου". Ο ιστοσυλλέκτης έχει μια βάση δεδομένων που αποθηκεύει μερικές τιμές των εννοιών και αναθέτει αυτές τις τιμές ως ερωτήματα στα

στοιχεία της φόρμας, όταν οι ετικέτες της φόρμας και οι ετικέτες της βάσης δεδομένων ταιριάζουν.

- *Τον Δημιουργό περιτυλίγματος* (wrapper generator): Ως είσοδος του δημιουργού περιτυλιγμάτων χρησιμοποιούνται οι ιστοσελίδες που συλλέχθηκαν από τον ιστοσυλλέκτη φορμών. Αρχικά ο δημιουργός περιτυλιγμάτων θεωρεί την ιστοσελίδα ως μια ακολουθία λέξεων που αποτελείται από ετικέτες HTML και μια ειδική λέξη "text" η οποία αναπαριστά οποιαδήποτε συμβολοσειρά μέσα στις ετικέτες HTML. Έπειτα εξάγει επαναλαμβανόμενες υπό-συμβολοσειρές από την ακολουθία λέξεων με αποτέλεσμα να οδηγείται σ' ένα περιτύλιγμα κανονικών εκφράσεων ακλουθώντας ιεραρχικές σχέσεις μεταξύ των επαναλαμβανόμενων υπό-συμβολοσειρών.



Εικόνα 18: Αρχιτεκτονική συστήματος DeLa

- *Το μηχανισμό ευθυγράμμισης δεδομένων* (Data aligner): Ο μηχανισμός ευθυγράμμισης δεδομένων (data aligner) εξάγει τα αντικείμενα δεδομένων από τις σελίδες ταιριάζοντας το περιτύλιγμα με την ακολουθία λέξεων κάθε σελίδας. Στη συνέχεια φιλτράρει τις ετικέτες HTML και επαναδιευθετεί τα στιγμιότυπα των δεδομένων σ' ένα πίνακα που μοιάζει με τον πίνακα που ορίζεται σε ένα σχεσιακό σχήμα βάσης δεδομένων. Οι γραμμές αναπαριστούν

στιγμιότυπα δεδομένων και οι στήλες γνωρίσματα. Τα εξαγόμενα αντικείμενα δεδομένων μπορεί να έχουν προαιρετικά γνωρίσματα ή με πολλαπλές τιμές, όπως για παράδειγμα το βιβλίο στην Εικόνα 16 που δεν είχε κανένα συγγραφέα και το βιβλίο που είχε 2 εκδόσεις.

- *Το Μηχανισμό ανάθεσης ετικετών* (Label assigner): Ο μηχανισμός ανάθεσης ετικετών (label assigner) είναι υπεύθυνος στο να αναθέτει ετικέτες στο πίνακα δεδομένων αντιστοιχίζοντας τις ετικέτες των φορμών που παρέχονται από τον ιστοσυλλέκτη με τις στήλες του πίνακα. Οι λέξεις του ερωτήματος που υποβλήθηκαν μέσω των στοιχείων της φόρμας είναι πιθανό να ξαναεμφανιστούν στα αντίστοιχα πεδία των αντικειμένων, αφού οι ιστοσελίδες συνήθως προσπαθούν να παρέχουν όσο το δυνατόν περισσότερο σχετικά δεδομένα στους χρήστες.

3.3.2 Παραγωγή περιτυλίγματος

3.3.2.1 Εξαγωγή τμήματος πλούσιου σε δεδομένα

Στη διαδικασία εξαγωγής δεδομένων από τις ιστοσελίδες πρέπει να αναγνωριστούν εκείνα τα μέρη της ιστοσελίδας τα οποία είναι πλούσια σε δεδομένα και αφορούν τους χρήστες, και συνεπώς πρέπει να αποφευχθούν τα δεδομένα θορύβου (όπως οι διαφημίσεις). Οι σελίδες οι οποίες βρίσκονται στον ίδιο δικτυακό τόπο έχουν παρόμοια δομή. Για την οργάνωση του περιεχομένου ακολουθείται ο αλγόριθμος DSE (Data-rich Section Extraction) για να μπορέσει να αναγνωρίσει τα μέρη εκείνα που είναι πλούσια σε δεδομένα συγκρίνοντας δυο σελίδες που ανήκουν στον ίδιο ιστοχώρο. Για το σκοπό αυτό κατασκευάζονται δέντρα DOM (Document Object Model) και για τις δυο σελίδες και κατόπιν γίνεται αναζήτηση πρώτα σε βάθος, συγκρίνοντας κόμβο προς κόμβο τις αντίστοιχες σελίδες και απορρίπτοντας τους κόμβους με διπλότυπα υπό-δέντρα του ίδιου βάθους.

3.3.2.2 C-επαναλαμβανόμενο πρότυπο

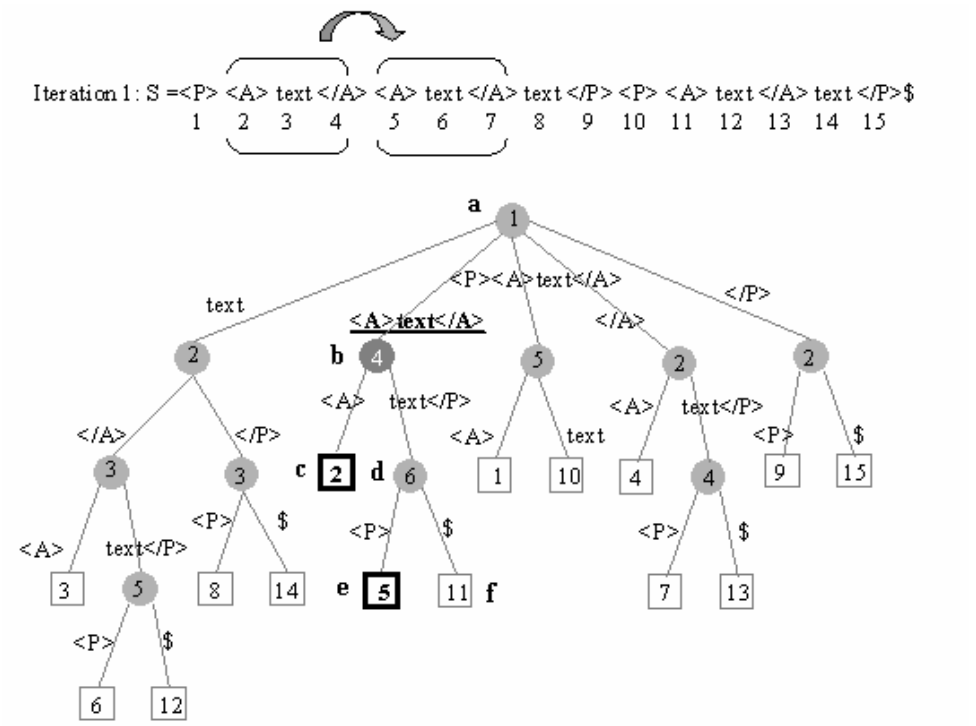
Οι σελίδες HTML περιέχουν δεδομένα τα οποία παράγονται από κοινά πρότυπα. Η δομή που αντιστοιχεί στα δεδομένα που ενσωματώνονται στη σελίδα μπορεί να εμφανιστεί επαναλαμβανόμενα όταν περιέχονται περισσότερα από ένα στιγμιότυπα αντικειμένων-δεδομένων. Συνεπώς, με επαναληπτικό τρόπο ανακαλύπτονται συνεχώς επαναλαμβανόμενα πρότυπα (continuously repeated ή C-επαναλαμβανόμενα πρότυπα) από τις ακολουθίες συμβόλων που αναπαριστούν σελίδες HTML και θεωρούνται υποψήφια περιτυλίγματα.

Ορισμός C-επαναλαμβανόμενου προτύπου: έχοντας μια συμβολοσειρά εισόδου S, η επαναλαμβανόμενη υπό-συμβολοσειρά του S που έχει τουλάχιστον ένα ζεύγος εμφανίσεων γειτονικό είναι η C-επαναλαμβανόμενη υπό-συμβολοσειρά (πρότυπο) του S.

Για την ανακάλυψη της εσωτερικής δομής μιας συμβολοσειράς χρησιμοποιείται μια δομή δεδομένων *δέντρου επιθεμάτων* (token suffix-tree). Στην Εικόνα 19 παρουσιάζεται μια ακολουθία στοιχείων (tokens) και το αντίστοιχο δέντρο

επιθεμάτων (token suffix-tree) όπου κάθε φύλλο παρουσιάζεται με κύκλο που έχει αριθμό την αρχική θέση ενός επιθέματος (suffix). Ένας «γεμάτος» κύκλος αναπαριστά κάθε εσωτερικό κόμβο και περιέχει έναν αριθμό που δείχνει την θέση στην οποία διαφέρουν οι κόμβοι-παιδιά. Όσοι κόμβοι μοιράζονται τον ίδιο πατέρα βρίσκονται σε αλφαβητική σειρά. Κάθε ακμή μεταξύ δυο εσωτερικών κόμβων έχει μια ετικέτα, που είναι η υπό-συμβολοσειρά που παρεμβάλλεται μεταξύ δύο κόμβων που περιέχουν στοιχεία (tokens). Κάθε κορυφή μεταξύ εσωτερικού κόμβου και φύλλου έχει μια ετικέτα όπου είναι το στοιχείο (token) στη θέση του εσωτερικού κόμβου του επιθέματος (suffix) που ξεκινά από το φύλλο.

Ένα δέντρο επιθεμάτων (suffix-tree) με στοιχεία (tokens) επιτυγχάνει καλύτερο χρόνο κατασκευής O(n).



Εικόνα 19: Ανακάλυψη C-επαναλαμβανόμενων προτύπων

Οι ετικέτες όλων των μονοπατιών των εσωτερικών κόμβων και τα προθέματα τους στο δέντρο επιθεμάτων (suffix-tree) με στοιχεία (tokens), είναι υποψήφια για την ανακάλυψη των C-επαναλαμβανόμενων προτύπων. Για κάθε υποψήφιο C-επαναλαμβανόμενο πρότυπο, έχουμε ένα C-επαναλαμβανόμενο πρότυπο αν δύο οποιαδήποτε στιγμιότυπα είναι γειτονικά, δηλαδή η απόσταση μεταξύ των δύο αρχικών θέσεων είναι ίση με το μήκος του συγκεκριμένου προτύπου. Για να ανακαλυφθούν οι εμφωλευμένες δομές από την ακολουθία συμβολοσειράς που αναπαριστά μια σελίδα HTML, χρησιμοποιείται ένα ιεραρχικό δέντρο προτύπων όπου στηρίζεται στην επανάληψη της δημιουργίας του δέντρου επιθεμάτων (suffix-tree) με στοιχεία (tokens) και στην ανακάλυψη C-επαναλαμβανόμενων προτύπων.

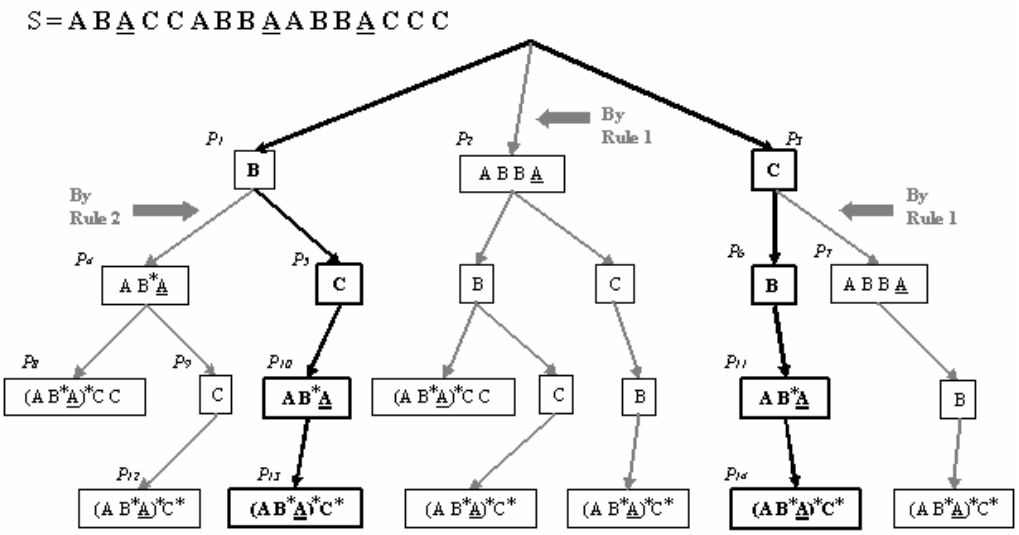
Το δέντρο προτύπων χρησιμοποιείται για την αναπαράσταση της εξάρτησης και μη, μεταξύ των C-επαναλαμβανόμενων προτύπων και την καταγραφή της επαναληπτικής ανακάλυψης.

Στα δέντρα προτύπων ισχύουν τα εξής:

- Το πρότυπο P_i είναι ένα παιδί του προτύπου P_j, στην περίπτωση που το P_i ανακαλυφθεί στην επαναληπτική φάση αμέσως μετά τη φάση όπου ανακαλύφθηκε το P_j.

- Το πρότυπο P_i είναι ένας αδερφός του προτύπου P_j, στην περίπτωση που το P_i ανακαλυφθεί στην ίδια επαναληπτική φάση όπου ανακαλύφθηκε και το P_j.

Αρχικά τοποθετείται ένα άδειο πρότυπο ως ρίζα του δέντρου επιθεμάτων (suffix-tree). Για κάθε πρότυπο διατηρείται το τελευταίο στιγμιότυπο σε κάθε επαναλαμβανόμενη περιοχή αυτού του προτύπου και ωθεί τα υπόλοιπα στιγμιότυπα της τρέχουσας ακολουθίας για τη δημιουργία μιας νέας. Στη συνέχεια δημιουργείται ένα δέντρο επιθεμάτων (suffix-tree) για τη νέα ακολουθία ανακαλύπτοντας τα νέα C-επαναλαμβανόμενα πρότυπα και εισάγοντας τα ως παιδιά του τρέχοντος προτύπου στο δέντρο προτύπων. Η διαδικασία επαναλαμβάνεται για κάθε νέο πρότυπο. Αφού τελειώσει η επεξεργασία των παιδιών του τρέχοντος προτύπου, ο αλγόριθμος γυρίζει ένα βήμα πίσω και συνεχίζει τη διαδικασία σε περίπτωση που υπάρχουν αδέρφια του τρέχοντος προτύπου. Στην περίπτωση επιστροφής στη ρίζα αυτό σημαίνει ότι η διαδικασία τελείωσε.



Εικόνα 20: Παράδειγμα δέντρου προτύπων

Στην Εικόνα 20 παρουσιάζεται ένα ολοκληρωμένο δέντρο προτύπων.

- Κάθε χαρακτήρας συμβολίζει ένα στοιχείο (token) στην ακολουθία HTML,
- Κάθε * δηλώνει ότι η υπό-συμβολοσειρά εμφανίζεται καμία ή περισσότερες φορές.

Στα δέντρα προτύπων όπως της Εικόνα 20 κάθε φύλλο αναπαριστά την εμφωλευμένη δομή που ανακαλύφθηκε και οι οποίες δίνονται από το μονοπάτι της ρίζας προς το φύλλο.

3.3.3 Ευθυγράμμιση δεδομένων

Ο μηχανισμός ευθυγράμμισης δεδομένων αποτελείται από δυο φάσεις:

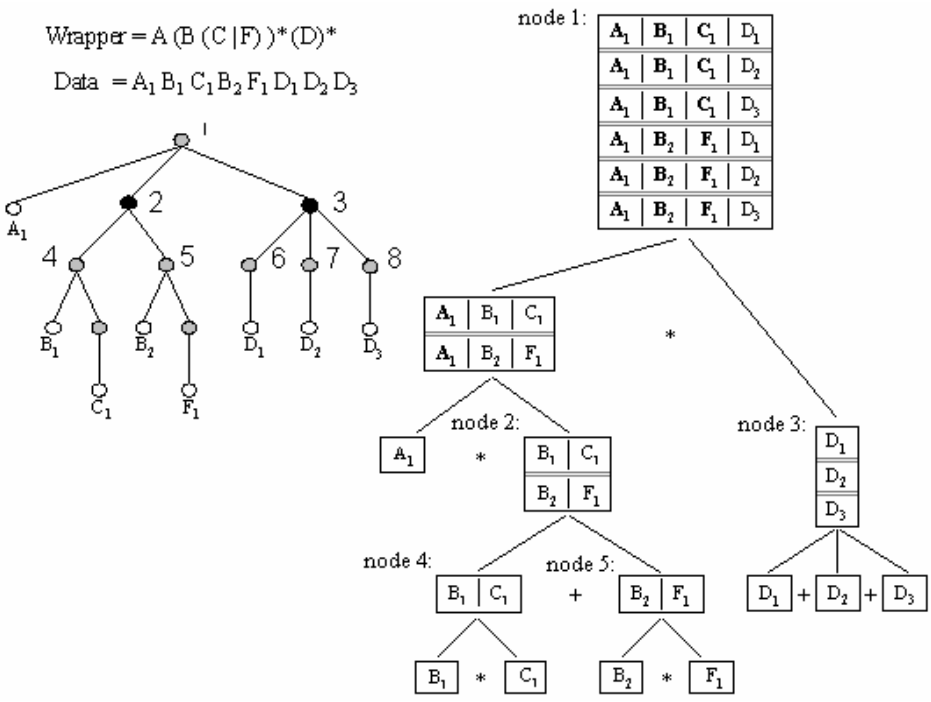
- Η *πρώτη φάση* αφορά την εξαγωγή δεδομένων, δηλαδή εξάγει πρώτα τα δεδομένα από τις ιστοσελίδες σύμφωνα με το περιτύλιγμα και γεμίζει ένα πίνακα με τα δεδομένα που έχει εξάγει.

- Η *δεύτερη φάση* είναι ο διαχωρισμός των χαρακτηριστικών γνωρισμάτων. Στη φάση αυτή εξετάζονται οι στήλες του πίνακα και έπειτα διαχωρίζονται τα γνωρίσματα που έχουν κωδικοποιηθεί σε μορφή συμβολοσειράς σε νέες στήλες οι οποίες θα προστεθούν στο πίνακα.

Στην *πρώτη φάση*, έχοντας ένα πρότυπο κανονικής έκφρασης και μια ακολουθία από στοιχεία (tokens) που απαρτίζουν την ιστοσελίδα, κατασκευάζεται ένα μη-ντετερμινιστικό και πεπερασμένων καταστάσεων αυτόματο (που από μια κατάσταση διαβάζοντας ένα σύμβολο εισόδου, μπορεί να μεταβεί σε μία ή και περισσότερες καταστάσεις) για να ταιριάζει με τις εμφανίσεις από τις ακολουθίες συμβολοσειρών που απεικονίζουν ιστοσελίδες.

Ένα δέντρο δεδομένων χρησιμοποιείται για μια κανονική έκφραση που έχει δοθεί ώστε να καταγραφεί η εμφάνιση αυτής της έκφρασης. Το δέντρο δεδομένων δίνεται με τον εξής τρόπο:

- Στην περίπτωση που η κανονική έκφραση είναι ατομική, το δέντρο δεδομένων είναι ένας απλός κόμβος και η εμφάνιση της έκφρασης είναι η ετικέτα του κόμβου αυτού.
- Στην περίπτωση που η κανονική έκφραση έχει τη μορφή $E_1E_2…E_n$, το δέντρο δεδομένων είναι ένας κόμβος με n παιδιά και το i-οστό παιδί ($1<i<n$) είναι ένα δέντρο δεδομένων το οποίο καταγράφει την εμφάνιση του E_i.
- Στην περίπτωση που η κανονική έκφραση είναι η $(E_1|E_2)$, το δέντρο δεδομένων είναι ένας κόμβος με ένα παιδί και καταγράφει την εμφάνιση του E_1 ή του E_2.
- Στην περίπτωση που η κανονική έκφραση είναι η $(E)^*$ και υπάρχουν m εμφανίσεις του E, το δέντρο δεδομένων είναι ένας κόμβος με m παιδιά και το i-οστό παιδί ($1<i<m$) είναι ένα δέντρο δεδομένων το οποίο καταγράφει την m-οστή εμφάνιση του E.



Εικόνα 21: Παράδειγμα δέντρου δεδομένων και συμπληρωματικός πίνακας

Στο αριστερό μέρος της Εικόνα 21 φαίνεται ένα δέντρο δεδομένων για την κανονική έκφραση "A(B(C|F))*D*", με την οποία καταγράφεται η εμφάνιση του "ABCBFDDD". Για το διαχωρισμό των διαφορετικών εμφανίσεων του ίδιου υπό-προτύπου, όπως είναι το "B_1" και το "D_3" χρησιμοποιούνται διαφορετικοί δείκτες. Στη ρίζα του δέντρου δεδομένων βρίσκονται 3 παιδιά γιατί το περιτύλιγμα είναι η συνένωση των "A", "(B(C|F))*" και "(D)*". Ο κόμβος 3 είναι το δέντρο δεδομένων του "(D)*" κι έχει 3 παιδιά επειδή υπάρχουν 3 εμφανίσεις του "D" στα δεδομένα. Έτσι, ο κόμβος 2 έχει 2 παιδιά για τις 2 εμφανίσεις "B_1C_1" και "B_2F_1" της έκφρασης "(B(C|F))*".

Στη *δεύτερη φάση* για τον διαχωρισμό των γνωρισμάτων, αφαιρούνται όλες οι ετικέτες HTML που περιέχονται στο πίνακα και οι στήλες του πίνακα που έχουν κενές συμβολοσειρές, και έτσι διασφαλίζεται ότι κάθε κελί του πίνακα δεδομένων περιέχει μια συμβολοσειρά ως περιεχόμενο. Το σημαντικό σημείο της φάσης διαχώρισης των χαρακτηριστικών γνωρισμάτων είναι ότι πρέπει να υπάρχουν κάποια ειδικά σύμβολα στη συμβολοσειρά ως διαχωριστές για την περίπτωση που κάποια χαρακτηριστικά γνωρίσματα κωδικοποιούνται σε μία συμβολοσειρά. Ο λόγος που πρέπει να υπάρχουν τα ειδικά σύμβολα είναι να είναι εφικτός ο διαχωρισμός των γνωρισμάτων από τους χρήστες, γι' αυτό πρέπει να βρεθούν οι σωστοί διαχωριστές για κάθε ιστοσελίδα. Έτσι, σε για κάθε στήλη γίνεται αναζήτηση σε όλα τα κελιά για χαρακτήρες που δε θα είναι γράμματα ούτε νούμερα και στη συνέχεια καταγράφονται οι εμφανίσεις και οι αντίστοιχες θέσεις στα κελιά. Στην περίπτωση που οι χαρακτήρες που ανακαλύπτονται έχουν και ίδιο αριθμό εμφανίσεων σε όλα τα κελιά μιας στήλης ο χαρακτήρας αναγνωρίζεται ως υποψήφιος διαχωριστής αυτής της στήλης.

Οι διαχωριστές που είναι υποψήφιοι μετρούνται με διάφορες ευρετικές μεθόδους. Οι χαρακτήρες "@" ή "$" για παράδειγμα δεν θεωρούνται ως έγκυροι διαχωριστές γιατί εμφανίζονται σε οποιαδήποτε συμβολοσειρά που αναπαριστά ηλεκτρονικό ταχυδρομείο (email) "@" ή αναπαριστά τιμές "$". Μόλις βρεθούν πολλοί διαχωριστές οι οποίοι είναι έγκυροι για κάποια στήλη τότε οι συμβολοσειρές της στήλης διαχωρίζονται από την αρχή μέχρι το τέλος σε διάταξη η οποία καθορίζεται από τις θέσεις που εμφανίζεται ο κάθε διαχωριστής.

3.3.4 Ανάθεση ετικετών

Για να ανατεθούν ετικέτες στις στήλες του πίνακα δεδομένων ακολουθούνται τέσσερεις ευρετικές μέθοδοι:

- Ευρετική μέθοδος 1: *ταίριασμα ετικετών στοιχείων της φόρμας με γνωρίσματα δεδομένων*. Η φόρμα αναζήτησης μέσα από την οποία υποβάλλονται τα ερωτήματα από τους χρήστες παρέχει μία περιγραφή της σχεσιακής βάσης δεδομένων της ιστοσελίδας. Σε κάθε στοιχείο φόρμας με τα ερωτήματα των λέξεων-κλειδιών, όταν οι λέξεις-κλειδιά εμφανίζονται σε μία συγκεκριμένη στήλη του πίνακα δεδομένων ανατίθεται ετικέτα του στοιχείου φόρμας στη στήλη.

- Ευρετική μέθοδος 2: *αναζήτηση για πιθανές ετικέτες στις κεφαλίδες του πίνακα*. Στην HTML ορίζονται κάποιες ετικέτες όπως είναι η <TH> και η <THEAD> για τους σχεδιαστές της σελίδας, ώστε να μπορούν να βρίσκουν τις κεφαλίδες των στηλών των πινάκων της HTML.

- Ευρετική μέθοδος 3: *αναζήτηση για πιθανές ετικέτες που είναι κωδικοποιημένες με χαρακτηριστικά γνωρίσματα δεδομένων*. Για κάθε στήλη του πίνακα δεδομένων, βρίσκεται το μέγιστο πρόθεμα (maximal-prefix) και το μέγιστο επίθεμα (maximal-suffix) που διαμοιράζονται όλα τα κελιά της στήλης. Υποθέτουμε ότι το πρόθεμα και το επίθεμα έχουν νόημα και ανατίθεται το πρόθεμα στη στήλη αυτή και το επίθεμα στην επόμενη στήλη.

- Ευρετική μέθοδος 4: *ανάθεση επιθεμάτων σε γνωρίσματα που βρίσκονται σε συμβατικές μορφές*. Ορισμένα δεδομένα όπως είναι η ημερομηνία που έχει τη μορφή "dd-mm-yy", το ηλεκτρονικό ταχυδρομείο (email) που περιέχει το χαρακτήρα "@", η τιμή που έχει το χαρακτήρα "$" έχουν συμβατική μορφή. Τέτοιου είδους πληροφορίες χρησιμοποιούνται στην αναγνώριση των αντίστοιχων χαρακτηριστικών γνωρισμάτων δεδομένων.

3.3.5 Πειραματική αξιολόγηση

Στην ενότητα αυτή συνοψίζεται ή αξιολόγηση των τριών στοιχείων του συστήματος DeLa. Εξετάζονται τρεις κατηγορίες ιστοσελίδων, τα καταστήματα βιβλίων, αγγελίες εύρεσης εργασίας καθώς και διαφημίσεις αυτοκινήτων, κι ακολούθως συλλέγονται εννέα ιστοσελίδες για κάθε κατηγορία από το http://www.invisible.com/.

Για την ανάθεση ετικετών, αντιστοιχίζονται οι ετικέτες των στοιχείων φορμών με χαρακτηριστικά γνωρίσματα δεδομένων. Για να εξεταστεί η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας για την εκχώρηση ετικετών, προσομοιάζονται οι ενέργειες του ιστοσυλλέκτη φορμών καθώς συλλέγονται οι ετικέτες των στοιχείων φορμών κι ακολούθως στέλνονται ερωτήματα μέσα από φόρμες HTML. Ο Πίνακας 5 δείχνει τον αριθμό των στοιχείων φορμών που περιέχονται σε κάθε ιστοσελίδα (στήλες με τίτλο "FE") και τον αριθμό των ερωτημάτων που στάλθηκαν (στήλες με τίτλο "Q"). Στέλνονται μόνο ερωτήματα για κάθε στοιχείο εισόδου "λίστα επιλογών" και "πλαίσια κειμένου", αγνοώντας άλλα στοιχεία, όπως τα "στοιχεία επιλογής" και "κουμπιά επιλογής". Για κάθε στοιχείο "λίστα επιλογών", το ερώτημα περιέχει μία από τις διαθέσιμες επιλογές και για κάθε στοιχείο "πλαίσιο κειμένου", το ερώτημα περιέχει μια απλή λέξη-κλειδί.

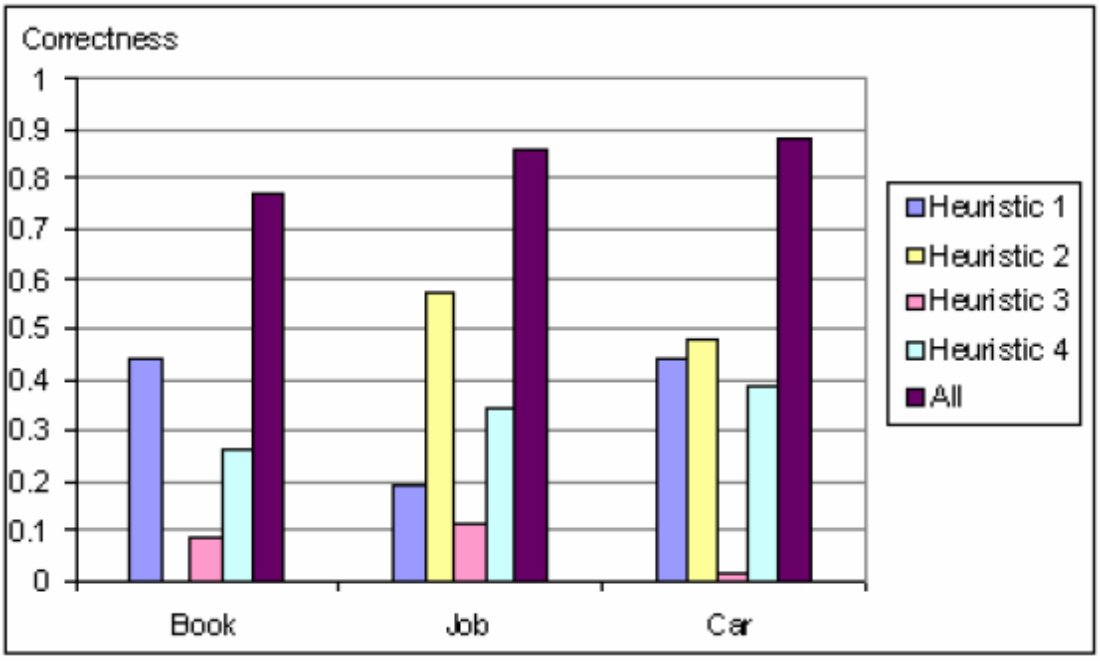
	Book		**Job**		**Car**	
	FE	**Q**	**FE**	**Q**	**FE**	**Q**
Site 1	13	7	6	3	2	2
Site 2	8	7	3	2	15	10
Site 3	4	3	2	2	6	5
Site 4	6	4	3	2	4	2
Site 5	5	5	8	2	8	8
Site 6	6	3	4	3	4	4
Site 7	15	8	13	3	15	5
Site 8	8	7	5	2	7	7
Site 9	6	5	4	2	9	7

Πίνακας 5: Αριθμός στοιχείων φορμών και ερωτημάτων

Στην Εικόνα 22 παρουσιάζεται η μέση ορθότητα (correctness) των τεσσάρων ευρετικών μεθόδων καθώς και ο συνδυασμός τους για τις τρεις κατηγορίες των ιστοσελίδων. Η ορθότητα μετράται με τον έλεγχο του αν η σημασία/έννοια της ετικέτας που ανατέθηκε ταιριάζει με τη σημασία/έννοια του αντίστοιχου χαρακτηριστικού γνωρίσματος.

Παρατηρείται ότι η ορθότητα της κάθε ευρετικής μεθόδου ποικίλλει ανάλογα με το τύπο της ιστοσελίδας. Για παράδειγμα, οι περισσότερες από τις ιστοσελίδες που αφορούν τις αγγελίες για θέσεις εργασίας ταξινομούν τα δεδομένα τους σε πίνακες HTML, χρησιμοποιώντας κεφαλίδες πίνακα, με συνέπεια η $2^{η}$ ευρετική μέθοδος να λειτουργεί καλά για τις συγκεκριμένες. Ομοίως, η $1^{η}$ ευρετική μέθοδος λειτουργεί καλά για τις ιστοσελίδες για τα καταστήματα βιβλίων και τις διαφημίσεις αυτοκινήτων, επειδή το βιβλία και τα αυτοκίνητα είναι δεδομένα που έχουν μια σαφή δομή δεδομένων. Η απόδοση της $4^{ης}$ ευρετικής μεθόδου είναι αρκετά σταθερή, επειδή

και οι τρεις κατηγορίες ιστοσελίδων έχουν χαρακτηριστικά γνωρίσματα, όπως η τιμή και η ημερομηνία.



Εικόνα 22: Ορθότητα των ευρετικών μεθόδων ανάθεσης ετικετών

4 Συνδυασμός πολλαπλών πηγών δεδομένων στον κρυμμένο ιστό

Σε μία συγκεκριμένη περιοχή, οι βάσεις δεδομένων του κρυμμένου ιστού συχνά δεν είναι ανεξάρτητες: για παράδειγμα τα αποτελέσματα που επιστρέφονται από μια βάση δεδομένων μπορεί να χρησιμοποιούνται για αναζήτηση σε μια άλλη βάση δεδομένων. Όταν ο χρήστης θέσει ένα ερώτημα, πολλές βάσεις δεδομένων χρειάζεται να ερωτηθούν με μια ευφυή σειρά ώστε να ανακτήσουν όλες τις πληροφορίες που ζητά ο χρήστης. Αυτό οδηγεί στην ανάγκη για τεχνικές που μπορούν να παράγουν σχέδια ερωτήσεων, υπολογίζοντας τη συσχέτιση μεταξύ των πηγών δεδομένων ανακτώντας τις πληροφορίες που θέλει ο χρήστης. Αυτό είναι μία πρόκληση για τα συστήματα κρυμμένου ιστού και στο παρόν κεφάλαιο εξετάζουμε τις προσεγγίσεις για το ζήτημα αυτό.

4.1 Σχεδιασμός ερωτήματος για αναζήτηση σε αλληλοεξαρτώμενες βάσεις δεδομένων του κρυμμένου ιστού

Στην Ενότητα 4.1 εξετάζεται ο σχεδιασμός ερωτήσεων σε ένα σύστημα ολοκλήρωσης του κρυμμένου ιστού [9]. Το σύστημα αυτό σχεδιάστηκε ώστε να παρέχει μια απλή διεπαφή ερωτήματος. Κάθε ερώτημα έχει ένα όρο-κλειδί και ένα σύνολο όρων-στόχων για τους οποίους ενδιαφέρεται ο χρήστης. Ο όρος-κλειδί είναι ένα όνομα και οι όροι-στόχοι αντανακλούν τις ιδιότητες ή το είδος των πληροφοριών που είναι επιθυμητοί για αυτό το όνομα. Για ένα τέτοιο σύστημα έχει αναπτυχθεί ένας δυναμικός σχεδιαστής ερωτημάτων, ο οποίος παράγει μία αποτελεσματική σειρά ερωτημάτων με βάση τις εξαρτήσεις των βάσεων δεδομένων του κρυμμένου ιστού. Ο σχεδιαστής ερωτήματος διαλέγει τα Κ βέλτιστα σχέδια ερωτήματος, έτσι ώστε να διασφαλίσει την ύπαρξη εναλλακτικών σχεδίων όταν το βέλτιστο σχέδιο δεν είναι εφικτό.

4.1.1 Η διατύπωση του προβλήματος

Το σύστημα ολοκλήρωσης του κρυμμένου ιστού που πραγματεύεται η εργασία [9] παρέχει ένα καθορισμένο σύνολο υποψήφιων όρων που μπορούν να χρησιμοποιηθούν στα ερωτήματα. Τέτοιοι όροι δηλώνονται ως *όροι-στόχοι* των ερωτημάτων. Ο χρήστης επιλέγει ένα υποσύνολο από τους όρους-στόχους και καθορίζει επίσης έναν *όρο-κλειδί*. Οι όροι-στόχοι καθορίζουν το είδος των πληροφοριών που επιθυμεί να λάβει ο χρήστης για τον όρο-κλειδί. Για ένα ερώτημα μπορεί να χρειάζεται να εξαχθούν από πολλαπλές βάσεις δεδομένων πολλά κομμάτια πληροφοριών. Μπορεί να υπάρχουν εξαρτήσεις μεταξύ των βάσεων δεδομένων, δηλαδή μια πληροφορία που μπορεί να ανακτηθεί από μια πηγή να απαιτείται για να διαμορφωθεί ένα ερώτημα προς μια άλλη πηγή. Στόχος είναι να υπάρχει στρατηγική σχεδίασης των ερωτημάτων όπου θα παρέχει ορθό και αποδοτικό σχεδιασμό ερωτημάτων για αναζήτηση σε σχετικές βάσεις δεδομένων.

Ο δυναμικός σχεδιαστής ερωτήματος παράγει ένα σύνολο που περιλαμβάνει τα Κ βέλτιστα σχέδια. Το σχέδιο με το *μικρότερο μήκος*, δηλ. το μικρότερο πλήθος βάσεων δεδομένων στις οποίες θα γίνει αναζήτηση, η μέγιστη κάλυψη των όρων που ο

χρήστης εισήγαγε αλλά και οι προτιμήσεις του χρήστη αποτελούν τους πιο σημαντικούς παράγοντες στην αξιολόγηση των σχεδίων. Με την παραγωγή Κ σχεδίων ερωτήματος, υπάρχουν εναλλακτικά πλάνα στη περίπτωση που το βέλτιστο σχέδιο δεν είναι εφικτό π.χ. λόγω της μη διαθεσιμότητας μιας βάσης δεδομένων.

Τυπικά, το πρόβλημα μπορεί να διατυπωθεί ως ακολούθως: δίνεται το καθολικό σύνολο $T = \{t_1, t_2, \dots, t_n\}$, όπου κάθε t_i είναι ένας όρος που μπορεί να ζητηθεί από τον χρήστη. Θεωρούμε το υποσύνολο $T' = \{t'_1, t'_2, \dots, t'_m\}$, $(t'_i \in T)$, το οποίο είναι το σύνολο των όρων που ο χρήστης ζητά σε ένα συγκεκριμένο ερώτημα. Επιπρόσθετα έχουμε το σύνολο $D = \{D_1, D_2, \dots, D_m\}$, όπου κάθε D_i είναι μια βάση δεδομένων στον κρυμμένο ιστό και κάθε D_i καλύπτει ένα σύνολο όρων $E_i = \{e_i^1, e_i^2, \dots, e_i^k\}$, με το E_i να είναι υποσύνολο του T. Κάθε βάση δεδομένων D_i χρειάζεται ένα σύνολο στοιχείων $\{r_i^1, r_i^2, \dots, r_i^k\}$ προκειμένου να γίνει αναζήτηση σε αυτή όπου $r_i^j \in T$.

Στόχος είναι να βρεθεί μια σειρά ερωτημάτων για τις βάσεις δεδομένων $D^* = \{D_1, D_2, \dots, D_k\}$, η οποία να καλύπτει το σύνολο T' με το *μεγαλύτερο όφελος* και παράλληλα να ελαχιστοποιεί το k. Το όφελος υπολογίζεται μέσω μιας συνάρτησης κόστους. Ονομάζουμε το πρόβλημα *δυναμικό σχεδιασμό ερωτήσεων,* διότι η σειρά των ερωτήσεων πρέπει να επιλεχθεί με βάση τους όρους που έχουν επιλεχθεί από τον χρήστη και δεν μπορεί συνακόλουθα να είναι προκαθορισμένος από το σύστημα ολοκλήρωσης.

4.1.1.1 Μοντελοποίηση συστήματος παραγωγής

Στην ενότητα αυτή παρουσιάζεται ένας αλγόριθμος, για την κατανόηση του οποίου είναι σκόπιμο να θεωρηθεί το πρόβλημα σχεδιασμού ερωτήματος ως *σύστημα παραγωγής*. Ένα σύστημα παραγωγής περιλαμβάνει τέσσερα στοιχεία που το χαρακτηρίζουν:

- τη μνήμη εργασίας,
- τον χώρο-στόχο,
- ένα σύνολο κανόνων παραγωγής,
- και έναν κύκλο ελέγχου *αναγνώρισης-ενέργειας*.

Η *μνήμη εργασίας* περιγράφει την επικρατούσα κατάσταση στη διαδικασία συλλογισμού. Ο *χώρος-στόχος* περιγράφει τον στόχο. Στην περίπτωση που η μνήμη εργασίας γίνει υπερσύνολο του χώρου-στόχου, τότε ολοκληρώνεται η λύση του προβλήματος. Ένας *κανόνας παραγωγής* είναι ένα ζεύγος (*συνθήκη δράση*) όπου η συνθήκη καθορίζει πότε εφαρμόζεται ο κανόνας ενώ η δράση καθορίζει το βήμα της λύσης του προβλήματος. Η μνήμη εργασίας αρχικοποιείται με την αρχική περιγραφή του προβλήματος και η κατάσταση συγκρίνεται με τις συνθήκες των κανόνων παραγωγής. Όταν ένας κανόνας παραγωγής πυροδοτείται, τότε η δράση του εκτελείται και η μνήμη εργασίας μεταβάλλεται ανάλογα. Η διαδικασία ολοκληρώνεται όταν το περιεχόμενο της μνήμης εργασίας γίνει υπερσύνολο του

χώρου-στόχου ή στη περίπτωση που δεν γίνεται να πυροδοτηθούν περαιτέρω κανόνες.

Το πρόβλημα σχεδίασης ερωτήματος αντιστοιχίζεται στα τέσσερα στοιχεία του συστήματος παραγωγής με τον ακόλουθο τρόπο: Η μνήμη εργασίας αποτελείται από όλα τα δεδομένα που έχουν ήδη εξαχθεί. Το σχέδιο ερωτήματος παράγεται βαθμιαία και όταν μια βάση δεδομένων προστίθεται στο σχέδιο ερωτήματος, τα στοιχεία που μπορούν να εξαχθούν από την βάση δεδομένων θεωρούνται ως αποθηκευμένα στη μνήμη εργασίας. Η μνήμη εργασίας αρχικά περιλαμβάνει μόνο τον όρο-κλειδί της ερώτησης. Ο χώρος-στόχος είναι ένα υποσύνολο των όρων-στόχων που επιλέχθηκαν από τον χρήστη.

Κάθε βάση δεδομένων διαθέτει ένα ή περισσότερα σχήματα ερωτημάτων. Τα σχήματα ορίζουν ποια είναι η είσοδος σε μια φόρμα υποβολής ερωτήματος προς τη βάση δεδομένων αλλά και ποια δεδομένα εξάγονται από τη βάση δεδομένων χρησιμοποιώντας τους όρους εισόδου. Τα σχήματα βάσεων δεδομένων παρέχονται από τους προμηθευτές πηγών δεδομένων του κρυμμένου ιστού είτε από τον υπεύθυνο ανάπτυξης της διεπαφής ερωτήματος. Οι κανόνες παραγωγής του συστήματος που περιγράφεται εδώ είναι τα σχήματα βάσεων δεδομένων, όπου μια βάση δεδομένων μπορεί να έχει πολλαπλά σχήματα. Αν συμβεί αυτό, το κάθε σχήμα έχει ως είσοδο διαφορετικά στοιχεία για να ανακτήσει τα διαφορετικά αποτελέσματα που παράγονται.

Οι όροι στη μνήμη εργασίας ταιριάζονται με το απαιρούμενο σύνολο εισόδου κάθε κανόνα παραγωγής. Με βάση την στρατηγική επιλογής κανόνων, επιλέγεται και πυροδοτείται ο κατάλληλος κανόνας. Στη μνήμη εργασίας, η αντίστοιχη βάση δεδομένων σημειώνεται ως ερωτηθείσα και η έξοδος του πυροδοτούμενου κανόνα προστίθεται σε αυτή. Οι κανόνες που επιλέγονται θεωρούνται ως επισκεφθέντες ώστε να αποφευχθούν οι περαιτέρω πυροδοτήσεις του ίδιου κανόνα.

Ακολούθως παρουσιάζονται δυο συνθήκες από τις οποίες αρκεί να ισχύει η μία προκειμένου να παραχθεί ένα πλήρες σχέδιο ερωτήματος. Στη πρώτη περίπτωση, η μνήμη εργασίας καλύπτει όλα τα στοιχεία στον χώρο-στόχο, δηλαδή όλοι οι όροι-στόχοι του ερωτήματος που έχουν ζητηθεί από τον χρήστη έχουν βρεθεί. Στη δεύτερη περίπτωση υπάρχουν ακόμη κάποιοι όροι στην κατάσταση-στόχο που δεν έχουν καλυφθεί στη μνήμη εργασίας, ωστόσο κανένας από τους κανόνες που έχουν απομείνει δεν μπορεί να καλύψει άλλα στοιχεία του χώρου-στόχοι. Με βάση αυτό δεν γίνεται να ανακτηθούν όλοι οι όροι αιτήματος χρησιμοποιώντας το τρέχον σύνολο βάσεων δεδομένων. Τυπικά, αυτό συμβαίνει όταν μερικές βάσεις δεδομένων δεν είναι διαθέσιμες.

4.1.2 Προσέγγιση και αλγόριθμος σχεδίασης ερωτήματος

Ο αλγόριθμος στην εργασία [9] βασίζεται σε μια δομή δεδομένων, τον *γράφο εξαρτήσεων*, ο οποίος εισάγεται ώστε να αναπαραστήσει τις εξαρτήσεις μεταξύ των

βάσεων δεδομένων. Ο αλγόριθμος σχεδίασης ερωτήματος και ο γράφος εξαρτήσεων παρουσιάζονται στη συνέχεια.

4.1.2.1 Γράφος εξαρτήσεων

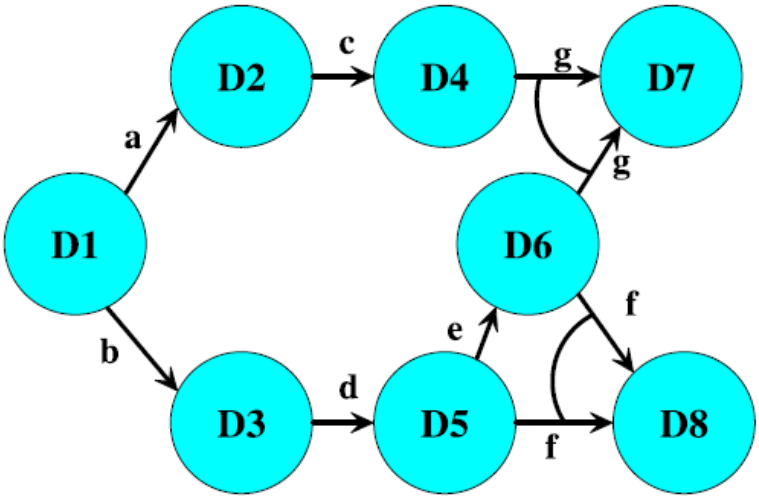
Μεταξύ των βάσεων δεδομένων υπάρχουν εξαρτήσεις: για παράδειγμα για να τεθεί ένα ερώτημα σε μια βάση δεδομένων D, πρέπει προηγουμένως να τεθούν ερωτήματα σε άλλες βάσεις δεδομένων ώστε να εξαχθούν τα απαραίτητα στοιχεία εισόδου για την ερώτηση στη βάση δεδομένων D. Χρησιμοποιείται αναπαράσταση κανόνων παραγωγής για να προσδιορισθούν οι εξαρτήσεις μεταξύ των βάσεων δεδομένων και να δημιουργηθεί ένας γράφος εξαρτήσεων που αναπαριστά τις σχέσεις μεταξύ των βάσεων δεδομένων.

Η σχέση εξάρτησης $DR, \prec_{DR} \subset 2^D \times D$, όπου 2^D είναι το δυναμοσύνολο (σύνολο όλων των υποσυνόλων) του συνόλου D. Εάν ισχύει η σχέση $\{D_i, D_{i+1}, \dots, D_{i+m}\} \prec_{DR} D_j$
τότε για να ερωτηθεί η πηγή δεδομένων D_j πρέπει πρώτα να ερωτηθούν οι πηγές δεδομένων $D_i, D_{i+1}, \dots, D_{i+m}$, προκειμένου να ληφθούν τα απαραίτητα στοιχεία ως είσοδος, προκειμένου να διαμορφωθεί η ερώτηση προς την πηγή δεδομένων D_j.

Για την αναπαράσταση της σχέσης εξάρτησης χρησιμοποιείται ένας υπεργράφος, ο οποίος αποτελείται από ένα σύνολο N κόμβων και ένα σύνολο υπερακμών. Το σύνολο των υπερακμών ορίζεται από διατεταγμένα ζεύγη, όπου το πρώτο στοιχείο του ζεύγους είναι ένα υποσύνολο του N και το δεύτερο στοιχείο είναι ένας κόμβος από το N. Το πρώτο στοιχείο του ζεύγους ονομάζεται *σύνολο γονέων,* και το δεύτερο *απόγονος*. Στην περίπτωση που το σύνολο γονέα δεν είναι μονοσύνολο, τα στοιχεία του συνόλου γονέων ονομάζονται κόμβοι "*AND*". Στο γράφο εξαρτήσεων, οι κόμβοι είναι βάσεις δεδομένων, και οι υπερακμές αναπαριστούν τις εξαρτήσεις ανάμεσα στις βάσεις δεδομένων. Σε μια υπερακμή, οι κόμβοι που ανήκουν στο σύνολο γονέων του ζεύγους είναι οι βάσεις δεδομένων που ερωτώνται για να συνεχιστεί το ερώτημα στη βάση δεδομένων που αντιπροσωπεύεται από τον κόμβο απόγονο του ζεύγους.

Ο γράφος εξαρτήσεων δημιουργείται με βάση τους κανόνες παραγωγής κάθε βάσης δεδομένων. Έστω δυο βάσεις δεδομένων D_i και D_j, όπου το D_i έχει ένα σύνολο κανόνων παραγωγής $R_i = \{r_{i1}, r_{i2}, \dots, r_{in}\}$ και το D_j έχει ένα σύνολο κανόνων παραγωγής $R_j = \{r_{j1}, r_{j2}, \dots, r_{jm}\}$. Εάν ένας κανόνας *στο* R_i έχει ως έξοδο ένα σύνολο το οποίο μπορεί να καλύψει *πλήρως* το σύνολο εισόδων οποιουδήποτε κανόνα παραγωγής στο R_j, τότε δημιουργείται μια ακμή μεταξύ του D_i και του D_j.Αντίθετα, αν ένας κανόνας στο R_i έχει ως έξοδο ένα σύνολο που μπορεί να καλύψει *μερικώς* την είσοδο οποιουδήποτε από τους κανόνες του συνόλου R_j, εξετάζονται οι κανόνες των άλλων βάσεων δεδομένων ώστε να βρεθεί ένα σύνολο βάσεων δεδομένων $C_{Dj}=\{D_a, \dots, D_i, \dots, D_x\}$ το οποίο να μπορεί πλήρως να καλύψει το σύνολο εισόδων των κανόνων στο R_j. Αν το σύνολο C_{Dj} υπάρχει, δημιουργείται μια υπερακμή από το C_{Dj} στο D_j . Σε περίπτωση που οι κανόνες παραγωγής μιας βάσης δεδομένων

ενημερωθούν, ο γράφος εξαρτήσεων μπορεί να ενημερωθεί αντίστοιχα. Στην Εικόνα 23 που ακολουθεί, παρουσιάζεται ένα παράδειγμα υπεργράφου.



Εικόνα 23: Παράδειγμα γράφου εξαρτήσεων

Στην Εικόνα 23 παρουσιάζονται 8 κόμβοι και 7 υπερακμές. Η πρώτη υπερακμή συμβολίζεται με a. Το στοιχείο-γονέας του διατεταγμένου ζεύγους του a είναι ο κόμβος D1 και το στοιχείο-απόγονος του ζεύγους είναι ο κόμβος D2. Η υπερακμή δηλώνει ότι μετά το ερώτημα στο D1 λαμβάνονται τα στοιχεία εισόδου που χρειάζονται για το ερώτημα στο D2. Οι υπόλοιπες υπερακμές δηλαδή από το b έως το e λειτουργούν με τον ίδιο τρόπο όπως η υπερακμή a. Στην υπερακμή f, το τόξο που ενώνει τις δυο ακμές του f δείχνει ότι η υπερακμή είναι σύνδεσμος βαθμού 2. Το στοιχείο-γονέας του f είναι ένα σύνολο που περιλαμβάνει το D5 και το D6, και το στοιχείο-απόγονος είναι το D8. Έτσι για να ερωτηθεί το D8 πρέπει να τεθεί το ερώτημα στα D5 και D6. Το ίδιο ισχύει και για την υπερακμή g.

Για ένα κόμβο D οι *γείτονες* του D είναι οι κόμβοι που έχουν μια ακμή που ξεκινά από το D. Οι *συνεργάτες* του D είναι οι κόμβοι που συνδέονται με το D μέσω μιας υπερακμής σε έναν κοινό απόγονο, για παράδειγμα το D5 έχει έναν συνεργάτη D6, επειδή το D5 και το D6 συνδέονται μέσω της υπερακμής f με το D8.

4.1.2.2 Αλγόριθμος σχεδίασης ερωτήματος

Ως πιο σημαντικές θεωρούμε τις βάσεις δεδομένων που είναι γονείς άλλων σημαντικών βάσεων δεδομένων, με βάση την ύπαρξη των εξαρτήσεων μεταξύ των βάσεων δεδομένων. Στην Εικόνα 23, ας υποθέσουμε ότι βρισκόμαστε στον κόμβο D1 και ο χρήστης ζητά όρους που μπορούν να ληφθούν μόνο από τη βάση δεδομένων D8. Όπως φαίνεται ωστόσο στο γράφημα, δεν μπορεί να τεθεί ερώτημα απ' ευθείας στο D8 με τον βασικό όρο ερωτήματος, αλλά πρέπει πρώτα να τεθούν ερωτήματα σε άλλες βάσεις δεδομένων όπως στο D3 έτσι ώστε να ληφθούν επιπλέον πληροφορίες, ακολουθώντας τις εξαρτήσεις των βάσεων δεδομένων, προκειμένου να είναι δυνατόν να τεθεί ερώτημα στο D8. Οι βάσεις δεδομένων οι οποίες παρεμβάλλονται μεταξύ της τρέχουσας (D1) και της D8 ονομάζονται *κρυμμένοι κόμβοι*.

Οι κρυμμένοι κόμβοι καθορίζονται βρίσκοντας όλους τους κόμβους που είναι προσβάσιμοι από τον αρχικό κόμβο μέσω του γράφου εξαρτήσεων. Για να επιτευχθεί αυτό, προσαρμόζεται η μέθοδος της αναζήτησης *πρώτα κατά πλάτος* έτσι ώστε να εφαρμοστεί σε έναν υπεργράφο.

Ο αλγόριθμος *Find_Reachable_Node (DG,s)* βρίσκει όλους τους προσβάσιμους κόμβους ξεκινώντας από έναν κόμβο εκκίνησης *s* στον γράφο εξαρτήσεων DG. Ο αλγόριθμος *Find_Reachable_Node (DG,s)* παρουσιάζεται ακολούθως όπου το Q1 είναι μια δομή δεδομένων τύπου ουράς η οποία αποθηκεύει όλους τους προσβάσιμους κόμβους ξεκινώντας από τον κόμβο *s,* ενώ το Q2 αποθηκεύει όλους τους προσβάσιμους κόμβους ξεκινώντας από το s με την βοήθεια των συνεργατών του. Η συνάρτηση PS(t) επιστρέφει το σύνολο συνεργατών του t.

Αλγόριθμος 4.1: FindReachableNodes(*DG, s*)

Initialize two queues Q1 and Q2

Add s to Q1, and mark s as visited

while Q1 is not empty

  Dequeue the first element in Q1, name it as t

  foreach n which is a neighbor of t

    if $n \in unvisited$ **and** $PS(t) = \Phi$ **and** rules match

      Add n to Q1 and mark n as visited

    else if $n \in unvisited$ **and** $PS(t) \neq \Phi$

      Add n to Q2

    while Q2 is not empty

      foreach $n \in Q2$

        Extract the partner set PS of n

        Denote each partner of n as p

        if $p \in Q1$

          foreach p of n **and** rules match

            Add n to Q1, and remove n from Q2

            Mark n as visited

return (*Q*1)

Εισάγεται επίσης η έννοια "*αναγκαιότητα των βάσεων δεδομένων" (Database Necessity).* Κάθε κανόνας παραγωγής σχετίζεται με ένα σύνολο όρων οι οποίοι μπορούν να εξαχθούν με την εκτέλεση του κανόνα. Υπάρχουν όροι οι οποίοι μπορούν

μόνο να παρασχεθούν από μία μόνο βάση δεδομένων, ενώ άλλοι όροι μπορούν να παρασχεθούν από πολλαπλές βάσεις δεδομένων. Ένας ζητούμενος όρος μπορεί μόνο να παρασχεθεί από έναν μόνο κανόνα, τότε εκείνος ο κανόνας πρέπει να έχει υψηλή προτεραιότητα ώστε να εκτελεστεί. Αντίθετα, εάν ο όρος μπορεί να παρασχεθεί από πολλούς κανόνες, σε αυτόν τον κανόνα μπορεί να οριστεί μια χαμηλότερη προτεραιότητα. Με βάση αυτό, ο κάθε όρος συνδέεται με μια τιμή *"αναγκαιότητας των βάσεων δεδομένων"*. Για έναν όρο *t,* εάν οι K βάσεις δεδομένων μπορούν να τον παρέχουν, η τιμή της αναγκαιότητας των βάσεων δεδομένων για το *t* είναι 1/*K*.

Οι κρυμμένοι κανόνες πρέπει να γίνουν μερικώς ορατοί στην επιφάνεια. Οι ιδιότητες ενός κρυμμένου αλλά χρήσιμου κανόνα είναι οι εξής:

- Πρέπει να εκτελεσθεί ώστε να εξαχθούν όλοι οι όροι που ζητήθηκαν από τον χρήστη.
- Τα απαραίτητα στοιχεία εισόδου του είναι κρυμμένα, δηλαδή είτε δεν είναι περιλαμβάνονται στους όρους-στόχους της ερώτησης, είτε μπορούν να εξαχθούν μόνο από τους κανόνες που βρίσκονται στο κρυμμένο επίπεδο.

Οι κρυμμένοι κανόνες γίνονται ορατοί με μία ανοδική προσέγγιση. Εξετάζονται όλοι οι όροι που δόθηκαν ως όροι-στόχοι της ερώτησης, και που αποτελούν τον αρχικό χώρο-στόχο. Αν υπάρχει κάποιος όρος στον αρχικό χώρο-στόχο με τιμή αναγκαιότητας βάσης δεδομένων ίση με 1 (δηλ. υπάρχει μόνο ένας κανόνας – έστω ο R που μπορεί να παρέχει αυτόν τον όρο), τότε ο κανόνας αυτός είναι ένας κρυμμένος κανόνας και πρέπει να πυροδοτηθεί. Προκειμένου να καταστεί ο R ορατός, προσθέτουμε τα αναγκαία στοιχεία εισόδου του R στο χώρο-στόχο, έτσι ώστε να επεκτείνουμε τον χώρο-στόχο. Ακολούθως, εξετάζεται ο πρόσφατα διευρυμένος χώρος-στόχος ώστε να εντοπισθούν νέοι κρυμμένοι κανόνες και να επεκταθεί περαιτέρω ο χώρος-στόχος. Η διαδικασία αυτή συνεχίζεται μέχρις ότου να μην υπάρχουν άλλοι κρυμμένοι κανόνες.

Στον αλγόριθμο τίθεται το θέμα που αφορά τη δυνατότητα απαλοιφής σχεδίων ερωτημάτων που είναι παρόμοια μεταξύ τους. Ο δυναμικός σχεδιαστής ερωτημάτων παράγει τα *K* κορυφαία σχέδια ερωτημάτων. Όταν ένα σχέδιο ερωτημάτων παράγεται, ο αλγόριθμος εκτελεί μια αναζήτηση από το τρέχον σημείο προς τα πίσω για να παράγει και άλλα σχέδια ερωτημάτων. Είναι πιθανό, δύο παραγόμενα σχέδια ερωτημάτων *QP*1 και *QP*2 να χρησιμοποιούν το ίδιο σύνολο βάσεων δεδομένων με ενδεχόμενη διαφορά στη σειρά που ερωτώνται δύο ή περισσότερες βάσεις δεδομένων που δεν έχουν εξαρτήσεις. Σε μια τέτοια περίπτωση τα σχέδια ερωτήματος *QP*1 και *QP*2 θεωρούνται ίδια και το τελευταίο από αυτά θα διαγραφεί.

Αλγόριθμος 4.2: Find Top K Query Plans(*PR,WS,TS*)

while enlargeable(WS)

  Enlarge WS

Initialize queue Q and P

while *size*(*Q*) ≤ *K*

  if (∃*e* ∈ *TS* **and** *e* ∉ *WS*)

    and (∃*r* ∈ *PR* **and** ∃*o* ∈ *O*(*r*) **and** *o* ∈ *TS*)

      Find candidate rule set CR

      foreach *r* ∈ *CR*

        Compute benefit score according to benefit model

      Select r opt, the rule with the highest benefit

      if !*prunable*(*P, r opt*)

        while *r opt* ≠ *null* **and** (∃*e* ∈ *TS* **and** *e* ∉ *WS*)

        and (∃*r* ∈ *PR* **and** ∃*o* ∈ *O*(*r*) **and** *o* ∈ *TS*)

          Add r opt to P, and update WS

          Select next r opt

        else Empty queue P

    else Add P to Q and re-order Q

  if *size*(*P*) > 0

    Remove the last rule of P, update WS, trace back

return (*Q*)

Κύριος αλγόριθμος. Ο δυναμικός αλγόριθμος σχεδίασης ερωτήματος χρησιμοποιεί τους όρους-στόχους της ερώτησης και τον όρο-κλειδί ερωτήματος, και παράγει με δυναμικό τρόπο *K* σχέδια ερωτήματος τα οποία καλύπτουν όσο το δυνατόν περισσότερους όρους από αυτούς που έχουν ζητηθεί. Αυτό το επιτυγχάνει μεγαλώνοντας τον χώρο-στόχο του χρήστη για να γίνουν ορατοί οι κρυμμένοι κανόνες και να υπολογισθεί ο διευρυμένος χώρος-στόχος. Η μνήμη εργασίας αρχικοποιείται με τον βασικό όρο του ερωτήματος και το σύστημα παραγωγής ξεκινά τη διαδικασία αναγνώριση-ενέργεια. Το σύστημα παραγωγής σε κάθε επανάληψη επιλέγει έναν κανόνα κατάλληλο ακολουθώντας το μοντέλο κέρδους και ενημερώνει την τρέχουσα μνήμη εργασίας. Όταν καλυφθούν όλοι οι όροι στον χώρο-στόχο ή δεν μπορούν να πυροδοτηθούν άλλοι κανόνες η διαδικασία ολοκληρώνεται.

Με βάση το στόχο και την τρέχουσα μνήμη εργασίας, το μοντέλο κέρδους επιλέγει τον καλύτερο κανόνα. Στον παραπάνω Αλγόριθμο 4.2 παρουσιάστηκε η δομή του αλγορίθμου σχεδίασης ερωτημάτων. Το *PR* δηλώνει το σύνολο κανόνων παραγωγής, το *WS* τη μνήμη εργασίας και το *TS* τον χώρο-στόχο. Το *Q* είναι μια ουρά που αποθηκεύει τα *K* κορυφαία σχέδια ερωτημάτων και το *P* είναι μια ουρά που αποθηκεύει τους κανόνες ενός σχεδίου ερωτημάτων. Η *O(r)* συνάρτηση επιστρέφει τα στοιχεία εξόδου ενός κανόνα παραγωγής *r*, ενώ η συνάρτηση *prunable()* εξετάζει αν μπορεί να παραλειφθεί ένα υποψήφιο σχέδιο ερωτήματος.

Ο αλγόριθμος είναι ένας άπληστος αλγόριθμος που επιλέγει τον κανόνα παραγωγής ανάλογα με το τοπικά μέγιστο όφελος ακολουθώντας το μοντέλο κέρδους. Κάθε άπληστος αλγόριθμος διαθέτει ένα λόγο προσέγγισης που μετρά την απόδοση του αλγορίθμου. Ο συμβολισμός $|R|$ αναπαριστά τον πληθάριθμο της συλλογής των κανόνων *R,* δηλαδή τον συνολικό αριθμό κανόνων παραγωγής. Στην εργασία [9] διατυπώνεται το θεώρημα ότι ο προσεγγιστικός αλγόριθμος που περιγράφεται στον Αλγόριθμο 4.2 έχει λόγο προσέγγισης $\frac{|R|+1}{2|R|}$.

4.1.3 Μοντέλο ωφέλειας

Το μοντέλο ωφέλειας επιλέγει ένα κατάλληλο κανόνα για κάθε επανάληψη του κύκλου αναγνώρισης-ενέργειας. Στον Αλγόριθμο που παρουσιάστηκε για την επιλογή του κανόνα χρησιμοποιήθηκαν οι τέσσερις μετρικές που ακολουθούν:

Διαθεσιμότητα βάσης δεδομένων (DA): Ένας κανόνας παραγωγής *R* εκτελείται όταν η αντίστοιχη βάση δεδομένων είναι διαθέσιμη. Για κάθε κανόνα στέλνεται ένα μήνυμα στη βάση δεδομένων ώστε να ελεγχθεί η διαθεσιμότητα της. Στη περίπτωση που δεν είναι διαθέσιμη ο κανόνας για την τρέχουσα επανάληψη αγνοείται.

Κάλυψη δεδομένων (DC): Το ποσοστό των ζητούμενων δεδομένων που μπορούν να παρασχεθούν από ένα μόνο κανόνα αντικατοπτρίζεται με τη μετρική της κάλυψης δεδομένων. Η κάλυψη δεδομένων του τρέχοντος κανόνα R_k σε σχέση με το TS υπολογίζεται λαμβάνοντας υπόψη τον κανόνα R_k, την κατάσταση του χώρου-στόχου *TS* και τους *k-1* κανόνες $R_1, R_2, \dots, R_{k-1}$ που έχουν επιλεγεί. Για τον ίδιο λόγο, χρησιμοποιούμε το πλήθος των όρων-στόχων στον χώρο-στόχο TS που καλύπτονται από τον κανόνα R_k αλλά που δεν έχουν εξαχθεί ακόμη από κάποιον από τους κανόνες που έχουν εφαρμοσθεί.

Προτίμηση χρηστών (UP): Μερικοί όροι μπορούν να εξαχθούν από περισσότερες της μίας βάσεις δεδομένων, και οι χρήστες είναι δυνατόν να έχουν εκφράσει τον βαθμό προτίμησής τους για κάθε μία από αυτές. Για κάθε όρο, μπορούμε να αντιστοιχίσουμε έναν βαθμό προτίμησης που δηλώνει την προτίμηση του χρήστη για κάθε μία από τις βάσεις δεδομένων σχετικά με την εξαγωγή του συγκεκριμένου όρου και να ορίσουμε μία συνάρτηση ωφέλειας που να λαμβάνει υπ' όψιν αυτόν τον βαθμό. Ας θεωρήσουμε έναν συγκεκριμένο όρο *t,* ο οποίος μπορεί να εξαχθεί από *r* βάσεις δεδομένων $D_1, D_2, \dots, D_r$. Σε κάθε μία από τις βάσεις δεδομένων $D_1, D_2, \dots, D_r$, μπορεί να αντιστοιχηθεί ένας αριθμός β_i με $0 \leq \beta_i \leq 1$, ως βαθμός προτίμησης για την

εξαγωγή του συγκεκριμένου όρου. Θα πρέπει $\sum_{i=1}^{r} \beta_i = 1$. Αν ο όρος t μπορεί να εξαχθεί μόνο από μια βάση δεδομένων, τότε θα πρέπει ο βαθμός αυτής της βάσης δεδομένων, σε σχέση με τον συγκεκριμένο όρο, να είναι ίσος με 1 και όλων των υπόλοιπων βάσεων δεδομένων να είναι ίσος με 0. Οι τιμές για τους βαθμούς θα δίνονται από έναν ειδικό περί το πεδίο.

Ας υποθέσουμε ότι εξετάζεται ο κανόνας παραγωγής R, ο οποίος συνδέεται με τη βάση δεδομένων D και έστω ότι οι k το πλήθος όρων $UF_1, UF_2, \dots, UF_k$ δεν έχουν βρεθεί. Θεωρούμε ότι η προτίμηση του χρήστη που αφορά τη βάση δεδομένων D για κάθε όρο UF_i,είναι UP_i. Χρησιμοποιούμε την τιμή αναγκαιότητας βάσεων δεδομένων για κάθε όρο (έστω ότι είναι DN_i για κάθε όρο UF_i) ως το βάρος της προτίμησης χρήστη για τη συγκεκριμένη βάση δεδομένων και υπολογίζουμε το σταθμισμένο άθροισμα όλων των όρων που δεν έχουν βρεθεί προκειμένου να υπολογίσουμε την τιμή προτίμησης χρήστη για τον κανόνα. Με βάση τα ανωτέρω η προτίμηση χρηστών του R να είναι:

$$\sum_{i=1}^{k} DN_i * UP_i$$

Πιθανή σπουδαιότητα (PI): Μερικές από τις βάσεις δεδομένων εμφανίζονται ως περισσότερο σημαντικές λόγω της σύνδεσή τους με άλλες σημαντικές βάσεις δεδομένων, θεωρώντας τις εξαρτήσεις των βάσεων δεδομένων. Στην Εικόνα **23**, υποθέτοντας ότι οι D2 και D3 έχουν την ίδια κάλυψη δεδομένων και προτιμήσεις χρήστη, η D3 έχει μεγαλύτερη *πιθανή σπουδαιότητα* επειδή μπορεί να βοηθήσει στην σύνδεση με τον τελικό στόχο D8. Συνακόλουθα, στο D3 ορίζεται μια μεγαλύτερη τιμή ωφέλειας, και με τον τρόπο αυτό ενσωματώνεται η πιθανή σπουδαιότητα στη συνάρτηση ωφέλειας.

Έστω ότι εξετάζουμε έναν κανόνα παραγωγής που αντιστοιχεί στη βάση δεδομένων *D*. Χρησιμοποιώντας τον Αλγόριθμο 4.1 βρίσκουμε ένα σύνολο βάσεων δεδομένων $D_{reachable} = \{D_1, D_2, \dots, D_m\}$, στο οποίο υποβάλλουμε ερωτήσεις με τη χρήση των δεδομένων που εξήχθησαν αποκλειστικά από τη βάση δεδομένων *D*. Έστω ότι έχουμε k όρους που δεν έχουν βρεθεί ακόμη και τους συμβολίζουμε με $UF_1, UF_2, \dots, UF_k$. Για τον όρο UF_i, η τιμή αναγκαιότητας των βάσεων δεδομένων της είναι DN_i, δηλ. ο όρος UF_i μπορεί να ληφθεί από $1/DN_i$ βάσεις δεδομένων. Αυτό το σύνολο βάσεων συμβολίζεται ως $NecessaryD_i$. Θέλουμε να γνωρίζουμε το πλήθος των αναγκαίων βάσεων δεδομένων του UF_i που μπορούν να προσπελαστούν από τον κανόνα παραγωγής *R*. Μετράται ο αριθμός των βάσεων δεδομένων στο $NecessaryD_i$, οι οποίες επίσης ανήκουν στο σύνολο $D_{reachable}$, δηλαδή υπολογίζεται ο πληθάριθμος του συνόλου $\{d | d \in NecessaryD_i, d \in D_{reachable}\}$. Ας υποθέσουμε ότι ο πληθάριθμος για τον όρο UF_i είναι r_i. Η πιθανή σπουδαιότητα για το UF_i όσον αφορά τον κανόνα *R* και την αντίστοιχη βάση δεδομένων *D* είναι:

$$\frac{r_i * \frac{1}{DN_i}}{|D_{reachable}|} = \frac{r_i}{m * DN_i}$$

Αντιστοίχως, η πιθανή σπουδαιότητα για τον κανόνα *R* είναι:

$$\sum_{i=1}^{k} \frac{r_i}{m * DN_i}$$

Η τιμή της μετρικής ωφέλειας για έναν υποψήφιο κανόνα R υπολογίζεται με βάση τις ακόλουθες τρεις μετρικές:

1. *την κάλυψη δεδομένων*,
2. *την προτίμηση χρηστών*,
3. *την πιθανή σπουδαιότητα*.

Οι τιμές των τριών μετρικών συσχετίζονται με τις τιμές αναγκαιότητας των βάσεων δεδομένων για τους όρους που δεν έχουν βρεθεί, κατά την εξέταση ενός κανόνα. Στην περίπτωση που ένας κανόνας εξετάζεται ως υποψήφιος πολλές φορές, κάθε φορά η τιμή της μετρικής ωφέλειας είναι διαφορετική κι' αυτό συμβαίνει διότι κάθε φορά το σύνολο των όρων που δεν έχουν βρεθεί είναι διαφορετικό. Επομένως η τιμή της μετρικής ωφέλειας ενός κανόνα παραγωγής συσχετίζεται δυναμικά με την τρέχουσα κατάσταση του χώρου εργασίας του συστήματος παραγωγής. Η συνάρτηση ωφέλειας ενός κανόνα *R*, σε σχέση με τον τρέχοντα χώρο εργασίας *WS* δηλώνεται ως εξής:

$$BF(R, WS) = DC * \alpha + UP * \beta + PI * \gamma, \qquad \alpha + \beta + \gamma = 1$$

όπου οι παράμετροι *α, β* και *γ* είναι συνδεμένες με κάθε μετρική.

4.1.4 Επεκτασιμότητα του συστήματος

Για ένα σύστημα εξόρυξης του κρυμμένου ιστού η επεκτασιμότητα είναι σημαντική επειδή μπορεί να προκύψουν νέες πηγές δεδομένων. Μια νέα πηγή δεδομένων ενσωματώνεται στο σύστημα, αφού πρώτα αναπαρασταθούν τα σχήματα ερωτημάτων προς τη βάση δεδομένων της νέας πηγής δεδομένων στη μορφή των κανόνων παραγωγής. Ακολούθως, ο ειδικός περί το πεδίο είναι αυτός ο οποίος ορίζει ή αλλάζει τις τιμές προτίμησης του χρήστη για τους όρους που εμφανίζονται στις πρόσφατα ενσωματωμένες πηγές δεδομένων. Για την αυτόματη ενσωμάτωση της νέας πηγής δεδομένων στον γράφο εξαρτήσεων και την ενημέρωση των τιμών αναγκαιότητας των βάσεων δεδομένων αναπτύχθηκαν απλοί αλγόριθμοι. Οι αλγόριθμοι αυτοί διαθέτουν ικανότητα κλιμάκωσης σε σχέση με το πλήθος των βάσεων δεδομένων. Ο αλγόριθμος σχεδίασης ερωτημάτων και η σχεδίαση του γράφου εξαρτήσεων στηρίζονται στα εγγενή χαρακτηριστικά των πηγών δεδομένων του κρυμμένου ιστού, όπως για παράδειγμα οι εξαρτήσεις βάσεων δεδομένων και τα σχήματα βάσεων δεδομένων. Το σύστημα είναι ανεξάρτητο από το πεδίο εφαρμογής, δηλαδή μπορεί να εφαρμοστεί σε οποιοδήποτε πεδίο εφαρμογής.

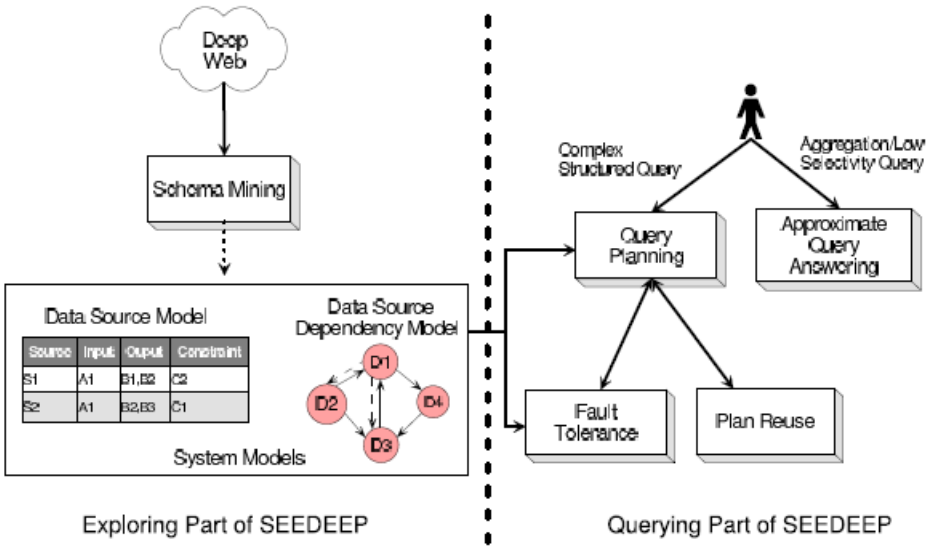
4.2 Το σύστημα SEEDEEP

Το SEEDEEP είναι ένα αυτόματο σύστημα εξερεύνησης και ερώτησης των πηγών δεδομένων του κρυμμένου ιστού [10]. Το σύστημα SEEDEEP ολοκληρώνει τις πήγες δεδομένων του κρυμμένου ιστού σε μια συγκεκριμένη περιοχή του δικτύου, και παρέχει στους χρήστες τη δυνατότητα αναζήτησης στηριζόμενη σε λέξεις-κλειδιά.

Το σύστημα SEEDEEP αποτελείται από τα ακόλουθα πέντε δομικά στοιχεία (modules):

1. την εξόρυξη σχήματος (schema mining),
2. τον σχεδιαστή ερωτήματος (query planning),
3. την προσεγγιστική απάντηση ερωτήματος (approximate query answering),
4. την επαναχρησιμοποίηση ερωτήματος (query reuse),
5. και την ανοχή σε σφάλματα (fault tolerance).

Στην Εικόνα 24 που ακολουθεί παρουσιάζεται η δομή του συστήματος SEEDEEP.



Εικόνα 24: Δομή συστήματος SEEDEEP

4.2.1 Εξόρυξη σχήματος

Η ανάκτηση μεταδεδομένων από κάθε πηγή δεδομένων είναι αναγκαία για την ολοκλήρωση δεδομένων από πολλαπλές πηγές δεδομένων του κρυμμένου ιστού. Η ανάκτηση των μεταδεδομένων, και ειδικότερα του σχήματος εξόδου, είναι ιδιαίτερα πολύπλοκη, διότι δοθέντος ενός ερωτήματος πολλές πηγές δεδομένων του κρυμμένου ιστού επιστρέφουν ένα υποσύνολο των γνωρισμάτων του σχήματος εξόδου για το συγκεκριμένο ερώτημα. Το δομικό στοιχείο εξόρυξης σχήματος έχει ως λειτουργία την αυτόματη εξόρυξη των μεταδεδομένων των πηγών δεδομένων του κρυμμένου ιστού και τη δημιουργία μοντέλων δεδομένων για τα υπόλοιπα δομικά στοιχεία του συστήματος.

Στο δομικό σχήμα εξόρυξης σχήματος για την αυτόματη εξόρυξη των μεταδεδομένων χρησιμοποιούνται δύο τεχνικές:

1. Η πρώτη τεχνική είναι η *προσέγγιση δειγματοληψίας σχήματος* (sampling model approach), η οποία βασίζεται σε ένα μοντέλο κατανομής των γνωρισμάτων εξόδου των πηγών δεδομένων του κρυμμένου ιστού. Με βάση αυτή την προσέγγιση, ένα δείγμα μετρίου μεγέθους από σελίδες εξόδου επιτυγχάνει υψηλή ανάκληση σε ό,τι αφορά τα γνωρίσματα εξόδου.
2. Η δεύτερη τεχνική είναι *το μεικτό μοντέλο* (mixture model approach), το οποίο βασίζεται στην παρατήρηση ότι είναι πιθανός ο πλεονασμός στις πηγές δεδομένων με αποτέλεσμα τα γνωρίσματα να μπορούν να διαμοιραστούν μεταξύ διαφορετικών πηγών δεδομένων. Έτσι, ένα γνώρισμα μπορεί να υποτεθεί ότι ανήκει στο σχήμα εξόδου μίας πηγής, με μία πιθανότητα που υπολογίζεται βάσει των εμφανίσεων του γνωρίσματος στο σχήμα εξόδου παρόμοιων πηγών.

4.2.2 Σχεδιαστής Ερωτήματος

Οι περισσότερες πηγές δεδομένων του κρυμμένου ιστού υποστηρίζουν απλές διεπαφές αναζήτησης, οι οποίες διαθέτουν ένα μοναδικό πλαίσιο κειμένου για εισαγωγή λέξεων-κλειδιών, όπως παρουσιάστηκε στο κεφάλαιο 4. Ένα από τα πλεονεκτήματα αυτών των διεπαφών είναι η απλότητα ενώ ένα μειονέκτημα τους είναι ότι υπολείπονται σε εκφραστική ικανότητα. Κατά συνέπεια, η επεξεργασία πολύπλοκων δομημένων ερωτήσεων σε πηγές του κρυμμένου ιστού είναι πολύπλοκη διαδικασία και απαιτεί σημαντική εργασία εκτός των πηγών και χειροκίνητη εργασία. Για τον λόγο αυτό, αναπτύχθηκαν τεχνικές οι οποίες επεξεργάζονται πολύπλοκα δομημένα ερωτήματα στις πηγές δεδομένων του κρυμμένου ιστού, με κυριότερες τις ακόλουθες:

1. *Ολοκλήρωση μέσω ανάδειξης* (surfacing integration): ένας τρόπος για την υποστήριξη πολύπλοκων δομημένων ερωτημάτων σε δεδομένα του κρυμμένου ιστού, είναι να αναδεικνύονται (surface) και να αποθηκεύονται τα δεδομένα σε μια σχεσιακή βάση δεδομένων, και στη συνέχεια να χρησιμοποιούνται οι κλασικές γλώσσες ερωτήσεων. Ωστόσο, τα δεδομένα σε πολλές πηγές δεδομένων του κρυμμένου ιστού ενημερώνονται συχνά, με αποτέλεσμα η τεχνική της ολοκλήρωσης μέσω ανάδειξης να οδηγεί σε προβλήματα συνέπειας των δεδομένων.
2. Εικονική ολοκλήρωση (virtual integration): η προσέγγιση αυτή δημιουργεί ένα ενδιάμεσο σχήμα χρησιμοποιώντας τεχνικές αντιστοίχισης σχήματος για τις πηγές δεδομένων. Ακολούθως ο χρήστης θέτει ερωτήματα στο ενδιάμεσο σχήμα, το οποίο παρουσιάζεται στους χρήστες ως διαδικτυακή φόρμα.

Στο σύστημα SEEDEEP το δομικό στοιχείο σχεδιαστή ερωτήματος υποστηρίζει πολύπλοκα δομημένα ερωτήματα στο σύνολο των πηγών δεδομένων του κρυμμένου ιστού που έχουν ολοκληρωθεί στο σύστημα. Οι κατασκευαστές του έχουν

δημιουργήσει ένα πλαίσιο για υποβολή ερωτήσεων σε ένα σύνολο ολοκληρωμένων πηγών δεδομένων του κρυμμένου ιστού.

Οι τρείς τύποι προηγμένων ερωτημάτων που υποστηρίζονται από το δομικό στοιχείο σχεδιαστή ερωτήματος είναι οι εξής:

1. τα ερωτήματα επιλογής-προβολής-συνένωσης SPJ (SPJ queries),
2. τα ερωτήματα συνάθροισης (aggregation queries),
3. και τα εμφωλευμένα ερωτήματα (nested queries).

Με βάση τα μεταδεδομένα που εξήχθησαν από τη χρήση του δομικού στοιχείου εξόρυξης σχήματος, δημιουργείται ένας κατευθυνόμενος γράφος αλληλεξάρτησης των πηγών δεδομένων, από τον οποίο φαίνονται οι εξαρτήσεις μεταξύ των συσχετιζόμενων πηγών δεδομένων. Ο τρόπος με τον οποίο υποστηρίζονται οι ανωτέρω τύποι ερωτημάτων έχει ως ακολούθως:

1. Εάν το ερώτημα περιλαμβάνει εμφωλευμένα ερωτήματα, τότε αποσυνθέτουμε το ερώτημα σε μια λίστα από απλά ερωτήματα, τα οποία δεν περιέχουν δομές εμφώλευσης. Ακολούθως για κάθε απλό ερώτημα, εξάγουμε τους όρους αναζήτησης και τα απαραίτητα κατηγορήματα για την κατανόηση του τύπου του ερωτήματος.
2. Οι νέοι αλγόριθμοι σχεδίασης ερωτήματος χρησιμοποιούνται για την παραγωγή ενός σχεδίου ερωτήματος για κάθε απλό ερώτημα.
3. Τα σχέδια ερωτήματος για όλα τα απλά ερωτήματα που προήλθαν από τη διάσπαση του αρχικού ερωτήματος, συνδυάζονται και ενώνονται για να διαμορφωθεί το τελικό σχέδιο ερωτήματος.
4. Στο τέλος εφαρμόζονται τεχνικές βελτιστοποίησης ώστε η διαδικασία εκτέλεσης να γίνει ταχύτερα.

4.2.3 Προσεγγιστική απάντηση ερωτήματος

Το δομικό στοιχείο προσεγγιστικής απάντησης ερωτήματος βρίσκει απαντήσεις για τα ερωτήματα συνάθροισης τα οποία χρειάζονται απαρίθμηση δεδομένων καθώς και ερωτήματα τα οποία δεν έχουν μεγάλη επιλεκτικότητα, στην περίπτωση περιορισμένης πρόσβασης δεδομένων στον κρυμμένο ιστό. Εκτιμάται ότι οι χρήστες θα είναι ικανοποιημένοι αν τους δοθεί μια προσεγγιστική μεν, επαρκώς ακριβής δε απάντηση. Για να γίνει αυτό χρειάζεται αποδοτική και αποτελεσματική δειγματοληψία.

Η τυχαία δειγματοληψία μπορεί να μην είναι αποτελεσματική για πολλές πηγές δεδομένων του κρυμμένου ιστού διότι έχει υψηλό κόστος δειγματοληψίας και χαμηλή ακρίβεια εκτίμησης σε δεδομένα που δεν είναι ομοιόμορφα κατανεμημένα. Στην εργασία [10], προτείνονται τεχνικές υποστήριξης προσεγγιστικών ερωτημάτων συνάθροισης στις πηγές δεδομένων του κρυμμένου ιστού με δεδομένα που δεν είναι ομοιόμορφα κατανεμημένα και παρουσιάζονται δύο προσαρμοστικοί αλγόριθμοι δειγματοληψίας:

1. ο προσαρμοστικός αλγόριθμος δειγματοληψίας γειτονιάς (Adaptive Neighborhood Sampling algorithm),
2. ο προσαρμοστικός αλγόριθμος δειγματοληψίας δύο φάσεων βασισμένος στην έννοια του υποχώρου (sub-space based Two Phase adaptive Sampling algorithm).

Οι δύο αλγόριθμοι παρουσιάζουν τα ακόλουθα πλεονεκτήματα σε σχέση με άλλους αλγορίθμους:

1. Παρέχουν ακριβείς εκτιμήσεις για ερωτήματα συνάθροισης σε δεδομένα που δεν είναι ομοιόμορφα κατανεμημένα, χωρίς να χρειάζεται να γνωρίζουν τη κατανομή των κρυμμένων δεδομένων ή άλλες στατιστικές πληροφορίες.
2. για τα δεδομένα με μικρή απόκλιση από την ομοιόμορφη κατανομή, ο δεύτερος αλγόριθμος επιτυγχάνει μικρότερο κόστος δειγματοληψίας.

Στα ερωτήματα χαμηλής επιλεκτικότητας χρησιμοποιείται η διαστρωμάτωση ώστε να εντοπιστούν οι ακριβείς απαντήσεις. Αυτό επιτυγχάνεται με τον διαχωρισμό του συνόλου των δεδομένων σε μικρότερα υποσύνολα, τα οποία ονομάζονται *στρώματα*, έτσι ώστε οι τιμές των γνωρισμάτων που χρησιμοποιούνται για επιλογή στην ερώτηση να γίνουν σχετικά ομοιογενείς μέσα στο κάθε στρώμα. Αυτό επιτυγχάνεται καλύτερα όταν η διακύμανση των γνωρισμάτων που χρησιμοποιούνται για επιλογή ελαχιστοποιείται εντός του στρώματος και μεγιστοποιείται μεταξύ διαφορετικών στρωμάτων. Για ορισμένα γνωρίσματα που κρίνονται πιο σημαντικά, χρησιμοποιείται αυξημένο ποσοστό δειγματοληψίας ώστε να προσεγγιστεί με μεγαλύτερη ακρίβεια το αποτέλεσμα της ερώτησης.

Με βάση τα ανωτέρω είναι σημαντικός ο προσδιορισμός της βέλτιστης διαστρωμάτωσης, δηλαδή να εντοπιστούν τα σημεία των γνωρισμάτων που χρησιμοποιούνται για επιλογή στα οποία μπορεί να εφαρμοστεί η διαστρωμάτωση.

Στο σύστημα SEEDEEP χρησιμοποιείται ένας προσαρμοστικός Bayesian αλγόριθμος διαστρωμάτωσης αρμονικής αναζήτησης. Χρησιμοποιώντας τον αλγόριθμο η διαστρωμάτωση που προκύπτει αντικατοπτρίζει με ακρίβεια την κατανομή του κρυμμένου γνωρίσματος πάνω στο οποίο γίνεται η επιλογή, και η ακρίβεια υπολογισμού, ακολουθώντας τη νέα μέθοδο, είναι υψηλότερη σε σχέση με την ακρίβεια άλλων μεθόδων.

4.2.4 Επαναχρησιμοποίηση ερωτήματος

Η εκτέλεση ενός σχεδίου ερωτήματος προϋποθέτει την πρόσβαση σε μια ακολουθία αλληλοεξαρτώμενων πηγών δεδομένων του κρυμμένου ιστού. Η απάντηση του ερωτήματος με τον τρόπο αυτό απαιτεί την αρκετό χρόνο λόγω του φόρτου εργασίας του εξυπηρέτη και την καθυστέρηση του δικτύου. Το δομικό στοιχείο επαναχρησιμοποίησης ερωτήματος παρέχει ένα αποδοτικό μηχανισμό προσωρινής αποθήκευσης αποτελεσμάτων ερωτήματος, ο οποίος επιταχύνει σε μεγάλο βαθμό την εκτέλεση ενός σχεδίου ερωτήματος, χρησιμοποιώντας τα πλεονεκτήματα των προσωρινά αποθηκευμένων δεδομένων.

Στο σύστημα SEEDEEP προτείνεται μια τεχνική προσωρινής αποθήκευσης αποτελεσμάτων ερωτημάτων που οδηγείται από τα σχέδια ερωτημάτων. Η τεχνική αυτή αποθηκεύει τα δεδομένα που έχουν εξαχθεί καθώς και τα σχέδια ερωτήματος για τα ερωτήματα. Συγκεκριμένα, παράγεται ένα σχέδιο ερωτήματος για το νέο ερώτημα επαναχρησιμοποιώντας τα προσωρινά αποθηκευμένα σχέδια ερωτήματος τα οποία έχουν χρησιμοποιηθεί ήδη σε προηγούμενα ερωτήματα και με τον τρόπο αυτό αυξάνεται η πιθανότητα επαναχρησιμοποίησης δεδομένων.

Η αποδοτική επαναχρησιμοποίηση δεδομένων μετατρέπει την απομακρυσμένη πρόσβαση πηγών δεδομένων σε πρόσβαση στον τοπικό δίσκο, και αυτό έχει ως αποτέλεσμα να μειώνεται ο χρόνος εκτέλεσης ενός σχεδίου ερωτήματος. Το δομικό στοιχείο επαναχρησιμοποίησης ερωτήματος δίνοντας ένα νέο ερώτημα NQ, ψάχνει μια λίστα από προηγούμενα ερωτήματα τα οποία δηλώνονται ως PQs = {PQ_1, …, PQ_n} και τα οποία είναι παρόμοια του NQ βάση μιας μετρικής συσχέτισης. Ακολούθως, ένας αλγόριθμος επιλογής Ψ επιλέγει από κάθε PQ_i ένα υπό-σχέδιο ερωτήματος $SubP_i$, το οποίο περιέχει όλα τα έγκυρα υπό-σχέδια του PQ_i. Το $SubP_i$, καλύπτει μεγιστικά (maximally covers) το NQ και έχει το μικρότερο δυνατό μέγεθος.

Το σχέδιο ερωτήματος για το NQ παράγεται με τη χρήση της λίστας των $SubP_i$, βασισμένο σε έναν τροποποιημένο αλγόριθμο σχεδίασης ερωτήματος, ο οποίος υλοποιήθηκε στο δομικό στοιχείο σχεδιαστή ερωτήματος. Ο τροποποιημένος αλγόριθμος σχεδίασης, όταν παράγεται το σχέδιο ερωτήματος για το νέο ερώτημα NQ προσπαθεί να εντάξει στο σχέδιο επαναχρησιμοποιήσιμα υπο-σχέδια ερωτήματος που έχουν αποθηκευθεί, έτσι ώστε να είναι εφικτή η επαναχρησιμοποίηση των προσωρινώς αποθηκευμένων δεδομένων.

4.2.5 Ανοχή σε σφάλματα

Η εκτέλεση ενός σχεδίου ερωτήματος απαιτεί πρόσβαση δεδομένων σε δίκτυα ευρείας κλίμακας, προκειμένου να προσπελαστεί μεγάλος αριθμός από απομακρυσμένες πηγές δεδομένων μέσω πολυάριθμων συνδέσεων επικοινωνίας. Οι εξυπηρέτες και οι διαδικτυακές συνδέσεις είναι επιρρεπείς σε κυκλοφοριακή συμφόρηση αλλά και σε αποτυχίες που έχουν ως αποτέλεσμα μια απροσδόκητη μη διαθεσιμότητα ή μη προσβασιμότητα των πληροφοριών που πρέπει να προσπελαστούν.

Στο σύστημα SEEDEEP χρησιμοποιείται μία επαυξητική αποτίμηση ερωτήσεων που βασίζεται στον πλεονασμό. Η απλούστερη προσέγγιση θα ήταν να ορίσουμε ότι αν στην επεξεργασία του αρχικού σχεδίου ερωτήματος κάποιες πηγές δεδομένων δεν είναι διαθέσιμες, τότε απορρίπτεται τελείως η εκτέλεση του σχεδίου και παράγεται νέο σχέδιο για το αρχικό ερώτημα. Αυτή η προσέγγιση έχει ωστόσο ως μειονέκτημα ότι σπαταλάται σημαντικός όγκος εργασίας ο οποίος έχει ήδη πραγματοποιηθεί.

Βάσει της προσέγγισης που υιοθετεί το σύστημα SEEDEEP, η επεξεργασία του ερωτήματος δεν θα τερματιστεί, αλλά αναστέλλεται η εκτέλεση του μέρους του σχεδίου ερωτήματος για το οποίο δεν υπάρχει διαθεσιμότητα των πηγών δεδομένων.

Η επεξεργασία του ερωτήματος προσαρμόζεται δυναμικά, αξιοποιώντας τον πλεονασμό που υπάρχει στις πηγές δεδομένων του κρυμμένου ιστού, και ένα επί μέρους νέο σχέδιο ερωτήματος παράγεται σταδιακά βρίσκοντας νέες πηγές δεδομένων οι οποίες δεν υπήρχαν στο αρχικό στάδιο ερωτήματος ώστε να αντικαταστήσουν το υποσχέδιο εντός του οποίου πηγές κατέστησαν μη προσβάσιμες.

5 Συμπεράσματα - προοπτικές του κρυμμένου ιστού

Στα κεφάλαια που προηγήθηκαν παρουσιάστηκε η έννοια του κρυμμένου ιστού, τα χαρακτηριστικά του γνωρίσματα αλλά και ειδικότερα θέματα όπως μέθοδοι ιστοσυλλογής για τον κρυμμένο ιστό, τρόποι για το πώς συλλέγονται πληροφορίες από τις ιστοσελίδες του κρυμμένου ιστού αλλά και για τον συνδυασμό πολλαπλών πηγών δεδομένων στον κρυμμένο ιστό.

Στην εργασία παρουσιάστηκε ο κρυμμένος ιστός, στον οποίο δεν υπάρχουν στατικοί σύνδεσμοι για τις σελίδες του, και αυτό με τη σειρά του κάνει τις μηχανές αναζήτησης ανίσχυρες στο να εντοπίσουν και να επιστρέψουν ως αποτέλεσμα αυτές τις σελίδες. Το πρόβλημα αυτό έχει απασχολήσει πλήθος ερευνητών, καθώς οι σελίδες αυτές περιέχουν πολλές φορές υψηλής ποιότητας περιεχόμενο. Αυτό συμβαίνει επειδή η πρόσβαση στον μεγάλο όγκο πληροφοριών του παγκόσμιου ιστού είναι συνήθως εφικτή μέσω διεπαφών αναζήτησης, στις οποίες ο χρήστης εισάγει ένα σύνολο λέξεων-κλειδιών, ενώ για την πρόσβαση σε δεδομένα του κρυμμένου ιστού ο στόχος αυτός δεν επιτυγχάνεται μέσω των διεπαφών των μηχανών αναζήτησης.

Αρχικά στο 1ο κεφάλαιο έγινε μια βασική αναφορά στις δύο κύριες κατηγορίες που υπάρχουν στον παγκόσμιο ιστό, τον επιφανειακό ιστό (surface web) και τον κρυμμένο ιστό (deep web). Ο επιφανειακός ιστός (surface web ή visible web ή indexable web) είναι εκείνο το μέρος του παγκόσμιου ιστού που δύναται να ευρετηριοποιηθεί από τις παραδοσιακές μηχανές αναζήτησης. Αντίθετα, το κομμάτι εκείνο του παγκόσμιου ιστού που δεν είναι προσβάσιμο από τις παραδοσιακές μηχανές αναζήτησης, καλείται κρυμμένος ιστός (deep web ή invisible web).

Το 2ο κεφάλαιο επικεντρώθηκε στον κρυμμένο ιστό (deep web). Η κατανόηση της πρακτικής του έννοιας δεν είναι εύκολα αντιληπτή, όμως αυτό που στην ουσία εννοείται είναι το περιεχόμενο εκείνο το όποιο υπάρχει στον ιστό αλλά δεν μπορεί να προσπελαστεί από τις μηχανές αναζήτησης γενικού σκοπού. Το περιεχόμενο αυτό μπορεί να αφορά αρχεία, σελίδες κειμένου και οποιαδήποτε άλλη πληροφορία η οποία δεν μπορεί να ανακτηθεί από τις γνωστές μηχανές αναζήτησης.

Ακόμη έγινε αναφορά στον κρυμμένο ιστό, ο οποίος περιέχει τις ακόλουθες κατηγορίες δεδομένων: μη συνδεδεμένες σελίδες, περιεχόμενο βάσεων δεδομένων, σελίδες που δεν έχουν μορφή HTML και περιέχουν κυρίως οπτικοακουστικό υλικό, σελίδες με εκτελέσιμα ή συμπιεσμένα αρχεία, περιεχόμενο περιορισμένης πρόσβασης αλλά και δυναμικό περιεχόμενο. Στο κεφάλαιο αυτό συμπεριλήφθηκαν και τα οφέλη από τη χρήση του κρυμμένου ιστού, που είναι η μη διαθεσιμότητα περιεχομένων στον ιστό, η εξειδίκευση, η εστίαση και η ποιότητα - αξιοπιστία.

Στο 3ο κεφάλαιο παρουσιάστηκε η μεθοδολογία συλλογής πληροφοριών για τις σελίδες του κρυμμένου ιστού. Το κεφάλαιο 3 χωρίζεται σε τρείς ενότητες όπου στην πρώτη παρουσιάζεται μια προσέγγιση με βάση την οποία ένας ιστοσυλλέκτης μπορεί να παράγει αυτόματα ερωτήματα, με σκοπό να εντοπίσει και να λάβει τις σελίδες του κρυμμένου ιστού. Ειδικότερα παρουσιάζεται μια μεθοδολογία κατασκευής ενός

αποδοτικού και λειτουργικού ιστοσυλλέκτη του κρυμμένου ιστού, ο οποίος είναι σε θέση να εντοπίζει και να λαμβάνει αυτόνομα σελίδες από τον κρυμμένο ιστό. Έμφαση δίνεται στις βάσεις δεδομένων κειμένου οι οποίες υποστηρίζουν ερωτήματα λέξεων-κλειδιών.

Εν συνεχεία, περιγράφηκε ένα σύστημα περιήγησης στο περιεχόμενο του κρυμμένου ιστού. Το σύστημα αυτό σχετίζεται με υποβολές προ-υπολογισμού κάθε φόρμας HTML αλλά και με προσθήκη των προκυπτουσών σελίδων HTML στα ευρετήρια μιας μηχανής αναζήτησης. Επιπροσθέτως, γίνεται παρουσίαση κάποιων αλγορίθμων που βοηθούν ώστε να αναδειχθεί (surface) ο κρυμμένος ιστός. Παρουσιάζεται ένας νέος αλγόριθμος που χρησιμεύει στην επιλογή έγκυρων τιμών εισόδου σε πλαίσια κειμένου τα οποία δέχονται λέξεις-κλειδιά, καθώς κι ένας αλγόριθμος για την επιλογή τιμών για πλαίσια που δέχονται τιμές ενός συγκεκριμένου τύπου. Τέλος, προτείνεται ένας αλγόριθμος με αποτελεσματική λειτουργία στον προσδιορισμό των πιθανών συνδυασμών εισόδου που παράγουν κατάλληλα URLs, τα οποία μπορούν να προστεθούν στα ευρετήρια μιας μηχανής αναζήτησης.

Στην τρίτη ενότητα του κεφαλαίου 3 παρουσιάστηκε το σύστημα DeLa, το οποίο επανακατασκευάζει ένα μέρος μιας βάσης δεδομένων του κρυμμένου ιστού. Αυτό επιτυγχάνεται μέσω της αποστολής ερωτημάτων με τη χρήση φορμών HTML, οι οποίες παράγουν αυτόματα περιτυλίγματα λογικών εκφράσεων, με σκοπό την εξαγωγή δεδομένων από τις ιστοσελίδες και την αποθήκευση αυτών σε έναν πίνακα με ετικέτες. Πιο συγκεκριμένα, αντιμετωπίζεται το πρόβλημα της αυτόματης εξαγωγής δεδομένων από μια ιστοσελίδα κι εν συνεχεία της ανάθεσης ετικετών στα δεδομένα αυτά. Αυτό επιτυγχάνεται με την επικέντρωση σε ιστοσελίδες που παρέχουν πολύπλοκες φόρμες HTML για την ερώτηση βάσεων δεδομένων από τους χρήστες και όχι μέσω αναζητήσεων με λέξεις-κλειδιά. Η επίλυση αυτού του προβλήματος επιτρέπει την εξαγωγή των δεδομένων από αυτές τις ιστοσελίδες, με απώτερο στόχο τον ευκολότερο χειρισμό αυτών των δεδομένων για την περαιτέρω ανάλυσή τους.

Στο 4ο κεφάλαιο παρουσιάστηκε ο συνδυασμός πολλαπλών πηγών δεδομένων στον κρυμμένο ιστό. Ειδικότερα, στην πρόκληση που σχετίζεται με τα συστήματα κρυμμένου ιστού, και συγκεκριμένα στο ότι οι βάσεις δεδομένων του κρυμμένου ιστού συχνά δεν είναι ανεξάρτητες: για παράδειγμα τα αποτελέσματα που επιστρέφονται από μια βάση δεδομένων χρησιμοποιούνται με τη σειρά τους για την αναζήτηση σε μια άλλη βάση δεδομένων. Για ένα συγκεκριμένο ερώτημα ενός χρήστη, πολλές βάσεις είναι πιθανόν να χρειάζεται να ερωτηθούν με μια ευφυή σειρά, δηλαδή υπάρχει ανάγκη για τεχνικές που μπορούν να παράγουν σχέδια ερωτήσεων υπολογίζοντας τη συσχέτιση μεταξύ των πηγών δεδομένων, ούτως ώστε να ανακτηθούν όλες οι πληροφορίες που ζητούνται.

Στην 1η ενότητα του κεφαλαίου 4 μελετήθηκε το πώς μπορεί να υπολογιστεί ο σχεδιασμός ερωτήσεων στα πλαίσια ενός συστήματος ολοκλήρωσης του κρυμμένου ιστού. Το σύστημα που περιγράφηκε, σχεδιάστηκε για να παρέχει μια πολύ απλή και

εύκολη διεπαφή ερωτήματος, όπου κάθε ερώτημα έχει ένα όρο-κλειδί κι ένα σύνολο όρων-στόχων για τους οποίους ενδιαφέρεται ο χρήστης. Ο όρος-κλειδί είναι ένα όνομα και οι όροι-στόχοι αντανακλούν τις ιδιότητες των πληροφοριών που επιθυμούνται για αυτό το όνομα. Στα πλαίσια ενός τέτοιου συστήματος, έχει αναπτυχθεί ένας δυναμικός σχεδιαστής ερωτημάτων με σκοπό την παραγωγή μιας αποτελεσματικής σειράς ερωτημάτων, βάση των εξαρτήσεων των βάσεων δεδομένων του κρυμμένου ιστού.

Στη συνέχεια παρουσιάστηκε ένα αυτόματο σύστημα εξερεύνησης και ερώτησης των πηγών δεδομένων του κρυμμένου ιστού, το οποίο καλείται "SEEDEEP" και το οποίο είναι σε θέση να ενσωματώσει πηγές δεδομένων του κρυμμένου ιστού και να παρέχει στους χρήστες μια λειτουργία αναζήτησης βασισμένη σε λέξεις-κλειδιά. Το σύστημα SEEDEEP αποτελείται από 5 δομικά στοιχεία, τα οποία είναι το σχήμα εξόρυξης, ο σχεδιαστής ερωτήματος, η προσεγγιστική απάντηση ερωτήματος, η επαναχρησιμοποίηση ερωτήματος και η ανοχή σε σφάλματα.

5.1 Παρόν και μέλλον του κρυμμένου ιστού

Ο παγκόσμιος ιστός σήμερα αποτελείται από δομημένα δεδομένα και σε αυτό το πλαίσιο τοποθετείται και ο κρυμμένος ιστός [11]. Οι εργασίες που αφορούν δομημένα δεδομένα ταξινομούνται στις 3 ακόλουθες κατηγορίες:

1. *Αδόμητα ερωτήματα*: Τα ερωτήματα συμπίπτουν με τις δημοφιλείς λειτουργίες αναζήτησης πληροφοριών στον παγκόσμιο ιστό, δηλαδή οι χρήστες θέτουν ερωτήματα με λέξεις-κλειδιά και λαμβάνουν ως αποτέλεσμα μια λίστα με διευθύνσεις (URLs) ιστοσελίδων.
2. *Δομημένα ερωτήματα μιας σελίδας*: Σε αυτή τη κατηγορία τίθενται πιο ακριβή ερωτήματα στη μηχανή αναζήτησης, δηλαδή μπορεί να χρησιμοποιηθεί μια διεπαφή που επιτρέπει στους χρήστες να υποβάλουν πιο δομημένα ερωτήματα. Παραδείγματα τέτοιων ερωτημάτων αποτελεί η αγορά κάποιου σπιτιού σε συγκεκριμένη τοποθεσία, η εύρεση συγκεκριμένης εργασίας, κ.λπ.
3. *Δομημένα ερωτήματα πολλαπλών σελίδων*: Στην κατηγορία αυτή τα ερωτήματα είναι πιο πολύπλοκα σε σχέση με την προηγούμενη κατηγορία, γιατί η μηχανή αναζήτησης αναλαμβάνει να διεκπεραιώσει πιο σύνθετα ερωτήματα, όπως ένας συνδυασμός δεδομένων από πολλαπλές δομημένες πηγές του Διαδικτύου.

Ο κρυμμένος ιστός αποτελεί ένα υποσύνολο της δεύτερης κατηγορίας δομημένων δεδομένων, δηλαδή των δομημένων ερωτημάτων μιας σελίδας. Συγκεκριμένα είναι υποσύνολο (και όχι ίσο) διότι σε πολλές περιπτώσεις υπάρχουν σύνδεσμοι HTML που οδηγούν στις ίδιες ιστοσελίδες, με αποτέλεσμα να μην χρειάζονται οι ιστοσυλλέκτες (crawlers) να χρησιμοποιήσουν κάποια επιπρόσθετη λειτουργία ανίχνευσης.

5.1.1 Προσεγγίσεις εικονικής ολοκλήρωσης και ανάδειξης

Οι δυο κύριες προσεγγίσεις για την παροχή πρόσβασης στο περιεχόμενο του κρυμμένου ιστού είναι: η εικονική ολοκλήρωση (virtual integration) και η ανάδειξη (surfacing). Η πρώτη προσέγγιση χρησιμοποιείται σε επιχειρήσεις και στον τομέα της έρευνας ενώ η δεύτερη άρχισε να χρησιμοποιείται και περιλαμβάνει ιστοσελίδες του κρυμμένου ιστού στη μηχανή αναζήτησης Google.

Η προσέγγιση της εικονικής ολοκλήρωσης είναι μια λύση από την περιοχή της ενσωμάτωσης δεδομένων με σκοπό την πρόσβαση σε περιεχόμενα του κρυμμένου Ιστού. Η προσέγγιση αυτή βασίζεται στη κατασκευή ενδιάμεσων συστημάτων. Για τη δημιουργία μιας εφαρμογής μέσω αυτής της προσέγγισης, χρειάζεται η ανάλυση φορμών και να προσδιοριστεί το πεδίο ορισμού/τομέας των συγκεκριμένων περιεχομένων κι εν συνεχεία να δημιουργηθούν σημασιολογικές αντιστοιχίσεις μεταξύ των εισόδων των φορμών και των δεδομένων των ενδιάμεσων συστημάτων.

Η προσέγγιση ανάδειξης επικεντρώνεται στον προ-υπολογισμό των πιο σχετικών υποβολών φορμών, δηλαδή των ερωτημάτων για όλες τις επιθυμητές φόρμες HTML. Οι διευθύνσεις URL που προκύπτουν από αυτές τις υποβολές παράγονται με εκτός σύνδεσης τρόπο (offline) και προστίθενται σε ένα πίνακα μηχανών αναζήτησης όπως γίνεται με οποιαδήποτε σελίδα HTML. Η προσέγγιση αυτή είναι επιθυμητή για τις μηχανές αναζήτησης μιας και δίνει τη δυνατότητα αξιοποίησης της υπάρχουσας υποδομής ως έχει και συνεχίζεται η απρόσκοπτη ανάδειξη των ιστοσελίδων του κρυμμένου ιστού.

Στην προσέγγιση αυτή υπάρχουν δυο σημαντικά τεχνικά προβλήματα. Πρώτον, πρέπει να προσδιοριστούν οι τιμές που είναι κατάλληλες για διάφορες εισόδους φορμών στα μενού. Δεύτερον, πρέπει να ελαχιστοποιηθεί ο αριθμός των ερωτημάτων σε κάθε φόρμα έτσι ώστε να μην δημιουργείται αδικαιολόγητος φόρτος εργασίας κατά τη διάρκεια της ανάλυσης εκτός σύνδεσης.

5.1.2 Σημασιολογία των εισόδων των φορμών

Η χρήση της σημασιολογίας στον κρυμμένο ιστό αφορά στο ποιες τιμές θα μπουν ως είσοδοι στις φόρμες ανάδειξης. Έχουμε δυο ειδών εισόδους στις φόρμες, τα πλαίσια κειμένου ελεύθερης εισόδου και τα πλαίσια κειμένου με συγκεκριμένο τύπο δεδομένων εισόδου.

Τα πλαίσια κειμένου σε μια συντριπτική πλειοψηφία των φορμών έχουν πεδίο για αναζήτηση, δηλαδή δέχονται λέξεις-κλειδιά και ανακτούν αρχεία και πληροφορίες που περιέχουν τους όρους αναζήτησης. Τα πλαίσια κειμένου με συγκεκριμένο τύπο δεδομένων εισόδου δεν δέχονται αυθαίρετες λέξεις-κλειδιά άρα είναι σημαντικότερα με την έννοια ότι πρέπει να κατανοηθεί ο τύπος δεδομένων της συγκεκριμένης εισόδου, γιατί ίσως αποφέρει καλύτερη κάλυψη του περιεχομένου πίσω από τη φόρμα και αποτρέπει να δίνονται ερωτήματα χωρίς νόημα στη φόρμα.

Προκύπτει το ερώτημα κατά πόσο είναι δυνατόν να εξαχθούν αυτόματα σχεσιακά δεδομένα από ιστοσελίδες του κρυμμένου ιστού, δηλαδή αν γίνεται να ανακτηθούν

ιστοσελίδες που έχουν παραχθεί από φόρμες κρυμμένου ιστού. Γι' αυτό γίνεται αναφορά στη μηχανική μάθηση, όπου τα περιτυλίγματα δημιουργούνται με βάση δεδομένα εκπαίδευσης τα οποία δημιουργούνται χειροκίνητα από τη σήμανση επιθυμητών δομημένων δεδομένων σε διαφορετικές σελίδες μιας συγκεκριμένης ιστοσελίδας. Η μεγάλη μεταβλητότητα στη μορφή των σελίδων HTML καθιστά τη διαδικασία ιδιαίτερα δύσκολη και η πρόκληση είναι η εξαγωγή γραμμών δεδομένων από σελίδες που παρήχθησαν από ιστοσελίδες του κρυμμένου ιστού, όπου οι είσοδοι που έχουν συμπληρωθεί ήταν γνωστές από την αρχή με σκοπό την δημιουργία ιστοσελίδων.

5.2 Κρυμμένος ιστός και σημασιολογικός ιστός

Στον κρυμμένο ιστό η άντληση πληροφοριών δεν είναι κάτι ασυνήθιστο, καθώς πραγματοποιείται μέσω φορμών αναζήτησης. Στις φόρμες αναζήτησης πραγματοποιείται μια αίτηση ή ένα ερώτημα για την εξόρυξη δεδομένων από τον κρυμμένο ιστό. Αυτή είναι μια συνηθισμένη τακτική για την πρόσβαση στο κρυμμένο ιστό. Είναι επίσης πιθανό μερικοί ιστοχώροι να χρησιμοποιούν μηχανικά ερωτήματα (robotic queries) για την εξαγωγή πληροφοριών από τις βάσεις δεδομένων τους είτε προγράμματα-ρομπότ που έχουν γραφεί για κάθε μια ιστοσελίδα.

Ο ρόλος του σημασιολογικού ιστού [12] έχει σκοπό την αυτοματοποίηση εργασιών και υπηρεσιών που πραγματοποιούνται από ανθρώπους στον παγκόσμιο ιστό. Η αυτοματοποίηση απαιτεί προγράμματα ευφυών πρακτόρων που βρίσκουν πόρους, δηλαδή σελίδες ή υπηρεσίες, στο διαδίκτυο. Οι ευφυείς πράκτορες μπορούν να αποδειχθούν αποτελεσματικοί μόνο εάν έχουν πρόσβαση σε πολλές πληροφορίες και γνώσεις, διαφορετικά η χρήση τους δεν θα είναι καθοριστική.

Οι πληροφορίες συγκεντρώνονται σε μία "αποθήκη γνώσεων" που ονομάζεται οντολογία. Οι οντολογίες που εμπεριέχουν τις κατάλληλες πληροφορίες μπορούν να μεταφράσουν σωστά το ερώτημα ενός χρήστη σε κάτι που θα μπορεί να αναγνωριστεί από μία βάση δεδομένων. Ο κρυμμένος ιστός μπορεί να συνδεθεί με το σημασιολογικό ιστό με τη χρήση οντολογιών.

Οι οντολογίες μπορούν να περιλαμβάνουν ευρετηριοποιημένες πληροφορίες από ιστοσελίδες που συνδέονται απευθείας σε μια οντολογία σημασιολογικού ιστού. Η κατασκευή οντολογιών δεν είναι εύκολη αν και πολλοί ερευνητές έχουν προσπαθήσει να τις κατασκευάσουν αυτόματα με χρήση τεχνικών φυσικής επεξεργασίας γλώσσας (natural language processing). Σε μακροπρόθεσμη βάση ίσως υπάρξει μια νέα γενιά προγραμμάτων περιήγησης στο διαδίκτυο ώστε να μπορούν να δημιουργούν οντολογίες. Με τις οντολογίες μπορεί στο μέλλον να ανακαλυφθούν νέες τεχνολογίες για την άντληση πληροφοριών του κρυμμένου ιστού σε μεγάλη κλίμακα και με γρήγορο τρόπο.

Η κατασκευή οντολογιών σύμφωνα με αυτά θα επιτευχθεί με το να επιλέγει αρχικά ο χρήστης μια περιοχή ενδιαφέροντός του κι εν συνεχεία το αντίστοιχο πρόγραμμα

περιήγησης θα φορτώνει όλες τις γνωστές οντολογίες που σχετίζονται με το θέμα που επιθυμεί, με απώτερο στόχο την αυτοματοποίηση όλης της διαδικασίας.

Πρέπει να προσεγγιστούν τεχνικές αλλά και μεθοδολογίες ώστε να χρησιμοποιηθούν για τη μοντελοποίηση, την εξαγωγή και το σχολιασμό των πόρων του κρυμμένου ιστού προκειμένου να επιτευχθεί η σύνδεσή του με τον σημασιολογικό ιστό. Αυτό μπορεί να πραγματοποιηθεί αναγνωρίζοντας αρχικά ο χρήστης τη θεματική περιοχή που τον ενδιαφέρει. Ακολούθως κι αφού εντοπιστούν οι σελίδες που περιέχουν βάσεις δεδομένων με περιεχόμενα κρυμμένου ιστού, πρέπει να γραφεί ένα μηχανικό πρόγραμμα (robot program) που θα διατρέχει όλα τα περιεχόμενα των ιστοσελίδων αυτών και θα εξάγει τα επιθυμητά αποτελέσματα. Δεν είναι όλες οι ιστοσελίδες "συνεργάσιμες" μιας και αρκετές επιστρέφουν μηνύματα λάθους ή άλλες μπορεί να μπλοκάρουν τέτοια ερωτήματα. Τα αποτελέσματα που συγκεντρώνονται θα περιληφθούν σε μια οντολογία. Τέλος, ο χρήστης απλά χρειάζεται να δηλώσει τον τομέα του ενδιαφέροντός του κι ακολούθως ένα πρόγραμμα, το οποίο συνεργάζεται με οντολογίες, θα ανακτά τις αντίστοιχες πληροφορίες, καθιστώντας ευκολότερη την πρόσβαση σε δεδομένα του κρυμμένου ιστού.

Στον σημασιολογικό ιστό προκύπτει το θέμα κατά πόσο και αν πρέπει κάποιες ιστοσελίδες να μένουν κρυφές και γενικότερα να μην υπάρχει απευθείας πρόσβαση στα περιεχόμενά τους. Με άλλα λόγια, τίθεται το θέμα κάποιες σελίδες να είναι σκόπιμα κλειδωμένες για τους απλούς χρήστες και αν αυτό –κατά την εκτίμηση των εταιρειών- είναι σημαντικό: καμία εταιρεία δεν θα ήθελε μια ανταγωνίστριά της να μπορεί να αντλήσει τη λίστα με τους πελάτες της μέσω της βάσης δεδομένων της που θα είναι διαθέσιμη μέσω του σημασιολογικού ιστού. Αλλά, από την άλλη πλευρά κάθε εταιρεία θέλει να υπάρχει ως αποτέλεσμα και κατ’ επέκταση να διαφημίζεται από την μηχανή αναζήτησης Google, και μάλιστα όσο πιο ψηλά γίνεται. Άρα το ζήτημα της αναγκαιότητας του απόρρητου σε κάποιες μεμονωμένες περιπτώσεις δεν είναι τόσο σημαντικό, αλλά αυτό που χρειάζεται είναι να εκπαιδευτεί το αντίστοιχο μέρος του διαδικτύου ώστε να μπορεί να επεκτείνει τη γενικότερη στάση που έχει για το διαδίκτυο και στις "απόρρητες" βάσεις δεδομένων του ή μόνο σε κάποια πεδία της κρυφής αυτής βάσης δεδομένων του.

Τελικά αναφερόμαστε για μια νέα κίνηση ανοικτών δεδομένων, όπως ακριβώς μια κίνηση ανοικτού κώδικα και το μόνο που θα χρειαστεί να κάνουν οι εταιρείες είναι να διαχωρίσουν τα δεδομένα τους σε ιδιωτικά και δημόσια, τότε απλά θα δίνουν τη δυνατότητα στους χρήστες να πραγματοποιούν αναζητήσεις στα δημόσια δεδομένα του κρυμμένου ιστού.

6 Βιβλιογραφία

1. Bergman, Michael K. White Paper: The Deep Web: Surfacing Hidden Value, Volume 7, Issue 1, August, 2001
2. G. Rathinasabapathy, Invisible Web and Knowledge discovery tools: A Study, 5th International CALIBER -2007, Panjab University, Chandigarh, 08-10 February, 2007
3. Jie Liang, Estimation Methods for the Size of Deep Web Textural Data Source: A Survey, ACM Transactions on Computational Logic, Vol. V, No. N, August 2008, Pages 1-17.
4. S. Lawrence and C.L. Giles, "Searching the World Wide Web," Science 80:98-100, April 3, 1998.
5. K. Bharat and A. Broder, "A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines," paper presented at the Seventh International World Wide Web Conference, Brisbane, Australia, April 14-18, 1998.
6. Ntoulas, P. Zerfos & J. Cho, Downloading Textual Hidden Web Content Through Keyword Queries, JCDL 2005.
7. J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen & A. Halevy, Google's Deep-Web Crawl, VLDB 2008.
8. J. Wang & F. H. Lochovsky, Data Extraction and Label Assignment for Web Databases, WWW 2003.
9. F. Wang, G. Agrawal & R. Jin, Query Planning for Searching Inter-dependent Deep-Web Databases, SSDBM 2008, pp 24-41.
10. Fan Wang, SEEDEEP: A SYSTEM FOR EXPLORING AND QUERYING DEEP WEB DATA SOURCES, Ohio State University, 2010
11. J. Madhavan, L. Afanasiev, L. Antova & A. Halevy, Harnessing the Deep Web - Present and Future, CIDR 2009.
12. J. Zaino, Why the Deep Web Needs the Semantic Web, http://semanticweb.com/why-the-deep-web-needs-the-semantic-web_b234, 2009.