ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΠΟΤΕΛΕΣΜΑΤΙΚΗ ΑΝΑΖΗΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

ΑΘΑΝΑΣΙΟΣ Ε. ΠΑΠΑΘΑΝΑΣΙΟΥ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ηράκλειο, Ιούνιος 1999

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

**ΑΠΟΤΕΛΕΣΜΑΤΙΚΗ
ΑΝΑΖΗΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ
ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ**

Εργασία, που υποβλήθηκε από τον
Αθανάσιο Ε. Παπαθανασίου
ως μερική εκπλήρωση των απαιτήσεων για
την απόκτηση

Μεταπτυχιακού Διπλώματος Ειδίκευσης
στην Επιστήμη Υπολογιστών

Συγγραφέας: _____
Αθανάσιος Ε. Παπαθανασίου
Τμήμα Επιστήμης Υπολογιστών

Εισηγητική Επιτροπή:_____
Βαγγέλης Μαρκάτος, Επίκουρος Καθηγητής, Επόπτης

_____
Απόστολος Τραγανίτης, Αναπληρωτής Καθηγητής, Μέλος

_____
Αικατερίνη Χούστη, Αναπληρώτρια Καθηγήτρια, Μέλος

Δεκτή: _____
Πάνος Κωνσταντόπουλος, Καθηγητής,
Πρόεδρος Επιτροπής Μεταπτυχιακών Σπουδών

Ηράκλειο, Ιούνιος 1999

ΑΠΟΤΕΛΕΣΜΑΤΙΚΗ ΑΝΑΖΗΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Αθανάσιος Ε. Παπαθανασίου

Μεταπτυχιακή Εργασία

Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

ΠΕΡΙΛΗΨΗ

Ο Παγκόσμιος Ιστός μεγαλώνει με εκρηκτικούς ρυθμούς, πράγμα που κάνει την ανάκληση πληροφορίας μία ιδιαίτερα δύσκολη διαδικασία. Οι παραδοσιακές μέθοδοι αναζήτησης, που στηρίζονται στις μηχανές αναζήτησης καταιγίζουν τους χρήστες με ένα υπερβολικά μεγάλο αριθμό URLs. Εάν ένας χρήσης επιθυμεί να παραμένει διαρκώς ενημερωμένος πάνω σε ένα θέμα, είναι αναγκασμένος να εκτελεί διαρκώς τις ίδιες επερωτήσεις στις ίδιες μηχανές αναζήτησης, οι οποίες απαντούν με ένα σχεδόν όμοιο σύνολο από URLs. Από το σύνολο αυτό μόνο ένα μικρό ποσοστό δείχνει σε νέα πληροφορία. Στο πρώτο μέρος της εργασίας αυτής παρουσιάζεται ένα εργαλείο, που στηρίζεται πάνω σε ήδη υπάρχουσες μηχανές αναζήτησης και ονομάζεται USEwebNET. Το USEwebNET αποθηκεύει τα ενδιαφέροντα κάθε χρήστη και εκτελεί περιοδικά τις επερωτήσεις του στις υποστηριζόμενες μηχανές αναζήτησης. Παράλληλα, παρακολουθεί τα URLs, που επισκέπτεται ο χρήστης, και σε κάθε χρήση παρουσιάζει μόνο τα νέα ή μη αναγνωσμένα URLs.

Το δεύτερο μέρος της εργασίας επικεντρώνεται σε μία επέκταση του USEwebNET για το χώρο των ψηφιακών βιβλιοθηκών, που ονομάζεται PaperFinder. Το PaperFinder έχει ως στόχο να βοηθήσει τον ερευνητή στην παρακολούθηση της πληροφορίας που τον ενδιαφέρει μέσα από το διαρκώς αυξανόμενο αριθμό των ερευνητικών άρθρων, που διατίθονται ηλεκτρονικά. Για το σκοπό αυτό, λειτουργεί πάνω από δημοφιλείς ψηφιακές βιβλιοθήκες, διαχωρίζει ενδιαφέροντα άρθρα και τα παραδίδει στους χρήστες. Υποστηρίζει δύο τρόπους αναζήτησης άρθρων. Ο πρώτος στηρίζεται στην εκτέλεση επερωτήσεων με τη χρήση λέξεων-κλειδιών - Keyword-based Mode - και ο δεύτερος στοχεύει στη γενίκευση των επερωτήσεων και αναταξινόμηση των αποτελεσμάτων ως προς ένα ή περισσότερα "χαρακτηριστικά άρθρα", ώστε να βρεθεί μεγαλύτερο πλήθος δημοσιεύσεων - Resource Discovery Mode.

Supervisor: Βαγγέλης Μαρκάτος
Επίκουρος Καθηγητής Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

EFFECTIVE RESOURCE DISCOVERY ON THE WORLD WIDE WEB

Athanasios E. Papathanasiou

Master of Science Thesis

Department of Computer Science
University of Crete

ABSTRACT

The World Wide Web grows at alarming rates, making information retrieval an increasingly difficult process. Traditional search methods based on search-engines usually flood the users with an overwhelming number of URLs. If a user wants to stay up-to-date on some issue and he repeatedly queries the above search engines, (s)he will be repeatedly flooded with (almost) the same set of URLs, out of which only a small percentage will point to new, previously unseen documents. In the first part of this thesis, a resource discovery tool built on top of traditional search engines and called USEwebNET. USEwebNET registers each user's interests and repeatedly queries several search engines for URLs matching a user's registered interests. USEwebNET keeps track of which URLs have been visited by each user. Thus, when a user invokes USEwebNET, he is presented only with new or "unvisited" URLs.

The second part of this thesis, focuses in an extension of USEwebNET implemented especially for the area of digital libraries and called PaperFinder. PaperFinder, aims in helping scientists to keep track of the articles, which they are interested in among the growing number of papers, that become available on-line. It operates on top of popular Digital Libraries of scientific publications, filters only relevant papers delivers them to the users. PaperFinder may operate in a Keyword-based Mode, where scientists present a list of keywords that describe their field of interest, and in a Resource-discovery Mode, where scientists present one or more "seed papers", that describe a field of interest. In the latter, PaperFinder searches for papers that are relevant to the seed paper and sorts the result by calculating their relevance to the seed paper.

Supervisor: Evangelos Markatos
Assistant Professor of Computer Science
University of Crete

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα εργασία δεν θα είχε πραγματοποιηθεί χωρίς τη βοήθεια συγκεκριμένων ατόμων, τους οποίους θα ήθελα να ευχαριστήσω.

Αρχικά, θα ήθελα να ευχαριστήσω τον επόπτη καθηγητή της εργασίας, Δρ. Βαγγέλη Μαρκάτο, ο οποίος έδωσε τη δυνατότητα της ανάπτυξης της και προσέφερε σημαντική βοήθεια σε όλα τα στάδια υλοποίησης της. Επίσης, ευχαριστώ το Δρ. Απόστολο Τραγανίτη και τη Δρ. Κατερίνα Χούστη, μέλη της εισηγητικής επιτροπής, για την εποικοδομητική κριτική της εργασίας.

Ακολούθως, θα ήθελα να ευχαριστήσω:

Το Σταύρο Παπαδάκη για τη βοήθεια, που προσέφερε κατά την υλοποίηση τμημάτων της εργασίας.

Τα μέλη της ομάδας "Λειτουργικών Συστημάτων Υψηλών Επιδόσεων" του Ινστιτούτου Τεχνολογίας και Ερευνας, Μιχάλη Φλουρή, Αντώνη Δανάλη και Ξένια Ασημακοπούλου για τη συνεργασία τους και το σχόλια τους κατά τη διάρκεια της υλοποίησης της εργασίας. Τα μέλη του Εργαστηρίου Ψηφιακών Συστημάτων του Ινστιτούτου Τεχνολογίας και Ερευνας, Νεραντζούλα Σεβασλίδου, Μιχάλη Λυγεράκη, Βαγγέλη Καραγιάννη και Κατερίνα Γιαλαμά για τη βοήθεια τους σε τμήματα της εργασίας.

Τη Μαριάννα Γιαλύτη για τα σχόλια και τις διορθώσεις, που έκανε στο παρόν κείμενο.

Τους γονείς μου, Ελευθέριο και Όλγα Παπαθανασίου, για την υποστήριξη και τη βοήθειά τους κατά τη διάρκεια των σπουδών μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΆΛΑΙΟ 1 **5**

1.1 Ο ΠΑΓΚΟΣΜΙΟΣ ΙΣΤΟΣ **5**
1.2 ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ **6**
1.3 ΤΟ ΠΡΟΒΛΗΜΑ ΤΟΥ ΚΑΤΑΙΓΙΣΜΟΥ ΠΛΗΡΟΦΟΡΙΑΣ **8**
1.4 ΑΠΟΤΕΛΕΣΜΑΤΙΚΗ ΑΝΑΚΛΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ **10**
1.5 USEwebNET **12**
1.6 ΣΥΝΕΙΣΦΟΡΕΣ (CONTRIBUTIONS) **13**
1.7 Η ΟΡΓΑΝΩΣΗ ΤΗΣ ΕΡΓΑΣΙΑΣ **13**

ΚΕΦΆΛΑΙΟ 2 **14**

2.1 Η ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ USEwebNET **14**
2.2 ΔΙΕΠΙΦΑΝΕΙΑ ΧΡΗΣΗΣ ΤΟΥ USEwebNET **15**
2.2.1 ΙΣΤΟΣΕΛΙΔΑ ΚΑΤΑΧΩΡΗΣΗΣ ΕΠΕΡΩΤΗΣΕΩΝ 17
2.2.2 ΙΣΤΟΣΕΛΙΔΑ ΠΑΡΟΥΣΙΑΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ 19
2.2.3 ΙΣΤΟΣΕΛΊΔΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΑΡΧΕΙΩΝ 23
2.2.4 ΙΣΤΟΣΕΛΙΔΑ ΡΥΘΜΙΣΗΣ ΠΡΟΣΩΠΙΚΟΥ ΠΟΡΤΡΑΙΤΟΥ 24
2.2.5 ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ 26
2.3 ΑΛΛΗΛΕΠΙΔΡΑΣΗ ΜΕ ΤΙΣ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ **29**
2.4 ΣΥΖΗΤΗΣΗ **30**
2.4.1 Ψηφιακές Βιβλιοθήκες (Digital Libraries) 31
2.4.2 Ηλεκτρονικό Εμπόριο (Electronic Commerce) 32
2.4.3 Επεκτάσεις 32
2.5 ΣΥΜΕΡΑΣΜΑΤΑ **33**

ΚΕΦΆΛΑΙΟ 3 **35**

3.1 ΕΙΣΑΓΩΓΗ **35**
3.2 ΥΛΟΠΟΙΗΣΗ ΤΗΣ ΔΙΕΠΙΦΑΝΕΙΑΣ ΧΡΗΣΗΣ **35**
3.2.1 Η Τεχνολογία του Παγκοσμίου Ιστού 36
3.2.2 Λειτουργία της Διεπιφάνειας Χρήσης 41
3.3 Η ''ΜΗΧΑΝΗ'' ΤΟΥ USEwebNET **42**
3.3.1 Συλλέκτης αποτελεσμάτων και Αναλυτής κειμένου 42
3.3.2 Τμήμα Ενημέρωσης της Βάσεως Δεδομένων 43
3.3.3 Πρόγραμμα Αποστολής ηλεκτρονικού μηνύματος 43
3.3.4 Τμήμα χειρισμού περιοδικών λειτουργιών (Crontab Manager) 44
3.3.5 Ελεγκτής εκδόσεων ιστοσελίδων 44

ΚΕΦΆΛΑΙΟ 4 **46**

4.1 ΕΙΣΑΓΩΓΗ **46**
4.2 ΨΗΦΙΑΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ **46**
4.3 PAPERFINDER: ΣΧΕΔΙΑΣΗ ΚΑΙ ΔΙΕΠΙΦΑΝΕΙΑ ΧΡΗΣΗΣ **49**
4.3.1 ΤΡΌΠΟΙ ΛΕΙΤΟΥΡΓΊΑΣ 50
4.3.2 ΔΙΕΠΙΦΆΝΕΙΑ ΧΡΉΣΗΣ 51
4.3.3 Η ΜΗΧΑΝΗ ΤΟΥ PAPERFINDER 60
4.3.4 ΑΝΑΓΝΩΡΙΣΗ ΒΙΒΛΙΟΓΡΑΦΙΚΏΝ ΑΝΑΦΟΡΏΝ 62
4.3.5 ΑΝΑΓΝΩΡΙΣΗ ΣΥΓΓΡΑΦΈΩΝ 63
4.4 ΘΕΜΑΤΑ ΥΛΟΠΟΙΗΣΗΣ **64**
4.4.1 ΔΙΕΠΙΦΆΝΕΙΑ ΧΡΉΣΗΣ 64
4.4.2 ΥΛΟΠΟΙΗΣΗ ΤΗΣ ΜΗΧΑΝΗΣ ΤΟΥ PAPERFINDER 65
4.5 ΣΥΜΠΕΡΑΣΜΑΤΑ **65**

ΚΕΦΆΛΑΙΟ 5 **67**

5.1 RESOURCE DISCOVERY MODE **67**
5.2 ΣΧΕΔΙΑΣΜΟΣ **67**
5.3 ΓΕΝΙΚΕΥΣΗ ΕΠΕΡΩΤΗΣΗΣ **70**
5.3.1 ΑΡΧΙΤΕΚΤΟΝΙΚΉ 73
5.3.2 ΥΛΟΠΟΙΗΣΗ 75
5.4 ΤΑΞΙΝΟΜΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ **75**
5.4.1 ΑΡΧΙΤΕΚΤΟΝΙΚΉ 77
5.4.2 ΥΛΟΠΟΙΗΣΗ 81
5.5 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ **82**
5.6 ΣΥΜΠΕΡΑΣΜΑΤΑ **87**

ΚΕΦΆΛΑΙΟ 6 **89**

6.1 ΕΙΣΑΓΩΓΗ **89**
6.2 Ο ΡΥΘΜΟΣ ΑΝΑΠΤΥΞΗΣ ΤΟΥ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ **89**
6.3 ΑΝΑΚΛΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΑΠΟ ΤΟΝ ΙΣΤΟ **91**
6.4 PAPERFINDER – ΨΗΦΙΑΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ **101**

ΚΕΦΆΛΑΙΟ 7 **105**

7.1 ΠΕΡΙΛΗΨΗ **105**

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Αριθμός Σχήματος *Σελίδα*
Σχήμα 1-1: *Το περιβάλλον του Παγκόσμιου Ιστου.* _____ *6*
Σχήμα 2-1: *Η φόρμα "Πιστοποίησης Ταυτότητας" του χρήστη.* _____ *16*
Σχήμα 2-2: *Η οθόνη δημιουργίας νέου λογαριασμού.*_____ *16*
Σχήμα 2-3: *Η ιστοσελίδα καταχώρησης νέων επερωτήσεων.* _____ *18*
Σχήμα 2-4: *Η ιστοσελίδα παρουσίασης αποτελεσμάτων.* _____ *20*
Σχήμα 2-5: *Ιστοσελίδα παρουσίασης προσωπικών αρχείων.* _____ *23*
Σχήμα 2-6: *Ο καθορισμός του προσωπικού πορτραίτου ενός χρήστη.* _____ *25*
Σχήμα 2-7: *Εισαγωγή μίας επερώτησης πάνω στο θέμα ''Loop Scheduling''.*_____ *27*
Σχήμα 2-8: *Το USEwebNET παρουσιάζει τέσσερις επερωτήσεις του χρήστη.* _____ *27*
Σχήμα 2-9: *Περιληπτική Παρουσίαση Αποτελεσμάτων.* _____ *28*
Σχήμα 2-10: *Τα αποτελέσματα της επερώτησης ''remote memory paging''.* _____ *28*
Σχήμα 2-11: *Τα αποτελέσματα της επερώτησης ''remote memory paging'' μετά από την αποθήκευση δύο ιστοσελίδων και την ανάγνωση τεσσάρων.* _____ *28*
Σχήμα 3-1: *Σχεδιάγραμμα της διαδρομής των αιτήσεων στο USEwebNET.*_____ *41*
Σχήμα 3-2: *Η αρχιτεκτονική της μηχανής του USEwebNET.* _____ *43*
Σχήμα 4-1: *Οθόνη καταχώρησης επερωτήσεων του PaperFinder.* _____ *52*
Σχήμα 4-2: *Δημιουργία μίας επερώτησης για το θέμα του "file caching".* _____ *53*
Σχήμα 4-3: *Ο πίνακας επερωτήσεων μετά τη δημιουργία της επερώτησης "file caching".* ___ *54*
Σχήμα 4-4: *Η παρουσίαση των αποτελεσμάτων. Ο χρήστης διαβάζει το άρθρο "Scheduling in Client-Server Systems".* _____ *55*
Σχήμα 4-5: *Η λίστα των επερωτήσεων μαζί με τον αριθμό των μη αναγνωσμένων άρθρων.* _ *56*
Σχήμα 4-6: *Ο πίνακας των αποτελεσμάτων της επερώτησης "file caching".* _____ *57*
Σχήμα 4-7: *Η οθόνη προσωπικών αρχείων.*_____ *59*
Σχήμα 5-1: *Παρουσίαση ενός αποτελέσματος από την ψηφιακή βιβλιοθήκη της ACM.* _____ *73*
Σχήμα 5-2: *Εικόνα από την ιστοσελίδα αποτελεσμάτων του PaperFinder. Το άρθρο "Serverless network file systems" χρησιμοποιείται ως "χαρακτηριστικό άρθρο" για τη συγκεκριμένη επερώτηση.* _____ *74*
Σχήμα 5-3: *Εικόνα από την ιστοσελίδα αποτελεσμάτων του PaperFinder. Το άρθρο "Serverless network file systems" χρησιμοποιείται ως "χαρακτηριστικό άρθρο" για την ταξινόμηση των αποτελεσμάτων της συγκεκριμένη επερώτηση.* _____ *81*
Σχήμα 5-4: *Λίστα συγκριτικών αποτελεσμάτων των 20 πρώτων άρθρων, που έδωσαν η ACM και το PaperFinder για το χαρακτηριστικό άρθρο "Serverless Network File Systems".* _____ *84*
Σχήμα 5-5: *Τα αποτελέσματα της επερώτησης "Distributed Systems" ταξινομημένα με βάση το άρθρο 3. Η ταξινόμηση αυτή έχει ως αποτέλεσμα το διαχωρισμό της ομάδας των ερευνητών του Πανεπιστημίου του Rochester.* _____ *86*

Κεφάλαιο 1

ΤΟ ΠΡΟΒΛΗΜΑ ΤΟΥ ΚΑΤΑΙΓΙΣΜΟΥ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

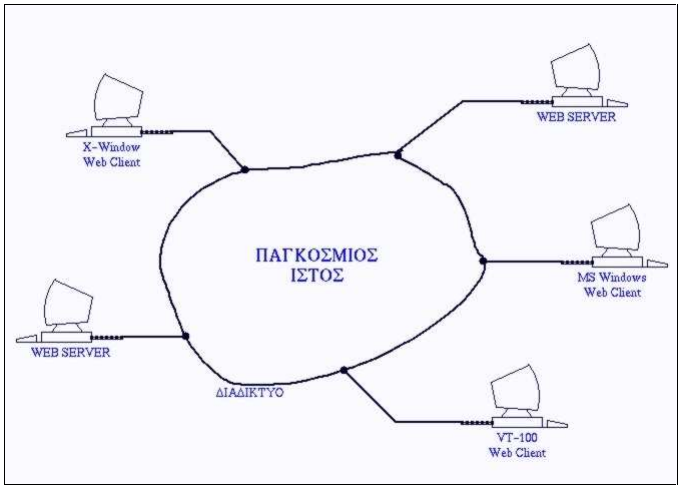
1.1 Ο ΠΑΓΚΟΣΜΙΟΣ ΙΣΤΟΣ

Ο Παγκόσμιος Ιστός (World Wide Web - WWW) - Σχήμα 1-1 - αποτελεί ένα περιβάλλον αποθήκευσης και ανάκτησης ηλεκτρονικής πληροφορίας. Η πληροφορία αυτή διατηρείται με τη μορφή μίας μεγάλης συλλογής συνδεδεμένων κειμένων, τα οποία είναι κατανεμημένα στο Διαδίκτυο (Internet) και χαρακτηρίζεται από τη συνεχή μεταβολή της και την αύξηση του όγκου της.

Η αύξηση αυτή της ηλεκτρονικής πληροφορίας ευνοείται από την αντίστοιχη παγκοσμίως ραγδαία εξάπλωση του Διαδικτύου τα τελευταία δέκα χρόνια. Η ταχύτατη ανάπτυξη της τεχνολογίας στους χώρους των υπολογιστών και των επικοινωνιών έδωσε τη δυνατότητα της αποθήκευσης και μεταφοράς και άλλων ειδών πληροφορίας εκτός του κειμένου, με αποτέλεσμα η υπάρχουσα ηλεκτρονική πληροφορία να εμπλουτιστεί σταδιακά με εικόνες, στη συνέχεια με ήχο και τέλος με video. Επιπροσθέτως, έδωσε τη δυνατότητα της απόκτησης ενός ηλεκτρονικού υπολογιστή σε ένα πολύ μεγάλο αριθμό καταναλωτών και κατά συνέπεια ευνόησε τη μεγάλη αύξηση των χρηστών του Παγκοσμίου Ιστού.

Η μεγάλη αύξηση των χρηστών έδωσε στο Διαδίκτυο τη δυνατότητα να αποκτήσει, πέρα από την αρχικά κυρίως ακαδημαϊκή του χρήση, νέες εφαρμογές, όπως εμπορικές, εκπαιδευτικές, ενημερωτικές και ψυχαγωγικές. Το γεγονός αυτό έχει ως αποτέλεσμα τη διαρκή και ραγδαία αύξηση του όγκου της πληροφορίας στον Παγκόσμιο Ιστό και ταυτόχρονα έκανε έντονα εμφανές το πρόβλημα της ευρέσεως

της "χρήσιμης" ή "επιθυμητής" πληροφορίας μέσα από ένα μεγάλο όγκο αδιάφορου υλικού.



Σχήμα 1-1: Το περιβάλλον του Παγκόσμιου Ιστου. Ενας μεγάλος αριθμός εξυπηρετητών και πελατών συνδεδεμένων πάνω στο Διαδίκτυο.

Παρά το γεγονός ότι μπορεί να υπάρχει μία μεγάλη ποσότητα ενδιαφέροντος υλικού στον Παγκόσμιο Ιστό, ο χρήστης συχνά χρειάζεται πολύ χρόνο και επίπονη προσπάθεια για να το εντοπίσει. Επιπροσθέτως, είναι αδύνατη η εξαντλητική αναδίφηση (browsing) του Ιστού με στόχο την εύρεση της επιθυμητής πληροφορίας. Συνοψίζοντας, η αναζήτηση υλικού στον Παγκόσμιο Ιστό μπορεί να γίνει αδύνατη εάν ο χρήστης δεν βοηθηθεί από τα κατάλληλα εργαλεία.

1.2 ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Για να διευκολυνθεί η ανάκτηση χρήσιμης πληροφορίας με αποτελεσματικό και ικανοποιητικό τρόπο, έχουν αναπτυχθεί πολλά εργαλεία. Πολλά από αυτά είναι ευρέως γνωστά με τον όρο "Μηχανές Αναζήτησης" (Search Engines) και μπορούν να διαχωριστούν σε δύο κατηγορίες: *Πράκτορας Τύπου Πελάτη (Client-based Agent -*

CBAs) και *Εργαλεία Αναζήτησης, που βασίζονται σε Ευρετήρια (Index-based Search Tools - IBSTs).*

Οι μηχανές της πρώτης κατηγορίας, που είναι γνωστές και ως *Internet Robots*, έχουν ως στόχο την αυτόματη εξερεύνηση του Ιστού. Δεδομένου ότι η δομή του Παγκοσμίου Ιστού είναι όμοια με αυτή ενός κατευθυνόμενου κυκλικού γράφου και βασίζεται στο Μοντέλο Πελάτη-Εξυπηρετητή (Client-Server), μπορεί να εξερευνηθεί από ένα πρόγραμμα, το οποίο εκτελείται σε ένα μόνο κόμβο (site). Η λειτουργία των προγραμμάτων αυτών στηρίζεται στην επαναλαμβανόμενη επίσκεψη κόμβων και σελίδων του Διαδικτύου. Ξεκινώντας από ένα μικρό σχετικά σύνολο Ομοιόμορφων Προσδιοριστών Πόρων (Uniform Resource Locator - URL), διασχίζουν τον Ιστό κατά-βάθος (depth-first) ή κατά-πλάτος (breadth-first) ανακαλύπτοντας διαρκώς νέα URLs. Ολοκληρώνοντας, τα εργαλεία της κατηγορίας αυτής χρησιμοποιούνται για πολλούς σκοπούς, όπως είναι η ανεύρεση λανθασμένων αναφορών (broken references), η ανακάλυψη νέων εξυπηρετητών, ο κατά προσέγγιση υπολογισμός του μεγέθους του Ιστού, η εύρεση και ανάκληση χρήσιμης πληροφορίας και η δημιουργία ευρετηρίων για τον Ιστό.

Από την άλλη πλευρά οι μηχανές της δεύτερης κατηγορίας στηρίζονται στην ύπαρξη ενός προδημιουργημένου ευρετηρίου. Το ευρετήριο είναι ένα αναζητήσιμο αρχείο ή βάση δεδομένων, το οποίο παρέχει δείκτες αναφοράς προς τη διαθέσιμη πληροφορία στον Ιστό. Το ευρετήριο δημιουργείται και ανανεώνεται από μία ομάδα ατόμων υπεύθυνη για τη συγκεκριμένη εργασία είτε αυτομάτως από ένα Internet Robot, το οποίο συνεργάζεται με τη μηχανή αναζήτησης.

Στην πρώτη περίπτωση οι χρήστες μπορούν να συμμετέχουν έμμεσα στη δημιουργία του ευρετηρίου μέσω της προτάσεως συγκεκριμένων σελίδων ή εξυπηρετητών. Παράλληλα, οι ιστοσελίδες ταξινομούνται σε θεματικά δένδρα. Κάθε θεματικό δένδρο περιέχει ένα σύνολο από URLs, τα οποία ανήκουν σε ένα συγκεκριμένο πεδίο ενδιαφέροντος. Τα θεματικά δένδρα και τα υπόδενδρά τους

σχεδιάζονται δια χειρός με τέτοιο τρόπο, ώστε να διευκολύνεται η προσπέλαση τους από τους χρήστες μέσω της τεχνικής της αναδίφησης (browsing). Ακολουθώντας δένδρα διαρκώς πιο σχετικά με το επιθυμητό θέμα, οι χρήστες μπορούν να εντοπίσουν πληροφορία συγκεκριμένη στο πεδίο, που τους ενδιαφέρει. Μία μηχανή αναζήτησης, που ανήκει στην κατηγορία αυτή είναι το Yahoo ([VIII]).

Στη δεύτερη περίπτωση, η διαδικασία γίνεται χωρίς την ανάμιξη ανθρώπινου δυναμικού. Το Internet Robot, το οποίο συνεργάζεται με τη μηχανή αναζήτησης, είναι υπεύθυνο να αναγνωρίσει λέξεις-κλειδιά μέσα στις ιστοσελίδες, που ανακαλύπτει και να δημιουργήσει το ευρετήριο. Μηχανές αναζήτησης, που ανήκουν στην κατηγορία αυτή, είναι οι Alta-Vista ([I]), HotBot ([II]), MetaCrawler ([III]), Excite ([IV]), Lycos ([V]), Infoseek ([VI]) και GoTo ([VII]).

Οι χρήστες επικοινωνούν με τη μηχανή αναζήτησης μέσω μίας διεπιφάνειας χρήσης, η οποία τους δίνει τη δυνατότητα να δημιουργούν επερωτήσεις (queries) με τη μορφή λέξεων-κλειδιών (keywords). Επίσης, πολλές μηχανές αναζήτησης δίνουν τη δυνατότητα του περιορισμού των επιστρεφόμενων δεικτών αναφοράς μέσω επιλογών, όπως είναι η ημερομηνία δημιουργίας ή τελευταίας ανανέωσης της ιστοσελίδας, η γλώσσα, στην οποία είναι γραμμένη το κείμενο της ιστοσελίδας, ο τύπος του υλικού, που περιέχει (εικόνα, ήχος) και ο τύπος του κόμβου, στον οποίο ανήκει (εμπορικός, ακαδημαϊκός, κυβερνητικός - ο διαχωρισμός αυτός αφορά τους κόμβους των Η.Π.Α. μόνο). Παρά τις διευκολύνσεις αυτές, το πρόβλημα της υπερφόρτωσης πληροφορίας και της γρήγορης, αποτελεσματικής ανάκλησης της δεν έχει λυθεί.

1.3 ΤΟ ΠΡΟΒΛΗΜΑ ΤΟΥ ΚΑΤΑΙΓΙΣΜΟΥ ΠΛΗΡΟΦΟΡΙΑΣ

Η ανάπτυξη και εξέλιξη των μηχανών αναζήτησης δεν έχει κατορθώσει να δώσει λύση στο πρόβλημα της αναζήτησης πληροφορίας. Αρχικά, το μέγεθος του Παγκοσμίου Ιστού σε συνδυασμό με τη συνεχή και εκρηκτική ανάπτυξη του δυσχεραίνουν διαρκώς περισσότερο τη στενή, καθημερινή παρακολούθηση της νέας

πληροφορίας καθώς γίνεται διαθέσιμη σε αυτόν. Η μεγάλη αύξηση των κόμβων του διαδικτύου αναγκάζει τα Internet Robots, που είναι υπεύθυνα για τη δημιουργία των ευρετηρίων, να επισκέπτονται κάθε εξυπηρετητή συνεχώς πιο σπάνια. Τα δημοφιλή Internet Robots, όπως της Alta-Vista, έχουν μέσο χρόνο επίσκεψης κάθε κόμβου περίπου δύο με τρεις μήνες και το χρονικό αυτό διάστημα τείνει να αυξηθεί. Κατά συνέπεια, η **νέα** και συνήθως πιο **ενδιαφέρουσα** πληροφορία να καθυστερεί περισσότερο να εμφανιστεί ως αποτέλεσμα μίας σχετικής επερώτησης.

Ακολούθως, το παραδοσιακό μοντέλο αναζήτησης πληροφορίας μέσω επερωτήσεων σε μία μηχανή αναζήτησης δυσχεραίνει την ανίχνευση νέας πληροφορίας πάνω σε ένα ορισμένο θέμα και το διαχωρισμό της από γνώση, η οποία έχει ήδη βρεθεί στο παρελθόν. Συγκεκριμένα, κάθε φορά που ένας χρήστης επιθυμεί να βρει νέες πληροφορίες για το πεδίο γνώσης, που τον ενδιαφέρει, χρησιμοποιεί τις ίδιες (ή περίπου τις ίδιες) λέξεις-κλειδιά. Η μηχανή αναζήτησης του επιστρέφει περίπου τα ίδια URLs, με αποτέλεσμα να τον πλημμυρίζει με ένα μεγάλο όγκο ανεπιθύμητης πληροφορίας, μεγάλο μέρος της οποία έχει ήδη επεξεργαστεί στο παρελθόν. Κατά συνέπεια, η ανεύρεση της νέας και πιο ενδιαφέρουσας πληροφορίας ανάμεσα στον τεράστιο αριθμό των URLs, τα οποία έχει ήδη επισκεφτεί ο χρήστης, γίνεται ιδιαίτερα επίπονη, χρονοβόρα και κουραστική διαδικασία.

Χαρακτηριστικό παράδειγμα αποτελεί η περίπτωση, που ένας χρήστης ενδιαφέρεται να παρακολουθεί διαρκώς τις τελευταίες εξελίξεις σε ένα συγκεκριμένο θέμα, όπως είναι τα "Κατανεμημένα Συστήματα". Η επερώτηση για το πεδίο αυτό σε μία μηχανή αναζήτησης, όπως το HotBot ([II]) πολλές εκατοντάδες URLs. Μόνο ένα πολύ μικρό ποσοστό από αυτό όμως θα έχει δημιουργηθεί ή ενημερωθεί μέσα στον τελευταίο μήνα. Συνεπώς, ο χρήστης θα αναγκαστεί να επεξεργαστεί ένα τεράστιο αριθμό από URLs, για να βρει μόνο ένα πολύ μικρό αριθμό από νέες ή ενημερωμένες ιστοσελίδες.

Φαινομενικά, το πρόβλημα του εντοπισμού πρόσφατης πληροφορίας θα μπορούσε να λυθεί με τον εμπλουτισμό της επερώτησης, ώστε να αναζητήσει ιστοσελίδες, οι οποίες έχουν δημιουργηθεί ή ενημερωθεί μετά από μία συγκεκριμένη ημερομηνία, μία λειτουργία που προσφέρεται από τις περισσότερες μηχανές αναζήτησης. Δυστυχώς, ακόμη και αυτή η προσέγγιση δεν επιλύει το πρόβλημα του κατακλεισμού της πληροφορίας. Όπως προαναφέρθηκε, εξαιτίας της ιλιγγιώδους ανάπτυξης του Διαδικτύου, τα Web Robots επισκέπτονται κάθε κόμβο περιοδικά, μία φορά κάθε δύο με τρεις μήνες. Καθώς η διαθέσιμη πληροφορία στον Παγκόσμιο Ιστό αυξάνεται το χρονικό αυτό διάστημα τείνει να αυξηθεί, πλησιάζοντας τους πέντε με έξι μήνες. Κατά συνέπεια, εάν ένας χρήστης επιθυμεί να βρει όλες τις σχετικές με το θέμα που τον ενδιαφέρει ιστοσελίδες, τις οποίες δεν έχει ξαναεπισκεφτεί, είναι υποχρεωμένος να αναζητήσει URLs, τα οποία έχουν δημιουργηθεί ή ανανεωθεί μέσα στους τελευταίους έξι μήνες. Ένας τέτοιος περιορισμός σε μία επερώτηση, που θα γίνει σε μία μηχανή αναζήτησης δεν μπορεί να περιορίσει αισθητά τον αριθμό των επιστρεφόμενων αποτελεσμάτων. Επομένως, η αναζήτηση μόνο πρόσφατων ιστοσελίδων δεν μπορεί να δώσει λύση στο πρόβλημα του καταιγισμού πληροφορίας.

1.4 ΑΠΟΤΕΛΕΣΜΑΤΙΚΗ ΑΝΑΚΛΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Η μη αποτελεσματική ανάκληση πληροφορίας των υπαρχόντων μηχανών αναζήτησης οφείλεται στο γεγονός ότι οι μηχανές αναζήτησης υλοποιούν ανεξάρτητες μεταξύ τους επερωτήσεις. Ένας χρήστης, ο οποίος αναζητά (**search**) πληροφορία με τις ίδιες λέξεις κλειδιά, κατακλείζεται διαρκώς από τα ίδια περίπου URLs, ανεξάρτητα από το αν τα έχει επισκεφτεί στο παρελθόν.

Η διαδικασία αυτή αναζήτησης έρχεται σε αντίθεση με τη σταδιακή και αυξητική ανακάλυψη γνώσεως, που στηρίζεται στην έρευνα (**re**search). Η έρευνα αποτελεί μία επαναληπτική διαδικασία, η οποία αποκλείει κάθε άχρηστη ή ήδη επεξεργασμένη πληροφορία και επικεντρώνεται στη μελέτη νέας, ανεξερεύνητης γνώσης.

Απαραίτητη ικανότητα για την αναγνώριση της νέας από τη γνωστή πληροφορία είναι αυτή της **μνήμης**. Η μνήμη είναι η λειτουργία, που επιτρέπει στον ερευνητή να αναγνωρίσει την ήδη αποκτημένη γνώση και να εστιάσει τις πνευματικές του ικανότητες στη μάθηση νέας.

Η ενσωμάτωση της ικανότητας της μνήμης στη λειτουργία των ήδη υπαρχόντων μηχανών αναζήτησης μπορεί να οδηγήσει στην επίλυση του προβλήματος του καταιγισμού της πληροφορίας και στην αποτελεσματική ανάκληση της. Δεδομένου ότι μία μηχανή αναζήτησης έχει την ικανότητα να αναγνωρίσει πληροφορία την οποία έχει παρουσιάσει στο παρελθόν ως αποτέλεσμα μίας ορισμένης επερώτησης ενός συγκεκριμένου χρήστη, θα μπορεί να αποκλείσει την επανεμφάνιση της σε μία πιθανή εκτέλεση της ίδιας επερώτησης από τον ίδιο χρήστη. Με τον τρόπο αυτό είναι δυνατόν να μειωθεί σημαντικά ο αριθμός των επιστρεφόμενων ιστοσελίδων για επερωτήσεις, που επαναλαμβάνονται περιοδικά. Συνεπώς, στην περίπτωση αυτή αντιμετωπίζεται το πρόβλημα του καταιγισμού πληροφορίας και διευκολύνεται ιδιαίτερη η ανάκληση της **πρόσφατης** πληροφορίας.

Επιπροσθέτως, η ενσωμάτωση της λειτουργίας της μνήμης σε μία μηχανή αναζήτησης μπορεί να ευνοήσει σημαντικά την αναγνώριση και επεξεργασία της ανανεωμένης πληροφορίας. Συχνά, το ενδιαφέρον σε ένα κείμενο επικεντρώνεται στις **μεταβολές**, που μπορεί να γίνουν σε αυτό περιοδικά, ως αποτέλεσμα μίας αλλαγής είτε ανανέωσης του περιγραφόμενου θέματος. Στις περιπτώσεις αυτές είναι χρήσιμη και επιθυμητή η άμεση αναγνώριση των διαφορών της νέας έκδοσης του κειμένου από τις παλαιότερες. Η ιδιότητα αυτή δίνει τη δυνατότητα στον ερευνητή να επικεντρώσει τη μελέτη του αποκλειστικά στην ανανεωμένη πληροφορία.

Ολοκληρώνοντας, οι προαναφερόμενες λειτουργίες μπορούν να ενσωματωθούν σε μία μηχανή αναζήτησης με τη μορφή ενός ανεξάρτητου επιπέδου λογισμικού, το οποίο αναλαμβάνει την προεπεξεργασία της επιστρεφόμενης πληροφορίας πριν την παρουσιάσει στο χρήστη.

1.5 USEwebNET

Μέρος της εργασίας αυτής αποτελεί η ανάπτυξη του USEwebNET. Το USEwebNET είναι ένα εργαλείο, το οποίο διευκολύνει την αναζήτηση πληροφορίας στον Παγκόσμιο Ιστό. Αποτελεί ένα ανεξάρτητο επίπεδο λογισμικού, που λειτουργεί πάνω από ήδη υπάρχουσες μηχανές αναζήτησης. Κύριος στόχος του είναι η παροχή μίας υπηρεσίας, που θα επιτρέπει στους χρήστες να ενημερώνονται εύκολα για τις τελευταίες εξελίξεις των θεμάτων που τους ενδιαφέρουν. Για να το πετύχει αυτό διατηρεί μία ιστορία με όλα τα πεδία ενδιαφέροντος του χρήστη και τις ιστοσελίδες, που επιστρέφονται από τις μηχανές αναζήτησης κατά την εκτέλεση των επερωτήσεων. Ιστοσελίδες, που κρίνονται αδιάφορες από το χρήστη δεν ξαναεμφανίζονται στα αποτελέσματα της συγκεκριμένης επερώτησης. Ενδιαφέρουσες ιστοσελίδες, που έχουν αναγνωστεί από το χρήστη δεν ξαναεμφανίζονται, παρά μόνο εάν ανανεωθούν. Κατά συνέπεια, το USEwebNET δίνει λύση στο πρόβλημα του καταιγισμού πληροφορίας και του διαχωρισμού της νέας γνώσης από την παλαιότερη.

Παράλληλα, επιτρέπει την εύκολη αναγνώριση των ανανεώσεων είτε μεταβολών σε ιδιαίτερα ενδιαφέρουσες ιστοσελίδες. Η λειτουργία αυτή επιτυγχάνεται δίνοντας τη δυνατότητα στο χρήστη να χαρακτηρίσει συγκεκριμένες σελίδες ως ιδιαίτερα ενδιαφέρουσες. Από το σημείο αυτό και έπειτα το USEwebNET φροντίζει να ελέγχει περιοδικά τις ιστοσελίδες αυτές, και εάν εντοπίσει μεταβολές τις υποδεικνύει στο χρήστη. Επομένως, διευκολύνει την αναγνώριση ανανεωμένης, νέας πληροφορίας και την εστίαση της μελέτης του ερευνητή σε αυτή.

Τέλος, προσφέρει μία σειρά χρήσιμων λειτουργιών, πάνω στις ενδιαφέρουσες ιστοσελίδες, όπως είναι η δυνατότητες της διατήρησης τους σε ξεχωριστά αρχεία (save) και της διαγραφής τους. Η επεξεργασία της επιστρεφόμενης πληροφορίας γίνεται διαμέσου μίας ευέλικτης διεπιφάνειας χρήσης εμπνευσμένη από αυτή, που χρησιμοποιείται για την ανάγνωση των USENET News.

1.6 ΣΥΝΕΙΣΦΟΡΕΣ (CONTRIBUTIONS)

Οι συνεισφορές της εργασίας αυτής είναι:

- Η ανάπτυξη ενός εργαλείου για τη διευκόλυνση της ανάκλησης πληροφορίας από τον Παγκόσμιο Ιστό.
- Η επέκταση του για την αποτελεσματική αναζήτηση ερευνητικών δημοσιεύσεων.
- Η δημιουργία μίας εύχρηστης και ευέλικτης διεπιφάνειας χρήσης για την επεξεργασία της ανακαλούμενης πληροφορίας.
- Η σχεδίαση και υλοποίηση αλγορίθμων για την ταξινόμηση των ανακαλούμενων δημοσιεύσεων ανάλογα με τα ενδιαφέροντα του χρήστη.
- Η σχεδίαση μεθόδων για τη διεύρυνση των επερωτήσεων και την επαύξηση του συνόλου των ενδιαφερόντων ανακαλούμενων αποτελεσμάτων.

1.7 Η ΟΡΓΑΝΩΣΗ ΤΗΣ ΕΡΓΑΣΙΑΣ

Στα Κεφάλαια 2 και 3 περιγράφεται η αρχιτεκτονική και η υλοποίηση του USEwebNET. Στο Κεφάλαιο 4 γίνεται αναφορά στο σχεδιασμό του PaperFinder καθώς και σε θέματα υλοποίησης του. Στο Κεφάλαιο 5 περιγράφονται οι λειτουργίες διεύρυνσης των επερωτήσεων και ταξινόμησης των άρθρων του PaperFinder. Τέλος στο Κεφάλαιο 6 γίνεται η επισκόπηση σχετικών εργασιών και στο Κεφάλαιο 8 δίνεται μία περίληψη και τα συμπεράσματα της εργασίας αυτής.

Στο επόμενο κεφάλαιο θα αναλυθεί λεπτομερώς η λειτουργία και ο σχεδιασμός του USEwebNET.

Κεφάλαιο 2

USEWEBNET: ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΚΑΙ ΔΙΕΠΙΦΑΝΕΙΑ ΧΡΗΣΗΣ

2.1 Η ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ USEwebNET

Στόχος του USEwebNET είναι να προσφέρει στους χρήστες ένα ευέλικτο εργαλείο, ικανό να απλοποιήσει την επίπονη διαδικασία της ανάκλησης πληροφορίας από τον Παγκόσμιο Ιστό. Για την επίτευξη του στόχου αυτού, το USEwebNET έχει σχεδιαστεί ως μία επιπρόσθετη υπηρεσία πάνω από ορισμένες πολύ διαδεδομένες μηχανές αναζήτησης, όπως είναι η Alta-Vista της Digital [I] και το HotBot [II]. Από τα χαρακτηριστικά του USEwebNET πιο σημαντικά είναι η δυνατότητα διατήρησης ενός προσωπικού πορτραίτου (user profile) για κάθε χρήστη και η ικανότητα του εύκολου διαχωρισμού της νέας από την παλαιότερη πληροφορία.

Η βασική ιδέα πίσω από το USEwebNET είναι η ικανότητα να διατηρεί στοιχεία για τα ενδιαφέροντα των χρηστών και να επερωτεί διαδεδομένες μηχανές αναζητήσεις σε τακτικά χρονικά διαστήματα, αναζητώντας νέα πληροφορία. Μία επερώτηση αποτελείται από ένα σύνολο λέξεων-κλειδιών, οι οποίες χαρακτηρίζουν το πεδίο ενδιαφέροντος του χρήστη, και μία ημερομηνία (την παλαιότερη ημερομηνία δημιουργίας ή ανανέωσης που μπορούν να έχουν οι επιστρεφόμενες ιστοσελίδες). Επιπροσθέτως, δίνεται η δυνατότητα εκτέλεσης ορισμένων χρήσιμων λειτουργιών, όπως είναι η αποθήκευση ιστοσελίδων σε ξεχωριστό αρχείο, η δυνατότητα αναγνώρισης και ελέγχου πιθανών αλλαγών και η απόρριψη ιστοσελίδων ως με σχετικές με το πεδίο ενδιαφέροντος του χρήστη.

Από την πλευρά της αρχιτεκτονικής, το USEwebNET μπορεί να διαχωριστεί σε δύο τμήματα, τα οποία λειτουργούν σε μεγάλο βαθμό ανεξάρτητα μεταξύ τους. Το

πρώτο τμήμα είναι υπεύθυνο για τον έλεγχο της διεπιφάνειας χρήσεως. Ευθύνεται για την καταγραφή των επερωτήσεων των χρηστών, την ορθή εκτέλεση των αιτήσεων τους και για την παρουσίαση των αποτελεσμάτων σε αυτούς. Το δεύτερο τμήμα αλληλεπιδρά με τις μηχανές αναζήτησης, μεταβιβάζοντας τις επερωτήσεις των χρηστών σε αυτές και συλλέγοντας τα αποτελέσματα, τα οποία μετά από κατάλληλη επεξεργασία αποθηκεύονται στη βάση δεδομένων του USEwebNET.

Στη συνέχεια του κεφαλαίου αυτού θα παρουσιαστεί η διεπιφάνεια χρήσης του USEwebNET και θα επεξηγηθεί ο τρόπος με τον οποίο αυτό επικοινωνεί με τις μηχανές αναζήτησης.

2.2 ΔΙΕΠΙΦΑΝΕΙΑ ΧΡΗΣΗΣ ΤΟΥ USEwebNET

Η διεπιφάνεια χρήσης του USEwebNET στηρίζεται στο περιβάλλον αναδίφησης του Παγκοσμίου Ιστού, δηλαδή μπορεί να χρησιμοποιηθεί από ένα κοινό αναδιφητή (web browser). Ο χρήστης αλληλεπιδρά με το σύστημα επιλέγοντας συνδέσμους είτε εικονίδια, που εμφανίζονται στις δυναμικά δημιουργούμενες ιστοσελίδες. Η χρήση του εργαλείου μπορεί να ξεκινήσει κατευθύνοντας ένα αναδιφητή του Παγκοσμίου Ιστού (web browser), όπως είναι το Netscape Navigator, σε ένα καθορισμένο εξυπηρετητή (web server). Ο τρέχον εξυπηρετητής του USEwebNET είναι ο: *cuckoo.ics.forth.gr:9002*. Από τη στιγμή, που θα δημιουργηθεί η σύνδεση με τον εξυπηρετητή, θα εμφανιστεί η "Σελίδα Εισαγωγής" (Login Page) στο παράθυρο του αναδιφητή. Από τη σελίδα αυτή, ο χρήστης μπορεί να πιστοποιήσει την ταυτότητα του και να ξεκινήσει τη χρήση του συστήματος, εάν έχει ένα λογαριασμό (account) στον εξυπηρετητή ή να δημιουργήσει ένα νέο λογαριασμό μέσω του συνδέσμου, που οδηγεί στην "Ιστοσελίδα Καταχώρησης Χρηστών" (Registration Page). Στα Σχήματα 2-1 και 2-2 παρουσιάζονται οι φόρμες πιστοποίησης της ταυτότητας των χρηστών και δημιουργίας νέου λογαριασμού αντίστοιχα.

Σχήμα 2-1: Η φόρμα "Πιστοποίησης Ταυτότητας" του χρήστη. Απο το σύνδεσμο "Create an account here" ο χρήστης οδηγείται στη σελίδα δημιουργίας λογαριασμού.

Σχήμα 2-2: Η οθόνη δημιουργίας νέου λογαριασμού. Ο χρήστης καθορίζει το όνομα του λογαριασμού του, το password του, την ηλεκτρονική του διεύθυνση, το όνομα του και την περιοχή, από την οποία χρησιμοποιεί το USEwebNET. Από τα στοιχεία αυτά, το όνομα και ο τόπος στον οποίο εργάζεται είναι προαιρετικά.

Μετά την πιστοποίηση της ταυτότητας του χρήστη εμφανίζεται η ιστοσελίδα με τις βασικές επιλογές του USEwebNET. Συγκεκριμένα, υποστηρίζονται τέσσερις δυνατότητες:

- ***Setup:*** Από την επιλογή αυτή δίνεται η δυνατότητα στο χρήστη να δημιουργήσει νέες επερωτήσεις προς τις μηχανές αναζήτησης, να απαιτήσει την εκτέλεση των επερωτήσεων και να προκαλέσει μεταβολές στα στοιχεία των ήδη υπάρχουσων επερωτήσεων.

- ***Results:*** Ανάγνωση και γενικότερα επεξεργασία των αποτελεσμάτων των επερωτήσεων.

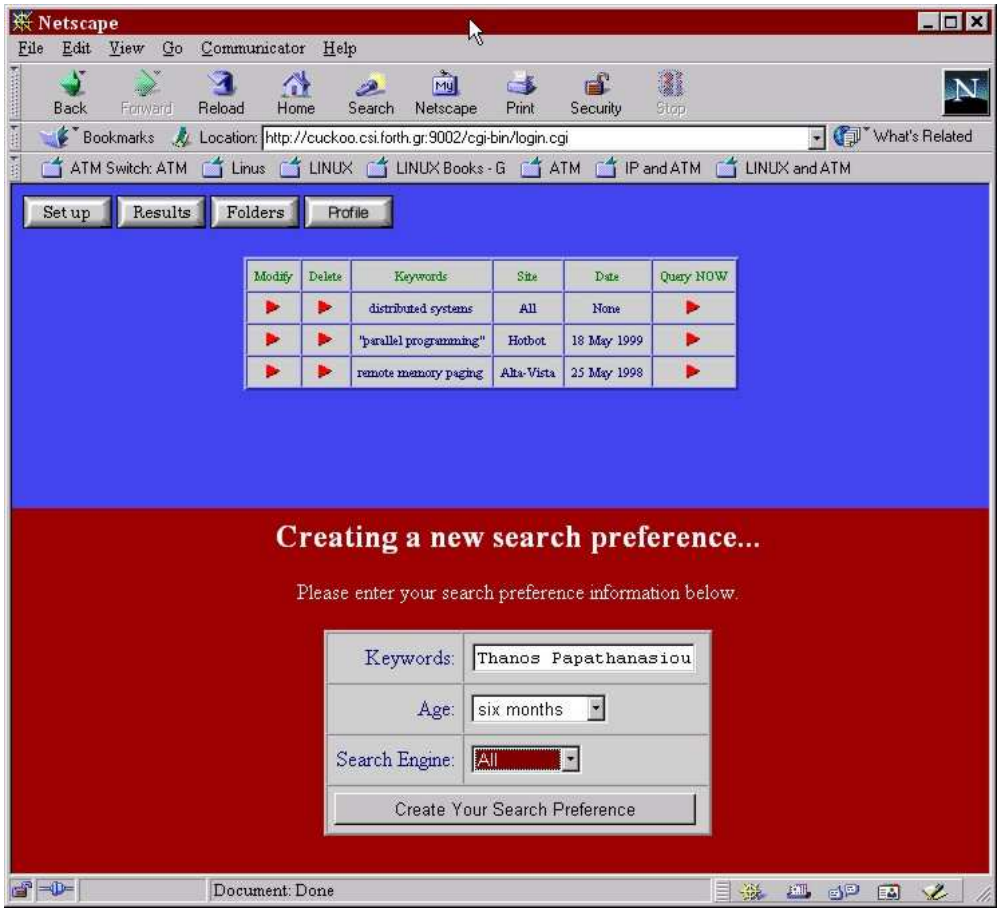
- ***Folders:*** Επεξεργασία των περιεχομένων των προσωπικών αρχείων, που διατηρεί ο χρήστης στον εξυπηρετητή του USEwebNET.

- ***Profile Definition:*** Καθορισμός του προσωπικού πορτραίτου του χρήστη. Καθορίζονται ορισμένες μεταβλητές, οι οποίες ελέγχουν τον τρόπο λειτουργίας του USEwebNET.

Στις επόμενες ενότητες επεξηγούνται αναλυτικά οι προαναφερόμενες λειτουργίες.

2.2.1 ΙΣΤΟΣΕΛΙΔΑ ΚΑΤΑΧΩΡΗΣΗΣ ΕΠΕΡΩΤΗΣΕΩΝ

Διαμέσου της ιστοσελίδας ρύθμισης επερωτήσεων (Σχήμα 2-3) ο χρήστης μπορεί να καταχωρήσει τα πεδία ενδιαφέροντος του είτε να διαγράψει – μεταβάλει μία προϋπάρχουσα επερώτηση.



Σχήμα 2-3: Η ιστοσελίδα καταχώρησης νέων επερωτήσεων. Ο χρήστης έχει ήδη δημιουργήσει τρεις επερωτήσεις και ετοιμάζεται να καταχωρήσει μία νέα. Για τις ήδη δημιουργημένες επερωτήσεις φαίνονται οι λέξεις-κλειδιά, που χρησιμοποιούνται, οι μηχανές αναζητήσεις, στις οποίες προωθούνται, και η ημερομηνία παλαιότερης δυνατής ενημέρωσης ή δημιουργίας αυτών.

Με την επιλογή του καταλόγου Καταχώρησης Επερωτήσεων (Setup Menu), η κεντρική ιστοσελίδα της εφαρμογής διαχωρίζεται σε δύο σκελετούς (frames). Ο ανώτερος σκελετός περιέχει πληροφορία σχετική με τα τρέχοντα ενδιαφέροντα του χρήστη. Η πληροφορία αυτή παρουσιάζεται με τη μορφή ενός πίνακα, του οποίου κάθε σειρά είναι αφιερωμένη στην περιγραφή μίας συγκεκριμένης επερώτησης. Οι

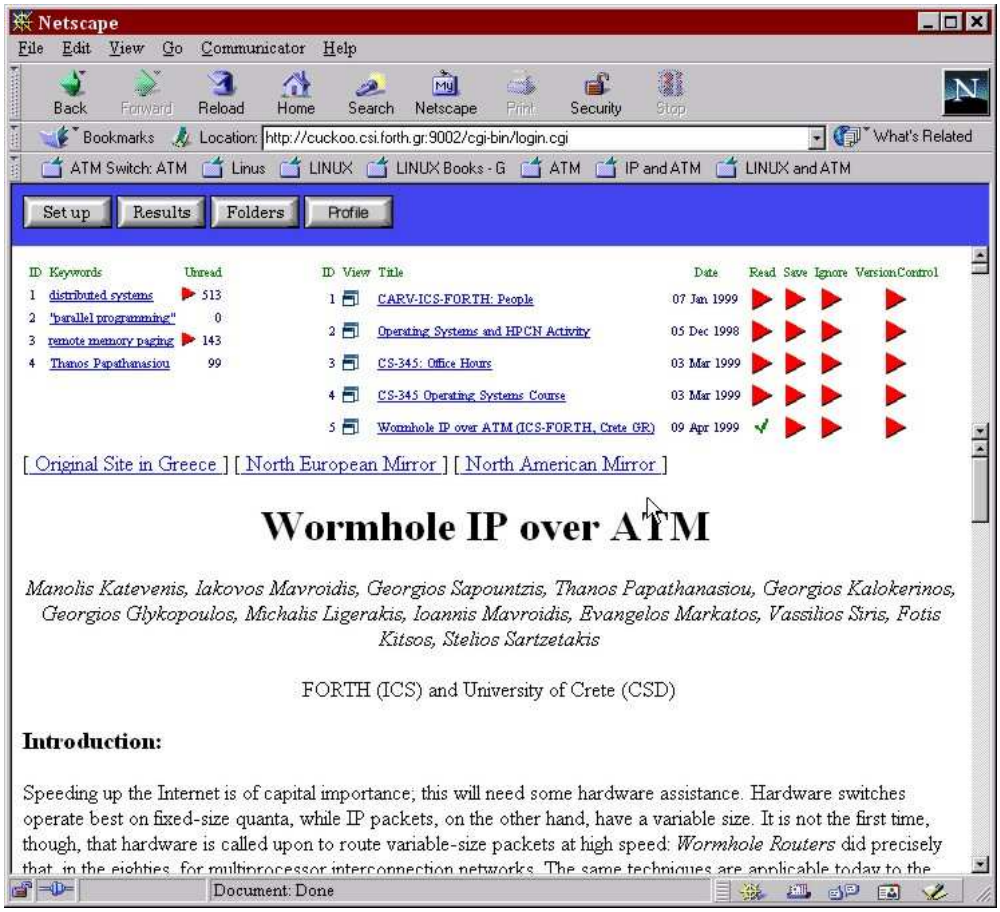
στήλες του πίνακα περιγράφουν ένα συγκεκριμένο χαρακτηριστικό της επερώτησης είτε χρησιμοποιούνται για να δώσουν στο χρήστη τη δυνατότητα εκτέλεσης μίας λειτουργίας. Συγκεκριμένα οι στήλες του πίνακα περιγραφής επερωτήσεων χρησιμοποιούνται, όπως περιγράφεται στην επόμενη παράγραφο.

Από αριστερά προς τα δεξιά (Σχήμα 2-3) οι δύο πρώτες στήλες περιέχουν συνδέσμους για την εκτέλεση των λειτουργιών χρήσης ``Τροποποίησης'' (Modify) και ``Διαγραφής'' (Delete). Η τρίτη στήλη περιέχει τις ``Λέξεις-κλειδιά'' (Keywords), που χαρακτηρίζουν την επερώτηση και η τέταρτη τη ``Μηχανές αναζήτησης'' (Site), στην οποία κατευθύνεται η επερώτηση. Η ένδειξη ``*None*'' στη στήλη αυτή δηλώνει ότι η επερώτηση είναι **απενεργοποιημένη,** δηλαδή δεν προωθείται σε καμία μηχανή αναζήτησης και δεν συλλέγονται αποτελέσματα για αυτήν, ενώ η ένδειξη ``*All*'' δείχνει ότι προωθείται σε όλες τις μηχανές αναζήτησης, που υποστηρίζει το USEwebNET. Ακολούθως, στην πέμπτη στήλη εμφανίζεται η παλαιότερη ημερομηνία (Date) δημιουργίας ή ανανέωσης, που μπορεί να έχει μία επιστρεφόμενη ιστοσελίδα για την επερώτηση αυτή. Η ένδειξη ``*None*'' **απενεργοποιεί** τον έλεγχο της ημερομηνίας κατά την ανάκληση πληροφορίας. Ολοκληρώνοντας, η τελευταία στήλη με την επικεφαλίδα Query περιέχει ένα σύνδεσμο για την **άμεση εκτέλεση** της επερώτησης.

Στον κατώτερο σκελετό του παραθύρου του αναδιφητή, εμφανίζεται η φόρμα δημιουργίας νέων επερωτήσεων. Διαμέσου της φόρμας αυτής ο χρήστης μπορεί να δηλώσει τα χαρακτηριστικά μιας επερώτησης (λέξεις κλειδιά, παλαιότερη ημερομηνία δημιουργίας ή ανανέωσης των επιστρεφόμενων ιστοσελίδων, μηχανές αναζήτησης) και να δημιουργήσει την επερώτηση.

2.2.2 ΙΣΤΟΣΕΛΙΔΑ ΠΑΡΟΥΣΙΑΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Με την επιλογή παρουσίασης αποτελεσμάτων (Results Menu), η κεντρική σελίδα του αναδιφητή διαχωρίζεται σε τρεις σκελετούς.



Σχήμα 2-4: Η ιστοσελίδα παρουσίασης αποτελεσμάτων. Ο χρήστης έχει ορίσει τέσσερις επερωτήσεις και ελέγχει τα αποτελέσματα της επερώτησης 4. Στο κάτω μέρος του παραθύρου διαβάζει την ιστοσελίδα με τίτλο *"Wormhole IP over ATM"*, η οποία εμφανίστηκε ως αποτέλεσμα της επερώτησης *"Thanos Papathanasiou"*. Το αποτέλεσμα αυτό (αύξων αριθμός 5) έχει σημειωθεί ως αναγνωσμένο στη λίστα των αποτελεσμάτων.

Όπως, φαίνεται στο Σχήμα 2-4, στο ανώτερο αριστερό τμήμα του παραθύρου παρουσιάζονται τα τρέχοντα πεδία ενδιαφέροντος του χρήστη, με τη μορφή των λέξεων-κλειδιών, που χρησιμοποιήθηκαν κατά τη δημιουργία των επερωτήσεων. Η πληροφορία αυτή βρίσκεται στο πεδίο με επικεφαλίδα ``*Keywords*''. Αμέσως δεξιά,

κάτω από την επικεφαλίδα *"Unread"* είναι καταγεγραμμένος ο αριθμός των άρθρων σχετικών με το αντίστοιχο πεδίο ενδιαφέροντος, που δεν έχουν αναγνωστεί.

2.2.2.1 ΛΕΙΤΟΥΡΓΙΑ ``CATCH UP''

Ανάμεσα στις λέξεις-κλειδιά κάθε επερώτησης και στον αριθμό των μη αναγνωσμένων αποτελεσμάτων είναι πιθανό να εμφανιστεί ένα μικρό κόκκινο εικονίδιο, εάν η ποσότητα των αδιάβαστων ιστοσελίδων είναι μεγάλη. Το τριγωνικό αυτό εικονίδιο δίνει τη δυνατότητα στο χρήστη να χαρακτηρίσει όλες ή την πλειοψηφία των ιστοσελίδων της συγκεκριμένης επερώτησης ως αναγνωσμένες. Η λειτουργία αυτή, ονομαζόμενη ``*Catch Up*'' είναι ιδιαίτερα χρήσιμη στην περίπτωση που ο χρήστης ενδιαφέρεται αποκλειστικά στην ανάκληση νέων ιστοσελίδων από μία χρονική στιγμή και έπειτα και θέλει να ακυρώσει κάθε είδους πληροφορία, που είχε συλλεχθεί στο παρελθόν.

2.2.2.2 ΚΑΤΑΣΤΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ

Οι λέξεις-κλειδιά κάθε πεδίου ενδιαφέροντος αποτελούν ένα σύνδεσμο, η επιλογή του οποίου έχει ως αποτέλεσμα την εμφάνιση των αποτελεσμάτων της συγκεκριμένης επερώτησης στο ανώτερο δεξιό τμήμα του παραθύρου του αναδιφητή. Τα αποτελέσματα παρουσιάζονται με τη μορφή πίνακα. Από αριστερά προς τα δεξιά το πρώτο πεδίο του πίνακα έχει την επικεφαλίδα ``*ID*'' και περιέχει τον αύξοντα αριθμό του αποτελέσματος. Ακολουθεί ένα μικρό εικονίδιο με το σχήμα δύο μικρών παραθύρων, το οποίο δίνει τη δυνατότητα στο χρήστη να διαβάσει την ιστοσελίδα που βρέθηκε σε ένα ξεχωριστό παράθυρο του αναδιφητή. Στη δεύτερη στήλη με ονομασία ``*Title*'' εμφανίζεται ο τίτλος της ιστοσελίδας, που ανακλήθηκε και στην τρίτη με τίτλο ``*Date*'' η ημερομηνία δημιουργίας ή τελευταίας ανανέωσης αυτής. Ακολουθούν τέσσερα πεδία με επικεφαλίδες: ``*Read*'', ``*Save*'', ``*Ignore*'' και ``*Version Control*'', τα οποία υποδεικνύουν την κατάσταση της συγκεκριμένης ιστοσελίδας. Τα πεδία αυτά μπορούν να περιέχουν ένα μικρό κόκκινο τρίγωνο ή ένα σημάδι επαλήθευσης (tick). Το κόκκινο τρίγωνο δηλώνει ότι η

αντίστοιχη λειτουργία δεν έχει γίνει για την συγκεκριμένη ιστοσελίδα και δίνει τη δυνατότητα πραγματοποίησης της, ενώ το σημάδι επαλήθευσης δείχνει ότι έχει πραγματοποιηθεί η αντίστοιχη λειτουργία και δίνει τη δυνατότητα ακύρωσης της.

Συγκεκριμένα η πληροφορία, που δίνουν τα τέσσερα πεδία για την κατάσταση της σελίδας είναι η εξής:

- ``**Read":** Η ιστοσελίδα έχει αναγνωσθεί.

- ``**Save":** Η ιστοσελίδα έχει αποθηκευτεί σε ένα προσωπικό αρχείο. Η ενεργοποίηση της διαδικασίας αποθήκευσης έχει ως αποτέλεσμα την εμφάνιση ενός καταλόγου στο κατώτερο τμήμα του παραθύρου του αναδιφητή, από το οποίο ο χρήστης μπορεί να επιλέξει το αρχείο, στο οποίο θα αποθηκευτεί η ιστοσελίδα. Στον κατάλογο αυτό εμφανίζονται τα ονόματα όλων των προσωπικών αρχείων του χρήστη και παράλληλα δίνεται η δυνατότητα για δημιουργία νέου αρχείου.

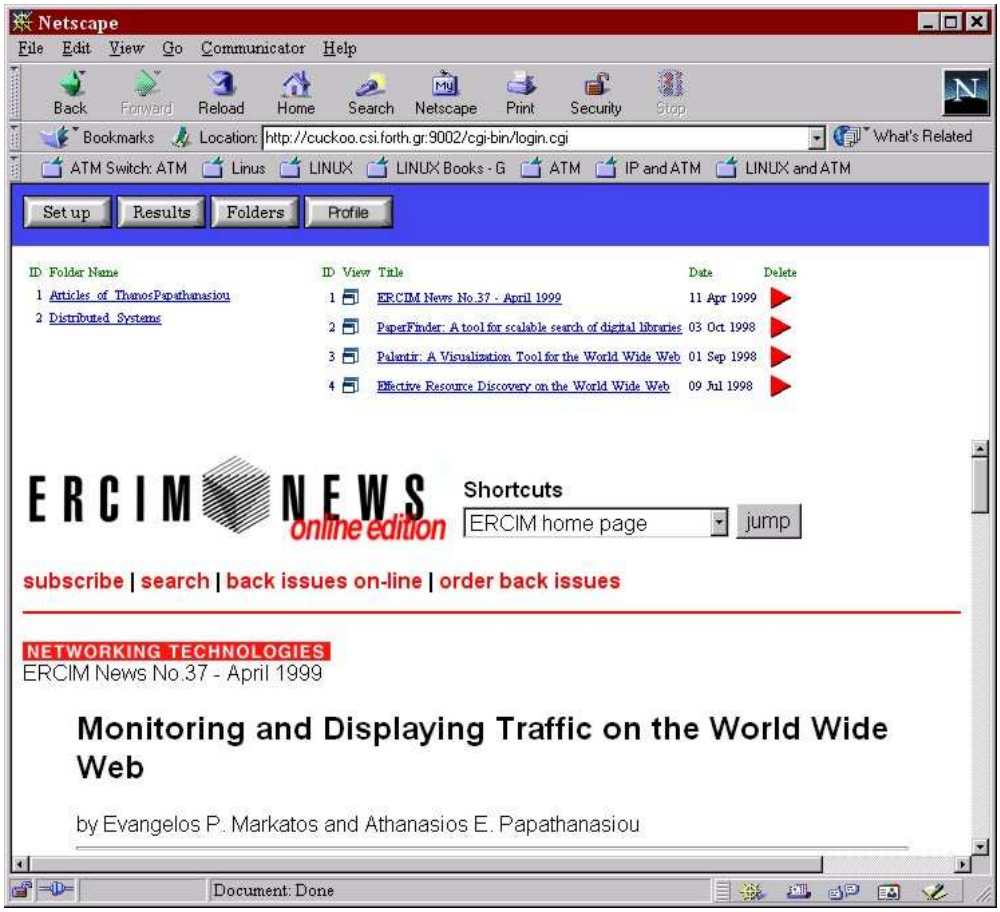
- ``**Ignore":** Η ιστοσελίδα δεν σχετίζεται με το πεδίο ενδιαφέροντος του χρήστη και δεν πρόκειται να εμφανισθεί ξανά ως αποτέλεσμα στη συγκεκριμένη επερώτηση.

- ``**Version Control":** Η ιστοσελίδα ελέγχεται περιοδικά για πιθανή ανανέωση της. Η ανακάλυψη μεταβολών σε αυτή θα έχει ως αποτέλεσμα την ενημέρωση του χρήστη και την υπόδειξη των αλλαγών.

Ολοκληρώνοντας, το κατώτερο τμήμα του παραθύρου του αναδιφητή χρησιμοποιείται για την ανάγνωση της ιστοσελίδας και την εμφάνιση μηνυμάτων σχετικών με τη λειτουργία του USEwebNET.

2.2.3 *ΙΣΤΟΣΕΛΊΔΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΑΡΧΕΙΩΝ*

Επιλέγοντας το εικονίδιο με τίτλο *"Folders"*, ο χρήστης οδηγείται στην ιστοσελίδα επεξεργασίας προσωπικών αρχείων.



Σχήμα 2-5: Ιστοσελίδα παρουσίασης προσωπικών αρχείων. Ο χρήστης οδηγείται στην οθόνη επεξεργασίας των προσωπικών του αρχείων μέσω της επιλογής *"Folders"*. Έχουν δημιουργηθεί δύο διαφορετικά αρχεία αποτελεσμάτων. Ο χρήστης έχει ανοίξει το αρχείο *"Articles_of_ThanosPapathanasiou"* και διαβάζει το άρθρο με τίτλο *"Ercim News No 37 - April 1999"*.

Όπως φαίνεται στο Σχήμα 2-5, το παράθυρο του αναδιφητή χωρίζεται σε τρία τμήματα. Στο ανώτερο αριστερά εμφανίζεται ένας κατάλογος με τα προσωπικά

αρχεία του χρήστη. Η επιλογή ενός από αυτά έχει ως αποτέλεσμα την εμφάνιση των περιεχομένων ιστοσελίδων στο ανώτερο δεξιά τμήμα του παραθύρου. Για κάθε ιστοσελίδα υποστηρίζονται οι ακόλουθες λειτουργίες:

- **Ανάγνωση σε νέο παράθυρο:** Ενεργοποιείται μέσω της επιλογής του εικονιδίου, που βρίσκεται από το πεδίο με επικεφαλίδα *"View"*.

- **Ανάγνωση (στο τρέχον παράθυρο):** Ενεργοποιείται μέσω της επιλογής του τίτλου της ιστοσελίδας. Τα περιεχόμενα της ιστοσελίδας εμφανίζονται στο κατώτερο τμήμα του παραθύρου του αναδιφητή.

- **Διαγραφή:** Διαγραφή μίας ιστοσελίδας από το αρχείο.

2.2.4 ΙΣΤΟΣΕΛΙΔΑ ΡΥΘΜΙΣΗΣ ΠΡΟΣΩΠΙΚΟΥ ΠΟΡΤΡΑΙΤΟΥ

Με την επιλογή του εικονιδίου με ονομασία *"Profile"* δίνεται στο χρήστη η δυνατότητα καθορισμού του προσωπικού του πορτραίτου (personal profile). Στην τρέχουσα έκδοση του USEwebNET προσφέρονται προς ρύθμιση τα ακόλουθα στοιχεία:

(α) Αντιμετώπιση προσπελασμένων ιστοσελίδων: Οι προσπελασμένες ιστοσελίδες μπορούν να παραμένουν στη λίστα των αποτελεσμάτων σημειωμένες ως διαβασμένες είτε να διαγράφονται και να ξαναεμφανίζονται μόνο σε περίπτωση, που ανανεωθούν.

(β) Ρύθμιση λειτουργίας "Catch Up": Όπως αναφέρθηκε στην Ενότητα 2.2.2.1, η λειτουργία *"Catch Up"* είναι αυτή, που επιτρέπει στο χρήστη να σημειώσει ως αναγνωσμένες όλες τις ιστοσελίδες, που έχουν ευρεθεί έως εκείνη τη χρονική στιγμή για μία επερώτηση. Με τον τρόπο αυτό, του δίνεται η δυνατότητα να παρακολουθεί μόνο ό,τι νεότερο υπάρχει για το συγκεκριμένο πεδίο ενδιαφέροντος. Η λειτουργία μπορεί να ρυθμιστεί με δύο τρόπους. Αρχικά, ο

χρήστης καθορίζει τον ελάχιστο αριθμό επιστρεφομένων ιστοσελίδων, για τις οποίες ενεργοποιείται η λειτουργία *"Catch Up"*. Στη συνέχεια καθορίζεται ο αριθμός των ιστοσελίδων, που παραμένουν σημειωμένες ως **μη** αναγνωσμένες μετά την εφαρμογή της λειτουργίας *"Catch Up"*.

Σχήμα 2-6: Ο καθορισμός του προσωπικού πορτραίτου ενός χρήστη. Ο χρήστης έχει δηλώσει ότι επιθυμεί: (α) την αυτόματη διαγραφή των αναγνωσμένων URLs, (β) την ενεργοποίηση της λειτουργίας *Catch Up* για επερωτήσεις με **100** ή περισσότερα αποτελέσματα, (γ) την αλλαγή της κατάστασης σε αναγνωσμένα όλων των άρθρων μετά από την εκτέλεση του *Catch Up* και (δ) την αποστολή ηλεκτρονικής ειδοποίησης σε περίπτωση, που βρεθούν πέντε τουλάχιστον νέα URLs για μία επερώτηση.

(γ) Ειδοποίηση διαμέσου ηλεκτρονικού μηνύματος (Mail Notification): Το USEwebNET δίνει τη δυνατότητα στους χρήστες να ενημερώνονται μέσω ηλεκτρονικού μηνύματος για την κατάσταση των επερωτήσεων τους. Συγκεκριμένα, οι χρήστες μπορούν να επιλέξουν να τους αποστέλλεται σχετικό ηλεκτρονικό μήνυμα σε περίπτωση, που ο αριθμός των νέων ιστοσελίδων για μία επερώτηση ξεπεράσει ένα ελάχιστο κάτω όριο. Το όριο αυτό καθορίζεται από κάθε χρήστη μέσα από τη ρύθμιση του προσωπικού του πορτραίτου.

Ο τρόπος με τον οποίο γίνεται η ρύθμιση του προσωπικού πορτραίτου του χρήστη φαίνεται στο Σχήμα 2-6.

2.2.5 ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ

Στην ενότητα αυτή, θα δοθεί ένα παράδειγμα λειτουργίας του USEwebNET. Ας υποτεθεί ότι ο χρήστης ενδιαφέρεται να βρει πληροφορία σχετική με τον τομέα του ``Loop Scheduling". Η εισαγωγή των προτιμήσεων του γίνεται, όπως φαίνεται στο Σχήμα 2-7.

Από τη στιγμή, που ο χρήστης θα ολοκληρώσει την εισαγωγή των προτιμήσεων, το USEwebNET θα δημιουργήσει μία περίληψη αυτών, όπως φαίνεται στο Σχήμα 2-8. Οι επερωτήσεις θα προωθηθούν στις μηχανές αναζήτησης αργότερα, σε βραδινή ώρα, εκτός εάν ο χρήστης υποδείξει ότι επιθυμεί την άμεση εκτέλεση τους μέσω του εικονιδίου *"Query Now"*.

Μετά την ολοκλήρωση της ανάκλησης πληροφορίας, ο χρήστης μπορεί να επισκεφθεί την οθόνη παρουσίασης αποτελεσμάτων μέσω του συνδέσμου *"Results"*, όπου θα εμφανιστεί μία εικόνα όμοια με αυτή του Σχήματος 2-10. Η Εικόνα αυτή πληροφορεί το χρήστη ότι έχουν βρεθεί **513** (μη αναγνωσμένες) ιστοσελίδες για την επερώτηση **1**, **0** για την επερώτηση **2**, **143** για την επερώτηση **3** και **763** για την επερώτηση **4**. Ο χρήστης μπορεί τώρα να επιλέξει μία από τις αναφερόμενες επερωτήσεις, οπότε θα εμφανιστεί μία εικόνα όμοια με αυτή του Σχήματος 2-11. Από την ιστοσελίδα αυτή, είναι δυνατή η ανάγνωση, διαγραφή και αποθήκευση ενός αποτελέσματος. Οι δυνατότητες αυτές δίνονται μέσω των συνδέσμων, που συνοδεύουν κάθε αποτέλεσμα. Στο Σχήμα 2-11 φαίνεται η οθόνη των αποτελεσμάτων της επιλεγμένης επερώτησης μετά από την εκτέλεση ορισμένων λειτουργιών αποθήκευσης, ανάγνωσης, μόνιμης διαγραφής και έναρξης ελέγχου εκδόσεων από το χρήστη.

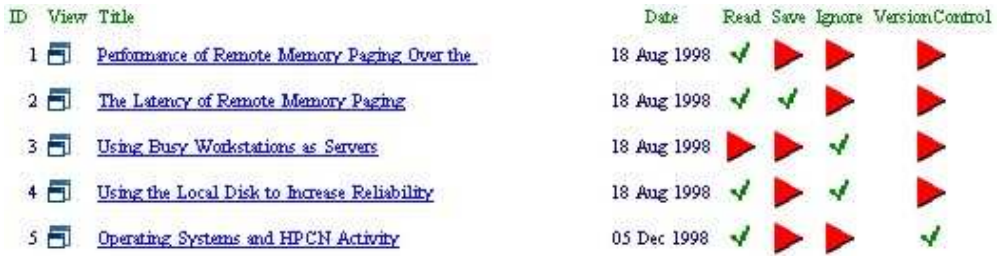
Σχήμα 2-7: Εισαγωγή μίας επερώτησης πάνω στο θέμα *"Loop Scheduling"*. Το όριο που δίνεται για την ηλικία των επιστρεφόμενων σελίδων είναι ένα έτος και η επερώτηση κατευθύνεται σε όλες τις υποστηριζόμενες μηχανές αναζήτησης *("All")*.

Modify	Delete	Keywords	Site	Date	Query NOW
▶	▶	distributed systems	All	None	▶
▶	▶	"parallel programming"	Hotbot	18 May 1999	▶
▶	▶	remote memory paging	Alta-Vista	25 May 1998	▶
▶	▶	loop scheduling	All	25 May 1998	▶

Σχήμα 2-8: Το USEwebNET παρουσιάζει τα στοιχεία των τεσσάρων επερωτήσεων του χρήστη.

Σχήμα 2-9: Μετά τη συλλογή των αποτελεσμάτων, το USEwebNET παρουσιάζει τα αποτελέσματα στο χρήστη σε περιληπτική μορφή. Για κάθε επερώτηση αναφέρεται ο αριθμός των "ακόμη μη αναγνωσμένων" (Unread) ιστοσελίδων.

Σχήμα 2-10: Τα αποτελέσματα της επερώτησης *"remote memory paging"*.



Σχήμα 2-11: Τα αποτελέσματα της επερώτησης *"remote memory paging"* μετά από την αποθήκευση δύο ιστοσελίδων και την ανάγνωση τεσσάρων.

2.3 ΑΛΛΗΛΕΠΙΔΡΑΣΗ ΜΕ ΤΙΣ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Το κύριο σύστημα λειτουργίας (back-end) του USEwebNET μπορεί να χωριστεί από την άποψη της αρχιτεκτονικής σε πέντε μεταξύ τους ανεξάρτητα τμήματα:

1. ***Συλλέκτης αποτελεσμάτων και Αναλυτής κειμένου (parser):*** Το τμήμα αυτό του USEwebNET είναι υπεύθυνο για την προώθηση μίας επερώτησης σε μία από της υποστηριζόμενες μηχανές αναζήτησης. Συλλέγει τις επιστρεφόμενες από αυτή ιστοσελίδες αποτελεσμάτων και στη συνέχεια ο αναλυτής κειμένου HTML εξάγει την απαιτούμενη πληροφορία. Για κάθε αποτέλεσμα η πληροφορία αυτή περιλαμβάνει το URL (Uniform Resource Locator) της αντίστοιχης ιστοσελίδας, τον τίτλο της, και την ημερομηνία δημιουργία της. Τα στοιχεία αυτά αποθηκεύονται με τη μορφή που χρησιμοποιεί η βάση δεδομένων του USEwebNET και προωθούνται στο πρόγραμμα, που ευθύνεται για την ενημέρωση της.

2. ***Τμήμα ενημέρωσης της βάσεως δεδομένων:*** Το πρόγραμμα αυτό χρησιμοποιεί ως είσοδο την έξοδο του αναλυτή κειμένου. Συγκρίνει το νέο σύνολο επιστρεφόμενων URLs με αυτά, που υπάρχουν ήδη στη βάση για την αντίστοιχη επερώτηση. Αποτελέσματα, τα οποία εμφανίζονται για πρώτη φορά, εισάγονται στη βάση. Για τα αποτελέσματα, που είναι ήδη γνωστά και υπάρχουν στη βάση του USEwebNET υπάρχουν τρεις δυνατές περιπτώσεις. Εάν έχουν ανανεωθεί από την προηγούμενη φορά, που ανακλήθηκαν, δηλαδή έχουν ημερομηνία δημιουργίας νεότερη από αυτή, με την οποία εμφανίζονται στη βάση του USEwebNET, σημειώνονται ως ανανεωμένα, ώστε να παρατηρηθούν από το χρήστη. Διαφορετικά, εάν έχουν σημειωθεί από το χρήστη ως αδιάφορα ή υπάρχουν ήδη στη βάση και δεν έχουν ανανεωθεί, αγνοούνται.

3. ***Πρόγραμμα αποστολής ηλεκτρονικού μηνύματος:*** Σε περίπτωση, που το τμήμα ενημέρωσης της βάσης διαπιστώσει ότι για μία επερώτηση έχουν βρεθεί αρκετά καινούργια ή ανανεωμένα αποτελέσματα (περισσότερα από όσα έχει

καθορίσει ο χρήστης κατά τη ρύθμιση του προσωπικού του πορτραίτου) καλεί το πρόγραμμα αποστολής ηλεκτρονικού μηνύματος για να στείλει μία σχετική ειδοποίηση στο χρήστη. Το μήνυμα, που αποστέλλεται αναφέρει την ή τις επερωτήσεις, για τις οποίες βρέθηκε ``αρκετή'' νέα πληροφορία καθώς και τον αριθμό των νέων αποτελεσμάτων για κάθε μία από αυτές.

4. ***Τμήμα χειρισμού περιοδικών λειτουργιών (Crontab Manager):*** Το πρόγραμμα αυτό είναι υπεύθυνο για την περιοδική εκτέλεση των επερωτήσεων. Για κάθε χρήστη και για κάθε επερώτηση προκαλεί κατά διαστήματα την έναρξη λειτουργίας του συλλέκτη αποτελεσμάτων και του προγράμματος ενημέρωσης της βάσης.. Συνεργάζεται με το δαίμονα ***Cron*** του λειτουργικού συστήματος Unix.

5. ***Ελεγκτής εκδόσεων ιστοσελίδων:*** Εκτελείται περιοδικά και ελέγχει ιστοσελίδες, για τις οποίες ο χρήστης επιθυμεί την στενή παρακολούθηση των εκδόσεων τους. Σε κάθε εκτέλεση του ανακαλεί από τον Παγκόσμιο Ιστό τις τρέχουσες εκδόσεις των ιστοσελίδων αυτών και τις συγκρίνει με αυτές, που είναι αποθηκευμένες στη βάση του USEwebNET. Εάν ανακαλύψει διαφορές δημιουργεί ένα αρχείο, στο οποίο τις υποδεικνύει. Ακολούθως τις σημειώνει ως ανανεωμένες και προαιρετικά ειδοποιεί το χρήστη με ηλεκτρονικό μήνυμα για να τις επισκεφτεί.

2.4 ΣΥΖΗΤΗΣΗ

Όπως έχει αναφερθεί το USEwebNET αποτελεί ένα επίπεδο λογισμικού, το οποίο εκτελείται πάνω από υπάρχοντες προμηθευτές πληροφορίας (information providers), όπως για παράδειγμα μηχανές αναζήτησης του Παγκοσμίου Ιστού. Στόχος του είναι η διευκόλυνση της ανακάλυψης και επεξεργασίας πληροφορίας πάνω σε ένα ορισμένο θέμα. Χάρη στον τρόπο, με τον οποίο έχει σχεδιαστεί είναι δυνατόν να τροποποιηθεί κατάλληλα, ώστε να συνεργάζεται με εξειδικευμένες βάσεις δεδομένων πιο αποτελεσματικά. Στην ενότητα αυτή θα μελετηθούν δύο

παραδείγματα αυτού του τύπου: ένα εργαλείο εύρεσης ερευνητικών άρθρων για ψηφιακές βιβλιοθήκες (digital libraries) και ένα εργαλείο εύρεσης προϊόντων για εφαρμογές ηλεκτρονικού εμπορίου (electronic commerce).

2.4.1 Ψηφιακές Βιβλιοθήκες (Digital Libraries)

Οι ερευνητές επιθυμούν να είναι διαρκώς ενημερωμένοι για τον τομέα της έρευνας, που τους ενδιαφέρει. Για το σκοπό αυτό, διαβάζουν επιστημονικά περιοδικά, παρακολουθούν συνέδρια και αλληλεπιδρούν με συνεργάτες. Για να μειώσουν την ποσότητα της πληροφορίας, που παραλαμβάνουν, χρησιμοποιούν μόνο ένα μικρό αριθμό περιοδικών και παρακολουθούν σχετικά λίγα συνέδρια. Δυστυχώς, ο αριθμός των επιστημονικών εκδόσεων αυξάνεται σημαντικά κάθε χρόνο, δυσχεραίνοντας το ήδη δύσκολο έργο της παρακολούθησης των νέων άρθρων, που εκδίδονται σε ένα τομέα της έρευνας. Κατά συνέπεια, ένα εργαλείο, το οποίο θα μπορούσε να παραδίδει στους ερευνητές μόνο τα ενδιαφέροντα ερευνητικά άρθρα που αφορούν τον τομέα τους, θα ήταν ιδιαίτερα χρήσιμο.

Το USEwebNET βρίσκεται πολύ κοντά στο ιδανικό αυτό εργαλείο. Οι περισσότερες επιστημονικές εκδόσεις προσφέρουν ηλεκτρονικές βάσεις δεδομένων, οι οποίες περιέχουν τους τίτλους, συγγραφείς, περιλήψεις (abstracts) και μερικές φορές το πλήρες κείμενο των άρθρων. Το USEwebNET είναι δυνατόν να χρησιμοποιηθεί ως ένα εργαλείο αναζήτησης άρθρων πάνω από υπάρχουσες ψηφιακές βιβλιοθήκες. Για παράδειγμα, ένας ερευνητής μπορεί να δημιουργήσει στο USEwebNET επερώτηση σχετική με το θέμα του "web caching". Το USEwebNET θα παρακολουθεί περιοδικά τις υποστηριζόμενες βάσεις δεδομένων για να βρει νέα άρθρα σχετικά με το συγκεκριμένο θέμα. Ο ερευνητής μπορεί να ελέγχει και να επεξεργάζεται τα άρθρα, που του παραδίδει το USEwebNET. Αναγνωσμένα άρθρα διαχωρίζονται από τα παλαιά με ευκολία. Αδιάφορα άρθρα δεν ξαναεμφανίζονται. Ανανεωμένα άρθρα υποδεικνύονται στο χρήστη.

2.4.2 Ηλεκτρονικό Εμπόριο (Electronic Commerce)

Είναι πολύ συνηθισμένο φαινόμενο, οι καταναλωτές να αναζητούν συγκεκριμένα προϊόντα για μεγάλες χρονικές περιόδους, επειδή τα αντικείμενα αυτά είναι σπάνια είτε ιδιαίτερα ακριβά. Για παράδειγμα ένα καταναλωτής είναι δυνατόν να επιθυμεί να αγοράσει ένα συγκεκριμένο μοντέλο αυτοκινήτου σε μία ορισμένη τιμή. Το USEwebNET μπορεί να βοηθήσει στην αναζήτηση αυτή, επερωτώντας βάσεις δεδομένων με νέα και διαφημίσεις και έχοντας ως στόχο την εύρεση ενός αυτοκινήτου, που ικανοποιεί τις απαιτήσεις του χρήστη. Από τη στιγμή που θα βρεθεί ένα τέτοιο αυτοκίνητο αποθηκεύεται στη βάση δεδομένων του USEwebNET και ειδοποιείται ο χρήστης για την ύπαρξη του.

2.4.3 Επεκτάσεις

Οι περισσότεροι χρήστες εσωκλείουν όλη την περιγραφή του θέματος που θέλουν να εξερευνήσουν μέσα ένα περιορισμένο αριθμό λέξεων-κλειδιών. Ας υποτεθεί για παράδειγμα ότι ένας χρήστης ενδιαφέρεται για μηχανισμούς "caching", που μειώνουν την καθυστέρηση ανάκλησης ιστοσελίδων από τον Παγκόσμιο Ιστό. Για να βρει τη διαθέσιμη ηλεκτρονική πληροφορία θα δημιουργούσε μία επερώτηση με πιθανή φράση-κλειδί την "web caching". Παρά το γεγονός ότι μία επερώτηση σαν αυτή θα επέστρεφε πολλά σχετικά αποτελέσματα, δεν θα τα ανακάλυπτε **όλα**. Ο λόγος είναι ότι δεν περιέχουν όλες οι σχετικές ιστοσελίδες την φράση κλειδί, που χρησιμοποιήθηκε. Πολλές από αυτές περιέχουν συνώνυμες φράσεις, όπως "www caching" είτε "caching in the web".

Η επιλογή του ορθού συνόλου λέξεων-κλειδιών είναι μία δύσκολη εργασία. Το USEwebNET θα μπορούσε να επεκταθεί, ώστε να αναζητά πληροφορία βασιζόμενο σε ένα σύνολο από **Χαρακτηριστικά Κείμενα (Seed Documents)**. Συγκεκριμένα, ο χρήστης θα μπορούσε να δίνει στο USEwebNET ένα σύνολο κειμένων αντιπροσωπευτικών για ένα τομέα ενδιαφέροντος και να ζητά την εύρεση πληροφορίας σχετικής με αυτό. Όμοια ή σχετικά κείμενα μπορούν να θεωρηθούν

αυτά, που έχουν όμοιους συγγραφείς, μεγάλο αριθμό όμοιων λέξεων κλειδιών, αναφέρονται σε πολλά κοινά άρθρα είτε έχουν δείκτες προς πολλές όμοιες ιστοσελίδες.

Το αρχικό σύνολο από **Χαρακτηριστικά Κείμενα** μπορεί να δημιουργηθεί μετά την αρχική ανάκληση πληροφορίας, δηλαδή να επιλεχθεί από τα επιστρεφόμενα αποτελέσματα της πρώτης αναζήτησης για ένα θέμα. Στη συνέχεια, καθώς ανακαλύπτονται περισσότερα αποτελέσματα, το σύνολο αυτό μπορεί να επαυξηθεί ή να μεταβληθεί, ώστε να αντικατοπτρίζει καλύτερα τον τομέα ενδιαφέροντος. Με τον τρόπο αυτό ο χρήστης θα μπορεί να βρει όμοια κείμενα, τα οποία δεν περιέχουν απαραίτητα μία συγκεκριμένη φράση, αλλά είναι εννοιολογικά κοντά σε ένα σύνολο καθοριστικών για την επερώτηση κειμένων.

Στο Κεφάλαιο 4 θα αναλυθεί μία τροποποίηση του USEwebNET, κατάλληλη για το χώρο των Ψηφιακών Βιβλιοθηκών και θα επεξηγηθεί η υλοποίηση και ο σχεδιασμός ορισμένων από των προαναφερόμενων επεκτάσεων.

2.5 ΣΥΜΕΡΑΣΜΑΤΑ

Ολοκληρώνοντας, το USEwebNET αποτελεί μία χρήσιμη, επιπρόσθετη υπηρεσία πάνω από υπάρχουσες μηχανές αναζήτησης και γενικότερα προμηθευτές ηλεκτρονικής πληροφορίας. Τα πλεονεκτήματα του USEwebNET είναι τα ακόλουθα:

- **Φιλτράρει την πληροφορία**, ώστε οι χρήστες να επικεντρώνονται μόνο στη **νέα** και κατά συνέπεια να παρακολουθούν εύκολα τις **τελευταίες εξελίξεις** του τομέα, που τους ενδιαφέρει.

- Διευκολύνει την ανακάλυψη **νέας** είτε **ανανεωμένης** πληροφορίας.

- Λειτουργεί σε ώρες χαμηλού φόρτου (off-line) κάθε νύχτα, που το κόστος επικοινωνίας στο δίκτυο είναι μικρό.

Στο επόμενο κεφάλαιο, θα επεξηγηθεί αναλυτικά ο τρόπος υλοποίησης του USEwebNET.

Κεφάλαιο 3

ΥΛΟΠΟΙΗΣΗ ΤΟΥ USEWEBNET

3.1 ΕΙΣΑΓΩΓΗ

Στόχος του κεφαλαίου αυτού είναι η αναλυτική επεξήγηση του τρόπου υλοποίησης του USEwebNET και η παρουσίαση της τεχνολογίας, που χρησιμοποιήθηκε. Το κεφάλαιο μπορεί να διαχωριστεί σε δύο κύριες ενότητες. Η πρώτη αφορά τη διεπιφάνεια χρήσης του USEwebNET, όπου δίνονται λεπτομέρειες για τον τρόπο χρήσης και το είδος τη τεχνολογίας, στην οποία στηρίζεται. Η δεύτερη αφορά την κύρια μηχανή του USEwebNET. Δίνει έμφαση στην εσωτερική επικοινωνία και κατασκευή των πέντε βασικών τμημάτων της κύριας μηχανής, που αναφέρθηκαν στο προηγούμενο κεφάλαιο.

3.2 ΥΛΟΠΟΙΗΣΗ ΤΗΣ ΔΙΕΠΙΦΑΝΕΙΑΣ ΧΡΗΣΗΣ

Η διεπιφάνεια χρήσης στηρίζεται στην τεχνολογία, που έχει αναπτυχθεί για τη μεταφορά δεδομένων στον Παγκόσμιο Ιστό. Συγκεκριμένα, η αλληλεπίδραση του χρήστη με το USEwebNET γίνεται διαμέσου ενός εξυπηρετητή του Παγκοσμίου Ιστού, ο οποίος εκτελεί το πρωτόκολλο "Hypertext Transfer Protocol" (HTTP) και μπορεί να επεξεργάζεται και να απαντάει σε αιτήσεις, που γίνονται τηρώντας το πρωτόκολλο αυτό.

Κάθε αίτηση στον εξυπηρετητή HTTP του USEwebNET ακολουθεί τη μορφή (format) του **Ομοιόμορφου Προσδιοριστή Πόρων (Uniform Resource Locator - URL)**. Οι αιτήσεις δημιουργούνται αυτόματα από τον αναδιφητή, που χρησιμοποιεί ο χρήστης, αφού πρώτα γίνει η επεξεργασία των δεδομένων, που εισάγει στις ιστοσελίδες του USEwebNET.

Οι ιστοσελίδες του USEwebNET είναι γραμμένες σε γλώσσα **HTML**. Ορισμένες από αυτές περιέχουν λίγες διαδικασίες σε γλώσσα JavaScript, ώστε να είναι δυνατή η δυναμική τροποποίηση τους ύστερα από συγκεκριμένες λειτουργίες, χωρίς να χρειάζεται επικοινωνία με τον εξυπηρετητή. Με εξαίρεση ορισμένες μόνο, όπως η ιστοσελίδα δημιουργίας νέων επερωτήσεων, οι περισσότερες ιστοσελίδες του USEwebNET παρουσιάζουν πληροφορία, η οποία δεν είναι στατικά διαθέσιμη. Κατά συνέπεια, δημιουργούνται δυναμικά από ένα σύνολο προγραμμάτων γραμμένων στη γλώσσα προγραμματισμού C. Τα προγράμματα αυτά, που χαρακτηρίζονται ως **CGI** (Common Gateway Interface) **Binaries** επεξεργάζονται τη βάση δεδομένων του USEwebNET. Ορισμένα από τα προγράμματα CGI του USEwebNET προκαλούν μεταβολή στην κατάσταση της βάσης, χωρίς να χρειάζεται η επιστροφή αποτελέσματος. Αλλά παράγουν ως έξοδο ιστοσελίδες σε γλώσσα HTML, οι οποίες περιέχουν την πληροφορία, που είχε απαιτήσει ο χρήστης.

Οι αιτήσεις και οι απαντήσεις σε αυτές μεταφέρονται σύμφωνα με το πρωτόκολλο **HTTP.** Οι αναδιφητές επικοινωνούν με το σύστημα του USEwebNET διαμέσου ενός HTTP εξυπηρετητή, ο οποίος ευθύνεται για την εκτέλεση των κατάλληλων προγραμμάτων CGI και την επιστροφή των απαντήσεων. Ο τρέχον εξυπηρετητής του USEwebNET είναι ο ***Apache HTTP Server*** [IX].

3.2.1 Η Τεχνολογία του Παγκοσμίου Ιστού

3.2.1.1 Ομοιόμορφος Προσδιοριστής Πόρων

Οι Ομοιόμορφοι Προσδιοριστές Πόρων (URLs) προσφέρουν ένα κοινό τρόπο αναφοράς στους πόρους του Παγκοσμίου Ιστού. Ακολουθούν την εξής σύμβαση για την ονομασία τους:

{protocol}://{host}:{port}/path?{arguments}#{reference}

Στη σύμβαση αυτή, μερικά από τα πιο γνωστά πρωτόκολλα (protocol), που μπορούν να χρησιμοποιηθούν είναι τα: HTTP, FTP, Gopher, Mailto, Telnet, News και Wais. Για το περιβάλλον του Παγκοσμίου Ιστού πιο διαδεδομένο είναι το πρωτόκολλο HTTP.

Το δεύτερο τμήμα της ονομασίας δίνει τη διεύθυνση του εξυπηρετητή, στον οποίο κατευθύνεται η αίτηση. Η διεύθυνση αποτελείται από δύο μέρη. Το πρώτο, με ονομασία *host,* αποτελεί το όνομα του υπολογιστή, στον οποίο λειτουργεί ο εξυπηρετητής. Το δεύτερο (port) ορίζει τη θύρα, στην οποία ακούει, δηλαδή δέχεται αιτήσεις ο εξυπηρετητής. Η τρέχουσα διεύθυνση του εξυπηρετητή του USEwebNET είναι η *cuckoo.ics.forth.gr:9002*.

Τέλος, ακολουθεί το όνομα και η θέση του πόρου (path). Σε περίπτωση που ο πόρος είναι κάποιο πρόγραμμα υπολογισμού, μετά το όνομα του μπορούν να μπουν προαιρετικά τα ορίσματα του προγράμματος.

3.2.1.2 Η γλώσσα HTML

Η γλώσσα **HTML (Hypertext Markup Language)** [XIII], έχει καθιερωθεί για τη διάδοση πληροφορίας στον Παγκόσμιο Ιστό. Ένα αρχείο τύπου HTML είναι ένα κοινό αρχείο κειμένου με τη διαφορά ότι συγκεκριμένα τμήματα του κειμένου έχουν ως στόχο να μορφοποιήσουν το υπόλοιπο. Η γλώσσα HTML περιλαμβάνει ένα σύνολο στοιχείων, τα οποία καθορίζουν τη δομή του κειμένου και οδηγούν τη μορφοποίηση του σε μία ποικιλία συστημάτων εξόδου, συνήθως το παράθυρο ενός αναδιφητή. Επιπροσθέτως, περιλαμβάνει μεθόδους για την εισαγωγή στοιχείων δεδομένων από το χρήστη. Τα δεδομένα αυτά μεταφέρονται από τον αναδιφητή στον εξυπηρετητή μέσω του πρωτοκόλλου HTTP. Κάθε στοιχείο έχει ένα όνομα (*tag*), το οποίο αντικατοπτρίζει ένα περιεχόμενο, όπως ότι το όνομα αρχείου που ακολουθεί είναι ένα αρχείο εικόνας, και χαρακτηρίζεται από συγκεκριμένες ιδιότητες, όπως το μέγεθος της εικόνας και το πλάτος του πλαισίου της.

Για παράδειγμα, η ακόλουθη γραμμή, που προέρχεται από μία ιστοσελίδα γραμμένη σε γλώσσα HTML, προκαλεί την ενσωμάτωση στη σελίδα μίας εικόνας.

<IMG SRC="/gif/SetupB.gif" ALIGN=LEFT>

Η γραμμή αυτή αποτελεί τη γραμμή δήλωσης μία εικόνας και καθορίζει τη μορφή της. Ο όρος ***IMG*** είναι ο τύπος-όνομα του στοιχείο (***tag***). Οι λέξεις ***SRC*** και ***ALIGN*** καθορίζουν τα χαρακτηριστικά της εικόνας. Συγκεκριμένα, ορίζονται τα εξής:

- ***SRC="/gif/SetupB.gif":*** Η εικόνα είναι αποθηκευμένη στο αρχείο "SetupB.gif", που βρίσκεται στο σχετικό μονοπάτι (relative pathname) "gif/".
- ***ALIGN=LEFT:*** Η εικόνα είναι ευθυγραμμισμένη στο αριστερό τμήμα της ιστοσελίδας.

3.2.1.3 Η γλώσσα JavaScript

Το μεγαλύτερο μειονέκτημα της γλώσσας HTML είναι ότι για κάθε αλλαγή είτε ανανέωση της τρέχουσας ιστοσελίδας είναι απαραίτητη η επικοινωνία με τον εξυπηρετητή και η μορφοποίηση από την αρχή της επιστρεφόμενης πληροφορίας. Συχνά, είναι επιθυμητό να αποφεύγεται η επικοινωνία αυτή για να μην επιβαρύνεται το δίκτυο και να μειώνεται η αναμονή του χρήστη. Παραδείγματα αποτελούν οι περιπτώσεις ανάγκης ελέγχου δεδομένων, που εισάγονται από το χρήστη πριν την αποστολή τους στον εξυπηρετητή, ενημέρωσης ορισμένων μετρητών, που βρίσκονται στην ιστοσελίδα και η αλλαγή εικόνων ύστερα από συγκεκριμένες ενέργειες του χρήστη.

Οι προαναφερόμενες περιπτώσεις μπορούν να αντιμετωπιστούν ικανοποιητικά με τη χρήση διαδικασιών (scripts) γραμμένων σε γλώσσα JavaScript [XII]. Οι διαδικασίες αυτές ενσωματώνονται σε ένα αρχείο HTML και μεταφράζονται από τους αναδιφητές κατά την επεξεργασία του αρχείου. Το χαρακτηριστικό, που δίνει τη

δυνατότητα αυτή στη γλώσσα JavaScript, είναι ότι κάθε αναδιφητής διατηρεί σε μία προκαθορισμένη δενδρική βάση δεδομένων, προσπελάσιμη από τις διαδικασίες, τη δομή κάθε ιστοσελίδας, που παρουσιάζει. Συνεπώς, για την ανανέωση οποιαδήποτε στοιχείο της ιστοσελίδας είναι αρκετή η τροποποίηση του καθορισμένου πεδίου της βάσης.

3.2.1.4 Τεχνολογία του Πρωτοκόλλου HTTP

Το πρωτόκολλο *"HyperText Transport Protocol" (HTTP)* είναι ένα πρωτόκολλο χωρίς τη δυνατότητα διατήρησης κατάστασης (stateless protocol), το οποίο έχει αναπτυχθεί ειδικά για την ανάκληση πληροφορίας από τον Παγκόσμιο Ιστό. Αρχικά, προοριζόταν για τη μεταφορά αρχείων HTML, αλλά σήμερα έχει εξελιχθεί με τέτοιο τρόπο, ώστε να χρησιμοποιείται για τη μεταφορά οποιουδήποτε είδους αρχείου.

Η επικοινωνία μέσω του HTTP γίνεται πάνω από το πρωτόκολλο TCP/IP (Transmission Control Protocol / Internet Protocol). Η ανταλλαγή δεδομένων ξεκινάει από τον πελάτη, συνήθως ένα αναδιφητή, ο οποίος αποστέλλει μία αίτηση προς ένα εξυπηρετητή πρωτοκόλλου HTTP. Ο εξυπηρετητής επιστρέφει την απάντηση στον πελάτη. Η απάντηση περιέχει το αποτέλεσμα της αίτησης, εάν δηλαδή ήταν επιτυχής, των τύπο των δεδομένων και τα δεδομένα.

3.2.1.5 Προγράμματα τύπου Common Gateway Interface – CGI

Το πρωτόκολλο HTTP δεν έχει τη δυνατότητα της δυναμικής παραγωγής ιστοσελίδων. Για την παραγωγή ιστοσελίδων, που βασίζονται σε δεδομένα εξωγενή ως προς τον εξυπηρετητή του πρωτοκόλλου χρησιμοποιείται ένας τύπος προγραμμάτων, που ονομάζονται **CGI Binaries**. Τα προγράμματα αυτά είναι γνωστά στον εξυπηρετητή και παραλαμβάνουν την απαραίτητη είσοδο για τη λειτουργία τους από αυτόν. Η είσοδος αυτή περιλαμβάνει τα ορίσματα, που δίνονται άμεσα από το χρήστη διαμέσου του URL, καθώς και αυτά που δημιουργούνται έμμεσα από τον αναδιφητή (συνήθως συνιστούν πληροφορία, που απαιτείται από το

HTTP για την ορθή μετάδοση των δεδομένων). Μετά την επεξεργασία των ορισμάτων εκτελούν τις απαιτούμενες από το χρήστη λειτουργίες, όπως ενημέρωση μίας βάσεως δεδομένων, και παράγουν δυναμικά την ιστοσελίδα με την επιθυμητή πληροφορία, όπως επιβεβαίωση της ενημέρωσης ή ειδοποίηση για την αποτυχία αυτής. Τη σελίδα αυτή την επιστρέφουν στον εξυπηρετητή, ο οποίος την προωθεί στον αναδιφητή.

3.2.1.6 Αντιμετώπιση του προβλήματος διατήρησης κατάστασης: Cookies

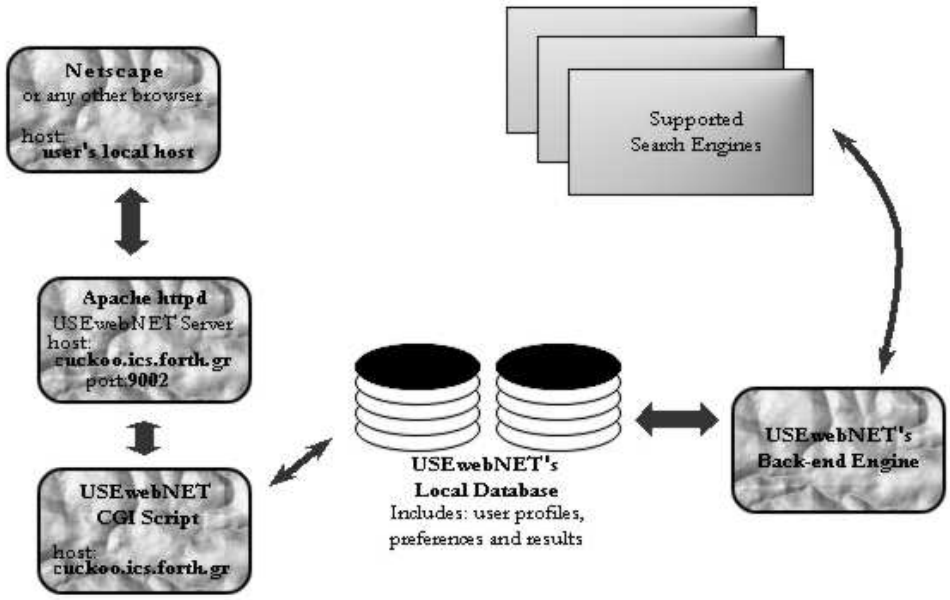
Όπως προαναφέρθηκε, το πρωτόκολλο HTTP δεν έχει τη δυνατότητα διατήρησης κατάστασης. Η έλλειψη αυτή εμφανίζεται ιδιαίτερα έντονη σε περιπτώσεις, που χρειάζεται να διατηρηθεί η κατάσταση μίας συναλλαγής. Ας υποτεθεί ότι ένας εξυπηρετητής HTTP (εμπλουτισμένος με CGI Binaries) διατηρεί διαφορετικές ιστοσελίδες για κάθε χρήστη. Για να αναγνωρίσει ένα χρήστη ζητάει από αυτόν την πιστοποίηση της ταυτότητας του και στη συνέχεια του παραδίδει την απαιτούμενη ιστοσελίδα. Με την παράληψη της απάντησης από τον αναδιφητή η διαδικασία της πιστοποίησης της ταυτότητας έχει ξεχαστεί και νέα αίτηση του χρήστη θα πρέπει να συνοδευτεί από νέα πιστοποίηση ταυτότητας, γεγονός ιδιαίτερα κουραστικό για τη χρήση των περισσότερων εφαρμογών.

Για την αντιμετώπιση του προβλήματος της αδυναμίας διατήρησης κατάστασης αναπτύχθηκε ο μηχανισμός των **Cookies**. Το Cookie μπορεί να οριστεί ως μία μεταβλητή αλφαριθμητικού περιεχομένου, η οποία έχει επιπροσθέτως μία ημερομηνία λήξης και τη διεύθυνση ενός εξυπηρετητή. Σε περίπτωση, που ο εξυπηρετητής χρειάζεται να διατηρήσει κάποιο είδος κατάστασης, όπως αυτό της πιστοποίησης της ταυτότητας , αποστέλλει στον αναδιφητή ένα Cookie με τιμή που χαρακτηρίζει την κατάσταση αυτή, την απαιτούμενη ημερομηνία λήξης και τη διεύθυνση του. Ο αναδιφητής παραλαμβάνει το Cookie, το αποθηκεύει εσωτερικά και υποχρεούται να συνοδεύει κάθε επόμενη αίτηση προς το συγκεκριμένο εξυπηρετητή με την τιμή και το όνομα του Cookie.

Στο προαναφερόμενο παράδειγμα, το Cookie θα μπορούσε να έχει το όνομα "UserId" και τιμή ένα μοναδικό ακέραιο αριθμό, προσεκτικά επιλεγμένο από το CGI Binary, που δημιούργησε το Cookie. Το Cookie θα έπαιρνε την τιμή του κατά τη διαδικασία πιστοποίησης της ταυτότητας και θα συνδεόταν με το συγκεκριμένο χρήστη. Από το σημείο αυτό και έπειτα κάθε αίτηση προς τον εξυπηρετητή θα συνοδευόταν από το Cookie, το οποίο θα λειτουργούσε ως στοιχείο ταυτότητας, χωρίς να χρειάζεται η επαναπιστοποίηση των στοιχείων του χρήστη. Με τον τρόπο αυτό λειτουργεί η διαδικασία πιστοποίησης ταυτότητας και ανάκλησης προσωπικών αρχείων στο USEwebNET.

3.2.2 Λειτουργία της Διεπιφάνειας Χρήσης

Στην ενότητα αυτή δίνεται ένα σχεδιάγραμμα του τρόπου λειτουργίας της διεπιφάνειας χρήσης του USEwebNET (Σχήμα 3-1).



Σχήμα 3-1: Σχεδιάγραμμα της διαδρομής των αιτήσεων στο USEwebNET.

Αρχικά, αποστέλλεται η αίτηση του χρήστη από τον αναδιφητή στον εξυπηρετητή. Σε περίπτωση, που η αίτηση αφορά μία στατική ιστοσελίδα ο εξυπηρετητής την επιστρέφει αμέσως στο χρήστη. Διαφορετικά, προκαλεί την εκτέλεση του κατάλληλου εκτελέσιμου CGI και προωθεί την αίτηση σε αυτό. Το πρόγραμμα CGI επεξεργάζεται τη βάση του USEwebNET και δημιουργεί δυναμικά την ιστοσελίδα με την επιθυμητή πληροφορία, την οποία επιστρέφει στον εξυπηρετητή. Σε περίπτωση, που η αίτηση του χρήστη απαιτεί την ανανέωση των δεδομένων της βάσης καλούνται από το πρόγραμμα CGI, τα κατάλληλα τμήματα της μηχανής του USEwebNET. Τέλος, ο εξυπηρετητής την προωθεί στον αναδιφητή, που πραγματοποίησε την αίτηση.

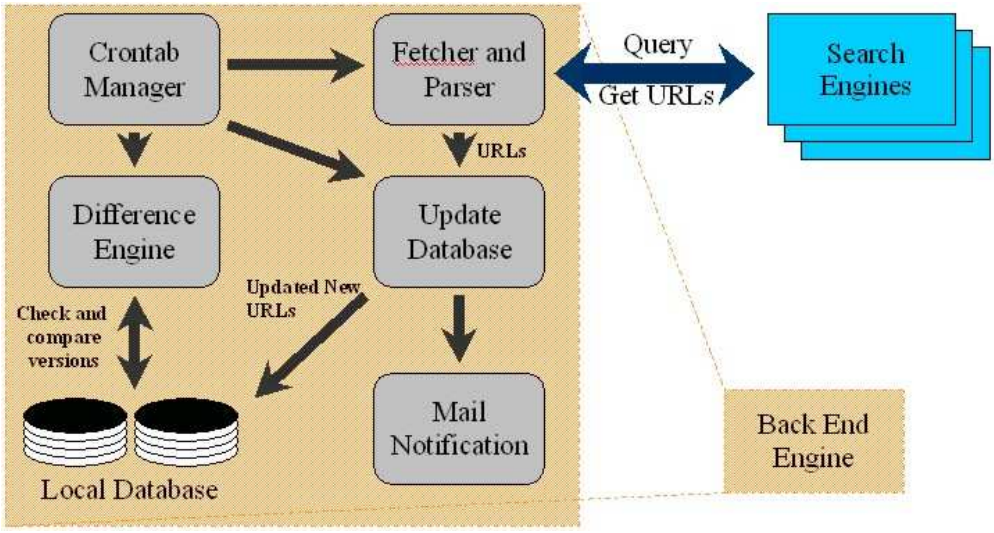
3.3 Η "ΜΗΧΑΝΗ" ΤΟΥ USEwebNET

Στην ενότητα αυτή θα αναλυθεί ο τρόπος, με τον οποίο λειτουργούν τα πέντε τμήματα, που αποτελούν τη "μηχανή" του USEwebNET (Παράγραφος 2.3). Στο Σχήμα 3-2 παρουσιάζεται η αρχιτεκτονική της μηχανής και ο τρόπος συνεργασίας των πέντε τμημάτων της.

3.3.1 Συλλέκτης αποτελεσμάτων και Αναλυτής κειμένου

Είναι γραμμένος στη γλώσσα προγραμματισμού **C**. Το όνομα του εκτελέσιμου είναι "usewebnetp". Ο αναλυτής κειμένου έχει υλοποιηθεί με τη βοήθεια των εργαλείων **Lex** και **Yacc**. Η λειτουργία του γίνεται ως εξής: Παίρνει ως ορίσματα το όνομα του αρχείου εξόδου, τις λέξεις-κλειδιά της επερώτησης, τη μηχανή αναζήτησης στην οποία θα προωθηθεί η επερώτηση και πιθανά μία πιστοποίηση ταυτότητας εάν αυτή απαιτείται από τη μηχανή αναζήτησης. Στη συνέχεια δημιουργεί αιτήσεις σε μορφή URL και ξεκινάει την επικοινωνία με τη μηχανή αναζήτησης. Η επικοινωνία γίνεται κάτω από το πρωτόκολλο HTTP. Κάθε ιστοσελίδα, που επιστρέφεται, αποθηκεύεται σε ένα προσωρινό αρχείο, το οποίο αναλύεται συντακτικά. Από την ανάλυση εξάγεται η απαιτούμενη πληροφορία για κάθε αποτέλεσμα και αποθηκεύεται στο τελικό αρχείο εξόδου στη μορφή, που χρησιμοποιεί εσωτερικά το USEwebNET.

Επίσης, διαπιστώνεται εάν υπάρχει επιπρόσθετη πληροφορία, που πρέπει να ζητηθεί από τη μηχανή αναζήτησης. Εάν υπάρχει δημιουργείται το νέο URL (βρίσκεται στην επιστρεφόμενη ιστοσελίδα), που εντοπίζει την πληροφορία αυτή και η διαδικασία επαναλαμβάνεται.



Σχήμα 3-2: Η αρχιτεκτονική της μηχανής του USEwebNET.

3.3.2 Τμήμα Ενημέρωσης της Βάσεως Δεδομένων

Είναι υλοποιημένο στη γλώσσα προγραμματισμού **C**. Το πρόγραμμα ονομάζεται ***"update_pref"***. Παίρνει ως είσοδο την έξοδο του αναλυτή κειμένου, το όνομα ενός χρήστη και την ταυτότητα μίας επερώτησης (για κάθε χρήστη μία επερώτηση καθορίζεται μοναδικά από το όνομα του αρχείου, στο οποίο αποθηκεύονται τα URLs). Συγκρίνει τα χαρακτηριστικά των νέων URLs με τα ήδη υπάρχοντα και ανανεώνει τη βάση.

3.3.3 Πρόγραμμα Αποστολής ηλεκτρονικού μηνύματος

Έχει επίσης υλοποιηθεί στη γλώσσα προγραμματισμού C. Είναι ενσωματωμένο στο εκτελέσιμο του τμήματος Ενημέρωσης της βάσεως δεδομένων (update_pref). Προκαλεί την εκτέλεση του προγράμματος Mail αφού πρώτα δημιουργήσει

δυναμικά το κείμενο του μηνύματος και παραλάβει από τη βάση του USEwebNET την ηλεκτρονική διεύθυνση του χρήστη, στον οποίο απευθύνεται το μήνυμα.

3.3.4 Τμήμα χειρισμού περιοδικών λειτουργιών (Crontab Manager)

Η υλοποίηση του έχει γίνει στη γλώσσα προγραμματισμού **C**. Το εκτελέσιμο ονομάζεται ***"cron_manager"***. Εκτελείται περιοδικά από τη λειτουργία **Cron** του λειτουργικού συστήματος **Unix**. Προκαλεί τη σειριακή εκτέλεση των τμημάτων *3.3.1* και *3.3.2* για κάθε επερώτηση κάθε χρήστη.

3.3.5 Ελεγκτής εκδόσεων ιστοσελίδων

Το τμήμα αυτό με όνομα ***"doc_versionControl"*** στηρίζεται σε ένα πρόγραμμα, το οποίο συγκρίνει δύο σελίδες γραμμένες σε HTML, βρίσκει τις διαφορές τους και παράγει μία ιστοσελίδα, στην οποία τις υποδεικνύει. Το εκτελέσιμο αυτό ονομάζεται **htmldiff** [10] και έχει δημιουργηθεί από το *Fred Douglis* της *AT&T Labs* [X]. Είναι υλοποιημένο στη γλώσσα **SML/NJ** (Standard ML of New Jersey).

Τη λειτουργία του **htmldiff** ελέγχει ένα μικρό πρόγραμμα γραμμένο σε γλώσσα **C**. Το πρόγραμμα αυτό ευθύνεται για την περιοδική εκτέλεση του htmldiff. Για όσες ιστοσελίδες έχει απαιτηθεί εκτελείται αρχικά το πρόγραμμα ανάκλησης πληροφορίας. Η ιστοσελίδα, που επιστρέφεται παραδίδεται μαζί την αντίστοιχη που υπάρχει ήδη στη βάση δεδομένων στο htmldiff για να συγκριθούν. Στη συνέχεια ελέγχεται η έξοδος του htmldiff και εάν βρεθούν διαφορές αντικαθίσταται στη βάση η υπάρχουσα έκδοση από τη νέα, η κατάσταση της σελίδας σημειώνεται ως ``ανανεωμένη'' και ο χρήστης ειδοποιείται για να δει τις διαφορές.

3.3.5.1 Η γλώσσα SML/NJ (Standard ML of New Jersey)

Η γλώσσα προγραμματισμού Standard ML [XI] αποτελεί μία τμηματική (modular), αυστηρή (strict), συναρτησιακή (functional), πολυμορφική γλώσσα προγραμματισμού. Υποστηρίζει έλεγχο τύπων κατά το χρόνο της μεταγλώττισης (compile-time type checking), διαχείριση εξαιρέσεων (exception handling) και

συλλογή μη χρησιμοποιούμενου χώρου μνήμης (garbage collection). Χρησιμοποιείται για την υλοποίηση μεγάλων συστημάτων στο χώρο της εφαρμοσμένης λογικής και επαλήθευσης καθώς και για την ανάλυση προγραμμάτων και την κατασκευή πολύπλοκων μεταγλωττιστών.

Κεφάλαιο 4

ΕΠΕΚΤΑΣΕΙΣ ΣΤΟ USEWEBNET: PAPERFINDER

4.1 ΕΙΣΑΓΩΓΗ

Στα προηγούμενα τρία κεφάλαια περιγράφτηκε το USEwebNET ως ένα επίπεδο λογισμικού, το οποίο λειτουργεί πάνω από υπάρχουσες μηχανές αναζήτησης. Στην Ενότητα 2.4 αναφέρθηκε ότι ο τρόπος, που έχει σχεδιαστεί το USEwebNET του επιτρέπει να λειτουργήσει ευεργετικά σε πολλές επιπρόσθετες κατηγορίες προμηθευτών ηλεκτρονικής πληροφορίας, όπως σε εφαρμογές ψηφιακών βιβλιοθηκών και ηλεκτρονικού εμπορίου. Επίσης, έγινε μία εισαγωγή στην ιδέα του "Χαρακτηριστικού Κειμένου" (Seed Document) μίας επερώτησης.

Στο κεφάλαιο αυτό θα παρουσιαστεί μία παραλλαγή του USEwebNET, με ονομασία **PaperFinder**, η οποία διευκολύνει την αναζήτηση και ανάκληση ερευνητικών άρθρων από δημοφιλείς ψηφιακές βιβλιοθήκες. ΤοPaperFinder έχει την ικανότητα να φιλτράρει την ανακαλούμενη πληροφορία, παραδίδοντας μόνο νέα άρθρα στο χρήστη, διαμέσου μίας φιλικής διεπιφάνειας χρήσης. Επιπροσθέτως, εισάγει ένα νέο αλγόριθμο ταξινόμησης των ανακαλώμενων άρθρων μίας επερώτησης, ο οποίος στηρίζεται στην έννοια του "Χαρακτηριστικού Κειμένου".

4.2 ΨΗΦΙΑΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ

Οι ερευνητές χρειάζονται να παρακολουθούν διαρκώς τις εξελίξεις στον τομέα της επιστήμης τους. Για να διευκολυνθεί η διαδικασία της έρευνας άρθρα, βιβλία και περιοδικά τείνουν να συγκεντρώνονται σε μεγάλες συλλογές, να διαχωρίζονται κατά θεματικές κατηγορίες και να ταξινομούνται. Επιπροσθέτως, χαρακτηρίζονται από ευρετήρια, τα οποία στηρίζονται στον τίτλο τους, στα ονόματα των συγγραφέων είτε

σε λέξεις-κλειδιά, που υποδεικνύουν το περιεχόμενο τους, ώστε να είναι δυνατός ο ταχύς εντοπισμός τους και η γρήγορη ανάκλησή τους. Συλλογές βιβλίων, περιοδικών και άρθρων με τα χαρακτηριστικά αυτά ονομάζονται **βιβλιοθήκες.**

Η διαδικασία αναζήτησης και ανάκλησης ενός άρθρου από μία παραδοσιακή βιβλιοθήκη επιβάλει την αναζήτηση του μέσω του ευρετηρίου, το οποίο μπορεί να είναι διαθέσιμο σε ηλεκτρονική μορφή, ώστε να επισπευθεί η διαδικασία και στη συνέχεια η επίσκεψη στο κτίριο, που στεγάζεται η βιβλιοθήκη για την παραλαβή του. Δυστυχώς, κατά τη διαδικασία αυτή συχνά ο ερευνητής συναντά προβλήματα, τα οποία τον αποπροσανατολίζουν από την εργασία του, δυσχεραίνοντας την ανάκληση της επιθυμητής πληροφορίας. Για παράδειγμα είναι πιθανό η βιβλιοθήκη να μην είναι σωστά ταξινομημένη και αρχειοθετημένη, με αποτέλεσμα η πληροφορία, που επιστρέφει το ευρετήριο για τη θέση του άρθρου να είναι λανθασμένη, ή η βιβλιοθήκη να στεγάζεται σε πολλά διαφορετικά κτίρια. Επιπροσθέτως, είναι αρκετές οι περιπτώσεις, που ένα άρθρο μπορεί να είναι δανεισμένο ή να έχει χαθεί.

Η εξάπλωση του Διαδικτύου και η ευκολία έκδοσης πληροφορίας στον Παγκόσμιο Ιστό έδωσαν πρόσφορο έδαφος για την έκδοση ηλεκτρονικού υλικού. Πολλοί συγγραφείς και ερευνητικοί οργανισμός συνηθίζουν να προσφέρουν τα άρθρα τους ηλεκτρονικά διαμέσου του Παγκοσμίου Ιστού. Επιπροσθέτως, οι οργανισμοί εκδόσεως ερευνητικών περιοδικών, όπως η ACM και η USENIX, άρχισαν να προσφέρουν το πλήρες κείμενο των άρθρων τους ηλεκτρονικά και δίνουν τη δυνατότητα εκτελέσεως επερωτήσεων στη βάση δεδομένων τους διαμέσου του περιβάλλοντος του Παγκοσμίου Ιστού με τρόπο όμοιο με αυτό, που προσφέρουν οι μηχανές αναζήτησης (Ενότητα 1.2). Οι συλλογές αυτές έχουν όλα τα χαρακτηριστικά μίας Παραδοσιακής Βιβλιοθήκης και ονομάζονται *"Ψηφιακές Βιβλιοθήκες"*. Το γεγονός αυτό διευκολύνει στις περισσότερες περίπτωσης την ανάκληση των επιθυμητών άρθρων, αλλά μειονεκτεί εξαιτίας της ανομοιογένειας, που χαρακτηρίζει την έκδοση πληροφορίας στον Παγκόσμιο Ιστό.

Κάθε ερευνητικός και εκδοτικός οργανισμός ακολουθεί το δικό του τρόπο έκδοσης άρθρων στον Παγκόσμιο Ιστό και χρησιμοποιεί διαφορετικές τεχνικές για την αρχειοθέτηση και την ταξινόμηση τους. Επιπροσθέτως, η μηχανές αναζήτησης των Ψηφιακών βιβλιοθηκών χαρακτηρίζονται από διαφορετικές διεπιφάνειες χρήσης με αποτέλεσμα να είναι αδύνατη η δημιουργία ενός ομοιογενούς τρόπου αναζήτησης και ανάκλησης πληροφορίας από αυτές. Τέλος, οι αναζητήσεις στις ψηφιακές βιβλιοθήκες υποφέρουν από προβλήματα όμοια με αυτά, που εντοπίζονται στο γενικότερο χώρο των μηχανών αναζητήσεως. Συγκεκριμένα, διαδοχικές όμοιες επερωτήσεις στην ίδια ψηφιακή βιβλιοθήκη επιστρέφουν πάντα το ίδια άρθρα, ανεξάρτητα από το εάν αυτά έχουν ήδη αναγνωσθεί από το χρήστη. Παράλληλα, ο ρυθμός με τον οποίο αυξάνεται η ηλεκτρονική έκδοση των επιστημονικών άρθρων και η αύξηση των χρηστών του Παγκοσμίου Ιστού. Το γεγονός αυτό εμφανίζει το πρόβλημα του καταιγισμού πληροφορίας και δυσχεραίνει την παρακολούθηση των τελευταίων εξελίξεων από τους ερευνητές.

Η αντιμετώπιση των προαναφερόμενων προβλημάτων απαιτεί τη χρήση ενός εργαλείου, το οποίο θα χαρακτηρίζεται από την ικανότητα της μνήμης, ώστε να διαχωρίζει τα νέα από τα παλαιότερα άρθρα, και επιπλέον θα προσφέρει μία κοινή διεπιφάνεια χρήσης για το σύνολο των ψηφιακών βιβλιοθηκών. Τα δύο αυτά στοιχεία χαρακτηρίζουν το **PaperFinder**. Το PaperFinder μπορεί να αντιμετωπιστεί ως ένα επίπεδο λογισμικού, το οποίο λειτουργεί πάνω από υπάρχουσες ψηφιακές βιβλιοθήκες. Προσφέρει ένα προσωπικό πορτραίτο χρήσης για κάθε χρήστη, δίνοντας τη δυνατότητα διατήρησης των επερωτήσεων του και των αποτελεσμάτων. Διευκολύνει το διαχωρισμό των νέων άρθρων από τα παλαιότερα και δίνει τη δυνατότητα μίας μεγάλης ποικιλίας χρήσιμων λειτουργιών πάνω στα επιστρεφόμενα άρθρα. Τέλος, προκαλεί την περιοδική εκτέλεση των επερωτήσεων, συλλέγοντας αποτελέσματα σε τακτικά χρονικά διαστήματα, και ειδοποιεί το χρήστη στην περίπτωση εύρεσης νέων άρθρων για τους τομείς του ενδιαφέροντος του. Επομένως,

το PaperFinder διευκολύνει την ερευνητική δραστηριότητα και αντιμετωπίζει το πρόβλημα του καταιγισμού πληροφορίας.

Ακολούθως, το PaperFinder υποστηρίζει εκτός από το συνηθισμένο τρόπο λειτουργίας, που στηρίζεται στη χρήση λέξεων-κλειδιών (**Keyword-Based Mode**), και ένα δεύτερο τρόπο ανάκλησης πληροφορίας, με ονομασία **Resource-Discovery Mode**. Κατά τη λειτουργία αυτή, το PaperFinder επιχειρεί να ανακαλύψει νέα άρθρα, τα οποία δεν χαρακτηρίζονται απαραιτήτως από ένα σύνολο προκαθορισμένων λέξεων-κλειδιών, αλλά είναι όμοια με ένα σύνολο άρθρων, που έχουν καθοριστεί από το χρήστη.

Στις επόμενες ενότητες του κεφαλαίου αυτού θα αναλυθεί ο τρόπος λειτουργίας και η κατασκευή του PaperFinder.

4.3 PAPERFINDER: ΣΧΕΔΙΑΣΗ ΚΑΙ ΔΙΕΠΙΦΑΝΕΙΑ ΧΡΗΣΗΣ

Ο στόχος του PaperFinder είναι να προσφέρει στους χρήστες ένα ευέλικτο εργαλείο για να απλοποιήσει το χρονοβόρο έργο της ανάκλησης πληροφορίας. Για την επίτευξη του στόχου αυτού, το PaperFinder συνεργάζεται με ορισμένες δημοφιλείς ψηφιακές βιβλιοθήκες, όπως αυτές που διατηρούν η ACM και η USENIX, αναζητώντας περιοδικά νέα άρθρα σχετικά με τους τομείς ενδιαφερόντων του χρήστη, και προσφέρει μία εύχρηστη διεπιφάνεια χρήσης για την προσπέλαση των αποτελεσμάτων, η οποία, όπως και στην περίπτωση του USEwebNET, βασίζεται σε αυτή, που χρησιμοποιείται για χρόνια από τα USENET News.

Η βασική ιδέα πίσω από το PaperFinder είναι η ικανότητα να διατηρεί πληροφορία για τα ενδιαφέροντα ενός χρήστη και να προωθεί επερωτήσεις στις υποστηριζόμενες ψηφιακές βιβλιοθήκες. Μία επερώτηση καθορίζεται από ένα σύνολο λέξεων-φράσεων, από τα ονόματα συγγραφέων, από μία ημερομηνία, που υποδεικνύει την παλαιότερη ημερομηνία έκδοσης των επιστρεφόμενων άρθρων, και από τις ψηφιακές βιβλιοθήκες, που θα χρησιμοποιηθούν για την εκτέλεση των επερωτήσεων. Το

PaperFinder αποθηκεύει εσωτερικά τα άρθρα, που επιστρέφονται για κάθε τομέα ενδιαφέροντος του ερευνητή. Με τον τρόπο αυτό άρθρα, τα οποία έχουν ήδη αναγνωσθεί ή απορριφθεί από το χρήστη δεν θα ξαναπαρουσιασθούν ανάμεσα στα αποτελέσματα της αντίστοιχης επερώτησης.

4.3.1 Τρόποι Λειτουργίας

Το PaperFinder υποστηρίζει δύο διαφορετικούς τρόπους λειτουργίας. Ο πρώτος στηρίζεται στη χρήση λέξεων-κλειδιών για την ανάκληση πληροφορίας και ονομάζεται ``**Keyword-based Mode**''. Ο δεύτερος στηρίζεται στην ιδέα των "Χαρακτηριστικών Κειμένων" (Seed Documents) και ονομάζεται ``**Resource-Discovery Mode**''.

- ***Keyword-Based Mode:*** Κατά τη λειτουργία αυτή οι χρήστες δίνουν στο PaperFinder λίγες λέξεις-κλειδιά, οι οποίες καθορίζουν το πεδίο ενδιαφέροντος τους, όπως ``digital libraries'' ή ``process scheduling''. Παράλληλα, καθορίζουν τις ψηφιακές βιβλιοθήκες, οι οποίες θα χρησιμοποιηθούν για την ανάκληση πληροφορίας. Στη συνέχεια το PaperFinder, δημιουργεί επερωτήσεις, τις οποίες προωθεί στις επιλεγμένες ψηφιακές βιβλιοθήκες και συλλέγει άρθρα, τα οποία παρουσιάζει στο χρήστη.

- ***Resource-Discovery Mode:*** Στο ``Resource-Discovery Mode'' το PaperFinder προσπαθεί να ανακαλύψει νέα άρθρα, τα οποία μπορούν να χαρακτηριστούν ως θεματικά όμοια με ένα σύνολο ``Χαρακτηριστικών Άρθρων'' (Seed Documents) καθορισμένων από το χρήστη. Ο ορισμός του βέλτιστου μέτρου ομοιότητας αποτελεί ένα μεγάλο και ενδιαφέρον τομέα έρευνας. Πιθανά κριτήρια ομοιότητας είναι τα εξής:

  - Η χρήση ενός όμοιου συνόλου αναφορών στη βιβλιογραφία τους αποτελεί ένδειξη ομοιότητας για τα δύο άρθρα.

- Οι πλειοψηφία των ερευνητών έχει ένα μικρό αριθμό πεδίων έρευνας, στο οποίο δημοσιεύει. Συνεπώς, άρθρα, τα οποία έχουν γραφεί από ένα όμοιο σύνολο ερευνητών θα ανήκουν στο ίδιο θεματικό τομέα.

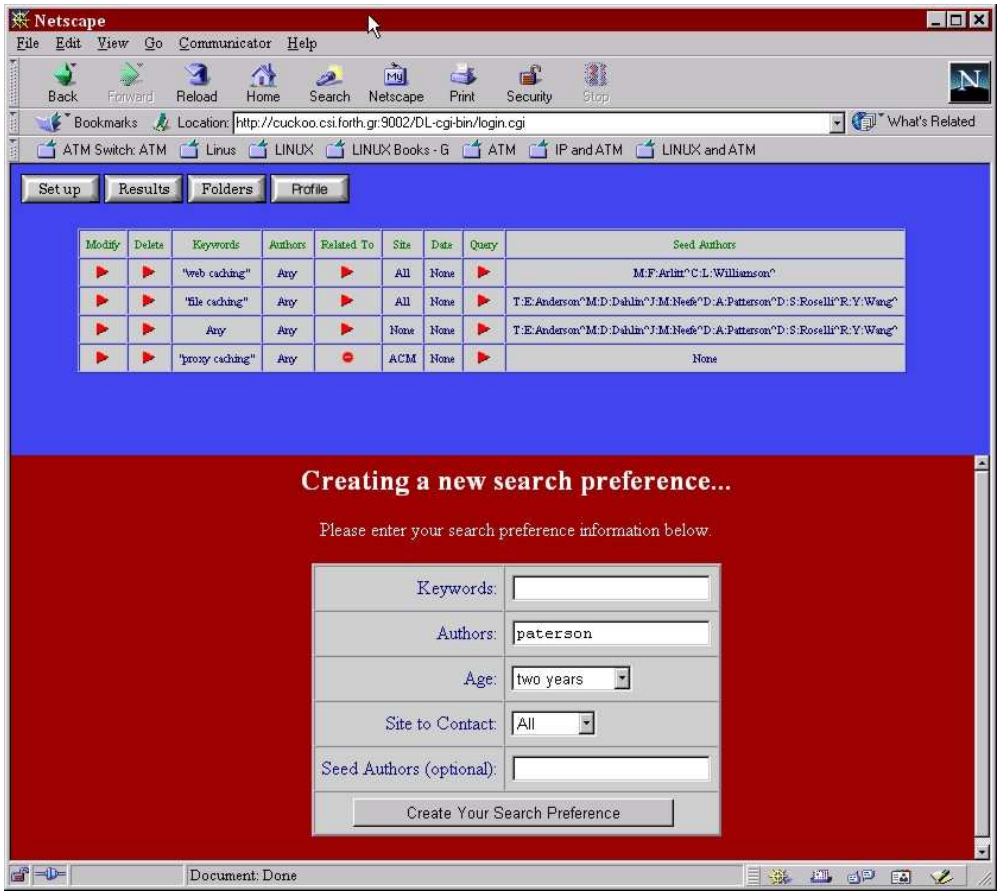
- Η χρήση ενός σημαντικού αριθμού όμοιων λέξεων-κλειδιών μεταξύ δύο άρθρων αποτελεί ένδειξη ότι τα άρθρα αυτά σχετίζονται θεματικά.

Σε επόμενη ενότητα, θα μελετηθεί αναλυτικά ο τρόπος αυτός λειτουργίας του PaperFinder.

4.3.2 Διεπιφάνεια Χρήσης

Το PaperFinder έχει σχεδιαστεί με τρόπο όμοιο με αυτό του USEwebNET. Για το λόγο αυτό η υλοποίηση και η αρχιτεκτονική του θα επεξηγηθούν με συντομία και θα δοθεί μεγαλύτερο βάρος στις διαφορές του από το USEwebNET και στις επεκτάσεις, που έχουν γίνει.

Όπως και στην περίπτωση του USEwebNET, το PaperFinder έχει σχεδιαστεί με βάση τη τεχνολογία, που έχει αναπτυχθεί για την προσπέλαση αρχείων και ανάκληση πληροφορίας από τον Παγκόσμιο Ιστό. Συγκεκριμένα, η χρήση του εργαλείου ξεκινάει κατευθύνοντας ένα αναδιφητή του Παγκοσμίου Ιστού σε ένα καθορισμένο εξυπηρετητή (*cuckoo.ics.forth.gr:9002)*. Από τη στιγμή, που δημιουργείται η σύνδεση, εμφανίζεται η σελίδα "Πιστοποίησης Ταυτότητας" του χρήστη. Από τη σελίδα αυτή, ο χρήστης μπορεί να ξεκινήσει τη χρήση του USEwebNET ή να δημιουργήσει ένα νέο λογαριασμό διαμέσου του συνδέσμου προς τη σελίδα "Καταχώρησης Χρηστών". Οι ιστοσελίδες "Πιστοποίησης Ταυτότητας" (Σχήμα 2-1) και "Καταχώρησης Χρηστών" (Σχήμα 2-2) είναι απολύτως όμοιες με αυτές του USEwebNET, που περιγράφονται στην Ενότητα 2.2.



Σχήμα 4-1: Με την επιλογή του πλήκτρου *"Setup"*, ο χρήστης οδηγείται στην οθόνη καταχώρησης επερωτήσεων. Το PaperFinder στο ανώτερο τμήμα του παραθύρου του αναδιφητή παρουσιάζει τις επερωτήσεις, που έχουν ήδη δημιουργηθεί. Μέσω της φόρμας, που βρίσκεται στο κατώτερο τμήμα του παραθύρου, ο χρήστης ετοιμάζεται να δημιουργήσει μία επερώτηση για το συγγραφέα με όνομα *"Paterson"*.

Από τη στιγμή, που θα γίνει η πιστοποίηση της ταυτότητας του χρήστη εμφανίζονται στο παράθυρο του αναδιφητή τέσσερις δυνατές επιλογές, οι οποίες λειτουργούν σε μεγάλο βαθμό, όμοια με τις αντίστοιχες επιλογές του USEwebNET. Οι επιλογές αυτές είναι:

- ***Setup:*** Πρόκειται για τη σελίδα δημιουργίας νέων επερωτήσεων.

Σχήμα 4-2: Ο χρήστης ετοιμάζεται να δημιουργήσει μία επερώτηση για το θέμα του *"file caching"*. Θέλει να βρει άρθρα, τα οποία έχουν εκδοθεί τα τελευταία δύο χρόνια Η επερώτηση θα προωθηθεί σε όλες τις ψηφιακές βιβλιοθήκες, που υποστηρίζει το PaperFinder.

- ***Results:*** Ο σύνδεσμος οδηγεί στη σελίδα επεξεργασία των ανακαλούμενων άρθρων.

- ***Folders:*** Από το σύνδεσμο αυτό ο χρήστης μπορεί να επεξεργαστεί τα προσωπικά του αρχεία, στα οποία φυλάει αποτελέσματα των επερωτήσεων του PaperFinder.

- ***Profile Definition:*** Πρόκειται για τη σελίδα καθορισμού του προσωπικού πορτραίτου του χρήστη.

4.3.2.1 Καταχώρηση επερωτήσεων

Διαμέσου του "Setup Menu", ο χρήστης οδηγείται στη σελίδα καταχώρησης νέων επερωτήσεων, η οποία είναι διαχωρισμένη σε δύο τμήματα (Σχήμα 4-1). Στο ανώτερο από αυτά εμφανίζονται οι ήδη δημιουργημένες επερωτήσεις. Στο κατώτερο

δίνεται η φόρμα δημιουργίας νέων επερωτήσεων είτε εμφανίζεται επιπρόσθετη πληροφορία για αυτές.

Modify	Delete	Keywords	Authors	Related To	Site	Date	Query	Seed Authors
▶	▶	"web caching"	Any	▶	All	None	▶	M:F:Arlitt^C:L:Williamson^
▶	▶	"proxy caching"	Any	⊖	ACM	None	▶	None
▶	▶	Any	paterson	⊖	All	25 May 1997	▶	None
▶	▶	file caching	Any	⊖	All	25 May 1997	▶	None

Σχήμα 4-3: Ο πίνακας επερωτήσεων μετά τη δημιουργία της επερώτησης *"file caching"*.

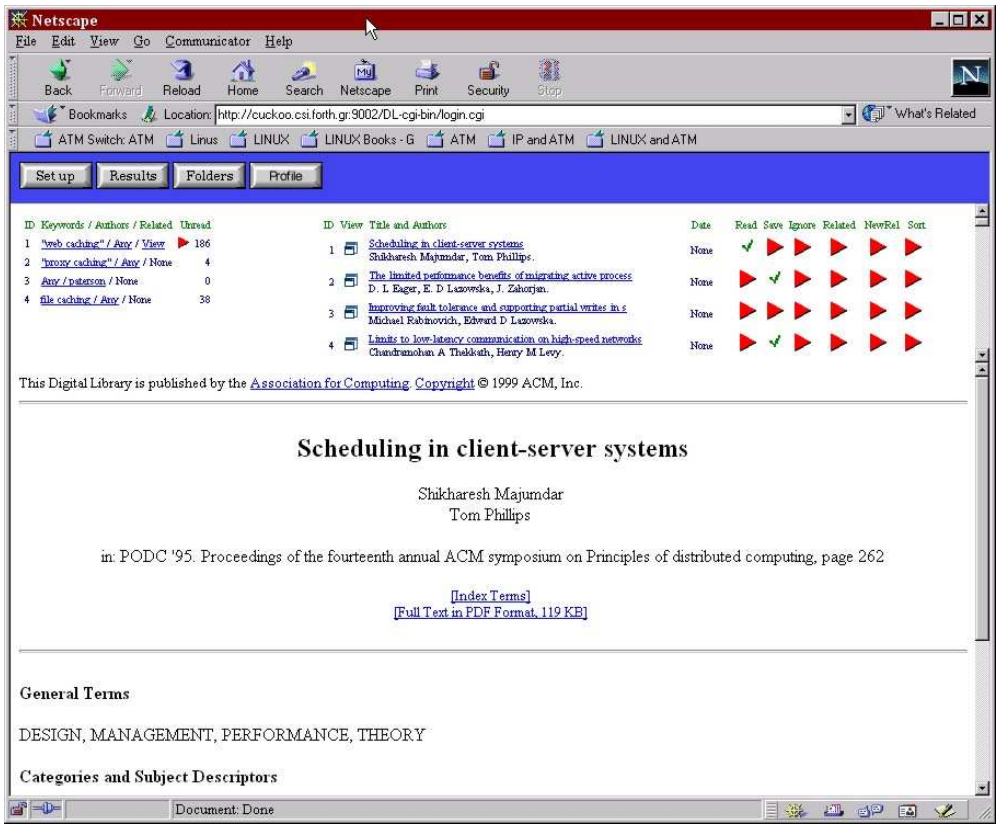
- *Φόρμα δημιουργίας και τροποποίησης επερωτήσεων:*

Διαμέσου της φόρμας αυτής (Σχήμα 4-2) δίνεται η δυνατότητα στο χρήστη να δημιουργήσει νέες επερωτήσεις ή να τροποποιήσει ήδη υπάρχουσες. Η φόρμα περιέχει πεδία για τον καθορισμό των λέξεων-κλειδιών, των συγγραφέων , της παλαιότερης δυνατής ημερομηνίας δημιουργίας των άρθρων, των ψηφιακών βιβλιοθηκών, στις οποίες θα προωθηθούν οι επερωτήσεις και των χαρακτηριστικού συνόλου συγγραφέων των επερωτήσεων.

- *Πίνακας επερωτήσεων:*

Όπως φαίνεται στο Σχήμα 4-3, αποτελείται από εννέα πεδία. Τα πεδία με ονόματα **"Modify"**, **"Delete"**, **"Query"**, **"Site"** και **"Date"** είναι όμοια με τα αντίστοιχα του USEwebNET. Τα τρία πρώτα προκαλούν την τροποποίηση, διαγραφή και άμεση εκτέλεσης της επερώτησης και στο πεδίο "Site" αναγράφεται η ψηφιακή βιβλιοθήκη, στην οποία προωθείται η επερώτηση, και στο πεδίο "Date" η παλαιότερη δυνατή ημερομηνία συγγραφής των ανακαλούμενων άρθρων. Στην τρέχουσα έκδοση του PaperFinder το "Site" μπορεί να είναι ένα από τα: **None** (η

επερώτηση είναι απενεργοποιημένη), **ACM**, **USENIX**, **DBLP** και **All** (η επερώτηση προωθείται σε όλες τις υποστηριζόμενες ψηφιακές βιβλιοθήκες).



Σχήμα 4-4: Η παρουσίαση των αποτελεσμάτων. Ο χρήστης διαβάζει το άρθρο *"Scheduling in Client-Server Systems"*.

Επίσης, έχουν προστεθεί τρία νέα πεδία. Τα πεδία αυτά ονομάζονται **"Authors"**, **"Seed Authors"** και **"Related To"**. Στο πεδίο "Authors" αναγράφονται οι συγγραφείς, που πρέπει να χαρακτηρίζουν τα ανακαλούμενα άρθρα. Το πεδίο "Seed Authors" χρησιμοποιείται από τον αλγόριθμο ταξινόμησης άρθρων του PaperFinder και παρουσιάζει το σύνολο των χαρακτηριστικών συγγραφέων με βάση, το οποίο θα ταξινομηθούν τα επιστρεφόμενα άρθρα.

Εκτός από τις επερωτήσεις, που στηρίζονται σε λέξεις-κλειδιά, το PaperFinder δίνει τη δυνατότητα δημιουργίας και εκτέλεσης επερωτήσεων με βάση ένα σύνολο χαρακτηριστικών άρθρων (**Seed Papers**). Η ύπαρξη ή όχι ενός τέτοιου συνόλου υποδεικνύεται στο πεδίο "Related To" με ένα μικρό τρίγωνο ή ένα σήμα απαγόρευσης αντίστοιχα. Η επιλογή του μικρού τριγώνου έχει ως αποτέλεσμα την εμφάνιση των χαρακτηριστικών άρθρων για τη συγκεκριμένη επερώτηση στο κατώτερο τμήμα του παραθύρου.

4.3.2.2 Παρουσίαση Αποτελεσμάτων

Όπως φαίνεται στο Σχήμα 4-4, κατά την παρουσίαση των αποτελεσμάτων το παράθυρο του αναδιφητή διαχωρίζεται σε τρία τμήματα. Το ανώτερο αριστερά χρησιμοποιείται για την περιληπτική αναφορά των τρεχόντων επερωτήσεων. Στο ανώτερο δεξιά παρουσιάζονται αποτελέσματα της επιλεγμένης επερώτησης με στοιχεία για την κατάσταση τους. Το κατώτερο τμήμα χρησιμοποιείται για την ανάγνωση των άρθρων, για την παρουσίαση επιπρόσθετης πληροφορίας πάνω στις επερωτήσεις και για την εμφάνιση μηνυμάτων ελέγχου του PaperFinder.

- *Περιληπτική αναφορά επερωτήσεων:*

ID	Keywords / Authors / Related	Unread
1	"web caching" / Any / View	184
2	"proxy caching" / Any / None	4
3	Any / paterson / None	0
4	file caching / Any / None	38

Σχήμα 4-5: Η λίστα των επερωτήσεων μαζί με τον αριθμό των μη αναγνωσμένων άρθρων.

Στο τμήμα αυτό της οθόνης (Σχήμα 4-5) αναγράφονται περιληπτικά τα χαρακτηριστικά των επερωτήσεων του χρήστη συνοδευμένα με τον αριθμό των μη αναγνωσμένων άρθρων. Τα χαρακτηριστικά, που δίνονται για κάθε επερώτηση, είναι οι λέξεις-κλειδιά, που τη χαρακτηρίζουν, οι συγγραφείς με βάση, τους οποίους

γίνεται η αναζήτηση, και η χρήση ή όχι χαρακτηριστικών άρθρων (Seed Papers). Η πιθανή εμφάνιση ενός μικρού τριγώνου ανάμεσα στα χαρακτηριστικά μίας επερώτησης και στον αριθμό των μη αναγνωσμένων άρθρων χρησιμοποιείται για τη λειτουργία "Catch Up" (Ενότητα 2.2.2.1). Η επιλογή μίας επερώτησης έχει ως αποτέλεσμα την εμφάνιση των άρθρων, που έχουν ευρεθεί για αυτή στο ανώτερο δεξιά τμήμα του παραθύρου. Για παράδειγμα, για την επερώτηση 1 του Σχήματος 4-5 αναφέρονται τα εξής:

- Χρησιμοποιεί τη φράση-κλειδί "web caching".
- Αναζητά άρθρα οποιουδήποτε συγγραφέα (Any).
- Η λέξη-σύνδεσμος "*View*" στο πεδίο Related δηλώνει ότι υπάρχει ένα τουλάχιστον χαρακτηριστικό άρθρο για την επερώτηση αυτή.
- Έχουν βρεθεί 184 μη αναγνωσμένα άρθρα.

➢ *Παρουσίαση άρθρων:*

Σχήμα 4-6: Ο πίνακας των αποτελεσμάτων της επερώτησης *"file caching"*.

Στην οθόνη αυτή (Σχήμα 4-6) παρουσιάζονται τα άρθρα, που έχουν ευρεθεί για μία συγκεκριμένη επερώτηση. Για κάθε άρθρο αναφέρεται ο τίτλος και οι συγγραφείς του κάτω από την επικεφαλίδα **"Title and Authors"**, και η ημερομηνία συγγραφής

του κάτω από την επικεφαλίδα **"Date"**. Επίσης, υποστηρίζονται όλες οι λειτουργίες, που προσφέρει το USEwebNET εκτός από την **"Version Control"**. Συγκεκριμένα παρέχονται οι επιλογές: **"Read"**, που προκαλεί ανάγνωση του άρθρου, **"Save"**, που αποθηκεύει το άρθρο σε ένα προσωπικό αρχείο, **"Ignore"**, που δηλώνει την έλλειψη ενδιαφέροντος για το συγκεκριμένο άρθρο. Επιπροσθέτως, παρουσιάζεται η κατάσταση των άρθρων, δηλαδή εάν έχουν αναγνωσθεί και εάν έχουν αποθηκευτεί. Ο λόγος που δεν προσφέρεται η λειτουργία εύρεσης νέων εκδόσεων του άρθρου και διαφορών από τις προηγούμενες είναι ότι τα άρθρα, που συγκεντρώνονται στις ψηφιακές βιβλιοθήκες που υποστηρίζονται δεν μεταβάλλονται από τη στιγμή, που θα εκδοθούν.

Εκτός από τις τρεις λειτουργίες, που προέρχονται από το USEwebNET, το PaperFinder έχει επεκταθεί, ώστε να προσφέρει τρεις νέες δυνατότητες. Οι τρεις αυτές λειτουργίες, που ονομάζονται **"Related"**, **"NewRel"** και **"Sort"** παρέχουν τρεις διαφορετικούς τρόπους με τους οποίους το επιλεγμένο άρθρο μπορεί να λειτουργήσει ως χαρακτηριστικό άρθρο (seed paper) μίας επερώτησης. Η πρώτη υποδεικνύει στο PaperFinder να αναζητήσει άρθρα σχετικά με το επιλεγμένο κατά την επομένη εκτέλεση της αντίστοιχης επερώτησης. Το δεύτερο προκαλεί τη δημιουργία μίας νέας επερώτησης, η οποία θα αναζητήσει μόνο άρθρα σχετικά με το επιλεγμένο. Τέλος, το τρίτο υποδεικνύει στο PaperFinder ότι τα αποτελέσματα θα πρέπει να ταξινομηθούν με βάση το συγκεκριμένο άρθρο.

4.3.2.3 Επεξεργασία προσωπικών αρχείων

Από το "Folders Menu" ο χρήστης οδηγείται στη σελίδα επεξεργασίας των προσωπικών αρχείων. Η σελίδα αυτή λειτουργεί ακριβώς όπως και η αντίστοιχη του USEwebNET και προσφέρει τις ίδιες λειτουργίες (Ενότητα 2.2.3). Στο Σχήμα 4-7 παρουσιάζονται τα άρθρα, που περιέχονται σε ένα αρχείο με όνομα "Web Servers". Εκτός από τα περιεχόμενα του αρχείου, που εμφανίζονται στο πάνω μισό του

παραθύρου, ο χρήστης διαβάζει ένα άρθρο με τίτλο "On disk caching Web objects in Proxy Servers" στο κάτω μέρος αυτού.

Σχήμα 4-7: Η οθόνη προσωπικών αρχείων. Ο χρήστης έχει δημιουργήσει πέντε προσωπικά αρχεία. Ελέγχει τα περιεχόμενα του αρχείου με τίτλο *"Servers"* και διαβάζει το άρθρο *"Internet Web Servers: Workload characterization and performance implications"*.

4.3.2.4 Καθορισμός προσωπικού πορτραίτου

Η ιστοσελίδα καθορισμού προσωπικού πορτραίτου χρήσης (personal profile) του PaperFinder προσφέρει δυνατότητες όμοιες με το USEwebNET. Συγκεκριμένα, υποστηρίζεται:

- *ο καθορισμός αντιμετώπισης των αναγνωσμένων ιστοσελίδων,*

- *η ρύθμιση της λειτουργίας "Catch Up" και*

- *η ειδοποίηση διαμέσου ηλεκτρονικού μηνύματος (Mail Notification).*

Τα προαναφερόμενα χαρακτηριστικά περιγράφονται αναλυτικά στην Ενότητα 2.2.4.

4.3.3 Η Μηχανή του PaperFinder

Η μηχανή του PaperFinder είναι σε μεγάλο βαθμό όμοια με αυτή του USEwebNET. Αποτελείται από τα ίδια τμήματα με εκείνα του USEwebNET με εξαίρεση τον Ελεγκτή εκδόσεων ιστοσελίδων και έχει επιπλέον ένα σύνολο προγραμμάτων, τα οποία είναι υπεύθυνα για τον αλγόριθμο ταξινόμησης των άρθρων και για τον υπολογισμό των παραμέτρων, που χρησιμοποιεί ο αλγόριθμος αυτός. Αναλυτικά, τα τμήματα, που συνιστούν τη μηχανή του PaperFinder είναι:

1. ***Συλλέκτης αποτελεσμάτων και Αναλυτής κειμένου (parser):*** Επικοινωνεί με τις υποστηριζόμενες ψηφιακές βιβλιοθήκες, εκτελώντας τις επερωτήσεις των χρηστών. Αφού συλλέξει τις ιστοσελίδες, που περιέχουν τα αποτελέσματα της επερώτησης σε μορφή HTML, ο αναλυτής κειμένου αναλαμβάνει να εξάγει την επιθυμητή πληροφορία από αυτές. Η πληροφορία αυτή περιλαμβάνει στοιχεία, όπως ο τίτλος κάθε άρθρου, οι συγγραφείς του, η ημερομηνία εκδόσεως του, η επίσημη βιβλιογραφική του αναφορά και το URL του κειμένου σε περίπτωση, που αυτό είναι διαθέσιμο. Τα στοιχεία αυτά αποθηκεύονται σε προσωρινό αρχείο και προωθούνται στο πρόγραμμα ενημέρωσης της εσωτερικής βάσης του PaperFinder.

2. ***Τμήμα ενημέρωσης της βάσεως δεδομένων:*** Παραλαμβάνει την έξοδο του αναλυτή κειμένου και ενημερώνει τη βάση του PaperFinder. Νέα άρθρα προστίθονται σε αυτή, ενώ ήδη υπάρχοντα αγνοούνται.

3. ***Πρόγραμμα αποστολής ηλεκτρονικού μηνύματος:*** Ευθύνεται για την αποστολή ηλεκτρονικού μηνύματος στο χρήστη, σε περίπτωση που βρεθεί ένας

επαρκής αριθμός νέων άρθρων, όπως αυτός έχει καθοριστεί από το προσωπικό πορτραίτο του.

4. ***Τμήμα χειρισμού περιοδικών λειτουργιών (Crontab Manager):*** Είναι υπεύθυνο για την περιοδική εκτέλεση των επερωτήσεων κάθε χρήστη. Ελέγχει την εκτέλεση όλων των υπόλοιπων τμημάτων του PaperFinder. Συνεργάζεται με τη λειτουργία **Cron** του λειτουργικού συστήματος Unix.

5. ***Πρόγραμμα ταξινόμησης αποτελεσμάτων:*** Κατά τον καθορισμό μίας επερώτησης, ο χρήστης έχει τη δυνατότητα να υποδείξει την επιθυμία για την ταξινόμηση των αποτελεσμάτων με βάση τον εσωτερικό αλγόριθμο ταξινόμησης του PaperFinder. Η υπόδειξη αυτή γίνεται με την επιλογή ενός συνόλου χαρακτηριστικών άρθρων διαμέσου της οθόνης παρουσίασης αποτελεσμάτων (Ενότητα 4.3.2.2 – Κατηγορία **"Sort"**). Το πρόγραμμα ταξινόμησης εκτελείται, εάν έχει ζητηθεί, κάθε φορά που ενημερώνεται η εσωτερική βάση του USEwebNET για τα αποτελέσματα μίας επερώτησης. Ο τρόπος λειτουργίας του αλγορίθμου ταξινόμησης θα εξηγηθεί αναλυτικά στο επόμενο κεφάλαιο.

6. ***Πρόγραμμα υπολογισμού φίλτρου:*** Ο εσωτερικός αλγόριθμος ταξινόμησης άρθρων του PaperFinder βασίζεται σε ένα φίλτρο, το οποίο περιέχει στοιχεία για την ερευνητική σχέση ενός μεγάλου πλήθους συγγραφέων. Το φίλτρο αυτό υπολογίζεται στατικά από τη βάση δεδομένων με τα άρθρα που έχει ανακαλύψει το PaperFinder και αποθηκεύεται, ώστε να χρησιμοποιείται από το πρόγραμμα ταξινόμησης όταν είναι απαραίτητο. Ο υπολογισμός του φίλτρου γίνεται από ένα ανεξάρτητο επίπεδο λογισμικού, το οποίο εκτελείται περιοδικά ενσωματώνοντας στο φίλτρο τη νέα γνώση, που έχει αποκτήσει το PaperFinder από την ανάκληση νέων βιβλιογραφικών αναφορών. Η αναλυτική περιγραφή του αλγόριθμου υπολογισμού του φίλτρου γίνεται στο επόμενο κεφάλαιο.

4.3.4 Αναγνώριση Βιβλιογραφικών Αναφορών

Ένα πρόβλημα, που εμφανίζεται στο σχεδιασμό του PaperFinder και γενικότερα κατά την αυτόματη ανάκληση βιβλιογραφικών αναφορών από τον Παγκόσμιο Ιστό αποτελεί η αναγνώριση όμοιων και ο διαχωρισμός διαφορετικών άρθρων. Τα στοιχεία, που χαρακτηρίζουν μία βιβλιογραφική αναφορά είναι:

(α) ο τίτλος της,

(β) τα ονόματα των συγγραφέων της,

(γ) το όνομα του περιοδικού ή συνεδρίου, στο οποίο έχει εκδοθεί,

(δ) οι σελίδες του περιοδικού, στο οποίο έχει εκδοθεί, και

(ε) η ημερομηνία έκδοσης.

Από τα προαναφερόμενα στοιχεία, κανένα δεν είναι επαρκές, ώστε η χρήση του να μπορεί να αναγνωρίσει μοναδικά μία βιβλιογραφική αναφορά. Για το λόγο αυτό στις περισσότερες περιπτώσεις χρησιμοποιείται ένα υποσύνολο αυτών, το οποίο συνήθως περιλαμβάνει: την τίτλο, το όνομα του περιοδικού, τις σελίδες, που έχουν χρησιμοποιηθεί από την έκδοση, τα ονόματα όλων ή μόνο του βασικού συγγραφέα και πιθανώς την ημερομηνία έκδοσης.

Στην περίπτωση του PaperFinder χρησιμοποιούνται δύο στοιχεία για τη μοναδική αναγνώριση των βιβλιογραφικών αναφορών: ο τίτλος και τα ονόματα των συγγραφέων. Οι λόγοι, που οδήγησαν στην επιλογή αυτή είναι:

(α) Η πληροφορία αυτή είναι επαρκής. Διαφορετικές βιβλιογραφικές αναφορές με όμοιους τίτλους και συγγραφείς εμφανίζονται στην περίπτωση τεχνικών αναφορών, οι οποίες στη συνέχεια εκδίδονται ως άρθρα σε περιοδικά ή

συνέδρια. Το PaperFinder συνεργάζεται με ψηφιακές βιβλιοθήκες, οι οποίες υποστηρίζουν μόνο άρθρα, που έχουν εκδοθεί, και όχι τεχνικές αναφορές.

(β) Λόγω του τρόπου, με τον οποίο επιστρέφονται οι βιβλιογραφικές αναφορές από τις υποστηριζόμενες ψηφιακές βιβλιοθήκες, είναι δύσκολος ο διαχωρισμός πεδίων, όπως οι σελίδες έκδοσης και τα ακριβή στοιχεία του περιοδικού ή συνεδρίου έκδοσης. Κατά συνέπεια, η επιπρόσθετη αυτή πληροφορία χρησιμοποιείται και παρουσιάζεται από το PaperFinder ως ένα μόνο πεδίο πληροφορίας, τύπου *"text"*.

4.3.5 Αναγνώριση Συγγραφέων

Ένα δεύτερο πρόβλημα, που συνοδεύει την αναγνώριση των βιβλιογραφικών αναφορών είναι αυτό της αναγνώρισης των συγγραφέων. Το όνομα ενός συγγραφέα μπορεί να θεωρηθεί ότι αποτελείται από τρία τμήματα:

(α) το **μικρό** του όνομα,

(β) ένα **μέσο αρχικό**, που μπορεί να είναι το αρχικό γράμμα ενός πιθανού δεύτερου μικρού ονόματος ή του ονόματος του πατέρα του συγγραφέα, και

(γ) το **επίθετο** του.

Για παράδειγμα, το όνομα *"Αθανάσιος Ε. Παπαθανασίου"* συνιστάται από το μικρό όνομα *"Αθανάσιος"*, το μέσο αρχικό *"Ε."* και το επίθετο *"Παπαθανασίου"*.

Στις βιβλιογραφικές αναφορές αναφέρεται απαραιτήτως το επίθετο του συγγραφέα και το μικρό του όνομα είτε ολόκληρο είτε με τη μορφή ενός αρχικού γράμματος. Δυστυχώς, τα δύο αυτά στοιχεία δεν επαρκούν για να αναγνωρίσουν μοναδικά ένα συγγραφέα, πράγμα που κάνει τη χρήση του μέσου αρχικού απαραίτητη.

Το PaperFinder για το διαχωρισμό των συγγραφέων χρησιμοποιεί το επίθετο, το αρχικό του ονόματος και το μέσο αρχικό. Η μέθοδος αυτή εμφανίζει το πρόβλημα ότι αναφορές του ίδιου συγγραφέα, από τις οποίες στη μία αναφέρεται το μέσο αρχικό, ενώ στην άλλη δεν εμφανίζεται θα διαχωριστούν ως διαφορετικοί συγγραφείς. Το πρόβλημα αυτό έχει επιπτώσεις στον υπολογισμό του φίλτρου, που χρησιμοποιείται από το PaperFinder για την ταξινόμηση των άρθρων, όπως θα φανεί στο επόμενο κεφάλαιο.

4.4 ΘΕΜΑΤΑ ΥΛΟΠΟΙΗΣΗΣ

Το PaperFinder αποτελεί μία εφαρμογή, η οποία λειτουργεί στο περιβάλλον του Παγκοσμίου Ιστού. Για το λόγο αυτό χρησιμοποιείται η τεχνολογία, που έχει αναπτυχθεί για την ανάκληση πληροφορίας από τον Παγκόσμιο Ιστό, όπως αυτή περιγράφεται στην Ενότητα 3.2.1.

4.4.1 Διεπιφάνεια Χρήσης

Για την υλοποίηση της διεπιφάνειας χρήσης έχει υλοποιηθεί ένα σύνολο από CGI Binaries, τα οποία επεξεργάζονται και τροποποιούν τη βάση δεδομένων του PaperFinder με βάση τις αιτήσεις του χρήστη. Οι αιτήσεις αυτές προωθούνται από τον αναδιφητή του χρήστη στον HTTP Server (Apache HTTP Server [IX]) του PaperFinder, ο οποίος τις μεταβιβάζει στο κατάλληλο πρόγραμμα CGI. Η έξοδος των προγραμμάτων CGI είναι μία ιστοσελίδα γραμμένη σε HTML με την προσθήκη ενός μικρού αριθμού ρουτινών σε γλώσσα JavaScript, η οποία μεταφέρεται από τον εξυπηρετητή στον αναδιφητή. Τα δεδομένα της ιστοσελίδας δημιουργούνται δυναμικά κατά την εκτέλεση του προγράμματος CGI σύμφωνα με τα τρέχοντα περιεχόμενα της εσωτερικής βάσης δεδομένων του PaperFinder. Η χρήση της γλώσσας JavaScript σε ορισμένες ιστοσελίδες στοχεύει στον ευέλικτο έλεγχο των γραφικών της σελίδας. Τα προγράμματα CGI έχουν γραφεί στη γλώσσα προγραμματισμού **C**.

4.4.2 Υλοποίηση της Μηχανής του PaperFinder

Με εξαίρεση τον αναλυτή κειμένου, ο οποίος περιέχει τμήματα γραμμένα στις γλώσσες **Lex** και **Yacc**, τα υπόλοιπα προγράμματα, που αποτελούν τη μηχανή του PaperFinder, έχουν υλοποιηθεί στη γλώσσα προγραμματισμού **C**. Το γεγονός αυτό τους προσφέρει το χαρακτηριστικό της εύκολης μεταφερσιμότητας σε άλλο τύπο υπολογιστή και ευνοεί τις επιδόσεις τους κατά την εκτέλεση τους.

Λεπτομέρειες για την υλοποίηση των προγραμμάτων ταξινόμησης των αποτελεσμάτων και υπολογισμού του φίλτρου, που χρησιμοποιείται από τον αλγόριθμο ταξινόμησης, δίνονται στο επόμενο κεφάλαιο. Ο τρόπος κατασκευής των υπόλοιπων τμημάτων λογισμικού, που συνιστούν τη μηχανή του PaperFinder, είναι όμοιος με των αντίστοιχων τμημάτων του USEwebNET και περιγράφεται στην Ενότητα 3.3.

4.5 ΣΥΜΠΕΡΑΣΜΑΤΑ

Το PaperFinder αποτελεί ένα χρήσιμο εργαλείο ανάκλησης ερευνητικών άρθρων, για τους εξής λόγους:

- Προσφέρει μία διεπιφάνειας χρήσης, η οποία στηρίζεται στη γνωστή και αποτελεσματική διεπιφάνεια χρήσης, που χρησιμοποιείται από τα USENET News.

- Διατηρεί προσωπικούς λογαριασμούς για κάθε χρήστη, οι οποίοι περιέχουν πληροφορία για τα άρθρα, που ο χρήστης ήδη ξέρει, και για τα θέματα, για τα οποία ενδιαφέρονται.

- Διατηρεί ένα προσωπικό πορτραίτο για κάθε χρήστη, το οποίο καθορίζει τον τρόπο αλληλεπίδρασης του χρήστη με τη διεπιφάνεια χρήσης.

- Εκμεταλλεύεται τις μηχανές αναζήτησης, διαδεδομένων Ψηφιακών Βιβλιοθηκών με στόχο την εύρεση νέων άρθρων.

- Αντιμετωπίζει το πρόβλημα του καταιγισμού πληροφορίας με την αποφυγή της παρουσίασης στο χρήστη άρθρων, τα οποία έχει ήδη διαβάσει.

- Μειώνει την αυξημένη κίνηση δεδομένων στο Διαδίκτυο και το φόρτο των εξυπηρετητών των ψηφιακών βιβλιοθηκών, επειδή διατηρεί τις επιστρεφόμενες βιβλιογραφικές αναφορές, που αφορούν μία επερώτηση τοπικά. Ανάκληση νέων άρθρων γίνεται περιοδικά.

- Αποφεύγει τη χρήση του δικτύου σε ώρες αιχμής. Η ανάκληση νέων άρθρων γίνεται σε βραδινές ώρες, που ο φόρτος του δικτύου είναι μειωμένος.

Κεφάλαιο 5

PAPERFINDER: RESOURCE DISCOVERY MODE

5.1 RESOURCE DISCOVERY MODE

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, το PaperFinder έχει τη δυνατότητα να αναζητά πληροφορία με δύο τρόπους λειτουργίας. Ο πρώτος είναι ο συνηθισμένος τρόπος αναζήτησης πληροφορίας με τη χρήση ορισμένων λέξεων-κλειδιών, που καθορίζουν το πεδίο ενδιαφέροντος του χρήστη (**Keyword-Based mode**). Ο δεύτερος τρόπος, που ονομάζεται **Resource Discovery Mode**, στηρίζεται στην ικανότητα του PaperFinder να αναζητά και να ανακαλύπτει νέα πληροφορία στηριζόμενο σε ένα σύνολο "χαρακτηριστικών άρθρων" (seed paper) και όχι απλώς σε ορισμένες λέξεις-κλειδιά.

5.2 ΣΧΕΔΙΑΣΜΟΣ

Κατά το Resource Discovery Mode ο χρήστης καθορίζει ένα ή περισσότερα χαρακτηριστικά άρθρα και το PaperFinder δημιουργεί και προωθεί επερωτήσεις στις υποστηριζόμενες ψηφιακές βιβλιοθήκες ανακαλώντας άρθρα όμοια με τα χαρακτηριστικά. Οι επερωτήσεις αυτές έχουν πιο γενική μορφή από την αρχική και στόχος τους είναι να ανακαλύψουν ένα μεγαλύτερο σύνολο δημοσιεύσεων. Για τη δημιουργία των γενικευμένων επερωτήσεων μπορούν να χρησιμοποιηθούν οι μέθοδοι, που περιγράφονται στη συνέχεια.

Ένας πρώτος τρόπος διεύρυνσης των αποτελεσμάτων μίας επερώτησης παρέχεται μέσω της έννοιας του "χαρακτηριστικού συγγραφέα". Με τρόπο όμοιο με αυτόν, που ένα γνωστό ερευνητικό άρθρο μπορεί να χαρακτηρίζει την περιοχή ή τις περιοχές έρευνας, στις οποίες ανήκει, ένας διάσημος ερευνητής μπορεί να αντιπροσωπεύει την ερευνητική περιοχή, στην οποία δημοσιεύει. Κατά συνέπεια, η

αρχική επερώτηση, που έγινε με τη χρήση λέξεων-κλειδιών, μπορεί να επαναληφθεί μετά την επιλογή ενός συνόλου "χαρακτηριστικών συγγραφέων", ώστε να δώσει ένα μεγαλύτερο πλήθος αποτελεσμάτων. Η γενικευμένη επερώτηση θα αποτελείται από ένα σύνολο επερωτήσεων, κάθε μία από τις οποίες θα ανακαλεί τα άρθρα ενός από του χαρακτηριστικούς συγγραφείς. Η μέθοδος αυτή υποστηρίζεται από την τρέχουσα έκδοση του PaperFinder, αλλά δεν γίνεται αυτοματοποιημένα, πράγμα που αποτελεί ένα σημείο μελλοντικής επέκτασης. Επιπροσθέτως, η έννοια του χαρακτηριστικού συγγραφέα χρησιμοποιείται έμμεσα από τον αλγόριθμο ταξινόμησης αποτελεσμάτων, όπως περιγράφεται στην Ενότητα 5.4.

Μία ακόμη μέθοδος διεύρυνσης μίας επερώτησης δίνεται μέσω των βιβλιογραφικών αναφορών, που περιέχει ένα άρθρο. Κάθε επιστημονική δημοσίευση προσφέρει μία επισκόπηση της τρέχουσας γνώσης για τον ερευνητικό τομέα, στον οποίο ανήκει μέσω των άρθρων, που παρουσιάζει στη βιβλιογραφία της. Παράλληλα, αποτελεί θεμελιώδες υλικό για τα άρθρα, τα οποία αναφέρονται σε αυτό. Κατά συνέπεια, οι βιβλιογραφικές αναφορές ενός άρθρου και τα άρθρα, που αναφέρονται σε αυτό, ανήκουν στο ίδιο πεδίο έρευνας και συνήθως βρίσκονταί πολύ κοντά θεματικά. Επομένως, είναι επιθυμητό η γενίκευση της επερώτησης να επιστρέφει τόσο τις βιβλιογραφικές αναφορές των χαρακτηριστικών άρθρων όσο και τα άρθρα, που το αναφέρουν. Δυστυχώς, η ανάκληση εργασιών μέσω βιβλιογραφικών αναφορών δεν υποστηρίζεται από την πλειοψηφία των ψηφιακών βιβλιοθηκών παρά τη χρησιμότητα, που μπορεί να έχει. Μία αξιοσημείωτη προσπάθεια για τη χρήση αυτή των βιβλιογραφικών αναφορών έχει γίνει από τους C. Lee Giles και Steve Lawrence με την υλοποίηση του **CiteSeer** ([22], [23]), το οποίο περιγράφεται στην Ενότητα 6.4.

Ακολούθως, η επέκταση μίας επερώτησης μπορεί να γίνει με την δημιουργία ενός νέου μεγαλύτερου συνόλου λέξεων-κλειδιών. Οι λέξεις-κλειδιά αυτές είναι δυνατόν να προκύψουν από τα χαρακτηριστικά άρθρα μέσω της λεξικογραφικής ανάλυσης των

τίτλων τους και του κειμένου τους. Κάθε λέξη-κλειδί αντιστοιχεί σε μία νέα επερώτηση, η οποία προωθείται στις υποστηριζόμενες ψηφιακές βιβλιοθήκες. Η σύζευξη των άρθρων, που ανακαλούνται από τις νέες επερωτήσεις, συνιστά το σύνολο των αποτελεσμάτων της γενικευμένης επερώτησης.

Τέλος, μία ακόμη μέθοδος για την επέκταση μίας επερώτησης μπορεί να δοθεί μέσω της ανάθεσης θεματικών προσδιοριστών σε κάθε άρθρο και την ανάκληση δημοσιεύσεων, που έχουν όμοιους θεματικούς προσδιοριστές με αυτούς του χαρακτηριστικού άρθρου. Ένας θεματικός προσδιοριστής ορίζεται ως ένα μικρό σύνολο από λέξεις-φράσεις είτε λέξεις-κλειδιά, το οποίο χαρακτηρίζει συνοπτικά ένα συγκεκριμένο πεδίο έρευνας. Η διαδικασία αυτή επεξηγείται στην Ενότητα 5.3.

Η επαύξηση του συνόλου των αποτελεσμάτων έχει ως ανεπιθύμητες συνέπειες τη μεγάλη αύξηση του όγκου της πληροφορίας, που παρουσιάζεται στον ερευνητή και πιθανώς τη λανθασμένη ανάκληση άρθρων, που δεν ανήκουν στο πεδία ενδιαφέροντος του. Για να αντιμετωπιστεί το πρόβλημα αυτό, το PaperFinder **φιλτράρει** (ταξινομεί) την πληροφορία, που ανακαλύπτει πριν την παρουσιάσει στο χρήστη. Τα άρθρα ταξινομούνται με βάση την ομοιότητας προς το χαρακτηριστικό άρθρο. Το μέτρο της ομοιότητας δύο άρθρων υπολογίζεται από ένα νέο αλγόριθμο, ο οποίος στηρίζεται στην έννοια της **"απόστασης συγγραφέων"**. Ο ορισμός της έννοια αυτής και λεπτομέρειες για τον τρόπο, με τον οποίο χρησιμοποιείται από το PaperFinder για την ταξινόμηση των αποτελεσμάτων, δίνονται στην Ενότητα 5.4.

Στην τρέχουσα υλοποίηση του PaperFinder το Resource Discovery Mode μπορεί να διαχωριστεί σε δύο τμήματα, τα οποία είναι σε σημαντικό βαθμό ανεξάρτητα μεταξύ τους

1. ***Γενίκευση Επερώτησης*** *(Query Generalization).*

2. ***Ταξινόμηση Αποτελεσμάτων*** *(Filtering).*

Στις επόμενες ενότητες περιγράφονται αναλυτικά οι δύο αυτές λειτουργίες και παρουσιάζονται συγκριτικά πειραματικά αποτελέσματα. Για τη σύγκριση του PaperFinder έχει επιλεχθεί η ψηφιακή βιβλιοθήκη της ACM, επειδή ήταν η μόνη από τις υποστηριζόμενες βιβλιοθήκες, που προσφέρει αντίστοιχες λειτουργίες.

5.3 ΓΕΝΙΚΕΥΣΗ ΕΠΕΡΩΤΗΣΗΣ

Ο στόχος της λειτουργίας γενίκευσης επερώτησης (Query Generalization) είναι να επεκτείνει το εύρος μίας επερώτησης, ώστε να ανακαλύψει περισσότερα ενδιαφέροντα άρθρα, τα οποία διαφορετικά δεν θα ήταν δυνατόν να βρεθούν. Για το λόγο αυτό, χρησιμοποιείται η έννοια του χαρακτηριστικού άρθρου (seed paper), όπως αυτή περιγράφηκε στα προηγούμενα κεφάλαια. Η γενίκευση της επερώτησης γίνεται με την αυτόματη δημιουργία επερωτήσεων, οι οποίες στο σύνολο των αποτελεσμάτων, που θα επιστρέψουν, πρέπει να περιλαμβάνουν και το χαρακτηριστικό άρθρο. Από τις προαναφερόμενες μεθόδους, με τις οποίες μπορεί να γενικευτεί μία επερώτηση, στην τρέχουσα έκδοση του PaperFinder χρησιμοποιείται αυτή, που στηρίζεται στους θεματικούς προσδιοριστές. Συγκεκριμένα, για κάθε άρθρο το PaperFinder χρησιμοποιεί τους προσδιοριστές θεμάτων, που δίνονται από τη ψηφιακή βιβλιοθήκη της ACM.

Η ACM διατηρεί μία βάση από προσδιοριστές θεμάτων (subject descriptors). Οι προσδιοριστές αυτοί σχηματίζουν μία δενδρική δομή, όπου κάθε παιδί ενός κόμβου αντιστοιχεί σε μία θεματική κατηγορία, που είναι υποσύνολο αυτής του πατέρα. Κάθε άρθρο, που εκδίδεται από την ACM, κατά την καταχώρησή του στην ψηφιακή βιβλιοθήκη, συνδέεται με ορισμένους προσδιοριστές θέματος. Συνεπώς, δημιουργείται μία βάση δεδομένων, στην οποία υπάρχει ένα σύνολο προσδιοριστών και κάθε ένας από αυτούς χαρακτηρίζει ένα πλήθος άρθρων. Αντιστρόφως, κάθε άρθρο χαρακτηρίζεται θεματικά από ένα μικρό αριθμό προσδιοριστών.

Η λειτουργικότητα των προσδιοριστών αυτών εμφανίζεται στην περίπτωση που ένας χρήστης επιθυμεί να ανακαλύψει άρθρα, τα οποία σχετίζονται με κάποιο από αυτά,

που επιστράφηκαν κατά την εκτέλεση μίας αρχικής επερώτησης. Για το σκοπό αυτό η βάση δεδομένων της ACM ανακαλεί όλα τα άρθρα, τα οποία ανήκουν σε θεματικούς προσδιοριστές όμοιους με αυτούς του επιλεγμένου άρθρου. Με τον τρόπο αυτό ο χρήστης κατορθώνει να βρει άρθρα με ιδιαίτερο για αυτόν ενδιαφέρον, τα οποία δεν ανήκουν απαραίτητα στο πεδίο έρευνας, που περιγράφει η πρωταρχική του ερώτηση και δεν περιέχουν τις λέξεις-κλειδιά, που χρησιμοποιήθηκαν. Η δυνατότητα αυτή παρέχεται μέσω ενός συνδέσμου με ονομασία *"Find Related Articles"*, ο οποίος συνοδεύει κάθε βιβλιογραφική αναφορά, που δίνεται από την ACM (Σχήμα 5-1). Ας υποτεθεί για παράδειγμα ότι ο χρήστης εκτελεί μία αρχική επερώτηση με τις λέξεις-κλειδιά *"file caching"*. Ένα από τα επιστρεφόμενα άρθρα είναι το:

> *Thomas E. Anderson, Michael D. Dahlin, Jeanna M. Neefe, David A. Patterson, Drew S. Roselli and Randolph Y. Wang:* ***"Serverless Network File Systems"****. ACM Transactions on Computer Systems, February 1996.*

Το άρθρο αυτό ανήκει στους ακόλουθους θεματικούς προσδιοριστές (όπως ακριβώς ορίζονται από την ACM):

(1) Software, OPERATING SYSTEMS, File Systems Management, Access methods.

(2) Software, OPERATING SYSTEMS, Storage Management, Allocation/deallocation strategies.

(3) Software, OPERATING SYSTEMS, Reliability, Checkpoint/restart.

(4) Software, OPERATING SYSTEMS, Performance, Measurements.

(5) Data, FILES, Organization/structure.

(6) Information Systems, INFORMATION STORAGE AND RETRIEVAL, Information Storage, File organization.

(7) Software, OPERATING SYSTEMS, Storage Management, Secondary storage.

(8) *Software, OPERATING SYSTEMS, File Systems Management, Directory structures.*

(9) *Software, OPERATING SYSTEMS, File Systems Management, Distributed file systems.*

(10) *Software, OPERATING SYSTEMS, File Systems Management, File organization.*

(11) *Software, OPERATING SYSTEMS, Reliability, Fault-tolerance.*

(12) *Software, OPERATING SYSTEMS, Performance, Simulation.*

(13) *Computer Systems Organization, COMPUTER-COMMUNICATION NETWORKS, Distributed Systems, Network operating systems.*

Από τους προσδιοριστές αυτούς μόνο οι (9) και (10) σχετίζονται με την αρχική επερώτηση. Οι υπόλοιποι δίνουν τη δυνατότητα στο χρήστη να βρει άλλα ενδιαφέροντα άρθρα, τα οποία δεν σχετίζονται με τον τομέα, που περιγράφεται από τις λέξεις-κλειδιά *"file caching"*. Συγκεκριμένα, ένα άρθρο, το οποίο βρίσκεται πολύ κοντά θεματικά στο προαναφερόμενο είναι το:

> *M. J. Feeley, W. E. Morgan, E. P. Pighin, A. R. Karlin, H. M. Levy, C. A. Thekkath:* ***"Implementing global memory management in a workstation cluster"****. ACM SIGOPS Operating Systems Review, Vol. 29, No. 5 (Dec. 3, 1995), Pages 201-212.*

Το άρθρο αυτό ανήκει στους θεματικούς προσδιοριστές:

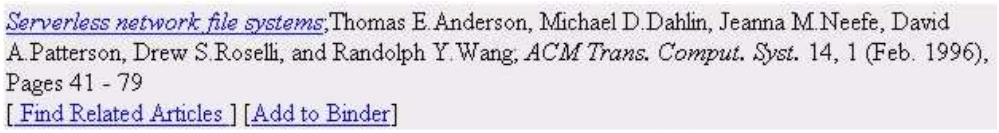
(1) *Software, OPERATING SYSTEMS, Storage Management, Virtual memory.*

(2) *Computer Systems Organization, COMPUTER-COMMUNICATION NETWORKS, Distributed Systems, Network operating systems.*

(3) *Software, OPERATING SYSTEMS, Performance.*

Η αρχική επερώτηση με τις λέξεις-κλειδιά *"file caching"* δεν επιστρέφει το προαναφερόμενο άρθρο. Το άρθρο αυτό όμως μπορεί να βρεθεί μέσω της

λειτουργίας *"Find Related Articles"* (Σχήμα 5-1) και των συσχετίσεων, που δημιουργούνται μέσω των θεματικών προσδιοριστών.


Serverless network file systems;Thomas E.Anderson, Michael D.Dahlin, Jeanna M.Neefe, David A.Patterson, Drew S.Roselli, and Randolph Y.Wang, *ACM Trans. Comput. Syst.* 14, 1 (Feb. 1996), Pages 41 - 79
[Find Related Articles] [Add to Binder]

Σχήμα 5-1: Η μορφή, παρουσίασης ενός αποτελέσματος από την ψηφιακή βιβλιοθήκη της ACM. Φαίνεται ο σύνδεσμος *"Find Related Articles"*, ο οποίος προκαλεί την ανάκληση σχετικών άρθρων.

Η γενίκευση της επερώτησης παρέχει τη δυνατότητα της εύκολης ανάκλησης χρήσιμης και ενδιαφέρουσας πληροφορίας. Χωρίς τη λειτουργία αυτή, ο χρήστης θα χρειαζόταν να εκτελέσει ένα μεγάλο αριθμό επερωτήσεων, με διαφορετικούς συνδυασμούς λέξεων-κλειδιών για να εντοπίσει τα επιθυμητά αποτελέσματα. Μία τέτοια διαδικασία είναι ιδιαίτερα επίπονη και χρονοβόρα. Επίσης, ένα άρθρο χαρακτηρίζει πολύ πιο καλά το πεδίο ενδιαφέροντος του χρήστη από ένα σύνολο λέξεων-κλειδιών. Κατά συνέπεια, μία μέθοδος εντοπισμού πληροφορίας, που στηρίζεται στην έννοια του *"χαρακτηριστικού άρθρου"* (seed paper) είναι περισσότερο αποτελεσματική και ευέλικτη από την παραδοσιακή αναζήτηση πληροφορίας, που στηρίζεται στον καθορισμό ενός μικρού συνόλου λέξεων-κλειδιών.

Στις επόμενες ενότητες θα αναλυθεί ο τρόπος, που έχει υλοποιηθεί η γενίκευση επερώτησης στο PaperFinder και θα παρουσιαστούν άλλες μέθοδοι με τις οποίες μπορεί να επιτευχθεί αυτή.

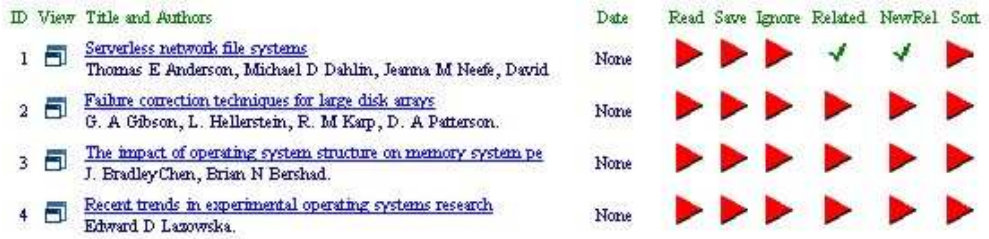
5.3.1 Αρχιτεκτονική

Όπως προαναφέρθηκε, η γενίκευση επερωτήσεων της τρέχουσας έκδοσης του PaperFinder, στηρίζεται στην αντίστοιχη λειτουργία, που προσφέρει η ψηφιακή βιβλιοθήκη της ACM. Η δυνατότητα εύρεσης "σχετικών" άρθρων παρέχεται από τη

μηχανή αναζήτησης της ACM διαμέσου της ιστοσελίδας των επιστρεφόμενων αποτελεσμάτων. Συγκεκριμένα, για κάθε επιστρεφόμενο αποτέλεσμα παρέχεται ένας σύνδεσμος με ονομασία *"Find Related Articles"* (Σχήμα 5-1), ο οποίος προκαλεί την εκτέλεση της επερώτησης που ανακαλεί όλα τα σχετικά άρθρα από τη βάση δεδομένων της ACM.

Κατά την ανάκληση αποτελεσμάτων από την ψηφιακή βιβλιοθήκη της ACM, ο αναλυτής κειμένου του PaperFinder αναγνωρίζει τους συνδέσμους με ονομασία *"Find Related Articles"* και τους αποθηκεύει στη βάση δεδομένων του μαζί με τα ανακαλούμενα άρθρα. Οι σύνδεσμοι αυτοί αποθηκεύονται με τη μορφή URL και χρησιμοποιούνται όταν απαιτήσει ο χρήστης να επιλεγεί ένα άρθρο ως "χαρακτηριστικό άρθρο" της επερώτησης. Κατά την εκτέλεση του προγράμματος ανάκλησης άρθρων αναγνωρίζονται τα *"χαρακτηριστικά άρθρα"* κάθε επερώτησης και λαμβάνονται οι ιστοσελίδες, που αντιστοιχούν στα αποθηκευμένα URLs.

5.3.1.1 Διεπιφάνεια Χρήσης



Σχήμα 5-2: Εικόνα από την ιστοσελίδα αποτελεσμάτων του PaperFinder. Το άρθρο *"Serverless network file systems"* χρησιμοποιείται ως *"χαρακτηριστικό άρθρο"* για τη συγκεκριμένη επερώτηση.

Η επιλογή ενός άρθρου ως χαρακτηριστικού μίας επερώτησης γίνεται μέσω της σελίδας αναφοράς των αποτελεσμάτων μίας επερώτησης. Όπως φαίνεται στο Σχήμα 5-2, το άρθρο *"Serverless network file systems"* έχει σημειωθεί ως χαρακτηριστικό της επερώτησης. Για το σκοπό αυτό μπορούν να χρησιμοποιηθούν δύο λειτουργίες:

1. ***Related:*** Τα άρθρο επιλέγεται ως χαρακτηριστικό της επερώτησης. Τα σχετικά με αυτό άρθρα προστίθονται ως αποτελέσματα της τρέχουσας επερώτησης.

2. ***NewRel:*** Δημιουργείται μία νέα επερώτηση, για την οποία ανακαλούνται τα άρθρα, που είναι θεματικά κοντά στο επιλεγμένο.

Η διαφορά των δύο λειτουργιών είναι ότι στη δεύτερη περίπτωση δημιουργείται μία νέα επερώτηση και τα νέα αποτελέσματα συλλέγονται χωριστά από αυτά της αρχικής επερώτησης., ενώ στην πρώτη περίπτωση τα νέα αποτελέσματα ενσωματώνονται σε αυτά της αρχικής.

5.3.2 Υλοποίηση

Η υλοποίηση και λειτουργία της διεπιφάνειας χρήσης περιγράφεται στην Ενότητα 4.4.1. Η υλοποίηση της μεθόδου γενικεύσεως των επερωτήσεων έχει ενσωματωθεί στο συλλέκτη αποτελεσμάτων και αναλυτή κειμένου (parser) του PaperFinder (Ενότητα. 4.4.2). Συγκεκριμένα, ο αναλυτής κειμένου κατά τη λεξικογραφική ανάλυση των ιστοσελίδων, που επιστρέφονται από την ψηφιακή βιβλιοθήκη της ACM, αναγνωρίζει τους συνδέσμους με όνομα *"Find Related Articles"* (Σχήμα 5-1) και αποθηκεύει το αντίστοιχο URL μαζί με τα υπόλοιπα χαρακτηριστικά κάθε δημοσίευσης, όπως τίτλο και συγγραφέα. Στη συνέχεια, για κάθε δημοσίευση, που έχει σημειωθεί ως χαρακτηριστική μίας επερώτησης, ο συλλέκτης αποτελεσμάτων ανακαλεί πληροφορία από το αντίστοιχο URL, που βρίσκει στη βάση του PaperFinder.

5.4 ΤΑΞΙΝΟΜΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Η λειτουργία "γενίκευσης επερώτησης" του PaperFinder, που επεξηγήθηκε στην προηγούμενη ενότητα, παρέχει τη δυνατότητα στο χρήστη να διευρύνει την επερώτηση του και να ανακαλύπτει με τον τρόπο αυτό ένα μεγαλύτερο αριθμό άρθρων, που τον ενδιαφέρουν. Δυστυχώς, η λειτουργία αυτή σε πολλές περιπτώσεις μπορεί να οδηγήσει στην ανάκληση ενός πολύ μεγάλου αριθμού άρθρων. Για

παράδειγμα, εάν ο χρήστης έχει χρησιμοποιήσει ένα σημαντικό αριθμό "χαρακτηριστικών άρθρων", τα οποία σχετίζονται με πολλούς τομείς έρευνας, τότε κατά η εκτέλεση της γενικευμένης επερώτησης θα έχει ως αποτέλεσμα τον καταιγισμό του χρήστη από νέα πληροφορία. Το πιθανά μεγάλο αυτό μέγεθος των επιστρεφόμενων δεδομένων μπορεί να κάνει την επεξεργασία των αποτελεσμάτων ιδιαίτερα κουραστική και επίπονη, πράγμα που εξαλείφει τα πλεονεκτήματα, που προσφέρει το PaperFinder.

Για το λόγο αυτό, κατά τη γενίκευση των επερωτήσεων κρίνεται απαραίτητη η ύπαρξη μίας μεθόδου αναγνώρισης της πιο χρήσιμης πληροφορίας. Μία τέτοια μέθοδο αποτελεί η ταξινόμηση των επιστρεφόμενων άρθρων, με τέτοιο τρόπο, ώστε τα πιο ενδιαφέροντα είτε πιο σχετικά με τον τομέα ενδιαφέροντος του χρήστη να παρουσιάζονται πρώτα. Η ταξινόμηση αυτή δίνει τη δυνατότητα στον ερευνητή να επικεντρώσει γρήγορα το ενδιαφέρον του στο πιο σημαντικό και χρήσιμο τμήμα της επιστρεφόμενης πληροφορίας.

Το PaperFinder υποστηρίζει ένα πρωτοποριακό αλγόριθμο ταξινόμησης αποτελεσμάτων. Ο αλγόριθμος αυτός προσπαθεί να εντοπίσει άρθρα, τα οποία βρίσκονται πιο κοντά στον τομέα ενδιαφέροντος του χρήστη και τα παρουσιάζει πρώτα στη λίστα των αποτελεσμάτων. Παράλληλα, διατηρεί στη λίστα των αποτελεσμάτων τα άρθρα, που χαρακτηρίζει λιγότερο σημαντικά. Με τον τρόπο αυτό, παρέχεται στο χρήστη η δυνατότητα να επεξεργαστεί μελλοντικά το σύνολο των επιστρεφόμενων άρθρων. Κατά συνέπεια, σε καμία περίπτωση δεν χάνεται σημαντική πληροφορία, αφού ακόμη και σε περιπτώσεις που ένα ενδιαφέρον άρθρο τοποθετήθηκε χαμηλά στη λίστα των αποτελεσμάτων θα παρουσιαστεί στο χρήστη.

Επομένως, ο αλγόριθμος ταξινόμησης του PaperFinder ευνοεί τη γρήγορη αναγνώριση της πιο χρήσιμης πληροφορίας και ταυτόχρονα αποτρέπει την απώλεια πληροφορίας, που λανθασμένα κρίθηκε ως μη σημαντική.

Στις επόμενες δύο ενότητες αναλύεται η αρχιτεκτονική και η υλοποίηση του αλγορίθμου.

5.4.1 Αρχιτεκτονική

5.4.1.1 Ορισμοί και περιγραφή λειτουργίας

Ο στόχος της ταξινόμησης (filtering) των αποτελεσμάτων είναι η εύρεση των πιο σημαντικών άρθρων από αυτά, που επιστράφηκαν από τη λειτουργία γενίκευσης της επερώτησης. Μία τέτοια διαδικασία είναι ιδιαίτερα δύσκολη και πολλές φορές μπορεί να καταλήξει σε νέο καταιγισμό μη σχετικής πληροφορίας.

Για την ταξινόμηση των αποτελεσμάτων το PaperFinder χρησιμοποιεί την έννοια του *"χαρακτηριστικού άρθρου"*. Συγκεκριμένα, εφαρμόζει ορισμένα κριτήρια ομοιότητας (similarity metrics) στα επιστρεφόμενα άρθρα, ώστε να διαπιστώσει πόσο όμοια είναι με το χαρακτηριστικό άρθρο.

Το κριτήριο ομοιότητας, που χρησιμοποιείται από το PaperFinder, στηρίζεται στην έννοια της *"απόστασης συγγραφέων" (author distance)*. Η έννοια αυτή είναι εμπνευσμένη από τον *αριθμό Erdos (Erdos Number)*[1] [8] και ορίζεται με τον ακόλουθο τρόπο. Δύο συγγραφείς έχουν απόσταση 1, εάν έχουν συγγράψει τουλάχιστον ένα άρθρο. Η απόστασή τους είναι δύο, εάν δεν έχουν κανένα κοινό άρθρο και υπάρχει τουλάχιστον ένας τρίτος συγγραφέας, με τον οποίο έχουν συγγράψει και οι δύο κάποιο άρθρο. Ο υπολογισμός των αποστάσεων συνεχίζεται με τον ίδιο τρόπο.

Χρησιμοποιώντας την έννοια της απόστασης συγγραφέων, το μέτρο ομοιότητας δύο άρθρων υπολογίζεται, όπως περιγράφεται στη συνέχεια:

---

[1] Ο Paul Erdos, ο πολυταξιδεμένος και πολύ γνωστός Ούγγρος μαθηματικός, έγραψε εκατοντάδες ερευνητικά άρθρα σε πολλές διαφορετικές περιοχές έρευνας. Τα περισσότερα από αυτά σε συνεργασία με άλλους επιστήμονες. Ο αριθμός Erdos του Paul Erdos ήταν 0. Άλλοι συγγραφείς, οι οποίοι έχουν συγγράψει ένα τουλάχιστον άρθρο μαζί του έχουν αριθμό Erdos 1. Συγγραφείς, οι οποίοι έχουν γράψει ένα κοινό άρθρο με κάποιον από αυτούς, που έχουν αριθμό Erdos 1, και όχι με τον Paul Erdos έχουν αριθμό Erdos 2. Η διαδικασία επιλογής των αριθμών Erdos συνεχίζεται με τον τρόπο αυτό.

1. Υπολογίζονται οι αποστάσεις κάθε ζεύγους συγγραφέων των δύο άρθρων. Δηλαδή για δύο άρθρα Χ και Υ με συγγραφείς X_1, X_2, X_3,… X_ν και Y_1, Y_2, Y_3,… Y_ν αντίστοιχα υπολογίζονται οι τιμές $AΣ(X_1,Y_1)$, $AΣ(X_1,Y_2)$,… $AΣ(X_1,Y_\nu)$,… $AΣ(X_\mu,Y_1)$, … $AΣ(X_\mu,Y_\nu)$, όπου ΑΣ η μέθοδος υπολογισμού της απόστασης συγγραφέων.

2. Υπολογίζεται το μέτρο ομοιότητας με βάση μία συγκεκριμένη συνάρτηση, η οποία παίρνει ως ορίσματα τις αποστάσεις των συγγραφέων. Πιθανές συναρτήσεις μπορούν να είναι: η συνάρτηση μέσου όρου, η συνάρτηση ελαχίστου, η συνάρτηση μεγίστου, κάποιος συνδυασμός αυτός.

   - *Συνάρτηση Μέσου Όρου (**Avg**):*
     Avg($AΣ(X_1,Y_1)$, $AΣ(X_1,Y_2)$,… $AΣ(X_1,Y_\nu)$,… $AΣ(X_\mu,Y_1)$, … $AΣ(X_\mu,Y_\nu)$**)**.

   - *Συνάρτηση Ελαχίστου (**Min**):*
     Min($AΣ(X_1,Y_1)$, $AΣ(X_1,Y_2)$,… $AΣ(X_1,Y_\nu)$,… $AΣ(X_\mu,Y_1)$, … $AΣ(X_\mu,Y_\nu)$**)**.

   - *Συνάρτηση Μεγίστου (**Max**):*
     Max($AΣ(X_1,Y_1)$, $AΣ(X_1,Y_2)$,… $AΣ(X_1,Y_\nu)$,… $AΣ(X_\mu,Y_1)$, … $AΣ(X_\mu,Y_\nu)$**)**.

3. Μικρό μέτρο ομοιότητας δηλώνει ότι τα δύο άρθρα βρίσκονται θεματικά κοντά.

5.4.1.2 Σχεδιασμός

Η διαδικασία ταξινόμησης των αποτελεσμάτων μπορεί να χωριστεί σε δύο τμήματα από την άποψη του σχεδιασμού. Το πρώτο τμήμα κατασκευάζει το φίλτρο, το οποίο χρησιμοποιείται από το δεύτερο, που είναι υπεύθυνο για την πραγματική ταξινόμηση των αποτελεσμάτων.

Το φίλτρο είναι μία μικρή βάση δεδομένων, η οποία περιέχει τις αποστάσεις όλων των συγγραφέων, που γνωρίζει το PaperFinder. Η διαδικασία υπολογισμού του είναι ιδιαίτερα χρονοβόρα. Για το λόγο αυτό γίνεται περιοδικά, κάθε φορά, που το

PaperFinder έχει ανακαλύψει ένα μεγάλο αριθμό νέων βιβλιογραφικών αναφορών και συγγραφέων.

Κατά την ταξινόμηση ενός συνόλου αποτελεσμάτων, αναγνωρίζονται οι συγγραφείς των βιβλιογραφικών αναφορών και ανακαλούνται από τη βάση δεδομένων (φίλτρο) οι αποστάσεις των συγγραφέων των αποτελεσμάτων από τους συγγραφείς των χαρακτηριστικών άρθρων. Στη συνέχεια, υπολογίζεται το μέτρο ομοιότητας κάθε άρθρου με το χαρακτηριστικό άρθρο. Σε περίπτωση, που το σύνολο των χαρακτηριστικών άρθρων περιέχει περισσότερα από ένα άρθρα υπολογίζεται ο μέσος όρος των μέτρων ομοιότητας με κάθε ένα από τα χαρακτηριστικά άρθρα. Τέλος, ταξινομούνται τα αποτελέσματα σε αύξουσα σειρά με βάση το τελικό μέτρο ομοιότητας.

5.4.1.3 Υπολογισμός φίλτρου

Ο υπολογισμός του φίλτρου γίνεται από ένα ανεξάρτητο πρόγραμμα του PaperFinder, το οποίο εκτελείται περιοδικά κάθε φορά, που έχει βρεθεί σημαντική ποσότητα νέας πληροφορίας, ώστε να είναι απαραίτητη η ανανέωση του. Ο υπολογισμός αυτός ξεκινάει αναλύοντας τις βιβλιογραφικές αναφορές, που γνωρίζει το PaperFinder. Για κάθε αναφορά αναγνωρίζονται οι συγγραφείς και εντοπίζονται οι περιπτώσεις ερευνητών, που έχουν γράψει μαζί τουλάχιστον ένα άρθρο, δηλαδή έχουν απόσταση ένα. Το σύνολο των συγγραφέων αποθηκεύεται εσωτερικά στο πρόγραμμα υπολογισμού με τη μορφή ενός *μη κατευθυνόμενου κυκλικού γράφου*, όπου κάθε συγγραφέας αποτελεί ένα κόμβο. Η ακμή μεταξύ δύο κόμβων δηλώνει ότι οι αντίστοιχοι συγγραφείς έχουν τουλάχιστον ένα κοινό άρθρο. Με τον τρόπο αυτό το πρόβλημα της εύρεσης των αποστάσεων ενός συνόλου συγγραφέων μετατρέπεται στο γνωστό πρόβλημα του υπολογισμού των ελάχιστων αποστάσεων μεταξύ όλων των κόμβων ενός γράφου. Το πρόβλημα αυτό λύνεται με τον αλγόριθμο του **Floyd** σε χρόνο, που φράζεται από τον κύβο του πλήθους των κόμβων του γράφου:

O(ν³), *όπου* ***ν*** *το πλήθος των κόμβων του γράφου.*

Με το πέρας του υπολογισμού, το φίλτρο με τις αποστάσεις των συγγραφέων αποθηκεύεται στη βάση δεδομένων του PaperFinder. Συγκεκριμένα, το φίλτρο περιέχεται σε δύο αρχεία. Το πρώτο περιέχει τα ονόματα όλων των συγγραφέων ταξινομημένα με αύξουσα σειρά μαζί με ένα μοναδικό αριθμό ταυτότητας. Το δεύτερο περιέχει όλα τα δυνατά ζεύγη συγγραφέων, που έχουν πεπερασμένη απόσταση και την τιμή της απόστασης αυτής. Το αρχείο αυτό είναι ταξινομημένο με βάση τους αριθμούς ταυτότητας των ζευγών.

Η ταξινόμηση των αρχείων γίνεται με στόχο τη γρήγορη προσπέλαση τους από το πρόγραμμα ταξινόμησης αποτελεσμάτων.

5.4.1.4 Ταξινόμηση Αποτελεσμάτων

Το πρόγραμμα ταξινόμησης των αποτελεσμάτων εκτελείται κάθε φορά, που ανακαλούνται νέες βιβλιογραφικές αναφορές για μία επερώτηση του χρήστη και έχει υποδειχθεί από αυτόν η επιθυμία να ταξινομούνται τα επιστρεφόμενα άρθρα. Για την ταξινόμηση ακολουθείται η εξής διαδικασία:

1. Αναλύονται οι βιβλιογραφικές αναφορές των επιστρεφόμενων και των χαρακτηριστικών άρθρων και βρίσκεται το σύνολο των συγγραφέων.

2. Αναζητούνται οι ταυτότητες των συγγραφέων από τη βάση δεδομένων του PaperFinder. Άγνωστοι από το φίλτρο συγγραφείς αγνοούνται.

3. Σχηματίζονται όλα τα δυνατά ζεύγη μεταξύ των συγγραφέων των αποτελεσμάτων και των χαρακτηριστικών άρθρων.

4. Ανακαλούνται οι αποστάσεις των σχηματιζόμενων από το αρχείο του φίλτρου. Άπειρες αποστάσεις εξισώνονται με το διπλάσιο της μέγιστης απόστασης, που υπάρχει στη βάση.

5. Υπολογίζεται το μέτρο ομοιότητας κάθε βιβλιογραφικής αναφοράς, το οποίο έχει οριστεί ως ίσο με το μέσο όρο των αποστάσεων ή την ελάχιστη απόσταση.

6. Οι βιβλιογραφικές αναφορές ταξινομούνται με βάση το μέτρο ομοιότητας και αποθηκεύονται στη βάση δεδομένων του PaperFinder.

5.4.1.5 Διεπιφάνεια Χρήσης