ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

ΜΕΛΕΤΗ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΜΙΑΣ ΜΗΧΑΝΗΣ ΜΕΤΑ-ΑΝΑΖΗΤΗΣΗΣ. (META-SEARCH ENGINE)

Λεύκιος Ματθαίου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Αναγνωστόπουλος Ιωάννης
Επίκουρος Καθηγητής

Λαμία, 2012

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Αφιερωμένο στους αγαπημένους μου γονείς.

Πρόλογος

Η παρούσα διπλωματική εργασία με θέμα «Μελέτη και υλοποίηση μιας Μηχανής Μετά-Αναζήτησης» πραγματοποιήθηκε εξ ολοκλήρου στο Τμήμα Πληροφορικής με εφαρμογές στην Βιοϊατρική του Πανεπιστημίου Στερεάς Ελλάδας, κατά τη διάρκεια του ακαδημαϊκού έτους, 2010-2011, υπό την επίβλεψη του Επίκουρου Καθηγητή, κ. Ιωάννη Αναγνωστόπουλου.

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου για την εμπιστοσύνη που μου έδειξε κατά την ανάθεση της διπλωματικής εργασίας, άλλα και για το μεγάλο ζήλο και όρεξη που έδειξε κατά όλη την χρονική διάρκεια διεκπεραίωσης της πτυχιακής μου εργασίας, παρέχοντας μου πολύτιμη βοήθεια, χρήσιμες πληροφορίες και συμβουλές.

Τέλος, ένα μεγάλο ευχαριστώ στην οικογένεια μου, στους φίλους μου και στη κοπέλα μου για την υποστήριξη και την αγάπη τους.

Περίληψη

Στις 25 Ιουλίου του 2008, οι μηχανικοί της Google Jesse Alpert και Nissan Haja ανακοίνωσαν ότι τα συστήματα επεξεργασίας συνδέσμων της Google ανακάλυψαν ένα τρισεκατομμύριο μοναδικές σελίδες στο διαδίκτυο.[1] Φανταστείτε λοιπόν το διαδίκτυο χωρίς καμία υπηρεσία αναζήτησης. Υποθέστε ότι θέλετε να βρείτε κάποιες πληροφορίες πάνω σε ένα θέμα χωρίς να γνωρίζεται κάποια ιστοσελίδα με αντίστοιχο περιεχόμενο. Όπως καταλαβαίνεται η αναζήτηση και εύρεση της σωστής πληροφορίας στο διαδίκτυο είναι σχεδόν αδύνατη εάν ο χρήστης δεν υποβοηθηθεί από τις μηχανές αναζήτησης. Ο χρήστης απλά θέτει το ερώτημα και η μηχανή αναζήτησης επιστρέφει μια λίστα από ιστοσελίδες που το προσεγγίζουν περισσότερο.

Είναι τέτοιο όμως το μέγεθος και η ραγδαία ανάπτυξη του παγκόσμιου ιστού που έχει ως αποτέλεσμα μια «έκρηξη» στις μηχανές αναζήτησης. Κάθε μηχανή αναζήτησης μπορεί να χειριστεί ένα μικρό ποσοστό πληροφοριών από το σύνολο του παγκοσμίου ιστού. Ο συνδυασμός μερικών μηχανών αναζήτησης μπορεί να οδηγήσει σε υψηλότερο ποσοστό κάλυψης και επομένως σε εύρεση ποιοτικότερων πληροφοριών. Αυτή ακριβώς είναι η λειτουργία μιας Μηχανής Μετά-Αναζήτησης. Είναι δηλαδή ένα σύστημα το οποίο παρέχει ολοκληρωμένη πρόσβαση σε πολλαπλές μηχανές αναζήτησης. Όταν ένα ερώτημα εκτελείται σε μια Μηχανή Μετά-Αναζήτησης το σύστημα περνά το ερώτημα σε πολλαπλές μηχανές αναζήτησης, συλλέγει τα επιμέρους αποτελέσματα και τα συγχωνεύει σε ένα ενιαίο κατάλογο κατάταξης. Η συγχώνευση των αποτελεσμάτων αποτελεί το βασικό συστατικό μιας Μηχανής Μετά-Αναζήτησης. Ο αλγόριθμος συγχώνευσης που χρησιμοποιείτε σχετίζεται άμεσα με το ποσό αποτελεσματική είναι μια μηχανή μετά-αναζήτησης.

Στη παρούσα διπλωματική, η Μηχανή Μετά Αναζήτησης που υλοποιήθηκε χρησιμοποιεί τα αποτελέσματα των τριών πιο διαδεδομένων μηχανών αναζήτησης, της Yahoo, της Google και της Bing και βασίζεται στην τεχνική συγχώνευσης του «Borda» [2]. Έχει υλοποιηθεί μια εφαρμογή μετά-αναζήτησης στο περιβάλλον της Java, κάνοντας χρήση της βιβλιοθήκης Jsoup [3] μιας διεπαφής προγραμματισμού (API) ανοικτού κώδικα (Open Source). Το πλαίσιο του API αυτού παρέχει τα κατάλληλα εργαλεία για εξαγωγή και χειρισμό δεδομένων από ένα HTML κώδικα.

Λέξεις Κλειδιά

Μηχανή Μετά Αναζήτησης, αλγόριθμος συγχώνευσης, τεχνική συγχώνευσης του Borda, βιβλιοθήκη Jsoup.

ABSTRACT

On July 25 of 2008, the search engineers of Google, Jesse Alpert and Nissan Haja, announced that the process links systems of the company had discovered one trillion unique URL's on the web at once. [1] Thus, can you imagine the Internet without any search engine service? Also, suppose that you want to find some information on an issue without knowing a website with relevant content; searching and finding accurate information on Internet is almost impossible if the user has none assistance from search engines. By using the search engines, the user simply asks the question and the engine returns a list of sites that are relevant and closer to the subject under investigation.

However, the size and the rapid growth of the web are such that it has resulted in an eruption of the search engines. Each search engine can handle a small amount of information from all of the World Wide Web. Consequently, the combination of some search engines may lead to higher coverage of the World Wide Web and subsequent to the discovering of more superior and precise information. This exactly, is the main function of a Meta-Search Engine; a system that provides integrated access to multiple search engines. When a query is performed on a Meta-Search Engine, the system passes the query to multiple search engines; it collects the individual results and merges them into a single ranking list. The combination of the results is the basic component of a Meta Search Engine. Therefore, the merger algorithm used, is directly related to how effective is a Meta-Search Engine.

The Meta-Search Engine occurred from this diplomatic paper, uses the results of the three most famous search engines, Yahoo, Google and Bing and is based on the merger technique of "Borda". It has implemented an application of meta-search in the environment of Java, using the library "Jsoup"; an open source application program

interface (API). The framework of the API thus, has provided the tools to export and manipulate data from an HTML code.

Key Words

Meta-Search Engine, merge algorithm, merger technique of “Borda”, Jsoup library.

ΚΕΦΑΛΑΙΟ 1 – ΕΙΣΑΓΩΓΗ

Η παρούσα διπλωματική εργασία μελετά τον τομέα της αναζήτησης πληροφορίας στο Διαδίκτυο (Κεφάλαιο 2), ξεκινώντας με μια ιστορική αναδρομή και συνεχίζοντας με την παρουσίαση των κυριότερων προβλημάτων κατά την αναζήτηση και ανάκτηση πληροφοριών. Στο ίδιο κεφάλαιο αναφέρονται τρόποι βελτίωσης της απόδοσης των υπηρεσιών αναζήτησης καθώς επίσης και κάποιες προηγμένες λειτουργίες που βοηθούν στην εύρεση καλύτερων αποτελεσμάτων.

Στη συνεχεία γίνεται μελέτη των περισσότερων διαδεδομένων μέσων για την εύρεση πληροφορίας στο διαδίκτυο, των Μηχανών Αναζήτησης (Κεφάλαιο 3). Το κεφάλαιο αυτό αρχίζει με μια ιστορική αναδρομή στην εξέλιξη των μηχανών αναζήτησης, συνεχίζοντας με τον τρόπο λειτουργίας τους και τα οφέλη που προσφέρει η χρήση τους. Το κεφάλαιο κλείνει με την μελέτη των τριών κυριότερων Μηχανών Αναζήτησης, της Google, Yahoo και Bing. Συγκεκριμένα, για κάθε μια μαθαίνουμε στοιχεία για την εξέλιξη και τη λειτουργιά της, καθώς επίσης για τις υπηρεσίες, τις επιλογές και τους περιορισμούς αναζήτησης που προσφέρει.

Το 4^{ο} κεφάλαιο ερευνά τις Μηχανές Μετά-Αναζήτησης αναφέροντας τις κατηγορίες στις όποιες χωρίζονται και τα σταδία λειτουργίας τους. Στη συνέχεια γίνεται περιγραφή των βασικών χαρακτηριστικών τους και των διαδικασιών που ενεργοποιούνται κατά την αναζήτηση πληροφοριών στο διαδίκτυο. Στο τέλος του κεφαλαίου γίνεται μια μικρή επισκόπηση στο χώρο των Μηχανών Μετά-Αναζήτησης με την παρουσίαση των βασικών χαρακτηριστικών μερικών δημοφιλών Μηχανών Μετά-Αναζήτησης

Τα κεφάλαια 5 και 6, ασχολούνται με την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε στα πλαίσια αυτής της διπλωματικής εργασίας. Στο Κεφάλαιο 5 αναλύεται η λειτουργικότητα της και τεκμηριώνεται ο κώδικας της. Στο Κεφάλαιο 6 γίνεται προσπάθεια αξιολόγησης της εφαρμογής, με χρήση των δεικτών ανάκλησης , ακρίβειας και παρουσίαση αρκετών διαγραμμάτων μέσα από τα οποία μπορούν να προκύψουν σημαντικά συμπεράσματα.

Στο τελευταίο Κεφάλαιο αναφέρονται τα συμπεράσματα που πρόεκυψαν από αυτή τη μελέτη. Τέλος, ακολουθεί ένα παράρτημα με τον πηγαίο κώδικα της εφαρμογής Μετά-Αναζήτησης που υλοποιήθηκε.

ΚΕΦΑΛΑΙΟ 2 - ΑΝΑΖΗΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

2.1 Διαδίκτυο και Αναζήτηση Πληροφορίας

Η ποσότητα πληροφορίας που υπάρχει καταχωρημένη στον παγκόσμιο ιστό είναι παρά πολύ μεγάλη. Η ταχύτατη ανάπτυξη του διαδικτύου και η συνεχώς αυξανόμενη ποσότητα πληροφορίας που συγκεντρώνεται σε αυτό οδηγούν στην ραγδαία αύξηση της ηλεκτρονικής πληροφορίας. Πως θα μπορούσατε μέσα σε αυτούς τους "τόνους" πληροφορίας να βρείτε αυτό που πραγματικά σας ενδιαφέρει την κάθε στιγμή;

«Χωρίς τα εργαλεία και τις μεθοδολογίες για τη συλλογή, την αξιολόγηση, τη διαχείριση και την παρουσίαση των πληροφοριών, η δυνατότητα του Παγκόσμιου Ιστού ως κόσμου γνώσης, θα μπορούσε να είχε χαθεί - John December ».

Οι βασικοί τρόποι αναζήτησης πληροφοριών και εντοπισμού των επιθυμητών ιστοσελίδων στο διαδίκτυο είναι:

- οι διαδικτυακοί κατάλογοι,
- οι ηλεκτρονικές βάσεις δεδομένων,
- οι δικτυακές πύλες,
- οι μηχανές αναζήτησης και
- οι μηχανές μετά-αναζήτησης.

Οι διαδικτυακοί κατάλογοι είναι ιστοσελίδες που περιέχουν καταλόγους με διευθύνσεις που έχουν ταξινομηθεί σε κατηγορίες και υποκατηγορίες. Η προσέγγιση της πληροφορίας που ενδιαφέρει γίνεται σταδιακά προχωρώντας μέσα στα επίπεδα μέχρι το τελευταίο οπού θα εμφανιστούν οι καταχωρημένες ιστοσελίδες (π.χ. «Εκπαίδευση» - «Δημόσια Εκπαίδευση» - «Τριτοβάθμια» - «Πανεπιστήμιο»). Οι ηλεκτρονικές Βάσεις Δεδομένων είναι συλλογές από λογικά συσχετιζόμενα και δομημένα δεδομένα προκειμένου να εξασφαλίζετε η εύκολη προσπέλαση τους, συνήθως ενός γνωστικού αντικείμενου (π.χ. Ιατρική, Ψυχολογία κ.α.). Οι δικτυακές πύλες είναι δικτυακοί τόποι που συγκεντρώνουν στις ιστοσελίδες τους ένα σύνολο από υπηρεσίες, επίκαιρες ειδήσεις και πληροφορίες, μηχανές αναζήτησης και θεματικούς καταλόγους με σκοπό ο απλός χρήστης να ξεκίνα την είσοδο του στο

διαδίκτυο από εκεί. Οι δικτυακές πύλες συχνά προσφέρουν δυνατότητα προσαρμογής του περιβάλλοντος στις προσωπικές ανάγκες του χρήστη και πάρα πολλές υπηρεσίες όπως δωρεάν ηλεκτρονικό ταχυδρομείο, πρόβλεψη καιρού, χάρτες, εφημερεύουσες υπηρεσίες κ.α. Τέτοιες ιστοσελίδες στον ελληνικό χώρο είναι www.in.gr και www.pathfinter.gr. Οι μηχανές αναζήτησης είναι ειδικά προγράμματα που συλλέγουν πληροφορίες από ιστοσελίδες και τις αποθηκεύουν σε μεγάλες βάσεις δεδομένων. Από αυτές τις ιστοσελίδες με βάση το τίτλο τους, το πλήρες κείμενο, το μέγεθος τους κ.α. δημιουργείται ένα ευρετήριο, στο οποίο οι χρήστες κάνουν αναζητήσεις με λέξεις κλειδιά και η βάση δεδομένων επιστρέφει τα αποτελέσματα που είναι πιο κοντά στο ερώτημα του χρήστη. Οι μηχανές μετά-αναζήτησης δεν διαθέτουν δική τους βάση δεδομένων και δικό τους κατάλογο ή ευρετήριο. Αντίθετα, λειτουργούν αναζητώντας ταυτόχρονα στις βάσεις δεδομένων άλλων μηχανών αναζήτησης και επιστρέφοντας όλα τα αποτελέσματα αφού πρώτα ταξινομηθούν κατάλληλα στους χρήστες. Δίνει με άλλα λόγια τη δυνατότητα στο χρηστή με ένα ερώτημα να ψάχνει σε περισσότερες από μια μηχανές αναζήτησης.

Το διαδίκτυο διαθέτει κάποια χαρακτηρίστηκα που δυσκολεύουν την αναζήτηση και την ανάκτηση της επιθυμητής πληροφορίας:

- Δυναμική αλλαγή. Το διαδίκτυο αλλάζει καθημερινά ενώ τα κλασσικά συστήματα ανάκτησης πληροφορίας είναι σχεδιασμένα για στατικές βάσεις δεδομένων.
- Όγκος πληροφοριών. Η ποσότητα της πληροφορίας στο διαδίκτυο αυξάνεται συνεχώς και είναι δύσκολο να προσδιοριστεί γεγονός που δυσκολεύει την αναζήτηση πληροφοριών.
- Ανομοιογένεια δεδομένων. Το διαδίκτυο διαθέτει πολλούς διαφορετικούς τύπους πληροφορίας όπως κείμενο, φωτογραφίες, αρχεία ήχου, βίντεο και άλλα.
- Πληθώρα γλωσσών. Στο διαδίκτυο χρησιμοποιούνται πολλές διαφορετικές γλώσσες, κάτι που καθίστα την αναζήτηση πληροφοριών πολυπλοκότερη.
- Διπλοτυπίες. Μεγάλο ποσοστό σελίδων στο διαδίκτυο υπάρχει περισσότερες από μια φορά κάτι που δυσκολεύει την αναζήτηση.
- Υψηλή συνδετικότητα. Σχεδόν κάθε σελίδα στο παγκόσμιο ιστό διαθέτει πολλές συνδέσεις προς άλλες σελίδες.

- Μη σωστά διαμορφωμένα ερωτήματα. Οι χρήστες συχνά δίνουν λάθος ερωτήματα στα συστήματα αναζήτησης πληροφοριών και ως συνέπεια δεν παίρνουν τις επιθυμητές πληροφορίες.
- Πληθώρα χρηστών. Διαφορετικοί χρήστες συνεπάγεται διαφορετικές ανάγκες, προσδοκίες και γνώσεις.
- Συγκεκριμένη συμπεριφορά. Έχει υπολογιστεί ότι το 85% των χρηστών του διαδικτύου, κοιτούν μονό την πρώτη σελίδα των επιστρεφόμενων αποτελεσμάτων μιας μηχανής αναζήτησης και το 78% δεν τροποποιούν την διατύπωση του αρχικού ερωτήματος.

Τα πιο πάνω χαρακτηριστικά δυσκολεύουν την αναζήτηση και την ανάκτηση της επιθυμητής πληροφορίας από το διαδίκτυο. Οι υπηρεσίες αναζήτησης πρέπει να βρουν αποτελεσματικούς τρόπους αντιμετώπισης αυτών των χαρακτηριστικών ούτως ώστε να βελτιώσουν τις υπηρεσίες που παρέχουν.

2.2 Ιστορική Εξέλιξη

Η αρχειοθέτηση πληροφοριών χρονολογείται από το 3000 π.Χ. Ο άνθρωπος από την αρχή της ύπαρξης του αντιλήφθηκε την σημασία της αρχειοθέτησης και αναζήτησης της πληροφορίας. Με την πάροδο των αιώνων και ιδιαίτερα όταν ανακαλύφθηκε το χαρτί και ο γραπτός λόγος, η ανάγκη για αποθήκευση και ανάκτηση πληροφοριών γινόταν ακόμα μεγαλύτερη. Όταν ανακαλύφθηκε ο υπολογιστής, οι άνθρωποι κατάλαβαν ότι μπορούσαν να τους χρησιμοποιήσουν για την αποθήκευση μεγάλου όγκου δεδομένων. Το 1945 ο Vannevar Bush δημοσίευσε ένα άρθρο με τίτλο «As we may think» [4], δηλαδή όσο μπορούμε να σκεφτόμαστε, το οποίο ήταν η αρχή για την ανάπτυξη της ιδέας της αυτόματης πρόσβασης σε μεγάλες ποσότητες αποθηκευμένης πληροφορίας. Αυτή η ιδέα, αποτέλεσε στη δεκαετία του ’50, αντικείμενο έρευνας για το πώς θα αναζητούνται αρχεία κειμένου με αυτόματο τρόπο. Οι περισσότερες έρευνες εμπνευστήκαν από την βασική ιδέα της αναζήτησης πληροφοριών με χρήση του υπολογιστή.

Στην επομένη δεκαετία και συγκεκριμένα το Δεκέμβριο του 1969 έκανε την πρώτη του εμφάνιση το διαδίκτυο, το οποίο αποτέλεσε ένα από τα πιο ραγδαία

αναπτυσσόμενα φαινόμενα και έγινε αποδεκτό από όλους. Τις επόμενες δυο δεκαετίες αναπτύχτηκαν διαφορετικά μοντέλα για την ανάκτηση πληροφοριών, τα όποια βοήθησαν σε όλες τις κατευθύνσεις τη διαδικασία της ανάκτησης πληροφοριών. Τα μοντέλα που δημιουργήθηκαν, επιδείκνυαν πειραματικά την αποτελεσματικότητα τους σε συλλογές από μερικές χιλιάδες άρθρα. Παρόλα αυτά λόγω της απουσίας συλλόγων από μεγάλα κείμενα, το ερώτημα για το αν τα μοντέλα αυτά ήταν ικανά για μεγάλου όγκου κείμενα, έμεινε αναπάντητο. Η απάντηση ήρθε το 1992 με την δημιουργία ενός οργανισμού με όνομα Text Retrieval Conference (TREC) [5]. Στόχος του οργανισμού ήταν να ενθαρρύνει την ερεύνα στο πεδίο ανάκτησης της πληροφορίας σε συλλογές από μεγάλα κείμενα.

Όταν πρώτο ξεκίνησε το διαδίκτυο η εύρεση συγκεκριμένων αρχείων ήταν δύσκολη αφού έπρεπε να γνωρίζεις την ακριβή διεύθυνση που βρίσκονταν. Αυτή η διαδικασία εύρεσης των αρχείων ήταν πολύ δύσκολη, χρονοβόρα και έπρεπε να έχεις αρκετή υπομονή. Αυτό γινόταν πριν ο μαθητής, Alan Emtage 1990, δημιουργήσει το πρώτο εργαλείο αναζήτησης (Search Engine Tool). Αυτό που έφτιαξε ήταν ένας κατάλογος από τα αρχεία που υπήρχαν στο διαδίκτυο και ονομάστηκε Archie (Archieve). Το 1991 ένας άλλος μαθητής ο Mark McCahill συνειδητοποίησε ότι αφού μπορείς να ψάχνεις τα αρχεία στο internet τότε μπορείς να ψάχνεις και τα κείμενα. Έτσι δημιούργησε το Gopher,[6] ένα πρόγραμμα που κατηγοριοποιούσε το απλό κείμενο των αρχείων που αργότερα έγιναν οι πρώτες ιστοσελίδες. Η πρώτη μηχανή αναζήτησης (Search Engine) με την μορφή που γνωρίζουμε τις μηχανές σήμερα, δημιουργήθηκε το 1993 από τον Mathew Gray, ο όποιος σήμερα εργάζεται στην Google, και ονομάστηκε Wandex.[7] Ήταν το πρώτο πρόγραμμα που δημιουργούσε και καταλόγους αλλά και έψαχνε τις ιστοσελίδες στο διαδίκτυο. Από το 1993 μέχρι το 1998 δημιουργήθηκαν όλες οι μεγάλες μηχανές αναζήτησης (Search Engines) που γνωρίζουμε και σήμερα.

Excite	1993
Lycos	1994
Yahoo	1994
WebCrawler	1994
Infoseek	1995
AltaVista	1995
Iktomi	1996
Askjevees	1997
Google	1997
Msn Search (Bing)	1998

Πίνακας 2.1: Χρονολογίες πρωτοεμφάνισής των μηχανών αναζήτησης.

Το διαδίκτυο αποτελεί πλέον ένα απαραίτητο εργαλείο στη ζωή των ανθρώπων και για αυτό το λόγο η ανάκτηση πληροφοριών δίνει έμφαση στην αναζήτηση πληροφοριών από το παγκόσμιο ιστό. Οπότε, η συνεχώς αυξανόμενη ποσότητα πληροφοριών που συσσωρεύεται σε αυτό, έχει κάνει την ανάκτηση πληροφοριών από το διαδίκτυο ένα αναγκαίο συστατικό.

2.3 Προβλήματα στην αναζήτηση πληροφορίας από το διαδίκτυο.

Οι μηχανές αναζήτησης κατά την αναζήτηση πληροφορίας από τον παγκόσμιο ιστό, αντιμετωπίζουν μια σειρά προβλημάτων σχετικά με τη διατήρηση άλλα και αύξηση της ποιότητας των υπηρεσιών που προσφέρουν. Κάποια από τα χαρακτηριστικά που αναφέρθηκαν πιο πάνω αποτελούν και αίτιες των προβλημάτων στην αναζήτηση πληροφορίας. Στη συνεχεία γίνεται περιγραφή κάποιων προβλημάτων στην αναζήτηση πληροφορίας από το διαδίκτυο.

2.3.1 Κακόβουλη Πληροφορία (spam)

Το 85.2% των χρηστών που χρησιμοποιούν τις μηχανές αναζήτησης, κοιτούν μονό τα αποτελέσματα της πρώτης σελίδας και το 7,5% και της δεύτερης. [8] Δηλαδή οι χρηστές όταν αναζητούν κάποιες ιστοσελίδες κοιτάζουν ως επί το πλείστο μονό τα πρώτα 10 επιστρεφόμενα αποτελέσματα. Αυτό έχει ως αποτέλεσμα οι ιστοσελίδες που εμφανίζονται στις πρώτες θέσεις των μηχανών αναζήτησης να γίνονται αρκετά δημοφιλείς και να προκαλείται μεγάλη κίνηση σε αυτές.

Στις ιστοσελίδες με εμπορικό περιεχόμενο, οι επισκέψεις σε αυτές όλο και περισσότερων χρηστών μπορεί να αποφέρει μεγάλα κέρδη για την επιχείρηση. Με άλλα λόγια όταν μια ιστοσελίδα με εμπορικό περιεχόμενο βρίσκεται στις πρώτες θέσεις των μηχανών αναζήτησης τότε έχει μεγάλη επισκεψιμότητα, άρα και περισσότερα κέρδη. Αυτό είναι κάτι που ενδιαφέρει αρκετά τις επιχειρήσεις άλλα και τις μηχανές αναζήτησης. Επομένως, αρκετοί ιδιοκτήτες ιστοσελίδων στο διαδίκτυο προσπαθούν σκόπιμα να παραπλανήσουν, ακόμα και με οικονομική βοήθεια τις γνωστότερες μηχανές αναζήτησης του ιστού για να τοποθετηθούν στις πρώτες θέσεις των επιστρεφόμενων αποτελεσμάτων. Από την άλλη πλευρά όμως και οι μηχανές αναζήτησης δύσκολα λένε όχι σε συμφέρουσες προσφορές για πλασάρισμα μιας ιστοσελίδας στις πρώτες θέσεις κατάταξης.

Η διαδικασία κατά την όποια ο ιδιοκτήτης μιας ιστοσελίδας προσπαθεί με χρήση κάποιων τεχνικών να «παραπλανήσει» τις μηχανές αναζήτησης ονομάζεται κακόβουλη πληροφορία των μηχανών αναζήτησης ή αλλιώς spamdexing (γνωστές και ως search engine spam, web spam ή Search Engine Poisoning). [9] Οι τεχνικές

αυτές μπορούν να κατηγοριοποιηθούν σε δυο γενικές κατηγορίες. Στην κακόβουλη πληροφορία περιεχομένου (context spam) και στην κακόβουλη πληροφορία συνδέσμου (link spam).

2.3.1.1 Κακόβουλη πληροφορία περιεχομένου (content spam) [10]

Οι τεχνικές αυτής της κατηγορίας προσπαθούν να μεταβάλουν την λογική άποψη που έχει μια μηχανή αναζήτησης σχετικά με το περιεχόμενο μιας ιστοσελίδας. Στοχεύουν όλες σε παραλλαγές του διανυσματικού χώρου για ανάκτηση πληροφοριών από συλλογές κειμένου. Η κατηγορία αυτή περιλαμβάνει τις παρακάτω τεχνικές:

- **Keyword stuffing (Γέμιση με λέξεις κλειδιά).** Τοποθέτηση συγκεκριμένων λέξεων κλειδιών μέσα σε μια ιστοσελίδα με σκοπό την αύξηση του αριθμού των λέξεων κλειδιών που υπάρχουν σε αυτή. Έτσι η ιστοσελίδα αυτή να μπορεί να βρεθεί πιο εύκολα από κάποιο ανιχνευτή ιστού (web crawler). Οι περισσότερες σύγχρονες μηχανές αναζήτησης έχουν την ικανότητα να αναλύουν μια ιστοσελίδα και να καθορίζουν αν η συχνότητα των λέξεων κλειδιών είναι συνεπής με άλλες ιστοσελίδες που έχουν σχεδιαστεί ειδικά για να προσελκύουν τις μηχανές αναζήτησης. Επίσης οι μεγάλες ιστοσελίδες περικόπτονται για να μην μπορούν μαζικές λίστες με λέξεις κλειδιά να υπάρχουν σε μια μονό ιστοσελίδα.
- **Hidden or invisible unrelated text (Κρυφό ή αόρατο κείμενο).** Απόκρυψη λέξεων-κλειδιά και φράσεων με χρήση ίδιου χρώματος με αυτό του φόντου (background) και χρησιμοποιώντας πολύ μικρό μέγεθος γραμματοσειράς. Οι μηχανές αναζήτησης μπορούν να μπλοκάρουν προσωρινά ή και μόνιμα ιστοσελίδες που περιέχουν αόρατο κείμενο. Ωστόσο το κρυφό κείμενο δεν είναι πάντα κακόβουλη πληροφορία (spamdexing), μπορεί επίσης να χρησιμοποιηθεί για την ενίσχυση της προσβασιμότητας.
- **Meta tag stuffing.** Επανάληψη λέξεων - κλειδιά στα Meta tags και χρησιμοποίηση λέξεων - κλειδιά που δεν σχετίζονται με το περιεχόμενο της σελίδας.

- **"Gateway" or doorway pages.** Είναι χαμηλής ποιότητας ιστοσελίδες που δημιουργούνται με πολύ μικρή περιεκτικότητα, άλλα είναι γεμισμένες με παρόμοιες λέξεις - κλειδιά και φράσεις. Είναι σχεδιασμένες να τοποθετούνται ψηλά στις μηχανές αναζήτησης, αλλά δεν εξυπηρετούν καθόλου τους επισκέπτες που αναζητούν πληροφορίες. Μια doorway page γενικά θα έχει ένα «κάντε κλικ για να μπείτε στην ιστοσελίδα».
- **Scraper sites.** Είναι επίσης γνωστά και ως «Made for Ad Sense sites» και αντιγράφουν τα περιεχόμενα άλλων ιστοσελίδων από το διαδίκτυο συχνά χωρίς αδεία. Τα περιεχόμενα των scraper sites μπορεί να είναι μοναδικά αλλά είναι απλώς μια συγχώνευση των δεδομένων που λαμβάνονται από άλλες πήγες. Ο σκοπός δημιουργίας τέτοιων σελίδων είναι η είσπραξη εσόδων από διαφημίσεις ή η «παραπλάνηση» των μηχανών αναζήτησης μέσω συνδέσεων με άλλες σελίδες με σκοπό να βελτιώσουν την κατάταξη τους. Η εμφάνιση τέτοιων σελίδων έχει πολλαπλασιαστεί τα τελευταία χρόνια.
- **Article spinning (Νηματοποίηση άρθρων).** Περιλαμβάνει την επανασυγγραφή υπάρχοντων άρθρων, σε αντίθεση με την απλή αντιγραφή περιεχομένου από άλλες σελίδες, για να αποφύγουν κυρώσεις που επιβάλλονται από τις μηχανές αναζήτησης για διπλό περιεχόμενο. Αυτή η διαδικασία γίνεται είτε με μίσθωση συγγραφέων είτε με αυτοματοποιημένο τρόπο, χρησιμοποιώντας μια βάση δεδομένων ή ένα νευρωνικό δίκτυο.

2.3.1.2 Κακόβουλη πληροφορία συνδέσμου (link spam) [9]

Κακόβουλη πληροφορία συνδέσμου ορίζεται ως συνδέσεις μεταξύ σελίδων που υπάρχουν για άλλους λόγους περά από την αξία τους. Τα link spam εκμεταλλεύονται τους αλγορίθμους αναζήτησης που ταξινομούν με βάση τις συνδέσεις μεταξύ των ιστοσελίδων (π.χ. Page Rank της Google) [11] και έχει ως αποτέλεσμα να δίνει σε αυτές τις σελίδες υψηλότερες βαθμολογίες από ότι αξίζουν. Οι τεχνικές αυτές αποσκοπούν επίσης στο να επηρεάσουν και άλλα συστήματα κατάταξης βασιζόμενα στις συνδέσεις όπως ο αλγόριθμος HITS. [12] Η κατηγορία αυτή περιλαμβάνει τις πιο κάτω τεχνικές.

- ***Link – building software*** - Χρήση λογισμικού που δημιουργεί συνδέσεις για αυτοματοποίηση της διαδικασίας βελτιστοποίησης της κατάταξης στις μηχανές αναζήτησης.
- ***Link farms*** - Είναι στενά συνδεδεμένες κοινότητες ιστοσελίδων που αναφέρεται η μια στην άλλη. Είναι γνώστες χιουμοριστικά και ως αμοιβαίες εταιρίες θαυμασμού.(mutual admiration societies).
- ***Hidden links*** - Τοποθέτηση υπερσυνδέσεων σε σημεία όπου οι επισκέπτες δεν μπορούν να τα δουν για να αυξήσουν την δημοτικότητα των συνδέσεων.
- ***Sybil attack*** - Δημιουργία πολλαπλών ιστοσελίδων με διαφορετικά ονόματα που όλα συνδέονται μεταξύ τους. Μια επίθεση «Sybil» είναι η «σφυρηλάτηση» πολλαπλών ταυτοτήτων για κακόβουλες προθέσεις και πήρε το όνομα του από τον διάσημο ασθενή με πολλαπλές διαταραχές προσωπικότητας «Sybil» (Shirley Ardell Mason).
- ***Spam blogs*** - Είναι ψεύτικα blocks που δημιουργούνται αποκλειστικά για εμπορική προώθηση και για πέρασμα συνδέσεων σε τοποθεσίες στόχους. Συχνά αυτά τα blogs σχεδιάζονται με παραπλανητικό τρόπο έτσι ώστε να δίνουν την εντύπωση μιας νόμιμης ιστοσελίδας, αλλά μετά από στενή επιθεώρηση τους, συχνά χρησιμοποιείται νηματοποιημένο λογισμικό (spinning software) ή πολύ κακή σύνταξη για να διαβαστεί το περιεχόμενο. Είναι παρόμοιας φύσης με τα Link farms.
- ***Page hijacking*** - Δημιουργία πειρατικών σελίδων που μοιάζουν πάρα πολύ με άλλες διάσημες ιστοσελίδες. Έχουν περιεχόμενο παρόμοιο με αυτό τον αυθεντικών ιστοσελίδων άλλα ανακατευθύνουν τους χρηστές σε κακόβουλες και μη σχετικές ιστοσελίδες.
- ***Buying expired domains*** - Μερικοί spammers παρακολουθούν τα ονόματα χώρου (domain names) τα οποία θα λήξουν σύντομα, στη συνεχεία τα αγοράζουν και τα αντικαθιστούν με σελίδες που ανακατευθύνουν σε άλλες κακόβουλες σελίδες. Η Google για να αντιμετωπίσει τέτοιες ενέργειες επαναφέρει τα δεδομένα για ονόματα χώρου που έχουν λήξει.
- ***Cookie stuffing (ή cookie dropping)*** - [13] Είναι μια απευθείας σύνδεση (online) τεχνική μάρκετινγκ η οποία χρησιμοποιείται για την παραγωγή παράνομων πωλήσεων. Συμβαίνει όταν ένας χρήστης επισκέπτεται μια

ιστοσελίδα και ως αποτέλεσμα αυτής της επίσκεψης λαμβάνει ένα τρίτο cookie (θυγατρική ιστοσελίδα στόχο) από μια εντελώς διαφορετική ιστοσελίδα, χωρίς ο χρήστης να το γνωρίζει αυτό. Εάν ο χρήστης επισκεφτεί την ιστοσελίδα στόχο και συμπληρώσει μια συναλλαγή τότε καταβάλετε στο cookie stuffer προμήθεια.

- ***Using world-writable pages*** - Είναι σελίδες του διαδικτύου που μπορούν να επεξεργαστούν οι χρηστές και χρησιμοποιούνται από spamdexers για να εισάγουν συνδέσεις σε τοποθεσίες ανεπιθύμητης αλληλογραφίας.
- ***Spam in blocks*** - Είναι η τοποθέτηση τυχαίων συνδέσμων σε άλλους δικτυακούς τόπους όπως βιβλία επισκεπτών (Guest books), φόρουμ, προσωπικά ιστολόγια και σε οποιαδήποτε ιστοσελίδα η οποία δέχεται σχόλια και γενικότερα εισαγωγή κειμένου από τους επισκέπτες.
- ***Comment spam*** - Είναι μια μορφή ανεπιθύμητης αλληλογραφίας που έχει προκύψει σε ιστοσελίδες που επιτρέπουν δυναμική επεξεργασία τους από χρήστες όπως wikis, blogs και βιβλία επισκεπτών. Οι spammers μέσω ειδικών προγραμμάτων βρίσκουν τέτοιες ιστοσελίδες και προσθέτουν σε αυτές κακόβουλους συνδέσμους.
- ***Wiki spam*** - Η χρησιμοποίηση της ανοικτής ηλεκτρονικής βιβλιοθήκης «Wikipedia» για εισαγωγή ανεπιθύμητων συνδέσμων προς κακόβουλες σελίδες.
- ***Referrer log spamming*** - Η τροποποίηση των συνδέσεων (logs) αναφοράς πολλών ιστοσελίδων έτσι ώστε να δείχνουν σε μια συγκεκριμένη ιστοσελίδα.

Κάποιες άλλες τεχνικές spamdexing που δεν έχουν κατηγοριοποιηθεί είναι:

- ***Mirror Websites*** - Μια ιστοσελίδα καθρέφτης είναι η φιλοξενία πολλαπλών ιστοσελίδων με εννοιολογικά παρόμοιο περιεχόμενο αλλά με διαφορετικές διευθύνσεις URL.
- ***URL redirection*** - Είναι η μεταφορά ενός χρήστη σε κάποια άλλη ιστοσελίδα χωρίς την παρέμβαση του.
- ***Cloaking*** - Η τεχνική «cloaking» (μανδύας) αναφέρεται σε οποιανδήποτε μέσα χρησιμοποιούνται για παραπλάνηση μιας μηχανής αναζήτησης σχετικά με το περιεχόμενο μιας ιστοσελίδας. Ωστόσο, μπορεί επίσης να χρησιμοποιηθεί για αύξηση της προσβασιμότητας σε χρήστες με αναπηρίες ή

να παρέχει στους χρήστες περιεχόμενο που οι μηχανές αναζήτησης δεν είναι σε θέση να αναλύσουν ή να επεξεργαστούν. Η ίδια η Google χρησιμοποιεί το «IP delivery», μια μορφή cloaking, για να παράγει τα αποτελέσματα της.

Η κακόβουλη πληροφορία είναι πολύ διαδεδομένη στο χώρο του διαδικτύου γι αυτό και οι μηχανές αναζήτησης πρέπει να συνεχίσουν να λαμβάνουν μέτρα έτσι ώστε να αναγνωρίζουν και να αγνοούν τις κακόβουλες πληροφορίες, βελτιώνοντας παράλληλα την ποιότητα των υπηρεσιών που προσφέρουν.

2.3.2 Ποιότητα περιεχομένου (Content Quality)

Το πρόβλημα της ποιότητας του περιεχομένου είναι κρίσιμο και επηρεάζει την ανάπτυξη και χρήση του διαδικτύου. Το μέγεθος τους προβλήματος της ποιότητας ανάλογα με το πεδίο διαφέρουν. Για παράδειγμα στο ηλεκτρονικό εμπόριο, η ποιότητα είναι συνώνυμη με την αξιοπιστία και ασφάλεια των συναλλαγών. Στην εκπαίδευση με την ακρίβεια και τη πιστότητα στη δομή και στο νόημα του περιεχομένου. Για το εκπαιδευτικό υλικό, η ποιότητα περιεχομένου άπτεται και της σημασιολογίας του περιεχομένου (semantic quality) και της μορφής παρουσίασης (syntactic quality). Ο ποιοτικός έλεγχος της ποιότητας του περιεχομένου κάθε σελίδας του παγκόσμιου ιστού είναι εξαιρετικά δύσκολο να εφαρμοστεί λόγω της φύσης του ίδιου του διαδικτύου που είναι ανοικτό περιβάλλον χωρίς κεντρικούς ρυθμιστικούς μηχανισμούς ελέγχου.

Ο παγκόσμιος ιστός δημιουργήθηκε στηριζόμενος στην αρχή της ισότητας όπου κάθε δικτυακός τόπος μπορεί να συνδεθεί με έναν οποιονδήποτε άλλο και κάθε ιστοσελίδα αντιμετωπίζεται με βάση την αρχιτεκτονική του συστήματος ως ισότιμη με οποιανδήποτε άλλη. Γι αυτούς του λόγους οποιοσδήποτε άνθρωπος, οποιουδήποτε μορφωτικού επίπεδου μπορεί να ανεβάσει πληροφορία στο διαδίκτυο με συνέπεια ο παγκόσμιος ιστός να γεμίσει από χαμηλής ποιότητας, αναξιόπιστο και πολλές φορές αντικρουόμενο περιεχόμενο. Συνεπώς ο έλεγχος της ποιότητας είναι αναγκαίος για την επιβίωση του διαδικτύου ιδιαίτερα ως εκπαιδευτικό εργαλείο.

Για τους πιο πάνω λόγους, ο ποιοτικός έλεγχος του περιεχομένου είναι αναγκαίος στο τρόπο που οι μηχανές αναζήτησης βαθμολογούν και παρουσιάζουν τα αποτελέσματα στο χρήστη. Οι διάφορες μηχανές αναζήτησης θα μπορούσαν να

χρησιμοποιήσουν ένα δείκτη ποιότητας περιεχομένου για κάθε ιστοσελίδα που να συνυπολογίζεται με τα υπόλοιπα κριτήρια για την κατάταξη μιας ιστοσελίδας. Αυτός ο δείκτης θα μπορούσε να προκύπτει από διαφορά στοιχειά όπως ο βαθμός αναγνωσιμότητας της ιστοσελίδας ή η αξιολόγηση της από τους χρήστες. Οι εμπορικές μηχανές αναζήτησης έχουν το πλεονέκτημα ότι μπορούν να αξιολογήσουν τα αποτελέσματα τους από τους ίδιους τους χρήστες και να εξάγουν έτσι χρήσιμα συμπεράσματα για την ποιότητα περιεχομένου κάθε ιστοσελίδας.

Οι μηχανές αναζήτησης στην προσπάθεια τους να βελτιώσουν ακόμη περισσότερο την ποιότητα των υπηρεσιών που προσφέρουν στους χρήστες, πρέπει να εκμεταλλευτούν όλους τους δυνατούς παράγοντες για ποιοτικό έλεγχο του περιεχομένου.

2.3.3 Κανόνες διαδικτύου (Web Conventions)

Στο χώρο του διαδικτύου γενικώς δεν υπάρχουν κανόνες οι όποιοι πρέπει να τηρούνται στην κατασκευή μιας ιστοσελίδας. Όπως προαναφέραμε, το διαδίκτυο δημιουργήθηκε στηριζόμενο στην αρχή της ισότητας οπόταν οποιοσδήποτε άνθρωπος μπορεί να κατασκευάσει και να ανεβάσει μια ιστοσελίδα όπως ο ίδιος επιθυμεί και χωρίς να τον περιορίζουν κάποιοι κανόνες. Παρόλα αυτά, οι περισσότεροι δημιουργοί ιστοσελίδων ακολουθούν κάποιους απλούς κανόνες-συμβάσεις χωρίς να είναι υποχρεωτικό από κάποιον. Δηλαδή πολλές σελίδες του διαδικτύου έχουν κάποια κοινά σημεία τα όποια θα αναφέρουμε ως κανόνες του διαδικτύου. Μερικοί από αυτούς τους «κανόνες» είναι:

1. Το λογότυπο (logo) κάθε ιστοσελίδας συνήθως τοποθετείται πάνω αριστερά Αυτό συμβαίνει για τον λόγο ότι ο χρήστης με το που επισκεφτεί μια ιστοσελίδα το πρώτο πράγμα που κοιτάζει είναι αν βρίσκετε στην ιστοσελίδα που ψάχνει, κάτι που φαίνεται από το λογότυπο.
2. Οι σημαντικότερες πληροφορίες πρέπει να τοποθετούνται στο πιο εμφανές σημείο της ιστοσελίδας, δηλαδή κάτω από το λογότυπο, κοντά στη κορυφή της ιστοσελίδας. Αυτό συμβαίνει διότι οι περισσότεροι χρήστες δεν έχουν την υπομονή να κοιτάξουν ολόκληρο το περιεχόμενο μιας ιστοσελίδας αλλά να «σαρώνουν» την ιστοσελίδα στα γρήγορα.

3. Οι υπογραμμισμένες λέξεις είναι συνήθως σύνδεσμοι. Επιπρόσθετα, σε ένα κείμενο, μια λέξη ή μια πρόταση που είναι σε διαφορετικό χρώμα συνήθως υποδηλώνει ότι είναι ένας σύνδεσμος προς μια άλλη σελίδα. Επομένως, οι δημιουργοί ιστοσελίδων πρέπει να χρησιμοποιούν τις υπογραμμισμένες λέξεις ή προτάσεις ή να τις τονίζουν με διαφορετικό χρώμα μόνο όταν πρόκειται για συνδέσμους, βοηθώντας έτσι τους χρήστες κατανοούν καλύτερα το περιεχόμενο κάθε ιστοσελίδας.
4. Σύνδεσμος προς την αρχική σελίδα. Τοποθετείτε σε εμφανές σημείο και συνήθως πάνω αριστερά, κάτω από το λογότυπο ή και στο ίδιο το λογότυπο Ένας τέτοιος σύνδεσμος είναι πολύ βοηθητικός στην πλοήγηση ενός χρήστη σε μία ιστοσελίδα.
5. Όταν μια ιστοσελίδα αποτελείται από πολλές σελίδες καλό είναι να έχουν την ίδια μορφή και εμφάνιση. Αυτό διευκολύνει και δεν μπερδεύει τους χρήστες καθώς μεταφέρονται από μια σελίδα σε άλλη.
6. Το μενού μιας ιστοσελίδας συνήθως βρίσκεται σε μια οριζόντια γραμμή στην κορυφή της σελίδας και κάτω από το λογότυπο ή σε μια κάθετη στήλη δεξιά ή αριστερά.

Οι μηχανές αναζήτησης με σκοπό την περεταίρω βελτίωση της ποιότητας των αποτελεσμάτων που προσφέρουν βασίζονται σε αυτούς τους κανόνες. Έτσι, οι δημιουργοί ιστοσελίδων όταν αθετούν αυτούς τους κανόνες, επηρεάζουν και τις μηχανές αναζήτησης. Το κύριο πρόβλημα είναι να αναγνωριστούν κάποιοι βασικοί κανόνες ούτως ώστε να υποβοηθούνται τόσο οι χρήστες κατά την πλοήγηση τους σε μια ιστοσελίδα, όσο και οι μηχανές αναζήτησης κατά την προσπάθεια τους να βρουν και να παρουσιάσουν ποιοτική πληροφορία στους χρήστες.

Εικόνα 2.1: Παράδειγμα εφαρμογής κανόνων διαδικτύου

2.3.4 Διπλότυποι Κόμβοι (Duplicate Hosts)

Ένα πολύ μεγάλο πρόβλημα στην αναζήτηση πληροφορίας στο διαδίκτυο είναι οι διπλότυποι κόμβοι, δηλαδή οι πανομοιότυπες ιστοσελίδες. Οι μηχανές αναζήτησης κατά την «σάρωση» του διαδικτύου για να εντοπίσουν ιστοσελίδες προσπαθούν να αποφύγουν την αποθήκευση στις βάσεις τους τέτοιων ιστοσελίδων, καθώς δεν προστίθεται νέα πληροφορία στα αποτελέσματα αλλά αντίθετα προκαλείται μεγαλύτερη ακαταστασία στα αποτελέσματα.

Οι περισσότερες μηχανές αναζήτησης ελέγχουν για διπλότυπες εγγραφές κατά την παράγωγη των αποτελεσμάτων και αφαιρούν την διπλή πληροφορία. Το πρόβλημα είναι ότι δεν υπάρχει εγγύηση για το ποια έκδοση της σελίδας θα εμφανιστεί στα αποτελέσματα και ποια δεν θα εμφανιστεί. Ακόμη, ο web crawler (ανιχνευτής ιστού) [14] κάθε μηχανής αναζήτησης μπορεί να σταματήσει την εισαγωγή σελίδων μιας ιστοσελίδας στα αποτελέσματα αν παρατηρήσει ότι υπάρχουν πολλά αντίγραφα ίδιων σελίδων σε διαφορετικά URLs (ενιαίοι εντοπιστές πόρων) [15].

Η αποφυγή διπλότυπων κόμβων από τις μηχανές αναζήτησης παρέχουν τα εξής πλεονεκτήματα στις μηχανές αναζήτησης:

- Αποφεύγεται η εμφάνιση δύο ή και περισσότερων ίδιων σελίδων στα αποτελέσματα, οπόταν βελτιώνεται η ποιότητα υπηρεσιών που παρέχει η μηχανή αναζήτησης.
- Εξοικονομούνται πόροι από την διαδικασία εισαγωγής στη βάση δεδομένων της μηχανής αναζήτησης σελίδων με πανομοιότυπο περιεχόμενο.

Ένα παράδειγμα όπου οι μηχανές αναζήτησης αντιλαμβάνονται την ύπαρξη πανομοιότυπου περιεχομένου είναι σε περιπτώσεις όπου κάποιες ιστοσελίδες πωλούν ένα προϊόν και για την περιγραφή του χρησιμοποιούν την περιγραφή από την ιστοσελίδα της εταιρίας κατασκευής τους προϊόντος. Ένα άλλο παράδειγμα είναι αυτό που πολλά διαφορετικά URLs οδηγούν στην ίδια ιστοσελίδα. Για παράδειγμα τα πιο κάτω URLs μπορεί να καταλήγουν στην ίδια ιστοσελίδα:

http://www.example.com

https://www.example.com

http://www.example.com/index.htm

https://www.example.com/index.htm

http://example.com

https://example.com

http://example.com/index.htm

https://example.com/index.htm

Σε αυτές τις περιπτώσεις, όπως και σε κάποιες άλλες [16], οι περισσότερες μηχανές αναζήτησης καταλαβαίνουν την ύπαρξη πανομοιότυπων ιστοσελίδων και τις αγνοούν.

2.3.5 Ασαφής καθορισμός δεδομένων (Vaguely-Structured Data)

Η δομή των πληροφοριών που υπάρχουν στο διαδίκτυο είναι από τα βασικότερα χαρακτηριστικά που επηρεάζουν τις τεχνικές που χρησιμοποιούνται από τις μηχανές αναζήτησης για ανάκτηση πληροφορίας.

Αρχικά οι ερευνητές των βάσεων δεδομένων ασχολούνταν με δεδομένα τα οποία ήταν δομημένα σε υψηλό επίπεδο, ενώ οι ερευνητές της επιστήμης ανάκτησης πληροφορίας μελετούσαν αδόμητες πληροφορίες. Πλέον, οι ερευνητές των βάσεων ελέγχουν αν η δομή είναι με βάση κάποια προκαθορισμένα κριτήρια, ενώ και οι αντίστοιχοι της επιστήμης ανάκτησης πληροφοριών άρχισαν να χρησιμοποιούν τη μετά – πληροφορία. (meta-data)[17]

Οι ιστοσελίδες οι οποίες είναι γραμμένες σε HTML [18] είναι κοντά τόσο στο ελεύθερο κείμενο όσο και στο καλά δομημένο. Η HTML δίνει περισσότερο έμφαση στην παρουσίαση της πληροφορίας παρά στη δομή της. Παρέχει ωστόσο και κάποια στοιχειά για την σημασιολογική πληροφορία. Οι μηχανές αναζήτησης ενδιαφέρονται περισσότερο για τη δομή παρά την εμφάνιση και για αυτό εκμεταλλεύονται αυτά τα στοιχεία για την καλύτερη παραγωγή αποτελεσμάτων.

Σε γλώσσες όπου δίνεται περισσότερη έμφαση στην δομή παρά στην εμφάνιση όπως για παράδειγμα η XML [19] τα αναμενόμενα κέρδη για τις μηχανές αναζήτησης είναι πολύ περισσότερα από γλώσσες τύπου HTML, όπως:

- Ταχύτερη εστιασμένη αναζήτηση.
- Μικρότερη συμφόρηση δικτύου.
- Οι μηχανές αναζήτησης αναζητούν συγκεκριμένες ετικέτες (tags) στον XML κώδικα κερδίζοντας σε ταχύτητα και ακρίβεια.

Πλέον η XML είναι μια γλώσσα που εξελίσσεται σε ένα παγκόσμιο πρότυπο για ανταλλαγή δεδομένων και θα βοηθήσει στο μέγιστο την αναβάθμιση της ποιότητας των υπηρεσιών που προσφέρουν οι μηχανές αναζήτησης.

2.4 Βελτίωση της απόδοσης των υπηρεσιών αναζήτησης.

Υπάρχουν τρεις κύριες κατευθύνσεις ως προς την βελτίωση της απόδοσης της αναζήτησης:

1. Βελτίωση της διεπαφής χρήστη. (user interface).
2. Καλύτερη επεξεργασία των αποτελεσμάτων.
3. Κατανόηση των αλγορίθμων ανάκτησης πληροφορίας.

Η πρώτη κατεύθυνση ασχολείται με το πρόβλημα της επικοινωνίας του χρήστη με τη διεπαφή που καλείται κατά την διαδικασία υποβολής του ερωτήματος. Το φιλικό περιβάλλον και η ταχύτητα επεξεργασίας και παρουσίασης των αποτελεσμάτων είναι από τα σημαντικότερα χαρακτηριστικά μιας μηχανής αναζήτησης. Οι αρχάριοι χρήστες πρέπει να είναι σε θέση να αλληλεπιδράσουν και να εξοικειωθούν γρήγορα με ένα σύστημα αναζήτησης. Σε μερικές μηχανές αναζήτησης ακόμα και έμπειροι χρήστες δεν μπορούν να εξοικειωθούν εύκολα με το τρόπο που λειτουργούν, αφού σε πολλές περιπτώσεις οι υπηρεσίες που παρέχουν είναι ασαφείς και μπερδεμένες.

Η δεύτερη κατεύθυνση ασχολείται με την περεταίρω επεξεργασία των αποτελεσμάτων. Τα αποτελέσματα που επιστρέφουν οι περισσότερες μηχανές αναζήτησης έχουν μικρό ποσοστό σχετικών πληροφοριών προκαλώντας δυσφορία στους χρήστες για τις υπηρεσίες που προσφέρουν. Οπόταν επιβάλλεται καλύτερη επεξεργασία των αποτελεσμάτων ούτως ώστε να αυξηθεί το ποσοστό των επιστρεφόμενων σχετικών αποτελεσμάτων. Ακόμη, σε πολλές μηχανές αναζήτησης οι πληροφορίες που αποθηκεύονται στις βάσεις τους δεν ενημερώνονται με αποτέλεσμα οι χρήστες να παίρνουν μη ανανεωμένες πληροφορίες.

Τέλος, η τρίτη κατεύθυνση ασχολείται με τη μελέτη των αλγορίθμων συλλογής, σύνταξης και παρουσίασης της πληροφορίας. Η κατανόηση αυτών των αλγορίθμων μπορεί να οδηγήσει σε καλύτερη επεξεργασία των επιστρεφόμενων αποτελεσμάτων επιτυγχάνοντας μεγαλύτερο ποσοστό σχετικών πληροφοριών, καλύτερη ενημέρωση των βάσεων δεδομένων καθώς και αύξηση του ποσοστού κάλυψης της συνολικής πληροφορίας.

2.5 Προηγμένες λειτουργιές αναζήτησης.

Οι μηχανές αναζήτησης πρόσφατα έχουν αναπτύξει κάποιες προηγμένες λειτουργίες αναζήτησης στην προσπάθεια τους να αναβαθμίσουν τις υπηρεσίες που προσφέρουν στους χρήστες. Αυτές οι λειτουργίες υιοθετούνται από όλο και περισσότερες μηχανές αναζήτησης, διότι διευκολύνουν τον χρήστη και αυξάνουν το ποσοστό των επιστρεφόμενων σχετικών αποτελεσμάτων. Αυτές οι λειτουργίες λαμβάνουν υπόψη τις ανάγκες και τα ενδιαφέροντα των χρηστών.

2.5.1 Εξατομίκευση

Καθώς ο παγκόσμιος ιστός αυξάνει σε μέγεθος και μάλιστα με εκθετικό βαθμό, νέες προκλήσεις εγείρονται και για την αναζήτηση πληροφοριών σε αυτόν. Μηχανές αναζήτησης που επιστρέφουν στο χρήστη σελίδες χρησιμοποιώντας ως κριτήριο λέξεις κλειδιά που εισάγονται από αυτόν δεν μπορούν να εξασφαλίσουν την ικανοποίησή του.

Χαρακτηριστικό παράδειγμα είναι η αναζήτηση ενός χρήστη για πληροφορίες σχετικά με το μήλο (apple στα αγγλικά), είναι αναπόφευκτο πως θα του επιστραφούν και αρκετές σελίδες σχετικές με την γνωστή εταιρία υπολογιστών Apple. Σε αυτό παίζει ρόλο και το γεγονός ότι ο παγκόσμιος ιστός παρουσιάζει μεγαλύτερη διασύνδεση και αναφορά από και προς σελίδες σχετικό με αυτό και γενικότερα τους υπολογιστές.

Ακόμη, το κόστος μπορεί να είναι και οικονομικό καθώς η πρόσβαση στο Παγκόσμιο ιστό από κινητά συνήθως χρεώνεται με βάση το μέγεθος πληροφοριών που προσπελαύνει ο χρήστης και επομένως απαιτείται η αποφυγή εμφάνισης μη χρήσιμων σελίδων και πολυπληθών κατηγοριών μέσα από τις οποίες θα πρέπει να διαλέξει αυτό που ανταποκρίνεται στην αναζήτησή του. Συνεπώς, ήταν απαραίτητη η ανάπτυξη μεθόδων, ημιαυτόματων και αυτόματων, που θα λαμβάνουν υπόψη την ταυτότητα του συγκεκριμένου χρήστη, τις προτιμήσεις και τις συνήθειές του ώστε να βελτιωθεί η διαδικασία και η ποιότητα αναζήτησης.

Κατά την διαδικασία της εξατομίκευσης επιδιώκεται διαφορετικό αποτέλεσμα ανάλογα με τον εκάστοτε χρήστη που έθεσε το ερώτημα. Λαμβάνονται υπόψη

στοιχεία από προηγούμενες αναζήτησης του και από τυχών προτιμήσεις του. Για παράδειγμα ένας χρήστης που δεν έχει αναζητήσει στο παρελθόν για οχήματα αλλά ενδιαφερόταν για εικόνες φυσικών τοπίων, θα ήταν πιθανότερο σε μια αναζήτηση με τη λέξη «jaguar» να τον ενδιαφέρει το γνωστό αιλουροειδές παρά η αντίστοιχη μάρκα αυτοκίνητου. Αυτός είναι και ένα τρόπος αντιμετώπισης του φαινόμενου της πολυσημίας, δηλαδή λέξεις κλειδιά που μπορούν να οδηγήσουν σε εντελώς ασύνδετα αποτελέσματα.

Εξατομίκευση ορίζεται [20] κάθε ενέργεια που προσαρμόζει την πληροφορία ή τις υπηρεσίες που παρέχονται από έναν ιστότοπο στις ανάγκες ενός συγκεκριμένου χρήστη ή συνόλου χρηστών, αξιοποιώντας τη γνώση που αποκτήθηκε μέσω της συμπεριφοράς - επιλογών πλοήγησης του χρήστη και τα ιδιαίτερα ενδιαφέροντά του, σε συνδυασμό με το περιεχόμενο και τη δομή του ιστοτόπου.

Η εξατομίκευση χωρίζεται σε 4 κατηγορίες:

- ***Χρήση μνημονικού*** - Πληροφορίες του χρήστη όπως το όνομα και το ιστορικό πλοήγησης αποθηκεύονται (π.χ. με χρήση cookies), για να χρησιμοποιηθούν αργότερα για την αναγνώρισή του. Ως μειονέκτημα είναι η παραβίαση του δικαιώματος του χρήστη στην ιδιωτικότητα.
- ***Παραμετροποίηση*** – Λαμβάνεται ως είσοδος το σύνολο των προτιμήσεων του χρήστη μέσω των σελίδων εγγραφής ώστε να παραμετροποιηθεί κατάλληλα το περιεχόμενο και η δομή μιας ιστοσελίδας.
- ***Συστήματα καθοδήγησης ή παροχής προτεινόμενων λύσεων*** - Είναι συστήματα τα οποία επιδιώκουν να προτείνουν αυτόματα υπερσυνδέσμους που δείχνουν να είναι σχετικοί με τα ενδιαφέροντα του χρήστη ώστε να παρέχουν καλύτερη πρόσβαση στην πληροφορία μέσα σε ένα μεγάλο ιστότοπο. [21, 22]
- ***Υποστήριξη συγκεκριμένων δραστηριοτήτων*** – Αυτά τα είδους συστήματα υλοποιούνται από τη πλευρά του χρήστη. Ένας προσωπικός βοηθός εκτελεί ενέργειες εκ μέρους του χρήστη ώστε να πετύχει πρόσβαση σε χρήσιμη πληροφορία. Αυτή η προσέγγιση απαιτεί μεγάλη συμμετοχή από τη πλευρά του χρήστη, συμπεριλαμβανομένων της απόκτησης, εγκατάστασης και συντήρησης του λογισμικού προσωπικής βοήθειας.

2.5.2 Κατηγοριοποίηση

Ο αριθμός των μη σχετικών αποτελεσμάτων μπορεί να μειωθεί σημαντικά με την κατηγοριοποίηση των αναζητήσεων. Ο περιορισμός της αναζήτησης σε κάποιο συγκεκριμένο θεματικό προσανατολισμό, βελτιστοποιεί τις λειτουργίες της μηχανής αναζήτησης.

Στο παράδειγμα που προαναφέραμε με την «jaguar» ο περιορισμός της αναζήτησης στο θεματικό προσανατολισμό «μηχανοκίνητα» θα μας επιστρέψει ως επί το πλείστον αποτελέσματα σχετικά με τα αυτοκίνητα της γνωστής εταιρίας «jaguar». Αντίθετα, ο περιορισμός της αναζήτησης στο θεματικό προσανατολισμό «ζώα» θα μας επιστρέψει ως επί το πλείστον αποτελέσματα σχετικά με το γνωστό αιλουροειδές.

Επιπλέον, ο χρήστης μπορεί να επιλέξει πληροφορίες που αφορούν μόνο ένα συγκεκριμένο διάστημα ή ακόμα και να ταξινομήσει τα αποτελέσματα κατά ημερομηνία.

Εύκολα συμπεραίνουμε ότι η κατηγοριοποίηση των αναζητήσεων βοηθά τις μηχανές αναζήτησης να επιστρέφουν σχετικότερα προς τα ενδιαφέροντα του χρήστη αποτελέσματα. Το μειονέκτημα αυτής της λειτουργίας είναι ότι ο χρήστης έκτος από τις λέξεις κλειδιά, πρέπει να εισάγει χειρονακτικά και τον θεματικό προσανατολισμό, τη χρονική περίοδο, τη γεωγραφική περιοχή κ.α. κάτι που απαιτεί περισσότερο χρόνο και εμπειρία από τη πλευρά του χρήστη.

2.5.3 Βοηθητικές πληροφορίες

Τέλος, μια λειτουργία που χρησιμοποιείται όλο και περισσότερο από τις μηχανές αναζήτησης είναι η εμφάνιση βοηθητικών πληροφοριών, πριν ή και κατά τη διάρκεια της αναζήτησης.

Για παράδειγμα στη μηχανή της Google κατά την εισαγωγή του ερωτήματος από τον χρήστη εμφανίζεται λίστα με προτεινόμενα ερωτήματα παρόμοια με το ερώτημα που εισάγει εκείνη την στιγμή ο χρήστης. Επίσης η προτεινόμενη διόρθωση του ερωτήματος είναι μια δυνατότητα που έχει εισαχθεί από πολλές υπηρεσίες αναζήτησης («*Did you mean:* »).

ΚΕΦΑΛΑΙΟ 3 – ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

3.1 Εισαγωγή

Ο παγκόσμιος ιστός είναι σήμερα μια φαινομενικά αστείρευτη πηγή πληροφοριών για όλα τα θέματα που μπορεί να μας απασχολήσουν και πάνω του στηρίζεται η διάδοση της παγκόσμιας γνώσης στο ευρύ κοινό. Το σημαντικότερο εμπόδιο στην εύρεση συγκεκριμένων πληροφοριών που ενδιαφέρουν τον χρηστή είναι το τεράστιο σε πλήθος πληροφοριών που εντοπίζεται στο παγκόσμιο ιστό. Επομένως, μια από τις ανάγκες που δημιούργησε ο εκθετικός ρυθμός αύξησης του παγκοσμίου ιστού είναι η ανάγκη για γρήγορη και με ακρίβεια αναζήτηση πληροφοριών.

Ο πιο ευέλικτος και δημοφιλής τρόπος αναζήτησης πληροφοριών στο παγκόσμιο ιστό είναι μέσω της μηχανής αναζήτησης. Μια μηχανή αναζήτησης ορίζεται ως [23] μια εφαρμογή που επιτρέπει την αναζήτηση κειμένων και αρχείων στο διαδίκτυο. Αποτελείται από ένα πρόγραμμα υπολογιστή που βρίσκεται σε έναν ή περισσότερους υπολογιστές στους οποίους δημιουργεί μια βάση δεδομένων με τις πληροφορίες που συλλέγει από το διαδίκτυο και το διαδραστικό περιβάλλον που εμφανίζεται στον τελικό χρήστη ο οποίος χρησιμοποιεί την εφαρμογή από άλλον υπολογιστή συνδεμένο στο διαδίκτυο. Μια μηχανή αναζήτησης είναι σχεδιασμένη να δέχεται τις ερωτήσεις του χρήστη, κυρίως μέσω λέξεων-κλειδιών και να του επιστρέφει σχετικές ιστοσελίδες και δεδομένα από το παγκόσμιο ιστό.

3.2 Ιστορική αναδρομή

Το πρώτο εργαλείο που χρησιμοποιήθηκε για την έρευνα στο παγκόσμιο ιστό όπως προαναφέραμε (κεφ. 2.2) ονομαζόταν «Archie» και δημιουργήθηκε το 1990 από τον Alan Emtage, ο όποιος ήταν φοιτητής στο πανεπιστήμιο McGill στο Μόντρεαλ.[24] Το πρόγραμμα που δημιούργησε, «κατέβαζε» τις λίστες καταλόγου όλων των αρχείων που βρισκόντουσαν στις δημόσιες ανώνυμες περιοχές FTP [25] δημιουργώντας μια εξερευνήσιμη βάση δεδομένων με βάση τα ονόματα των αρχείων. Ενώ το εργαλείο Archie αποτελούσε ένα ευρετήριο των αρχείων ενός υπολογιστή, ο Mark McCahill του πανεπιστημίου της Μινεσότας, δημιούργησε το 1991 το «Gopher», για την καταχώρηση συντεταγμένων αρχείων κειμένου. [26] Ακολούθως 2 άλλα προγράμματα το «VERONICA» (Very Easy Rodent-Oriented Net-wide Index Computerized Archives) [27] και το «JUDHEAD» (Jonzy's Universal Gopher Hierarchy Excavation And Display) [28] έψαχναν τα αρχεία που υπήρχαν στον Gopher. Το πρώτο πραγματοποιούσε μια έρευνα λέξεων στους τίτλους από τις λίστες που είχε το Gopher και το δεύτερο ήταν ένα εργαλείο για την προμήθεια πληροφοριών από τους εξυπηρετητές του Gopher.

Τον Ιούνιο του 1993, ο σπουδαστής του MIT, Matthew Gray, δημιούργησε το αποκαλούμενο «World Wide Web wanderer» (περιπλανώμενος στο Διαδίκτυο) [29], το οποίο θεωρείται ο πρώτος Web Crawler. Αρχικά χρησιμοποιήθηκε για την μέτρηση των κεντρικών υπολογιστών του δικτύου και λειτουργούσε μηνιαία από 1993 έως 1995. Χρησιμοποιήθηκε πιο πρόσφατα για τη λήψη URLs, διαμορφώνοντας την πρώτη βάση δεδομένων των ιστοχώρων, αποκαλούμενη ως "Wandex" [30]. Ο web crawler (ή spider) είναι στην ουσία ένα πρόγραμμα – ρομπότ, το οποίο επισκέπτεται κάποιες περιοχές για να κάνει αίτηση των εγγράφων που περιέχουν αυτές οι περιοχές.

Την ίδια χρονιά, ο Martijn Koster δημιούργησε τη δεύτερη χρονικά μηχανή αναζήτησης το "ALIWEB" (Archie-Like Indexing of the Web), το οποίο δεν χρησιμοποιούσε κάποιο Web Crawler αλλά βασιζόταν στους ίδιους τους διαχειριστές των ιστοσελίδων να καταχωρήσουν την σελίδα τους [31]. Αυτό έδινε την δυνατότητα στους διαχειριστές των ιστοσελίδων να καθορίσουν τους όρους που θα οδηγούσαν

τους χρήστες στην ιστοσελίδα τους. Λόγω του ότι λίγοι διαχειριστές υπόβαλλαν τις ιστοσελίδες τους, το ALIWEB δεν χρησιμοποιήθηκε ευρέος.

Η JumpStation [32], ήταν η πρώτη WWW (World Wide Web) μηχανής αναζήτησης που συμπεριφέρθηκε και εμφανίστηκε στο χρηστή με το τρόπο που είναι οι σημερινές μηχανές αναζήτησης. Ξεκίνησε τη δημιουργία του ευρετηρίου της, στις 12 Δεκεμβρίου του 1993 με έδρα της το Πανεπιστήμιο του Stirling της Σκωτίας και δημιουργήθηκε από τον Jonathon Fletcher. Το JumpStation χρησιμοποιούσε τους τίτλους και τις επικεφαλίδες των εγγράφων για την ευρετηριοποίηση των ιστοσελίδων που έβρισκε, χρησιμοποιώντας μια απλή γραμμική αναζήτηση που δεν περιείχε καμία κατάταξη αποτελεσμάτων. Η ανάπτυξη του JumpStation διακόπτεται το 1994, όταν ο Jonathon έφυγε από το πανεπιστήμιο έχοντας αποτύχει να βρει επενδυτές για να στηρίξουν οικονομικά την ιδέα του. Σε εκείνο το σημείο η βάση δεδομένων είχε 275000 εγγραφές που εκτείνονταν σε 1500 εξυπηρετητές.(Servers)

Το 1994 εμφανίζεται η WebCrawler, η όποια ήταν από τις πρώτες μηχανές αναζήτησης που βασίζονταν σε ένα ανιχνευτή ιστού πλήρους κειμένου ("full text" crawler). Σε αντίθεση με τους προκάτοχους της, επέτρεπε στους χρήστες την αναζήτηση οποιασδήποτε λέξης μέσα σε μια ιστοσελίδα και έγινε πρότυπο για όλες τις μεγάλες μηχανές αναζήτησης από τότε και έπειτα. Ήταν επίσης η πρώτη μηχανή αναζήτησης που έγινε ευρέως γνωστή στο κοινό. Την ίδια χρόνια, ξεκίνησε την λειτουργία της η Lycos [33] (η όποια ξεκίνησε στο Carnegie Mellon University) και σύντομα εξελίκτηκε σε μια μεγάλη εμπορική δύναμη.

Λίγο αργότερα, αρκετές μηχανές αναζήτησης εμφανιστήκαν και συναγωνίστηκαν για τη δημοτικότητα. Μηχανές όπως η Magellan [34], Excite [35], Infoseek [36], Inktomi [37], Northern Light [38] και AltaVista [39]. Σε αυτή την περίοδο το Yahoo! [40] ήταν από τους πιο δημοφιλής τρόπους αναζήτησης για τους χρήστες. Ωστόσο, η λειτουργία αναζήτησης του Yahoo! λειτουργούσε στο κατάλογο ιστοσελίδων που διέθετε, αντί σε πλήρες κείμενα αντίγραφα των ιστοσελίδων. Οι χρηστές που αναζητούσαν πληροφορίες μπορούσαν να περιηγηθούν στο κατάλογο, αντί να κάνουν μια αναζήτηση βασιζόμενη σε λέξεις κλειδιά.

Το 1996 η Netscape [41], μια Αμερικανική εταιρία υπηρεσιών πληροφορικής γνωστή για τον ομώνυμο περιηγητή ιστού της (web browser), έψαχνε μία μηχανή αναζήτησης για να τη χρησιμοποιήσει αποκλειστικά στον περιηγητή ιστού της. Το

ενδιαφέρον ήταν τόσο μεγάλο που αντί για μια συμφωνία, η Netscape συμφώνησε με 5 από τις μεγαλύτερες μηχανές αναζήτησης. Για 5 εκατομμύρια ευρώ ετησίως, κάθε μηχανή αναζήτησης θα ήταν με περιστροφή στη σελίδα αναζήτησης του Netscape browser. Οι 5 μηχανές αναζήτησης της συμφωνίας ήταν το Yahoo!, το Maggelon, η Lycos, το Infoseek και το Excite.

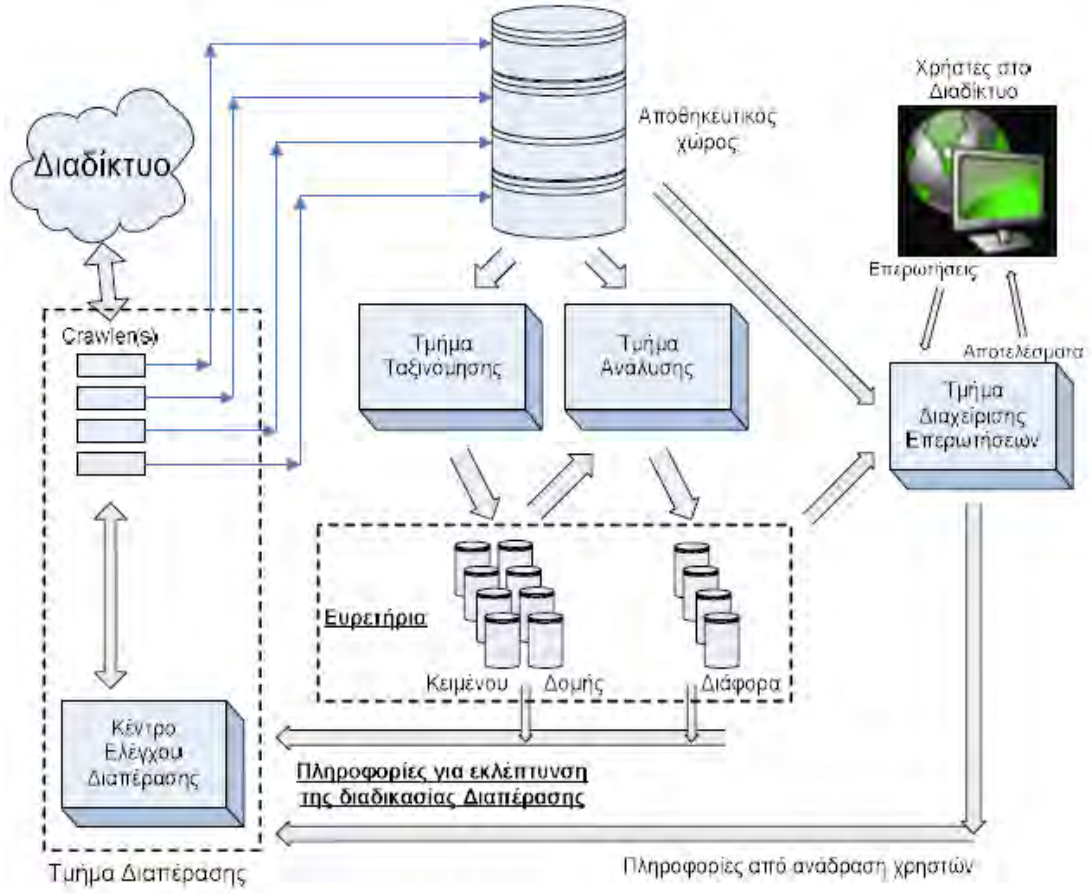
Το 1997 έκανε την εμφάνιση της η μηχανή αναζήτησης Google που δημιουργήθηκε από τον Sergay Brin και τον Larry Page ως τμήμα ενός ερευνητικού προγράμματος του πανεπιστήμιου του Stanford. Η Google χρησιμοποίησε τότε τις εισερχόμενες συνδέσεις (inbound links) [42] για να κατατάσσει τις σελίδες.

Το 1998 έκαναν την εμφάνιση τους οι MSN search [43] και Open Directory [44]. Το Open Directory αποτελεί συμφώνα με στοιχεία του διαδικτυακού του χώρου, ο μεγαλύτερος και περιεκτικότερος κατάλογος του Διαδικτύου.

Από το 1990 έχουν γίνει πολλές προσπάθειες για δημιουργία μηχανών αναζήτησης, λίγες όμως παρείχαν χρήσιμα αποτελέσματα για το χρήστη. Σε αυτό το διάστημα μέχρι και σήμερα οι μηχανές αναζήτησης, που λειτουργούσαν είτε με λέξεις - κλειδιά είτε με καταλόγους, προχωρούσαν στην αγορά η μια της άλλης. Σήμερα υπάρχουν πολλές μηχανές αναζήτησης αλλά αυτές που κατέχουν «την μερίδα του λέοντος» στην παγκόσμια αγορά είναι η Google, το Yahoo! και η Bing (πρώην Msn search).

3.3 Αρχιτεκτονική μιας μηχανής αναζήτησης

Οι μηχανές αναζήτησης εμφανίζονται στους χρήστες μέσα από μια πολύ απλή διεπαφή χρήστη. Πίσω όμως από αυτό το απλό περιβάλλον βρίσκεται μια ιδιαιτέρα σύνθετη εφαρμογή, απαιτητική στη δημιουργία αλλά και στη συντήρηση της. Τα κυριότερα μέρη που έχουν συνήθως οι μηχανές αυτές παρουσιάζονται στην εικόνα 3.1 και αναλύονται στη συνεχεία.



Εικόνα 3.1: Τα τμήματα μιας Μηχανής Αναζήτησης

Τμήμα Διαπέρασης – Επιτελεί την «κοπιαστική» διαδικασία να επισκεφτεί και να συλλέξει ιστοσελίδες και άλλα δεδομένα από το Παγκόσμιο ιστό. Συνήθως αποτελείται από περισσότερους του ενός crawlers (ανιχνευτές ιστού), οι οποίοι λειτουργούν παράλληλα εξοικονομώντας χρόνο μέχρι την επίτευξη του τελικού στόχου. Οι crawlers ξεκινάνε από ένα σύνολο αρχικών πηγών και ακλουθώντας τους υπερσυνδέσμους ανακαλύπτουν νέες ιστοσελίδες και έγγραφα στο παγκόσμιο ιστό προσέχοντας πάντα να μην επισκέπτονται τις ίδιες ιστοσελίδες και πηγές

πληροφοριών. Οι νέες σελίδες και έγγραφα που ανακαλύπτονται αποθηκεύονται στον διατιθέμενο χώρο ή λόγω οικονομίας χώρου αποθηκεύουν απλά τις διαδρομές που οδηγούν σε αυτά. Το κέντρο έλεγχου διαπέρασης συντονίζει την λειτουργιά των crawlers και καθορίζει τους αμέσους επόμενους υπερσυνδέσμους προς επίσκεψη. Η διαδικασία αυτή συνεχίζεται μέχρι να εξαντληθούν οι πόροι του συστήματος (πληρότητα αποθηκευτικού χώρου) πάνω στο οποίο «τρέχει» η εφαρμογή ή μέχρι να ικανοποιηθούν οι στόχοι που έχουν τεθεί αρχικά. Για παράδειγμα, η διαπέραση θα μπορούσε να έχει ως στόχο:

- Να ανακαλυφθούν όσο το δυνατό πιο πολλές αρχικές σελίδες ιστοτόπων.
- Να ανακαλυφθούν όσο πιο πολλά έγραφα και πηγές, τερματίζοντας μόνο όταν εξαντληθούν οι πόροι του συστήματος.
- Να γίνει διαπέραση μόνο των πηγών που εντοπίζονται εντός κάποιο συγκεκριμένου domain, π.χ. του πανεπιστήμιου της Λαμίας ή του δικτύου μιας εταιρίας.
- Να εντοπίζουν μόνο όσες ιστοσελίδες ή έγγραφα σχετίζονται με κάποιο ή κάποια συγκεκριμένα θέματα.
- Ακόμη και συνδυασμός των πιο πάνω ή όποιων άλλων επιθυμούν οι δημιουργοί της μηχανής αναζήτησης.

Η διαδικασία της διαπέρασης ανάλογα με τις πολίτικες που ακολουθεί η κάθε μηχανή αναζήτησης, μπορεί να επαναλαμβάνεται ανά κάποια χρονικά διαστήματα ολόκληρη ή εν μέρει ώστε τα αποτελέσματα να συμβαδίζουν με την διαρκή αλλαγή και εξέλιξη του παγκοσμίου ιστού. Στην επιλογή του επόμενου υπερσυνδέσμου που θα ακολουθηθεί από έναν crawler μπορεί να συμβάλει και η προϊστορία από προηγούμενες διαπεράσεις.

Τμήμα αποθηκευτικού χώρου – Αποθηκεύονται συνήθως προσωρινά, μέχρι να αξιοποιηθούν οι πληροφορίες τους, οι ιστοσελίδες που εντοπίζονται κατά τη διαπέραση. Για παράδειγμα, η μηχανή αναζήτησης Google αποθηκεύει και ιστοσελίδες παρέχοντας έτσι την δυνατότητα στους χρηστές του να βλέπουν την αποθηκευμένη έκδοση μιας ιστοσελίδας ακόμα και αν αυτή για κάποιο λόγο δεν είναι πλέον προσβάσιμη από το διαδίκτυο. Ωστόσο το μέγεθος του παγκόσμιου ιστού είναι τέτοιο που δεν είναι δυνατόν να χωρέσει σε οποιονδήποτε τοπικό χώρο και επομένως η αποθήκευση των πληροφοριών του απαιτεί ειδική αντιμετώπιση για να

αντιμετωπιστούν σημαντικά θέματα. Κάποια από αυτά είναι η εξοικονόμηση χώρου, η γρήγορη εύρεση των αποθηκευμένων πληροφοριών που απαιτούνται από τους χρήστες, η ανανέωση των αποτελεσμάτων που αποθηκευτήκαν (αν η λειτουργία αυτή πραγματοποιηθεί μόνο μια φορά στην αρχή λειτουργίας της μηχανής αναζήτησης τότε μετά από κάποιο χρονικό διάστημα οι πληροφορίες που αποθηκεύτηκαν μπορεί να μην είναι πλέον έγκυρες) κ.ο.κ.

Τμήμα ταξινόμησης – Εξάγει λέξεις από όλα τα έγγραφα που έχουν ανακαλυφθεί από το τμήμα της διαπέρασης και δημιουργεί ένα πίνακα για την αντιστοίχιση των εμφανίσεων των λέξεων στα εκάστοτε έγγραφα και στις εκάστοτε διευθύνσεις στο παγκόσμιο ιστό. Το μέγεθος του παραγόμενου πίνακα λόγω και του μεγέθους του διαδικτύου είναι τεράστιο και η διαδικασία της ταξινόμησης είναι ιδιαίτερα δύσκολη και απαιτεί μεγάλους πόρους από το σύστημα. Από τα στοιχειά αυτά μπορούν να δημιουργηθούν ευρετήρια, π.χ. ευρετήρια δομής που αποθηκεύουν πληροφορίες για τη συνδεσμολογία μεταξύ των εγγράφων και αξιοποιούνται από το κέντρο ελέγχου διαπέρασης.

Τμήμα ανάλυσης – Συνδυάζει πληροφορίες από τον αποθηκευτικό χώρο και τα απλά ευρετήρια που προαναφέρθηκαν με σκοπό την ανάλυση τους ώστε να προκύψουν ειδικά ευρετήρια συγκεκριμένου σκοπού. Για παράδειγμα, ευρετήρια ιστοσελίδων με βάση το πλήθος ή το είδος των εικόνων που περιέχουν, ευρετήρια ιστοσελίδων με προκαθορισμένο ελάχιστο όριο σημαντικότητας με βάση τις πολίτικες βαθμονόμησης τους από τις εκάστοτε μηχανής αναζήτησης ή ευρετήρια ιστοσελίδων μιας συγκεκριμένης γλώσσας ή γεωγραφικής περιοχής.

Τμήμα Διαχείρισης Επερωτήσεων – Είναι το τμήμα που αναλαμβάνει την επικοινωνία με το χρήστη. Δέχεται τα ερωτήματα που θέτει ο χρήστης, πραγματοποιεί την αναζήτηση στα υπάρχοντα ευρετήρια και παρουσιάζει τα αποτελέσματα στο χρήστη. Συνήθως λόγο του μεγάλου πλήθους των αποτελεσμάτων που απαντούν στο ερώτημα του χρήστη, απαιτείται ειδική αντιμετώπιση στην παρουσίαση τους. Είναι πολύ σημαντικό οι μηχανές αναζήτησης να παρουσιάζουν στο χρήστη μια σειρά κατάταξης με τα σχετικότερα αποτελέσματα. Δηλαδή όσο πιο σχετικό είναι ένα αποτέλεσμα με το ερώτημα του χρήστη τόσο πιο νωρίς πρέπει να εμφανίζονται στο χρήστη. Αυτό απαιτεί την βαθμολόγηση των αποτελεσμάτων από της μηχανές αναζήτησης με σκοπό την κατάταξη τους από το σχετικότερο στο λιγότερο σχετικό

αποτέλεσμα. Κάθε μηχανή αναζήτησης χρησιμοποιεί διαφορετικούς αλγόριθμους για βαθμολόγηση των αποτελεσμάτων και είναι το κυριότερο χαρακτηριστικό που ξεχωρίζει τις μηχανές αναζήτησης μεταξύ τους.

3.4 Οφέλη χρήσης Μηχανών Αναζήτησης

Όταν δεν μπορούμε να θυμηθούμε το τηλέφωνο ενός φίλου μας, δεν έχουμε παρά να ψάξουμε στο τηλεφωνικό κατάλογο για να το βρούμε. Το ίδιο ισχύει και στην περίπτωση που θέλουμε να μάθουμε την διεύθυνση του. Αν χρειαζόμαστε επειγόντως ένα ηλεκτρολόγο, ανοίγουμε το «Χρυσό Οδηγό» και με κλειδί αναζήτησης το «ηλεκτρολόγος» ψάχνουμε τον πρώτο διαθέσιμο ηλεκτρολόγο που μπορεί να μας βοηθήσει. Ακριβώς το ρολό του τηλεφωνικού καταλόγου και του «Χρυσού Οδηγού» παίζουν οι μηχανές αναζήτησης στο χώρο του διαδικτύου.

Η χρήση των μηχανών αναζήτησης επιφέρει πολλαπλά οφέλη τόσο στον απλό χρήστη όσο και στον επιχειρηματία.

Χρήστης – Η ύπαρξη μηχανών αναζήτησης καθίστα την αναζήτηση πληροφορίας στο διαδίκτυο μια εξαιρετικά εύκολη διαδικασία. Ο χρήστης πλέον το μόνο που χρειάζεται να γνωρίζει είναι τον σύνδεσμο που οδηγεί στην ιστοσελίδα της μηχανής αναζήτησης, να εισάγει τους όρους (με όσο το δυνατόν περισσότερη σαφήνεια) που περιγράφουν το θέμα που τον ενδιαφέρει και μέσα σε λίγα δευτερόλεπτα η μηχανή αναζήτησης θα του επιστρέψει μια λίστα με τις σχετικότερες ιστοσελίδες που απαντούν στο ερώτημα του. Επομένως ο χρήστης με την χρήση των μηχανών αναζήτησης εξυπηρετείται γρηγορότερα, ευκολότερα, πληρέστερα και δωρεάν.

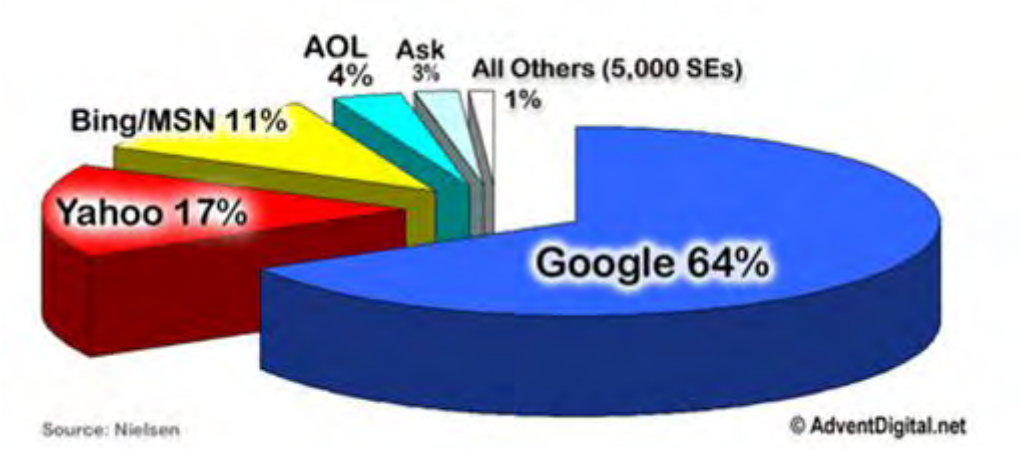
Επιχείρηση – Εξίσου σημαντικά είναι τα οφέλη μιας επιχείρησης που διαθέτει δικτυακό τόπο από την ύπαρξη των μηχανών αναζήτησης. Στο προηγούμενο παράδειγμα μας με τον «Χρυσό Οδηγό» η εγγραφή ενός επαγγελματία ή επιχείρησης σε αυτόν όπως είναι προφανές αυξάνει την πελατεία του. Τον ίδιο ακριβώς ρόλο αλλά με πολύ μεγαλύτερη βαρύτητα λόγω του μεγαλύτερου αριθμού χρηστών που χρησιμοποιούν το διαδίκτυο έχει μια μηχανή αναζήτησης. Η παρουσία του δικτυακού τόπου μιας επιχείρησης στη λίστα αποτελεσμάτων μιας μηχανής αναζήτησης συνεπάγεται σε αύξηση του αριθμού των χρηστών - πελατών που θα επισκεφτούν την ιστοσελίδα της. Επόμενος η επιχείρηση επιτυγχάνει να προσελκύσει ευκολότερα και

γρηγορότερα περισσότερους πελάτες - ενδιαφερομένους για τα προϊόντα / υπηρεσίες της και μάλιστα χωρίς κανένα επιπρόσθετο κόστος για την επιχείρηση.

Συνεπώς, οι μηχανές αναζήτησης αποτελούν ένα πανίσχυρο εργαλείο τόσο για τους χρήστες όσο και για τις επιχειρήσεις μέσα στο διαδίκτυο.

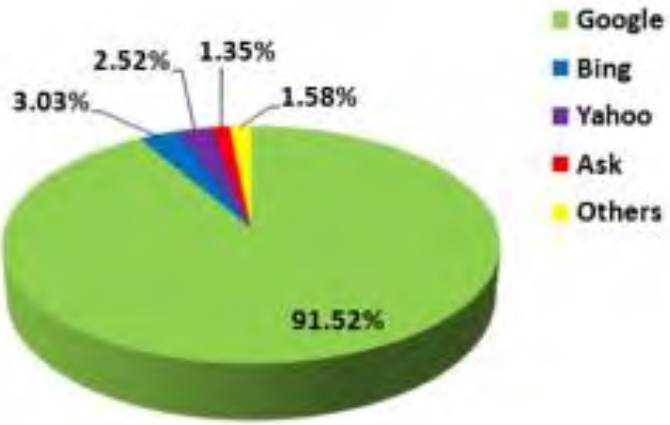
3.5 Δημοφιλέστερες μηχανές αναζήτησης

Οι τελευταίες μετρήσεις από την Nielsen Online [45] για το πρώτο εξάμηνο του 2011 δείχνουν ότι η μηχανή αναζήτησης που χρησιμοποιείτε περισσότερο στις Η.Π.Α. είναι η Google με ποσοστό 64%. Ακολουθούν η Yahoo με ποσοστό 17% ,η Bing με 11%, η AOL [46] με 4%, η ASK [47] με 3% και όλες οι υπόλοιπες (περίπου 5000 μηχανές αναζήτησης) κατέχουν το υπόλοιπο 1%. Στο πιο κάτω σχήμα φαίνονται και γραφικά τα πιο πάνω ποσοστά χρήσης των μηχανών αναζήτησης.



Εικόνα 3.2: Ποσοστά χρήσης των δημοφιλέστερων μηχανών αναζήτησης στις Η.Π.Α.

Σύμφωνα με την Experian Hitwise [48] για τον Ιούλιο του 2011 η μηχανή αναζήτησης που χρησιμοποιείται περισσότερο στο Ηνωμένο Βασίλειο είναι η Google με το ποσοστό της σχέση με τις Η.Π.Α να εκτοξεύετε στο 91.52%. Ακολουθούν η Bing με 3.3%, η Yahoo με 2.52%, η ASK με 1.35% και οι υπόλοιπες μηχανές αναζήτησης με 1.58%. Στο πιο κάτω σχήμα φαίνονται και γραφικά τα πιο πάνω ποσοστά χρήσης των μηχανών αναζήτησης για το Ηνωμένο Βασίλειο.



Εικόνα 3.3: Ποσοστά χρήσης των δημοφιλέστερων μηχανών αναζήτησης στο Η. Β.

3.5.1 Μηχανή Αναζήτησης Google

Εικόνα 3.4: Λογότυπο Google

Η Google είναι σήμερα μια από τις μεγαλύτερες εταιρείες διαδικτυακών υπηρεσιών. Η λειτουργία της ξεκίνησε τον Σεπτέμβριο του 1998. Ο στόχος της είναι να οργανώσει όλες τις πληροφορίες του κόσμου και να τις κάνει παγκόσμια διαθέσιμες. Οι ιδρυτές του Google ο Larry Page και ο Sergey Brin, γνωστοί και ως «Google Guys», όντας και οι δυο ισχυρές προσωπικότητες και παθιασμένοι με τους υπολογιστές διαφωνούσαν με έντονο τρόπο για όλα την πρώτη φορά που συναντήθηκαν το 1995. Σπουδαστές στην επιστήμη των υπολογιστών και οι δυο, σε μια συνάντηση ξενάγηση στο Stanford, διαφωνούσαν σε όλα μέχρι που βρήκαν κοινό τόπο στον τρόπο που σκέφτονταν πως έπρεπε να αντιμετωπιστεί το ζήτημα της άντλησης σχετικών μεταξύ τους πληροφοριών από μια τεραστία βάση δεδομένων.

Η συνεργασία τους δεν άργησε να ξεκινήσει, έτσι κατά τη διάρκεια του 1996 συνεργάστηκαν για να αναλύσουν τις διασυνδέσεις, ως θέμα των διδακτορικών τους

διατριβών. Δεδομένου ότι η διατριβή διήρκησε περισσότερο από τον αρχικό σχεδιασμό, ο Larry διατύπωσε την θεωρία πως ο υπολογισμός του αριθμού των διασυνδέσεων σε έναν ιστοχώρο θα μπορούσε να είναι ένας τρόπος μέτρησης των διασυνδέσεων αυτού του ιστοχώρου. Στις αρχές του 1997 αναπτύχτηκε η πρωτόγονη μηχανή αναζήτησης BackRub. Έχοντας έλλειψη πόρων, δοκίμασαν να στήσουν ένα δίκτυο υπολογιστών με φθηνά μηχανήματα μιας και ο Larry είχε ευχέρεια με τις επισκευές μηχανημάτων. Το φθινόπωρο του 1997 μετονομάστηκε σε Google, από το googol έναν μαθηματικό όρο, ο οποίος σημαίνει ένας αριθμός ίσος με το 1 ακολουθούμενο από 100 μηδενικά και γράφεται ως 10 στην 100ή δύναμη. Με τον όρο αυτό η Google επιθυμεί να υποδηλώσει την αποστολή της εταιρίας να οργανώσει το τεράστιο πλήθος πληροφοριών του Διαδικτύου.

Ένα χρόνο αργότερα ήδη είχε ξεκινήσει να μεγαλώνει η φήμη για την καινούργια τεχνολογία αναζήτησης που εφάρμοζαν. Το 1998 συνέχιζαν να καλυτερεύουν την τεχνολογία που αποτέλεσε την καρδιά του google και έψαχναν για πιθανούς συνεργάτες παρόλο που δεν είχαν σκεφτεί ακόμα να δημιουργήσουν μια εταιρεία για την τεχνολογία τους. Μεταξύ αυτών που απευθύνθηκαν ήταν και ο David Filo, φίλος κι ο ένας εκ των συνιδρυτών του Yahoo, ο οποίος τους ενθάρρυνε και τους είπε όταν θα ανέπτυσσαν πλήρως την ιδέα τους να τον ειδοποιούσαν, τους συμβούλευσε μάλιστα να ιδρύσουν μια δικιά τους εταιρεία με τη μηχανή αναζήτηση τους. Όμως δεν ήταν όλα θετικά, ο πρόεδρος ενός μεγάλου portal τους είπε ότι "όσο είμαστε στο 80% ισάξιοι του ανταγωνισμού, είμαστε μια χαρά και ότι οι χρήστες του Internet δεν ενδιαφέρονται για αναζητήσεις". Έχοντας αποτύχει να τραβήξουν το ενδιαφέρον των μεγάλων portal της εποχής αποφάσισαν ότι κάνουν να το κάνουν μόνοι τους, έτσι άρχισαν την αναζήτηση μετρητών καταρχάς για να ξεπληρώσουν τις πιστωτικές τους κάρτες που τις είχαν φτάσει στα όρια για να αγοράσουν ένα terabyte μνήμης για το δίκτυο τους. Τότε μίλησαν στον Andy Bechtolsheim έναν από τους ιδρυτές της Sun Microsystems μόνο που τον είδαν για σύντομο χρονικό διάστημα, ίσα ίσα που πρόλαβαν τα του αναπτύξουν τις ιδέες τους. Ο Bechtolsheim, άνθρωπος που έβλεπε μακριά κατάλαβε το πόσο σημαντικό είναι το εγχείρημα τους. Όμως ήταν βιαστικός και έπρεπε να φύγει. "Αντί να μου εξηγήσετε όλες τις λεπτομέρειες γιατί να μη σας δώσω κατευθείαν μια επιταγή;" τους είπε και τους την έδωσε. 100000 δολάρια πληρωτέες στην εταιρεία Google. Όμως εταιρεία Google δεν υπήρχε ακόμα για να

την εισπράξει. Η επιταγή περίμενε μερικές βδομάδες στο συρτάρι τους. Όταν μετά από επίμονες παρακλήσεις σε φίλους, συγγενείς και γνωστούς κατάφεραν να μαζέψουν τελικά γύρω στο ένα εκατομμύριο δολάρια, τον Σεπτέμβριο του 1998, η εταιρεία Google ήταν πια επίσημο γεγονός και ήταν η εταιρεία που σε λίγα χρόνια ανέτρεψε τα πάντα στην τεχνολογία των μηχανών αναζήτησης.

Εικόνα 3.5: Η λιτή εμφάνιση του Google είναι αυτή που οδήγησε τη μηχανή αναζήτησης σε τεράστια δημοτικότητα.

Σήμερα η μηχανή αναζήτησης Google είναι η δημοφιλέστερη και οι φράσεις «κάνω google», «γκουγκλάρω» και «γκουγκλίζω» είναι συνώνυμες με το «ψάχνω για πληροφορίες στο Διαδίκτυο». Αντίστοιχα, στην αγγλική γλώσσα το ρήμα "to google" έχει αποκτήσει πλέον ταυτόσημη έννοια με το ρήμα «αναζητώ», και πρόσφατα, το ίδιο ρήμα προστέθηκε στο αγγλικό λεξικό Merriam-Webster [49] με όλα τα παράγωγά του (to google > googling > googled).

Το Googleplex είναι η κεφαλή της Google, αφού όλες οι λειτουργίες της μηχανής αναζήτησης εμπνέονται και υλοποιούνται μέσα σε αυτό το κτήριο. Το όνομα έχει διπλό συμβολισμό.

1. Αποτελεί λογοπαίγνιο της αγγλικής μονάδας μέτρησης googol (10^{100})
2. Είναι συνθετικό των λέξεων Google και complex (=συγκρότημα), σε μια πιο γλωσσολογική ερμηνεία.

Το καινούργιο κτήριο που παραδόθηκε στην Google ειδικά διαμορφωμένο για τις ανάγκες της εταιρίας, βρίσκεται στο Mountain View της Καλιφόρνιας. Οι συνθήκες εργασίας έχουν χαρακτηριστεί από τις καλύτερες που έχουν εφαρμοστεί σε εταιρίες, αφού η Google έχει προνοήσει για όλες τις ανάγκες των εργαζομένων, όπως η

ξεκούραση, τα τακτικά διαλείμματα σε ειδικά διαμορφωμένους χώρους ή ακόμα και την ύπαρξη παιδότοπου για την φύλαξη των παιδιών των εργαζομένων. Το 2010 η εταιρία Google εργοδοτούσε 24.400 ανθρώπους.

3.5.1.1 Αλγόριθμος PageRank

Ο αλγόριθμος PageRank είναι η μέθοδος με την οποία η Google προσδιορίζει την σημαντικότητα μιας ιστοσελίδας και αποτελεί την καρδιά του λογισμικού της που την έκανε την κορυφαία μηχανή αναζήτησης στην κόσμο. Πείρε το όνομα της από τον Larry Page, συνιδρυτή της Google.

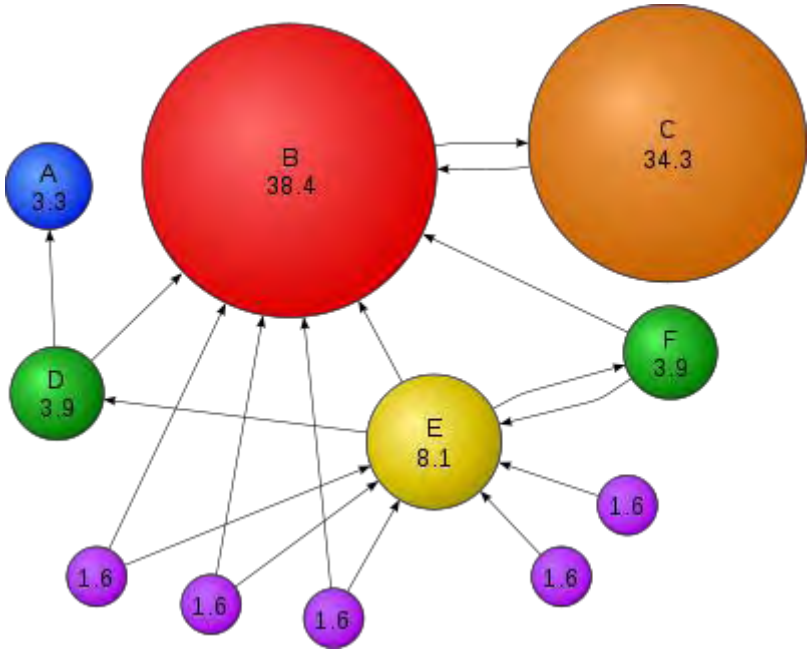
Το PageRank είναι μία αριθμητική τιμή, η οποία αντιπροσωπεύει το πόσο σημαντική και σπουδαία είναι μία ιστοσελίδα στο διαδίκτυο, κατά την αντικειμενική κρίση της μηχανής αναζήτησης Google. Το PageRank επηρεάζει σημαντικά, μαζί με κάποιους επιπλέον παράγοντες, την κατάταξη στα αποτελέσματα αναζήτησης. Για τον υπολογισμό και προσδιορισμό του PageRank , η Google χρησιμοποιεί ένα πολύπλοκο αλγόριθμο με εκατομμύρια μεταβλητές και όρους και ο υπολογισμός του γίνετε με αυτοματοποιημένο τρόπο.

Κύρια λογική της όμως στην αξιολόγηση είναι η σύνδεση (link) π.χ. μιας Ιστοσελίδας Α με μία Ιστοσελίδα Β. Θεωρεί τον σύνδεσμο αυτό ότι είναι θετική ψήφος για την Ιστοσελίδα Β από την Ιστοσελίδα Α. Συνεπώς οι διασυνδέσεις των ιστοσελίδων και εσωτερικά και εξωτερικά λαμβάνοντα θετικά υπόψη, διότι όσο πιο πολλά links έχει μία ιστοσελίδα από άλλες, καθώς και η σπουδαιότητά τους, θεωρούνται πολλές θετικές ψήφοι, άρα η ιστοσελίδα χαρακτηρίζετε σημαντική και σπουδαία. Εάν δηλαδή οι ψήφοι προς την ιστοσελίδα προέρχονται από σημαντικές ιστοσελίδες μεγάλης σπουδαιότητας, η αξία η οποία προσδίδετε είναι ακόμα μεγαλύτερη.

Πιο συγκεκριμένα έστω ότι Α, η ιστοσελίδα που πρόκειται να βαθμολογηθεί και $T_1, T_2, \ldots, T_n$ οι ιστοσελίδες που έχουν σύνδεσμο προς την Α. Έστω επίσης ότι ο C(A) είναι ο αριθμός των εξωτερικών συνδέσμων της ιστοσελίδας Α. Τότε η βαθμολόγια PageRank της ιστοσελίδας υπολογίζεται ως:

$$\mathbf{PR(A)=(1-d) + d\ (PR(T_1)/C(T_1) + \ldots + PR(T_n) / C(T_n)),}$$

όπου d είναι μια παράμετρος απόσβεσης με τιμή μεταξύ του 0 και 1. Εάν αθροίσουμε όλες της βαθμολογίες των ιστοσελίδων παίρνουμε 1, δηλαδή η PR(A) είναι μια κανονικοποιημένη κατανομή πιθανότητας. Η συνάρτηση PR(A) είναι μια πολιτικοποιημένη έκφραση της συμπεριφοράς ενός τυχαίου χρήστη του ιστού, ο οποίος ξεκινώντας από μια τυχαία αρχική σελίδα, ακολουθεί κάποιους συνδέσμους και πηγαίνει σε άλλες σελίδες μέχρι να βαρεθεί και να σταματήσει. Η πιθανότητα αυτός ο χρήστης να επισκεφτεί την ιστοσελίδα A είναι PR(A). Ο παράγοντας απόσβεσης d είναι η πιθανότητα να σταματήσει ο χρήστης σε κάθε ιστοσελίδα.

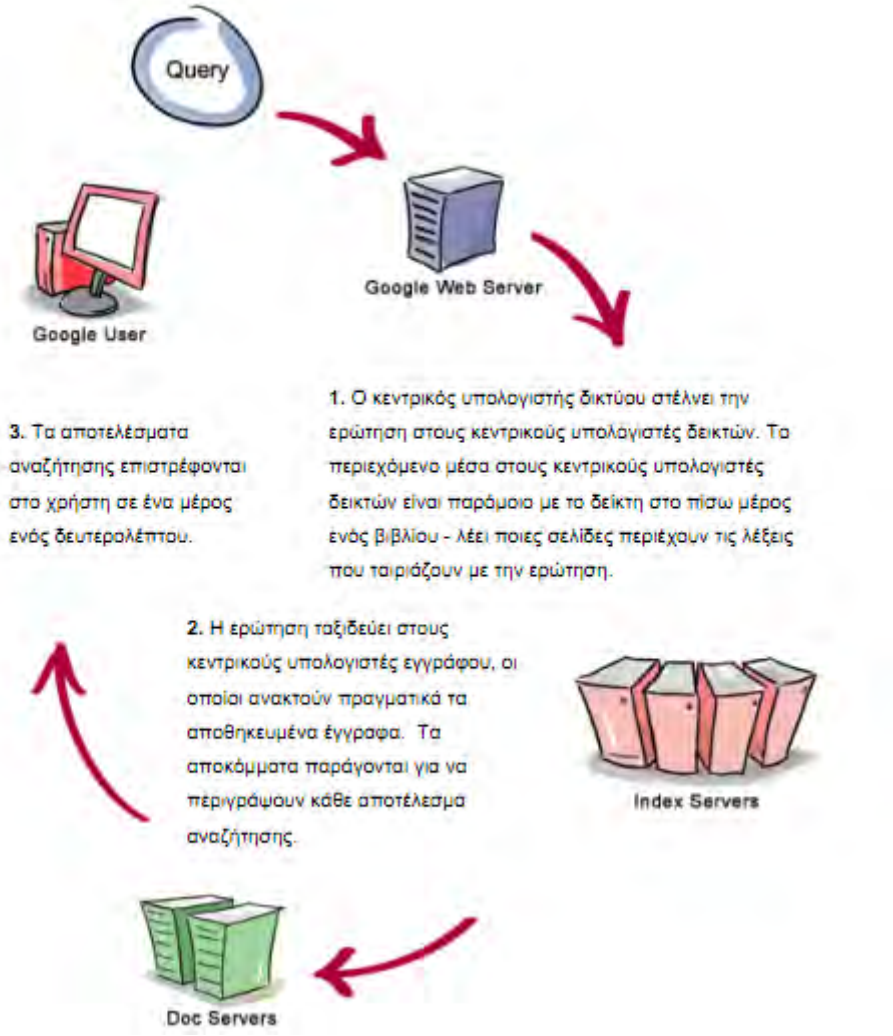


Εικόνα 3.6: Γραφικό παράδειγμα PageRank: Η ιστοσελίδα B έχει το υψηλότερο PageRank διότι «δείχνεται» τις περισσότερες φορές σε σχέση με τις άλλες ιστοσελίδες. Ακολουθεί η ιστοσελίδα C η οποία δείχνεται μόνο από μια ιστοσελίδα (B) αλλά λόγω του υψηλού PageRank που έχει η B παίρνει και αυτή υψηλό PageRank.

Άλλος σημαντικός παράγοντας αξιολόγησης είναι το περιεχόμενο των σελίδων. Η Google αναλύει όχι μόνο τα meta-tags [50] αλλά και το σύνολο του περιεχομένου της ιστοσελίδας. Λαμβάνει υπόψη της πολλές παραμέτρους, όπως το μέγεθος και χρωματισμό των γραμματοσειρών, τους τίτλους, τις παραγράφους, τις λέξεις και τη συνάφειά τους με το κείμενο αλλά και το υπόλοιπο περιεχόμενο των άλλων ιστοσελίδων.

3.5.1.2 Διάρκεια ζωής των ερωτημάτων

Η διάρκεια ζωής μια ερώτησης στη μηχανή της Google διαρκεί υπό κανονικές συνθήκες λιγότερο από μισό δευτερόλεπτο. Ωστόσο υπάρχουν διάφορα στάδια που πρέπει να ολοκληρωθούν προτού να μπορέσουν τα αποτελέσματα να παρουσιαστούν στο χρήστη. Στη πιο κάτω εικόνα παρουσιάζονται τα κυριότερα στάδια μιας αναζήτησης στη μηχανή της Google.



Εικόνα 3.7: Κύκλος ζωής μιας αναζήτησης στη Google [51]

Αφού ο χρήστης εισάγει το ερώτημα του, ο κεντρικός υπολογιστής δικτύου (web server) στέλνει τις λέξεις κλειδιά στους κεντρικούς υπολογιστές δεικτών (index servers) στους οποίους είναι αποθηκευμένοι οι κατάλογοι – ευρετήρια της Google. Έπειτα η ερώτηση μεταφέρεται στους κεντρικούς υπολογιστές εγγράφου (doc

servers) οι οποίοι ανακτούν τις αποθηκευμένες ιστοσελίδες και δημιουργούν τα αποτελέσματα. Τέλος, τα αποτελέσματα εμφανίζονται στην οθόνη του χρήστη.

3.5.1.3 Βάσεις δεδομένων της Google

Οι google μέσα στο πυρήνα της περιέχει αρκετές βάσεις δεδομένων. Οι κυριότερες είναι:

- **Διαδικτύου**: Καταχωρημένες οι ιστοσελίδες με δείκτες και προσθετοί τύποι αρχείων όπως .pdf, .xls, .ps, .txt, .doc, .ppt, .rtf και άλλα.
- **Διαφημίσεων – Αγγελιών**: Πληρωμένες διαφημίσεις που εμφανίζονται σε περίοπτη θέση (πάνω δεξιά της οθόνης με το πρόθεμα ads).
- **Εικόνων**: Βάση δεδομένων από εικόνες.
- **Ειδήσεων**: Αποθηκεύονται ιστοσελίδες ειδήσεων των τελευταίων 30 ημερών.
- **Βιβλίων**: Περιέχει ολόκληρα βιβλία κειμένου.
- **Μελετών (Scholar)**: Περιέχει ακαδημαϊκές μελέτες, δημοσιεύσεις από επιστημονικά συνέδρια, διατριβές κτλ.
- **Προϊόντων (Froogle)**: Βάση για αγορά και εξεύρεση προϊόντων.

Υπάρχουν αρκετές ακόμη βάσεις δεδομένων πιο εξειδικευμένες, όπως βάση δεδομένων μιας κυβέρνησης ή του στρατού (με κατάληξη .gov, .mil αντίστοιχα), βάση για αναζήτηση πανεπιστήμιων, όπως και βάση για Linux, Apple και Microsoft.

Οι βάσεις δεδομένων της Google χρησιμοποιούνται και από άλλες μηχανές αναζήτησης και από διάφορα site (AOL, iWon, Netscape κ.α.).

3.5.1.4 Υπηρεσίες αναζήτησης της Google

Η Google προσφέρει τις ακόλουθες υπηρεσίες αναζήτησης στους χρήστες του διαδικτύου**:** [52]

- Google Search - Γενική αναζήτηση (http://www.google.com/).
- iGoogle – Προσωπική σελίδα (http://www.google.com/ig?aig=0&reason=1).
- Google Image – Αναζήτηση εικόνων (http://images.google.com/).
- Google Video – Αναζήτηση βίντεο (http://video.google.com).
- Google Maps – Αναζήτηση Χαρτών (http://maps.google.com/).

- Google News – Αναζήτηση Ειδήσεων (http://news.google.com/).
- Google Products Search – Αναζήτηση προϊόντων προς πώληση (http://www.google.com/prdhp).
- Google Blog Search – Αναζήτηση σε blogs (http://www.google.com/blogsearch).
- Google Book Search – Αναζήτηση βιβλίων (http://books.google.com/).
- Google Scholar – Αναζήτηση μελετών (http://scholar.google.com/).
- Special Search - Αναζήτηση σε ειδικά θέματα (http://www.google.com/help/features.html).
- Special Features - Βιβλία, ταινίες, μουσική και άλλα (http://www.google.com/intl/en/help/features.html).
- Google Patents – Αναζήτηση Ευρεσιτεχνιών (http://www.google.com/patents).
- Google Finance – Υπηρεσία πληροφόρησης οικονομικών νέων και ζωντανή μετάδοση του χρηματιστήριου (http://www.google.com/finance).
- Google Alerts – Υπηρεσία ειδοποιήσεων (http://www.google.com/alerts).
- Google Desktop – Αναζήτηση στον υπολογιστή (http://desktop.google.com/).
- Google Earth – Προβολή του πλανήτη σε 3D δορυφορικές εικόνες. (http://www.google.com/earth/index.html).

3.5.1.5 Περιορισμοί αναζήτησης

Η Google δίνει τη δυνατότητα αναζήτησης με χρήση κάποιων περιορισμών:

- **Γλώσσα** – Επιστροφή αποτελεσμάτων συγκεκριμένης γλώσσας.
- **Ημερομηνία** – Επιστροφή αποτελεσμάτων κάποιας χρονικής διάρκειας (τελευταίες 24 ώρες, τελευταία εβδομάδα, τελευταίο μήνα ή για το τελευταίο έτος).
- **Τύπους αρχείων** – Επιστροφή συγκεκριμένων αρχείων τύπου. Γίνεται με χρήση της εντολής «filetype:» ακολουθουμένη από την κατάληξη του αρχείου. (Π.χ. filetype: doc).
- **Δικτυακός χώρος** – Επιστροφή αποτελεσμάτων συγκεκριμένου δικτυακού χώρου (domain).

- **Περιεχόμενα Ενηλίκων**: Οικογενειακό φίλτρο το οποίο εξαιρεί περιεχόμενα ενηλίκων.

3.5.1.6 Επιλογές αναζήτησης

Η Google παρέχει κάποιες εξειδικευμένες επιλογές αναζήτησης για εξαγωγή πιο συγκεκριμένων αποτελεσμάτων: [53]

- « **OR** » Τελεστής ή. Αναζήτηση για το ένα ή το άλλο. Για παράδειγμα η αναζήτηση «Greece OR Cyprus» θα επιστρέψει αποτελέσματα σχετικά με την Ελλάδα ή την Κύπρο.
- « - » Τελεστής αφαίρεσης. Αναζήτηση αποκλείοντας μια συγκεκριμένη λέξη. Για παράδειγμα η αναζήτηση για «apple - tree» δεν θα επιστρέψει αποτελέσματα που έχουν σχέση με το δέντρο (=tree)
- « + » Τελεστής πρόσθεσης. Η google σε μια απλή αναζήτηση ψάχνει και για τα συνώνυμα της λέξης / φράσης που δίνεται από το χρήστη. Με την χρήση του τελεστή + αναζητεί ακριβώς όπως δοθεί η λέξη / φράση. Για παράδειγμα η αναζήτηση «πανεπιστήμια +Ελλάδας +πληροφορικής», απαιτεί στα αποτελέσματα να περιέχονται οι λέξεις «Ελλάδας» και «πληροφορικής».
- « ***** » Τελεστής πολλαπλασιασμού. Με τη χρήση του αστερίσκου μέσα σε μια φράση, η αναζήτηση γίνεται για οποιανδήποτε λέξη στη συγκεκριμένη θέση. Για παράδειγμα η αναζήτηση «Google *» θα επιστρέψει μια πληθώρα αποτελεσμάτων σχετικά με όλες τις υπηρεσίες και προϊόντα που παρέχει η Google.
- « « » » Αναζήτηση για μια συγκεκριμένη φράση ακριβώς όπως δοθεί από τον χρήστη.
- **Define –** Μια αναζήτηση με αυτό το πρόθεμα, παρέχει ορισμούς των λέξεων που ακολουθούν. Για παράδειγμα η αναζήτηση «define water» θα επιστρέψει αποτελέσματα σχετικά με τον ορισμό του νερού.
- **Stocks –** Παρέχει τιμές των μετοχών που ακολουθούν (π.χ. stocks bank of Cyprus).
- **Site –** Περιορίζει τα αποτελέσματα της αναζήτησης μόνο από τη συγκεκριμένη ιστοσελίδα που δίνεται.

- **Allintitle** – Αναζητεί μόνο στους τίτλους των ιστοσελίδων και όχι στο περιεχόμενο.
- **Intitle** – Απαιτεί να υπάρχει μια συγκεκριμένη λέξη στον τίτλο μιας ιστοσελίδας και μια άλλη στο υπόλοιπο κείμενο. Για παράδειγμα στην αναζήτηση «intitle:google search», θα εμφανίσει ιστοσελίδες που περιέχουν στο τίτλο την λέξη «google» και οπουδήποτε στο υπόλοιπο κείμενο τη λέξη «search».
- **Allinurl** – Αναζητεί μόνο στο κείμενο της διεύθυνσης μιας ιστοσελίδας αγνοώντας το περιεχόμενο της.
- **Inurl** – Απαιτεί να υπάρχει μια συγκεκριμένη λέξη στη διεύθυνση μιας ιστοσελίδα και μια άλλη στο περιεχόμενο της ιστοσελίδας.
- **Cache –** Εμφανίζει την αποθηκευμένη έκδοση της ιστοσελίδας.
- **Link** – Εμφανίζει όλες τις ιστοσελίδες που έχουν σύνδεσμο προς την συγκεκριμένη ιστοσελίδα που δίνεται από τον χρήστη. (π.χ. link:www.google.com)
- **Related** – Εμφανίζει ιστοσελίδες παρόμοιες με την ιστοσελίδα που δίνεται από τον χρήστη.
- **Info –** Εμφανίζει πληροφορίες για μια συγκεκριμένη ιστοσελίδα.

3.5.2 Μηχανή αναζήτησης Yahoo! Search

Εικόνα 3.8: Λογότυπο Yahoo!

Η Yahoo! είναι σήμερα μια από τις μεγαλύτερες και δημοφιλέστερες εταιρίες διαδικτυακών υπηρεσιών. Η αρχή έγινε τον Ιανουάριο του 1994 από τους τότε τελειόφοιτους του πανεπιστήμιου του Stanford, David Flo και Jerry Yang. Οι δυο τους ήταν μεταπτυχιακοί σπουδαστές στο τμήμα των ηλεκτρονικών υπολογιστών και μέλη στην ιδία ομάδα ερευνάς του τμήματος. Μια από τις αγαπημένες τους ασχολίες ήταν να κρατούν σημειώσεις από κάθε σελίδα του διαδικτύου που τους κέντριζε το

ενδιαφέρον. Τον Φεβρουάριο του 1994 οι δυο φίλοι αποφάσισαν πως έπρεπε να οργανώσουν την λίστα τους.

Η λίστα στήθηκε γρήγορα στον προσωπικό υπολογιστή David και την χρησιμοποιούσαν από κοινού. Λίγο αργότερα οργάνωσαν την λίστα με τις αγαπημένες τους ιστοσελίδες, χωρίζοντας τα σε κατηγορίες και την δημοσίευσαν στο διαδίκτυο. Η ευκολία χρήσης της λίστας και η πληρότητα της την έκανε αμέσως γνωστή στους κύκλους του πανεπιστήμιου. Μέσα στο πρώτο μήνα λειτουργιάς της η λίστα δεχόταν πάνω από 100 επισκέπτες τη ημέρα, χωρίς οι ίδιοι να την έχουν διαφημίσει.

Έξι μήνες αργότερα η σελίδα είχε γίνει πασίγνωστη, παρόλα αυτά η λίστα δεν είχε μεγαλώσει αρκετά. Έτσι οι δυο άντρες έδωσαν το δικαίωμα στους επισκέπτες της ιστοσελίδας τους να προσθέτουν τις δίκες τους αγαπημένες ιστοσελίδες έχοντας όμως τον τελευταίο λόγο στην κατάταξη τους.

Τον Απρίλιο του 1995, η εταιρία Sequoia Capital αποφάσισε να χρηματοδοτήσει την προσπάθεια των δυο φίλων με το ποσό των 2 εκατομμύριων δολαρίων. Χωρίς χρονοτριβή οι δυο άντρες προσέλαβαν άμεσα προγραμματιστές και λοιπό προσωπικό για να οργανώσουν την εταιρία τους. Η Yahoo! Inc. με 49 άτομα προσωπικό μόλις είχε γεννηθεί.

Σήμερα, 11 χρόνια μετά, ένα μεγάλο χαρτόκουτο γεμάτο σημειώσεις εξελίχθηκε σε μια από τις μεγαλύτερες εταιρίες παροχής διαδικτυακών υπηρεσιών.

Αν και ξεκίνησε ως θεματικός κατάλογος, αργότερα εξελίχτηκε σε μια πανίσχυρη μηχανή αναζήτησης. Η επίσημη ερμηνεία για την προέλευση του ονόματος «Yahoo» είναι: «Yet Another Hierarchical Oracle». Συμφώνα με τους δυο ιδιοκτήτες το «Yahoo» επιλέχθηκε επειδή μια μοιάζει με επιφώνημα χαράς, αλλά κι εξαιτίας της σημασίας της λέξης: αγενής, άξεστος, «τσαμπουκάς».

3.5.2.1 Εξέλιξη της τεχνολογίας του Yahoo!

Αρχικά, η διαδικασία της αναζήτησης, ανάκτησης και αποθήκευσης πληροφοριών στο διαδίκτυο δεν γινόταν από την ιδία την Yahoo. Τα ερευνήσιμα ευρετήρια προέρχονταν από άλλες εταιρίες. Μέχρι το 2001 παρέχονταν από την Inktomi, μια εταιρία παροχής αποτελεσμάτων αναζήτησης που τα αποτελέσματα της όμως εμφανίζονταν σε ιστοσελίδες άλλων εταιριών.

Το 2002 αγόρασε την Inktomi και το 2003 την «Overture Services, Inc.», η οποία κατείχε τις μηχανές αναζήτησης «AlltheWeb» και «AltaVista». Αρχικά αν και η Yahoo κατείχε πολλές μηχανές αναζήτησης δεν τις χρησιμοποίησε στην ιστοσελίδα της. Συνεργαζόταν με την Google και χρησιμοποιούσε τα αποτελέσματα της μέχρι τις αρχές του 2004.

Το 2004, η Yahoo αποφάσισε να δημιουργήσει την δίκη της τεχνολογία, έτσι ώστε να γίνει ανεξάρτητη μηχανή αναζήτησης. Κάτι που έκανε με τη χρήση του crawler «*Yahoo! Slurp*» [54] που δημιουργήθηκε με βάση την τεχνολογία που χρησιμοποιούσε η Inktomi. Έτσι, η Yahoo συνδυάζοντας τις δυνατότητες των μηχανών αναζήτησης που είχε αγοράσει, δημιούργησε την δική της μηχανή αναζήτησης.

Η Yahoo αποθηκεύει σε ευρετήρια κοινές σελίδες HTML καθώς επίσης και γνωστούς τύπους αρχείων όπως excel, pdf, powerpoint, word και απλά αρχεία κειμένου. Τα αποτελέσματα της ταξινομούνται με χρήση ενός αλγόριθμου σχετικότητας. Όσο πιο σχετικό είναι το ερώτημα του χρήστη με κάποια ιστοσελίδα του ευρετηρίου τόσο πιο ψηλά στη κατάταξη των αποτελεσμάτων θα εμφανιστεί η ιστοσελίδα.

Το 2007, η αρχική σελίδα της Yahoo επανασχεδιάστηκε και απέκτησε μια πιο μοντέρνα εμφάνιση. Επίσης, πρόσθεσε την υπηρεσία Yahoo Assist, η οποία παρέχει σε πραγματικό χρόνο προτάσεις για το ερώτημα και σχετικές ιδέες ενώ πληκτρολογεί ο χρήστης το ερώτημα.

Το 2008, ανακοινώθηκε η εισαγωγή μιας νέας υπηρεσίας με όνομα «Build Your Own Search Service» ή «BOSS». Αυτή η υπηρεσία επιτρέπει στους σχεδιαστές ιστοσελίδων να χρησιμοποιήσουν το σύστημα της Yahoo και να δημιουργήσουν μια δική τους προσαρμοζόμενη μηχανή αναζήτησης για την ιστοσελίδα τους.

Το Φεβρουάριο η Microsoft έκανε πρόταση εξαγοράς της Yahoo με το ποσό των 44.6 δισεκατομμυρίων. Η Yahoo απέρριψε την προσφορά θεωρώντας την πολύ κατώτερη της αξίας της εταιρίας. Δυο μήνες αργότερα, η Microsoft ανέβασε την προσφορά της κατά 5 δισεκατομμύρια, περίπου 33 δολάρια για κάθε μετοχή. Οι διαπραγματεύσεις δεν καρποφόρησαν και ένα μήνα αργότερα Microsoft ανακοινώσε την απόσυρση της προσφοράς της για εξαγορά ολόκληρης της Yahoo.

Ένα χρόνο αργότερα, τον Ιούλιο του 2009, ανακοινώθηκε μια δεκάχρονη συμφωνία μεταξύ Yahoo και Microsoft. Η συμφωνία προβλέπει σταδιακά την πλήρη πρόσβαση της Microsoft στη μηχανή αναζήτησης της Yahoo και χρησιμοποίηση της σε μελλοντικά σχεδία της Microsoft για την μηχανή αναζήτησης της, Bing. Στο πλαίσιο της συμφωνίας η Microsoft δεν υποχρεούται να υποβάλει οποιαδήποτε χρήματα εκ τον πρότερων στην Yahoo. Μια μέρα μετά από τη συμφωνία η μετοχή της Yahoo μειώθηκε περισσότερο από 10% έως 15.14 δολάρια ανά μετοχή, περίπου 60% χαμηλότερη από τη προσφορά εξαγοράς της Microsoft ένα χρόνο πριν.

3.5.2.2 Βάσεις δεδομένων της Yahoo

- **Ιστοσελίδων**: Βάση δεδομένων με ιστοσελίδες του διαδικτύου από όπου αναζητούνται τα αποτελέσματα.
- **Διαφημίσεων – Αγγελιών**: Πληρωμένες διαφημίσεις που εμφανίζονται σε περίοπτη θέση.
- **Καταλόγου αρχείων Yahoo**: Η αρχική μορφή του Yahoo με καταλόγους.
- **Βάση εικόνων, βίντεο, ήχων**
- **Χαρτών**
- **Αγορών – προϊόντων**
- **Ειδήσεων**
- **Απαντήσεων**: Βάση δεδομένων με απαντήσεις για πολλά ερωτήματα.

Υπάρχουν αρκετές ακόμη άλλες βάσεις δεδομένων οι οποίες περιέχουν ένα μεγάλο μέρος των πληροφοριών από την κανονική ιστοσελίδα της Yahoo.

3.5.2.3 Υπηρεσίες αναζήτησης της Yahoo

Το Yahoo! προσφέρει στους χρήστες του μεταξύ άλλων τις παρακάτω υπηρεσίες:

- **Δωρεάν Υπηρεσίες**: Οι δωρεάν υπηρεσίες του Yahoo! καλύπτουν τα έξοδά τους από την διαφημιστική καμπάνια και οι βασικότερες είναι οι εξής:
    - **Yahoo! Mail:** Υπηρεσία ηλεκτρονικού ταχυδρομείου
    - **Yahoo! Search:** Μηχανή αναζήτησης
    - **Yahoo! Radio:** Ραδιόφωνο (κυρίως αμερικάνικη μουσική)
    - **Yahoo! Music Videos**: Μουσικά βίντεο
    - **Yahoo! News**: Ειδήσεις από όλο τον κόσμο
    - **Yahoo! Weather**: Καιρός σε χιλιάδες περιοχές.
    - **Yahoo! Finance**: Χρηματιστηριακές ειδήσεις και παρακολουθήσεις χαρτοφυλακίων.
    - **Yahoo! Games**: Παιχνίδια για έναν παίκτη ή ομαδικά με διαδικτυακή σύνδεση
    - **Yahoo! Movies**: Βάση δεδομένων αμερικάνικων ταινιών
    - **Yahoo! kids**: Ιστοσελίδα για παιδιά
    - **Yahoo! Messenger**: Πρόγραμμα ανταλλαγής άμεσων μηνυμάτων
    - **Yahoo! Maps**: Χάρτες από όλο τον κόσμο.
    - **Yahoo! Groups**: Ομάδες συζητήσεων.
    - **Yahoo! 360**: Δημιουργία προσωπικής σελίδας.
    - **Flickr**: Δημοσίευση φωτογραφιών.
- **Υπηρεσίες Επί πληρωμή**
    - **Yahoo! Mail Plus**: Ηλεκτρονικό ταχυδρομείο με επιπλέον χώρο και επιπλέον προστασία.
    - **Yahoo! Music Unlimited**: Μουσικό κατάστημα πώλησης νόμιμης μουσικής
    - **Yahoo! TiVo**: Προγραμματισμός του TiVo μέσω του Yahoo!
    - **Flickr Pro**: Το Flickr αλλά με δυνατότητα ανεβάσματος απεριόριστων φωτογραφιών, δημιουργία απεριόριστων άλμπουμ και δυνατότητα για ανέβασμα βίντεο.
    - **MyBlogLog Pro**: Δίνει ζωντανά (real-time) στατιστικά.

3.5.2.4 Περιορισμοί αναζήτησης

Η Yahoo! δίνει τη δυνατότητα αναζήτησης με χρήση κάποιων περιορισμών:

- **Γλώσσα** – Επιστροφή αποτελεσμάτων συγκεκριμένης γλώσσας.
- **Ημερομηνία** – Επιστροφή αποτελεσμάτων κάποιας χρονικής διάρκειας (τελευταίους 3 μήνες / 6 μήνες / 1 χρόνο).
- **Τύπους αρχείων** – Επιστροφή συγκεκριμένων αρχείων τύπου. Γίνεται με χρήση της εντολής «originurlextension:» ακολουθουμένη από την κατάληξη του αρχείου. (Π.χ. originurlextension: doc).
- **Δικτυακός χώρος** – Επιστροφή αποτελεσμάτων συγκεκριμένου δικτυακού χώρου (domain).
- **Περιεχόμενα Ενηλίκων**: Οικογενειακό φίλτρο το οποίο εξαιρεί περιεχόμενα ενηλίκων.

3.5.2.5 Επιλογές αναζήτησης

Η Yahoo, όπως και η Google, παρέχει κάποιες εξειδικευμένες επιλογές αναζήτησης για εξαγωγή πιο συγκεκριμένων αποτελεσμάτων:

- **«AND»** Αναζήτηση και για το ένα και για το άλλο. Για παράδειγμα η αναζήτηση «Greece AND Cyprus» θα επιστρέψει μόνο αποτελέσματα που είναι σχετικά και με την Ελλάδα και την Κύπρο.
- **«OR»** Αναζήτηση για το ένα ή το άλλο. Για παράδειγμα η αναζήτηση «Greece OR Cyprus» θα επιστρέψει μόνο αποτελέσματα που είναι σχετικά ή με την Ελλάδα ή με την Κύπρο.
- **« NOT »** Αναζήτηση αποκλείοντας μια συγκεκριμένη λέξη. Για παράδειγμα η αναζήτηση για «apple NOT tree» δεν θα επιστρέψει αποτελέσματα που έχουν σχέση με το δέντρο (=tree)
- **« * »** Τελεστής πολλαπλασιασμού. Με τη χρήση του αστερίσκου μέσα σε μια φράση, η αναζήτηση γίνεται για οποιανδήποτε λέξη στη συγκεκριμένη θέση. Για παράδειγμα η αναζήτηση «Yahoo *» θα επιστρέψει μια πληθώρα αποτελεσμάτων σχετικά με όλες τις υπηρεσίες και προϊόντα που παρέχει η Yahoo.

- **«intitle»** – Απαιτεί να υπάρχει μια συγκεκριμένη λέξη στον τίτλο μιας ιστοσελίδας.
- **«site»** – Περιορίζει τα αποτελέσματα της αναζήτησης μόνο από τη συγκεκριμένη ιστοσελίδα που δίνεται.
- **«hostname»** – ίδια λειτουργιά με site.
- **«link»** – Εμφανίζει όλες τις ιστοσελίδες που έχουν σύνδεσμο προς την συγκεκριμένη ιστοσελίδα που δίνεται από τον χρήστη.
- **«url»** – Βρίσκει ακριβώς μια URL διεύθυνση στη βάση δεδομένων.
- **«Inurl»** – Απαιτεί να υπάρχει μια συγκεκριμένη λέξη στη διεύθυνση μιας ιστοσελίδας.
- « » - Αναζήτηση για μια συγκεκριμένη φράση ακριβώς όπως δοθεί από τον χρήστη.

3.5.3 Μηχανή αναζήτησης Bing (Msn Search)

Εικόνα 3.9: Λογότυπο της Bing

Η μηχανή αναζήτησης Bing είναι η νέα μηχανή αναζήτησης της Microsoft. Αποτελεί εξέλιξη της live Search και ξεκίνησε την λειτουργία της τον Ιούνιο του 2009. Αυτή τη στιγμή είναι η δεύτερη σε χρήση μηχανή αναζήτησης στο κόσμο, αφού τον Φεβρουάριο του 2011 ξεπέρασε για πρώτη φορά την μηχανή αναζήτησης της Yahoo. Η Bing κερδίζει συνεχώς έδαφος σε σχέση με τις άλλες μηχανές αναζήτησης. Το 2010 είχε 29% αύξηση στο ποσοστό των αναζητήσεων της σε σχέση με το 2009. Παρόλα αυτά η Google συνεχίζει να κατέχει το συντριπτικό ποσοστό στο

τομέα της αναζήτησης στο διαδίκτυο και θα χρειαστεί πολύ μεγαλύτερη προσπάθεια από πλευράς της Bing για να καταστεί ισχυρός ανταγωνιστής της Google.

3.5.3.1 Η εξέλιξη της μηχανής αναζήτησης Bing

Το φθινόπωρο του 1998 ξεκίνησε την λειτουργία της με την ονομασία «MSN Search» και χρησιμοποιούσε τα αποτελέσματα της Inktomi. Το 2009 άρχισε να χρησιμοποιεί καταλόγους αρχείων από την LookSmart [55] σε συνδυασμό με τα αποτελέσματα της Inktomi. Από τότε η Microsoft αναβάθμισε την μηχανή αναζήτησης της έτσι ώστε να παρέχει τα δικά της αποτελέσματα αναζήτησης. Εμφάνιζε διευθύνσεις διαδικτύου με δείγμα από το περιεχόμενο κάθε σελίδας που ήταν σχετικό με το ερώτημα του χρήστη και οι κατάλογοι της ανανεώνονταν εβδομαδιαίος ή ακόμα και καθημερινός.

Το 2006 η Microsoft αποφάσισε να αντικαταστήσει την «MSN Search» με την «Windows Live Search». Η νέα μηχανή αναζήτησης προσέφερε στους χρήστες τη δυνατότητα να αναζητούν συγκεκριμένους τύπους πληροφορίας όπως ειδήσεις, εικόνες, μουσική κτλ. Σκοπός της ήταν να παρέχει πιο χρήσιμες πληροφορίες στις ερωτήσεις των χρηστών.

Τον Μάρτιο του 2007, η Microsoft ανακοινώσε ότι διαχωρίζει την μηχανή αναζήτησης από τις υπηρεσίες Windows Live, αλλάζοντας το εμπορικό της σήμα σε Live Search. Κάτω από αυτή την νέα ονομασία, έγινε μια σειρά από αναδιοργανώσεις των προσφορών αναζήτησης. Τον Μάιο του 2008,[56] η Microsoft ανακοινώσε την διακοπή λειτουργίας του Live Search Book και του Live Search Academic, ενσωματώνοντας τα αντίστοιχα αποτελέσματα στην κανονική αναζήτηση. Ακλούθησε η διακοπή των υπηρεσιών Windows Live Expro, Live Search Macros, Live Product Upload και τέλος της Live Search QnA.

Η Microsoft διαπίστωσε ότι η παραμονή της λέξης «Live» στο εμπορικό σήμα της μηχανής αναζήτησης ήταν πλέον αχρείαστη. Έτσι στις 3 Ιουνίου του 2009 η μηχανή αναζήτησης Live Search αντικαταστάθηκε επίσημα από την Bing.

Η λέξη Bing είναι ονοματοποιητική, δηλαδή η προφορά της υπενθυμίζει τον ήχο που αντιπροσωπεύει. Συμφώνα με την Microsoft ο ήχος αυτός σχετίζεται με τον ήχο κατά την διάρκεια μιας ανακάλυψης ή της λήψης μιας απόφασης, Επίσης η

ονομασία Bing μνημονεύεται εύκολα, είναι μικρή, εύκολη στη προφορά και επομένως κατάλληλη για να λειτουργήσει ως URL παγκοσμίως.

3.5.3.2 Αποτελέσματα Αναζήτησης

Η Bing για την εύρεση των αποτελεσμάτων της χρησιμοποιεί αλγόριθμους clustering [57]. Τα περισσότερα αποτελέσματα της αναζήτησης παρουσιάζουν σελίδες βασιζόμενες κυρίως στο περιεχόμενο των σελίδων. Η ταξινόμηση των αποτελεσμάτων γίνεται κατά σχετικότητα. Δηλαδή, όσο πιο σχετικό είναι ένα αποτέλεσμα με την ερώτηση του χρήστη, τόσο πιο ψηλά τοποθετείτε στην κατάταξη των αποτελεσμάτων. Προκαθορισμένα εμφανίζονται 2 αποτελέσματα ανά ιστοσελίδα, κάτι το οποίο μπορεί να αλλάξει μέσα από τις ρυθμίσεις σε ένα ή τρία αποτελέσματα ανά ιστοσελίδα.

Ένα σημαντικό χαρακτηριστικό της Bing είναι η δυνατότητα που παρέχει κατά την προηγμένη αναζήτηση να ταξινομούνται οι σχετικές ιστοσελίδες κατά ημερομηνία, βάθος ή και τίτλο. Επίσης μετά από μια αναζήτηση ο χρήστης έχει τη δυνατότητα να προσθέσει και άλλες λέξεις κλειδιά στην υπάρχουσα αναζήτηση, να διαλέξει χώρα προέλευσης της ιστοσελίδας και γλώσσα που είναι γραμμένη η ιστοσελίδα. Ένα άλλο σημαντικό χαρακτηριστικό της μηχανής Bing είναι η προσπάθεια να αναγνωριστεί η τοποθεσία του χρήστη για την τοπική αναζήτηση χωρίς να την εισάγει ο χρήστης.

3.5.3.3 Βάσεις δεδομένων

- **Ιστοσελίδων**: Βάση δεδομένων με ιστοσελίδες του διαδικτύου από όπου αναζητούνται τα αποτελέσματα.
- **Βάση εικόνων.**
- **Ειδήσεων.**
- **Local:** Βάση δεδομένων με τοπικές πληροφορίες για επιχειρήσεις.
- **Βίντεο.**
- **Ακαδημαϊκών θεμάτων.**

Οι διαφημίσεις που εμφανίζονται στα αποτελέσματα δεν έχουν δίκη τους βάση δεδομένων αλλά προέρχονται από το Microsoft AdCenter.

3.5.3.4 Υπηρεσίες Αναζήτησης

Η Bing παρέχει τις πιο κάτω υπηρεσίες αναζήτησης:

- Web: Αναζήτηση ιστοσελίδων στο διαδίκτυο (www.bing.com)
- Images: Αναζήτηση εικόνων στο διαδίκτυο (images.bing.com)
- Videos: Αναζήτηση βίντεο (videos.bing.com)
- Shopping: Αναζήτηση προϊόντων (shopping.bing.com)
- News: Αναζήτηση ειδήσεων (news.bing.com)
- Maps: Αναζήτηση χαρτών (maps.bing.com)
- Travel: Αναζήτηση ταξιδιωτικών πληροφοριών (travel.bing.com)
- Local: Αναζήτηση τοπικών επιχειρήσεων (local.bing.com)
- xRank: Έλεγχος αναζητήσεων που πραγματοποιούνται περισσότερο (xrank.bing.com)

3.5.3.5 Περιορισμοί αναζήτησης

Η Bing δίνει τη δυνατότητα αναζήτησης με χρήση κάποιων περιορισμών:

- **Γλώσσα** – Επιστροφή αποτελεσμάτων συγκεκριμένης γλώσσας. (41 διαθέσιμες γλώσσες)
- **Ημερομηνία** – Επιστροφή αποτελεσμάτων κάποιας χρονικής διάρκειας (τελευταίους 24 ώρες / 1 εβδομάδα / 1 μήνα).
- **Τύπους αρχείων** – Επιστροφή συγκεκριμένων αρχείων τύπου. Γίνεται με χρήση της εντολής «filetype:» ακολουθουμένη από την κατάληξη του αρχείου. (Π.χ. filetype: doc).
- **Δικτυακός χώρος** – Επιστροφή αποτελεσμάτων συγκεκριμένου δικτυακού χώρου (Εντολή: «domain:»).
- **Περιεχόμενα Ενηλίκων**: Οικογενειακό φίλτρο τριών επιπέδων το οποίο εξαιρεί περιεχόμενα ενηλίκων. Προκαθορισμένα το φίλτρο αυτό είναι ενεργοποιημένο στην μεσαία κατάσταση.

3.5.3.6 Επιλογές αναζήτησης

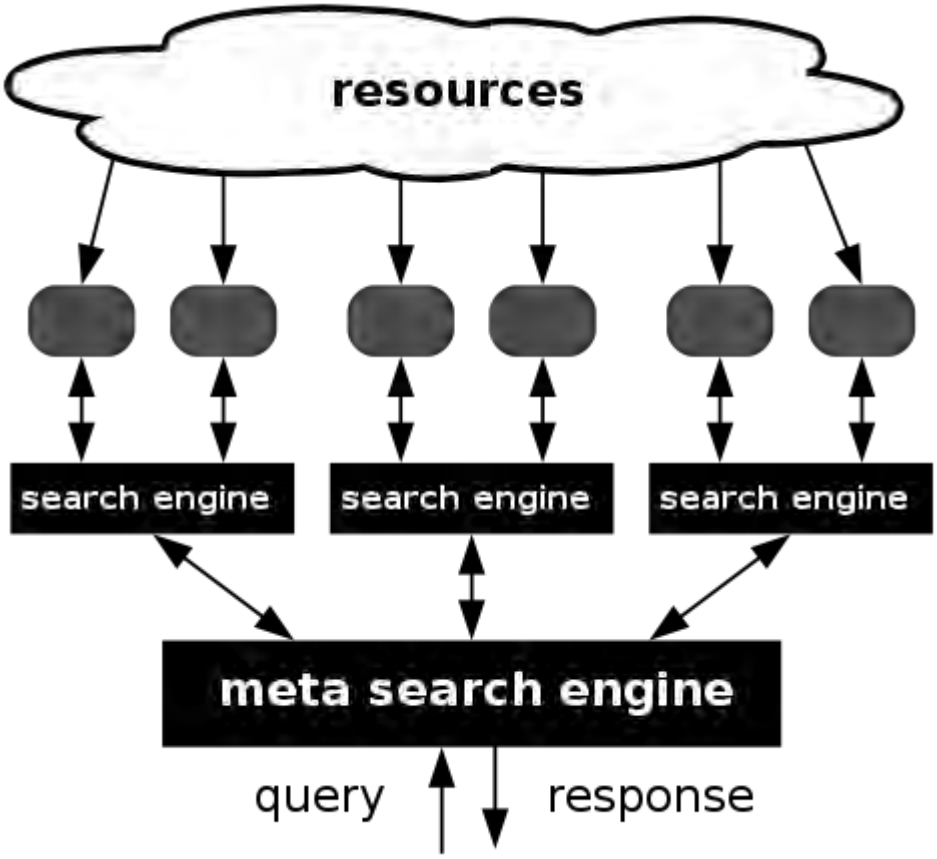
Η Bing όπως και η Google με τη Yahoo, παρέχει κάποιες εξειδικευμένες επιλογές αναζήτησης για εξαγωγή πιο συγκεκριμένων αποτελεσμάτων:

- **«AND» ή «&»** - Αναζήτηση και για το ένα και για το άλλο. Για παράδειγμα η αναζήτηση «Greece AND Cyprus» θα επιστρέψει μόνο αποτελέσματα που είναι σχετικά και με την Ελλάδα και την Κύπρο.
- **«OR» ή «** | **»** - Αναζήτηση για το ένα ή το άλλο. Για παράδειγμα η αναζήτηση «Greece OR Cyprus» θα επιστρέψει μόνο αποτελέσματα που είναι σχετικά ή με την Ελλάδα ή με την Κύπρο.
- **« NOT »** - Αναζήτηση αποκλείοντας μια συγκεκριμένη λέξη. Για παράδειγμα η αναζήτηση για «apple NOT tree» δεν θα επιστρέψει αποτελέσματα που έχουν σχέση με το δέντρο (=tree)
- **« + » Τελεστής πρόσθεσης.** Με την χρήση του τελεστή + αναζητεί ακριβώς όπως δοθεί η λέξη / φράση. Για παράδειγμα η αναζήτηση «πανεπιστήμια +Ελλάδας +πληροφορικής», απαιτεί στα αποτελέσματα να περιέχονται οι λέξεις «Ελλάδας» και «πληροφορικής».
- **« * » Τελεστής πολλαπλασιασμού.** Με τη χρήση του αστερίσκου μέσα σε μια φράση, η αναζήτηση γίνεται για οποιανδήποτε λέξη στη συγκεκριμένη θέση. Για παράδειγμα η αναζήτηση «Bing *» θα επιστρέψει μια πληθώρα αποτελεσμάτων σχετικά με όλες τις υπηρεσίες και προϊόντα που παρέχει η Bing.
- **«site»** – Περιορίζει τα αποτελέσματα της αναζήτησης μόνο από τη συγκεκριμένη ιστοσελίδα που δίνεται.
- **«intitle»** – Απαιτεί να υπάρχει μια συγκεκριμένη λέξη στον τίτλο μιας ιστοσελίδας.
- **«inbody»** – Αναζητεί ιστοσελίδες που περιέχουν μια συγκεκριμένη λέξη στο κυρίως σώμα τους. (Μέσα στην ετικέτα <body>)
- **«inanchor»** - Αναζητεί ιστοσελίδες που περιέχουν μια συγκεκριμένη λέξη μέσα σε ένα σύνδεσμο. (Ετικέτα <href>)
- **«inurl»** - Αναζητεί ιστοσελίδες που περιέχουν μια συγκεκριμένη λέξη στην URL διεύθυνση.

- **«language»** - Ενεργοποιεί τον περιορισμό της γλώσσας των αποτελεσμάτων. Για παράδειγμα language:gr.
- **«loc» ή «location»** - Καθορίζει μια συγκεκριμένη χώρα ή περιοχή για αναζήτηση αποτελεσμάτων όταν ακολουθείται από τους δυο χαρακτήρες της χώρας ή τον κωδικό της περιοχής.
- **«link»** – Εμφανίζει όλες τις ιστοσελίδες που έχουν σύνδεσμο προς την συγκεκριμένη ιστοσελίδα που δίνεται από τον χρήστη.
- **«linkfromdomain»** - Βρίσκει τις εξωτερικές συνδέσεις μιας ιστοσελίδας.
- **«contains»** - Βρίσκει αποτελέσματα που περιέχουν ένα συγκεκριμένο τύπο αρχείων. Για παράδειγμα: mp3.
- « » - Αναζήτηση για μια συγκεκριμένη φράση ακριβώς όπως δοθεί από τον χρήστη.

ΚΕΦΑΛΑΙΟ 4 – ΜΗΧΑΝΕΣ ΜΕΤΑ-ΑΝΑΖΗΤΗΣΗΣ

Οι μηχανές μετά-αναζήτησης (Μ.Μ.Α.) αποτελούν ένα συνδυαστικό εργαλείο πολλών μηχανών αναζήτησης, οπού η έρευνα διεξάγεται παράλληλα, δίνοντας έτσι τη δυνατότητα στο χρήστη με μια ερώτηση να ψάχνει ταυτόχρονα σε περισσότερες από μια μηχανές αναζήτησης. Οι Μ.Μ.Α. δεν διατηρούν δική τους βάση δεδομένων, ευρετήρια ή καταλόγους. Είναι στην ουσία μηχανές αναζήτησης των μηχανών αναζήτησης (Εικόνα 4.1).



Εικόνα 4.1: Δομή λειτουργίας μιας μηχανής Μετά-αναζήτησης.

Κάθε μηχανή αναζήτησης μπορεί να χειριστεί ένα μικρό ποσοστό πληροφοριών από το σύνολο του παγκοσμίου ιστού. Ο συνδυασμός μερικών μηχανών αναζήτησης μπορεί να οδηγήσει σε υψηλότερο ποσοστό κάλυψης και επόμενος σε εύρεση ποιοτικότερων πληροφοριών. Αυτή ακριβώς είναι η λειτουργία μιας Μ.Μ.Α.. Είναι δηλαδή ένα σύστημα το οποίο παρέχει ολοκληρωμένη πρόσβαση σε πολλαπλές

μηχανές αναζήτησης. Όταν ένα ερώτημα εκτελείται σε μια Μ.Μ.Α. το σύστημα περνά το ερώτημα σε πολλαπλές μηχανές αναζήτησης, συλλέγει τα επιμέρους αποτελέσματα και αφού διαγράψει τις διπλοεγγραφές (duplicate entries) τα συγχωνεύει σε ένα ενιαίο κατάλογο κατάταξης. Η συγχώνευση των αποτελεσμάτων αποτελεί το βασικό συστατικό μιας Μ.Μ.Α. Ο αλγόριθμος συγχώνευσης που χρησιμοποιείτε σχετίζεται άμεσα με το ποσό αποτελεσματική είναι μια Μ.Μ.Α.

4.1 Βασικά χαρακτηριστικά μιας Μ.Μ.Α.

Οι Μ.Μ.Α. αποτελούν τη βασική παραλλαγή μιας μηχανής αναζήτησης. Στην ουσία οι Μ.Μ.Α είναι ένας μηχανισμός που συνδυάζει δυο ή περισσότερες μηχανές αναζήτησης και αναλαμβάνει να ψάξει τους ζητούμενους όρους που έθεσε ο χρήστης στα περιεχόμενα όλων των μηχανών επιδιώκοντας με αυτό το τρόπο την καλύτερη δυνατή κάλυψη του θέματος που αναζητείται.

Τα πλεονεκτήματα από τη χρήση μιας Μ.Μ.Α. είναι σημαντικά:

- Ο χρήστης κερδίζει χρόνο εισάγοντας το ερώτημα μόνο μια φορά.
- Επιτυγχάνει καλύτερη κάλυψη, αφού τα ευρετήρια που ερευνούνται δεν ανήκουν σε μια αλλά σε δυο ή περισσότερες βάσεις δεδομένων.
- Με εφαρμογή φίλτρων ή βελτιωμένων αλγορίθμων προς τα ανακτώμενα αποτελέσματα επιτυγχάνεται μεγαλύτερη ακρίβεια.
- Τέλος οι Μ.Μ.Α δεν διαθέτουν δική τους βάση δεδομένων ούτε ευρετήριο, αντί αυτού αναζητούν στα περιεχόμενα των ευρετηρίων των βάσεων δεδομένων που διαθέτουν οι μηχανές αναζήτησης στις οποίες διεξάγεται η έρευνα.

Από την άλλη, οι Μ.Μ.Α είναι τις περισσότερες φορές αργές στην αναζήτηση και στην παρουσίαση των αποτελεσμάτων αφού ερευνούν σε πολλαπλάσιο αριθμό ευρετηριασμένων σελίδων. Επίσης σημαντικό μειονέκτημα των Μ.Μ.Α είναι ότι το ποσοστό των ανακτώμενων πληροφοριών που χρησιμοποιούν από κάθε μηχανή αναζήτησης είναι περίπου 10% της συνολικής πληροφορίας που διαθέτει η εκάστοτε μηχανή αναζήτησης που ερευνάτε. Για παράδειγμα μια Μ.Μ.Α χρησιμοποιεί μόνο τα πρώτα 100 αποτελέσματα του ευρετηρίου της Google για ένα συγκεκριμένο ερώτημα από τα 1000 αποτελέσματα που διαθέτει η Google το ερώτημα αυτό. Επιπρόσθετα οι

Μ.Μ.Α. δεν προσφέρουν τις ευκολίες για παραμετροποίηση της έρευνας που μπορεί να βρει ο χρήστης σε μια μεμονωμένη μηχανή αναζήτησης.

Τα προβλήματα αυτά είναι τα βασικότερα που καλούνται να ξεπεράσουν οι Μ.Μ.Α. με σκοπό την βελτίωση των μετά-αποτελεσμάτων. Στο ερώτημα εάν τελικά αξίζει να χρησιμοποιούμαι τις Μ.Μ.Α. η απάντηση είναι απλή και προφανής. Στις περισσότερες περιπτώσεις, μια Μ.Μ.Α. συγκεντρώνει κάτω από μια διεπαφή χρήστη (interface) μια σειρά από μηχανές αναζήτησης, αλληλοκαλύπτοντας έτσι τα πλεονεκτήματα και τα μειονεκτήματα της καθεμίας, προς όφελος φυσικά του τελικού χρήστη. Έτσι με την χρήση μιας Μ.Μ.Α. ο χρήστης δεν πρόκειται πότε να καταλάβει ότι ο ανιχνευτής ιστού (web crawler) της Excite είναι από τα πιο αργά που υπάρχουν, αφού χρειάζεται ένα μήνα να ολοκληρώσει το κύκλο του, ούτε ότι η Bing ευρετηριάζει περίπου 15 δισεκατομμύρια σελίδες, σε σχέση με την Google που ευρετηριάζει περίπου 44 δισεκατομμύρια σελίδες και τη Yahoo που καλύπτει περίπου 15 δισεκατομμύρια σελίδες. Το σημαντικότερο είναι ότι διαμορφώνοντας μια και μόνο φορά το ερώτημα του, ο χρήστης μπορεί ταυτόχρονα να αναζητήσει πληροφορίες σε όλες τις διαθέσιμες μηχανές αναζήτησης που περιλαμβάνει η μηχανή μετά αναζήτησης που έχει επιλέξει.

Όπως και στις μηχανές αναζήτησης, έτσι και στις Μ.Μ.Α υπάρχουν ορισμένα χαρακτηρίστηκα που συναντά κάποιος στις περισσότερες Μ.Μ.Α. και τείνουν να καθιερωθούν και στις υπόλοιπες. Ακολούθως συνοψίζονται μερικά από τα πιο σημαντικά χαρακτηριστικά που πρέπει να διαθέτει μια Μ.Μ.Α.

- Εύρος αναζήτησης (full text indexing, phrase search).
- Τελεστές Boolean.
- Παρουσίαση αποτελεσμάτων.
- Δυνατότητα εντοπισμού και απαλοιφής διπλών εγγράφων.
- Δυνατότητα ενοποίησης και ταξινόμησης αποτελεσμάτων από διαφορετικές μηχανές.
- Δυνατότητες βελτίωσης αποτελεσμάτων.
- Δυνατότητες περιορισμού.
- Οδηγίες χρήσης και βοήθειας.

Σε περίπτωση που κάποια από τις μηχανές αναζήτησης που χρησιμοποιεί η Μ.Μ.Α διαθέτει full text indexing ή phrase search ενώ κάποιες άλλες όχι, η Μ.Μ.Α. θα

πρέπει να παρέχει την ιδία δυνατότητα στο χρήστη άσχετα αν οι υπόλοιπες μηχανές αναζήτησης δεν διαθέτουν αυτή τη δυνατότητα. Σημαντικό στοιχείο είναι η συγκέντρωση των αποτελεσμάτων και η απαλοιφή των διπλοεγγραφών. Το σημαντικότερο όμως χαρακτηριστικό μιας Μ.Μ.Α. είναι η ενοποίηση και ταξινόμηση αποτελεσμάτων από διαφορετικές πήγες. Η ποιότητα των υπηρεσιών που προσφέρει μια Μ.Μ.Α. εξαρτάτε κυρίως από το ποσό καλή είναι η μέθοδος (αλγόριθμος) που χρησιμοποιείτε για την ενοποίηση, ταξινόμηση και τέλος την παρουσίαση των αποτελεσμάτων.

4.2 Κατηγορίες μηχανών μετά-αναζήτησης

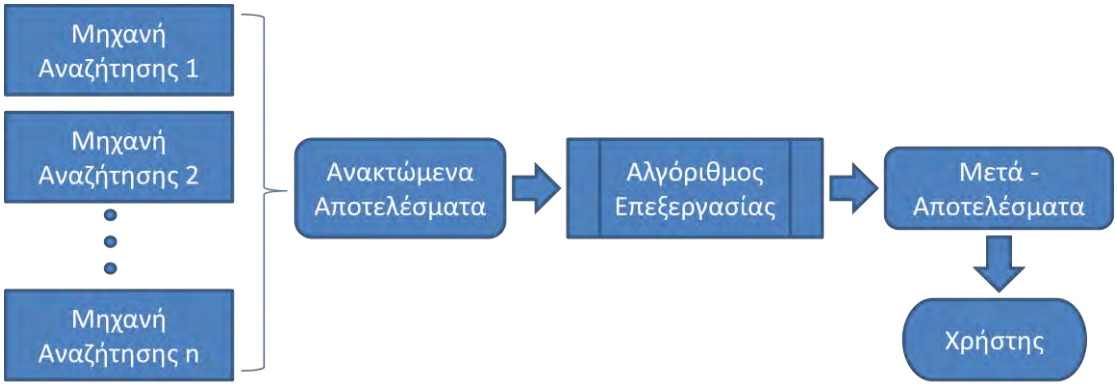
Οι Μ.Μ.Α. μπορούν να κατηγοριοποιηθούν με βάση τις μεθόδους αναζήτησης που χρησιμοποιούν σε 3 κατηγορίες. Μια Μ.Μ.Α. μπορεί να αναζητά τους ζητούμενους όρους ενεργοποιώντας την μια μηχανή μετά την άλλη (σειριακή αναζήτηση), ενεργοποιώντας όλες τις μηχανές ταυτόχρονα (παράλληλη αναζήτηση) ή παραθέτοντας απλά τη λίστα με όλες τις διαθέσιμες μηχανές και τα εργαλεία αναζήτησης, προκειμένου ο χρήστης να επιλέξει αυτά που επιθυμεί (Μ.Μ.Α. με λειτουργία μεσολάβησης). Καθεμία από αυτές τις μεθόδους όπως είναι φυσικό έχει τα πλεονεκτήματα και τα μειονεκτήματα της.

4.2.1 Σειριακή αναζήτηση

Οι περισσότερες Μ.Μ.Α. χρησιμοποιούν την σειριακή αναζήτηση. Ο χρήστης εισάγει στο πεδίο αναζήτησης που υπάρχει το ερώτημα του και ακολούθως ενεργοποιεί (είτε με pull down menus είτε με check boxes) τις μηχανές στις οποίες επιθυμεί να απευθύνει το ερώτημα του. Αφού πραγματοποιηθεί η αναζήτηση, επεξεργάζονται τα ανακτώμενα αποτελέσματα συμφώνα με τον αλγόριθμο επεξεργασίας της εκάστου Μ.Μ.Α. και επιστρέφονται στην οθόνη του χρηστή τα μετά αποτελέσματα.

Το κύριο μειονέκτημα αυτής της μεθόδου είναι ότι, καθώς η αναζήτηση είναι σειριακή, η μια μηχανή ενεργοποιείται μετά την άλλη και μονό όταν η τελευταία επιστρέψει τα αποτελέσματα της θα παρουσιαστεί το σύνολο των αποτελεσμάτων.

Έτσι μπορεί ο χρήστης να εισήγαγε μόνο μια φορά το ερώτημα του, ωστόσο χάνει πολύτιμο χρόνο περιμένοντας όλες της μηχανές να επιστρέψουν τα αποτελέσματα τους.

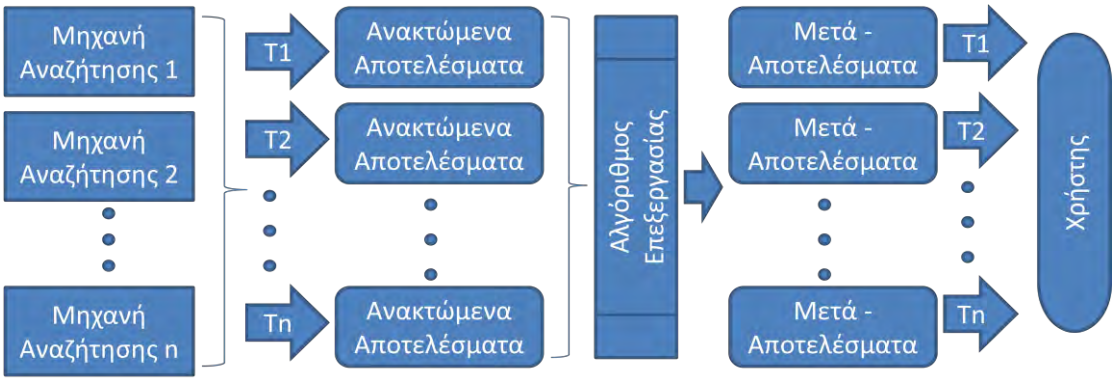


Εικόνα 4.2: Λειτουργία σειριακής αναζήτησης

4.2.2 Παράλληλη αναζήτηση

Οι Μ.Μ.Α. με παράλληλη αναζήτηση μοιάζουν πολύ με αυτές που χρησιμοποιούν την σειριακή αναζήτηση. Η διάφορα είναι ότι η ανάκτηση, επεξεργασία και παρουσίαση των αποτελεσμάτων πραγματοποιείται σταδιακά μέχρι κάποιο ορισμένο χρονικό διάστημα, χωρίς δηλαδή να είναι απαραίτητο να έχουν ανακτηθεί όλα τα αποξέσματα από κάθε ξεχωριστή χρησιμοποιούμενη υπηρεσία αναζήτησης. Αυτό έχει ως αποτελέσματα τα επιστρεφόμενα μετά-αποτελέσματα να είναι πολύ πιο γρήγορα σε σύγκριση με την σειριακή αναζήτηση.

Παρόλα αυτά με αυτή η προσέγγιση, χάνεται ένα σημαντικό ποσοστό ακρίβειας όσον αφορά την σχετικότητα των μετά αποτελεσμάτων. Αυτό συμβαίνει διότι ο αλγόριθμος της κάθε Μ.Μ.Α. επεξεργάζεται και ταξινομεί τα αποτελέσματα που ανακτηθήκαν στο προκαθορισμένο χρονικό διάστημα και όχι επί της συνολικής πληροφορίας που για να ανακτηθεί πιθανώς να χρειάζεται μεγαλύτερο χρονικό διάστημα.

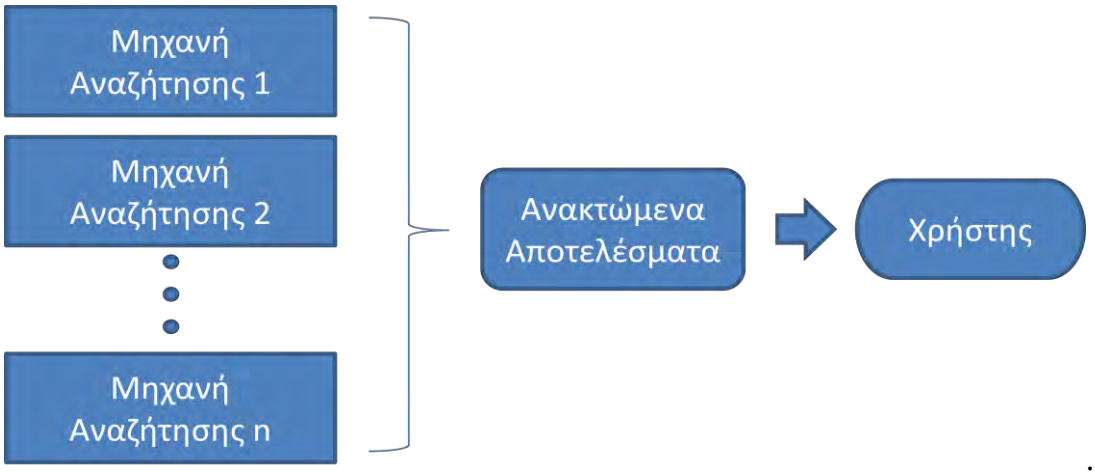


Εικόνα 4.3: Λειτουργία παράλληλης αναζήτησης

4.2.3 Μηχανή Μετά-Αναζήτησης με λειτουργία μεσολάβησης

Οι Μ.Μ.Α. με λειτουργία μεσολάβησης είναι ιστοσελίδες όπου συγκεντρώνουν μια σειρά από μηχανές και εργαλεία αναζήτησης στα οποία ο χρήστης μπορεί να εισάγει το ερώτημα του σε καθεμία από αυτές.

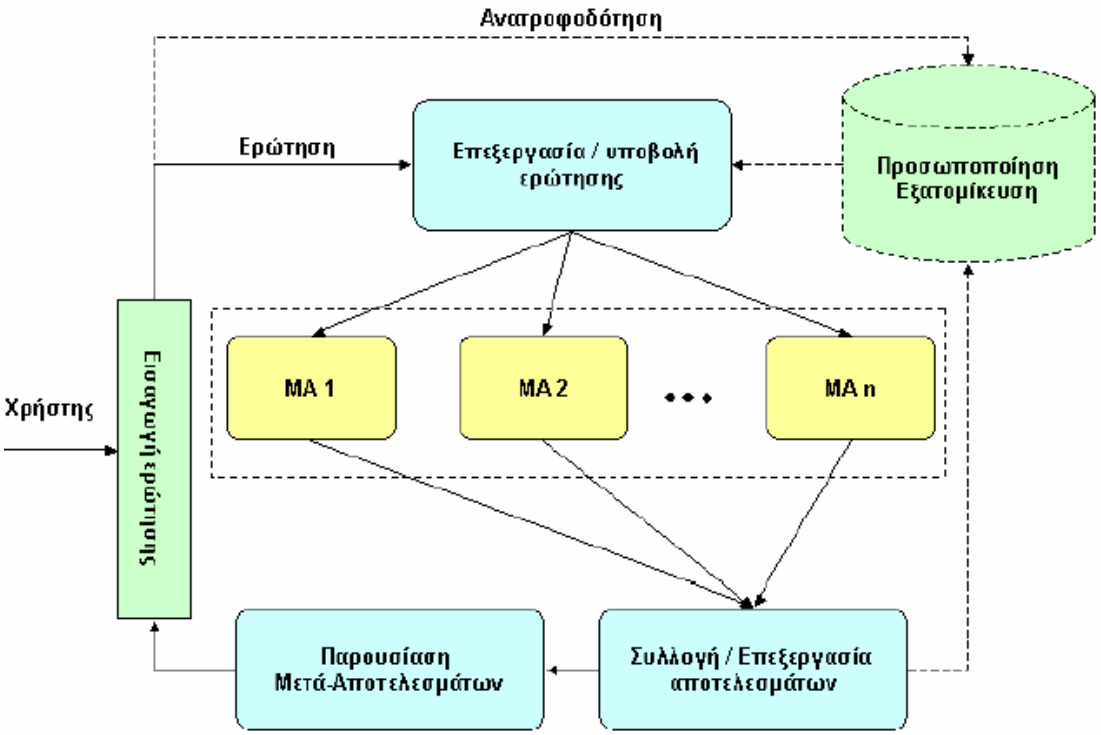
Το πλεονέκτημα της προσέγγισης αυτής είναι ότι ο χρήστης απαλλάσσεται από την ανάγκη να μεταβαίνει από μια ιστοσελίδα μηχανής αναζήτησης σε άλλη προκειμένου να χρησιμοποιήσει όλες αυτές τις μηχανές. Επιπλέον μπορεί να χρησιμοποιήσει και μηχανές αναζήτησης που πιθανώς να μην γνώριζε. Βασικό μειονέκτημα είναι ότι στην ουσία τέτοιες ιστοσελίδες δεν συνιστούν Μ.Μ.Α. αλλά μια συγκέντρωση συστημάτων αναζήτησης προκειμένου ο χρήστης να τα βρίσκει συγκεντρωμένα



Εικόνα 4.4: Λειτουργία Μηχανής Μετά-Αναζήτησης με λειτουργία μεσολάβησης

4.3 Λειτουργιά των Μηχανών Μετά-Αναζητησης

Η λειτουργιά των Μ.Μ.Α. πραγματοποιείται μέσα από 4 στάδια (Εικόνα 4.5). Κατά την διάρκεια κάθε σταδίου ενεργοποιείται και ένα υποσύστημα (Πίνακας 4.1).



Εικόνα 4.5: Στάδια λειτουργίας Μηχανής Μετά-Αναζήτησης

Στάδιο Λειτουργιάς	Υποσύστημα
1. Εισαγωγή ερωτήματος	**Διεπαφή χρήστη**
2. Επεξεργασία και υποβολή ερωτήματος σε πολλαπλές υπηρεσίες αναζήτησης	**Διεκπεραιωτής ή Αποστολέας του ερωτήματος.**
3. Συλλογή και επεξεργασία των αποτελεσμάτων.	**Συλλέκτης αποτελεσμάτων**
4. Παρουσίαση των μετά-αποτελεσμάτων	**Διεπαφή χρήστη**

Πίνακας 4.1: Λειτουργιές και υποσυστήματα των Μηχανών Μετά-Αναζήτησης.

4.3.1 Εισαγωγή ερωτήματος

Μέσω της διεπαφής του χρήστη γίνεται η εισαγωγή του ερωτήματος που αρχικοποιεί την λειτουργία του συστήματος. Σε αυτό το στάδιο ο χρήστης καθορίζει το ερώτημα που θα αποστείλει στην Μ.Μ.Α., τις υπηρεσίες αναζήτησης που θα εμπλακούν και το είδος της πληροφορίας που επιθυμεί να επιστραφεί από την Μ.Μ.Α. (εικόνα, βίντεο, άρθρα κτλ). Επιπλέον, ανάλογα με τις δυνατότητες που παρέχει η κάθε Μ.Μ.Α. ενδέχεται να ορίζει τον αριθμό των αποτελεσμάτων που θα επιστέφονται, την γλωσσά, τη χώρα προέλευσης ή να χρησιμοποιούνται προηγούμενες αναζητήσεις που έχουν ολοκληρωθεί με σκοπό την περαιτέρω εξατομίκευση του ερωτήματος, με την βοήθεια διαδικασιών ανατροφοδότησης και πιθανώς μια τοπικής βάσης δεδομένων.

4.3.2 Επεξεργασία και υποβολή ερωτήματος σε πολλαπλές υπηρεσίες αναζήτησης

Αφού ολοκληρωθεί το πρώτο στάδιο, η Μ.Μ.Α. επεξεργάζεται το ερώτημα διαμορφώνοντας το κατάλληλα έτσι ώστε να είναι κατανοητό σε κάθε υπηρεσία αναζήτησης που θα χρησιμοποιηθεί. Το στάδιο αυτό είναι πολύ σημαντικό διότι ενεργοποιεί την διαδικασία αναζήτησης των αποτελεσμάτων. Κάθε μια από τις μηχανές αναζήτησης που θα χρησιμοποιηθούν πιθανόν να έχουν διαφορετικό τρόπο σύνταξης όσον άφορα το υποβαλλόμενο ερώτημα από το χρήστη. Γι’ αυτό και είναι σημαντικό το ερώτημα να επεξεργάζεται και να διαμορφώνεται κατάλληλα με βάση τις ιδιαιτερότητες της κάθε μηχανής αναζήτησης προτού αποσταλεί σε αυτές. Ακόμη λόγο των συχνών αλλαγών που παρατηρείται στα χαρακτηριστικά των μηχανών αναζήτησης, απαιτείται η συνεχής παρακολούθηση τους και η αντίστοιχη ανανέωση του συστήματος που είναι υπεύθυνο για την λειτουργία αυτή. Το σύστημα το οποίο ενεργοποιείται στο στάδιο αυτό ονομάζεται διεκπεραιωτής ή αποστολέας του ερωτήματος.

4.3.3 Συλλογή και επεξεργασία των αποτελεσμάτων.

Το στάδιο αυτό είναι υπεύθυνο για την ανάκτηση των αποτελεσμάτων από τις υπηρεσίες αναζήτησης που καθορίστηκαν στο προηγούμενο στάδιο με σκοπό την επεξεργασία τους και παραγωγή των μετά-αποτελεσμάτων. Μετά την υποβολή του ερωτήματος στις υπηρεσίες αναζήτησης που χρησιμοποιούνται, ένα υποσύστημα ανακτά τα αποτελέσματα και τα στέλνει πίσω στην Μ.Μ.Α. Το υποσύστημα αυτό ονομάζεται συλλέκτης αποτελεσμάτων και σκοπός του είναι να ανακτά την απαιτούμενη πληροφορία σε όποια μορφή και αν εντοπίζεται στις εκάστοτε υπηρεσίες που χρησιμοποιούνται.

Μετά την ολοκλήρωση της συλλογής των αποτελεσμάτων ακολουθεί η επεξεργασία τους. Οι περισσότερες Μ.Μ.Α. χρησιμοποιούν τοπικές βάσεις δεδομένων για την διαδικασία αυτή. Αρχικά αντιμετωπίζεται το πρόβλημα των διπλότυπων εγγράφων μεταξύ διαφορετικών υπηρεσιών αναζήτησης. Κατά την χρησιμοποίηση διαφορετικών υπηρεσιών αναζήτησης ο χρήστης λαμβάνει πολλές φορές διπλότυπα αποτελέσματα. Το πρόβλημα αυτό αντιμετωπίζεται επαρκώς με χρήση αλγόριθμων που εντοπίζουν τις διπλές εγγραφές και τις διαγράφουν. Παρόλα αυτά, τα διπλότυπα πεδία που παρουσιάζονται στην ίδια υπηρεσία αναζήτησης είναι δύσκολο να αντιμετωπιστούν και συνήθως λαμβάνονται ως ξεχωριστά αποτελέσματα.

Στην συνέχεια αντιμετωπίζεται το πρόβλημα της συγχώνευσης των αποτελεσμάτων. Αφού ολοκληρωθεί η συγχώνευση των αποτελεσμάτων, ενεργοποιείται ο μηχανισμός επεξεργασίας, κατάταξης και παρουσίασης των μετά-αποτελεσμάτων στο χρήστη. Όπως είναι γνωστό, οι αλγόριθμοι ταξινόμησης των αποτελεσμάτων που χρησιμοποιούν οι διάφορες μηχανές αναζήτησης είναι άγνωστοι και μη διαθέσιμοι στους χρήστες. Αυτό έχει ως συνέπεια το πρόβλημα της συγχώνευσης αποτελεσμάτων από διαφορετικές μηχανές αναζήτησης, που χρησιμοποιούν διαφορετικούς αλγόριθμους ταξινόμησης, να είναι ιδιαίτερα δύσκολο. Ακόμα και μηχανές αναζήτησης που χρησιμοποιούν την ίδια προσέγγιση όσο αφορά τη μεθοδολογία των αλγόριθμων ταξινόμησης που χρησιμοποιούν συνήθως παρουσιάζουν διαφορετική κατάταξη για ένα δεδομένο ερώτημα.

Το υποσύστημα το οποίο είναι υπεύθυνο για την επεξεργασία και την ταξινόμηση των ανακτώμενων αποτελεσμάτων, ονομάζεται ταξινομητής ή χειριστής

μετά-αποτελεσμάτων. Το υποσύστημα αυτό κατατάσσει τα ανακτώμενα αποτελέσματα σύμφωνα με της διαδικασίες και τεχνικές που ακολουθεί η κάθε Μ.Μ.Α. όσο αφορά τη συγχώνευση και ταξινόμηση των αποτελεσμάτων.

4.3.4 Παρουσίαση των μετά-αποτελεσμάτων

Η λειτουργιά μιας μηχανής Μ.Μ.Α. ολοκληρώνεται με το στάδιο που είναι υπεύθυνο για την παρουσίαση των μετά-αποτελεσμάτων. Όπως και στο πρώτο στάδιο της εισαγωγής του ερωτήματος, το υποσύστημα που είναι υπεύθυνο για την παρουσίαση των μετά-αποτελεσμάτων είναι η διεπαφή του χρήστη. Γενικά υπάρχει μεγάλη ανομοιογένεια στο τρόπο παρουσίασης των αποτελεσμάτων μεταξύ των διαφόρων μηχανών αναζήτησης. Για παράδειγμα μερικές μηχανές αναζήτησης παραθέτουν απλά τις διευθύνσεις των δικτυακών τόπων που έχουν βρεθεί, ενώ άλλες παρέχουν επιπρόσθετες πληροφορίες ή και προτάσεις. Γενικά οι Μ.Μ.Α. παρουσιάζουν τα μετά-αποτελέσματα σύμφωνα με τις δικές τους προτιμήσεις στο δικό τους περιβάλλον.

4.4 Επισκόπηση στο χώρο των Μηχανών Μετά-Αναζήτησης

Σήμερα, υπάρχουν πολλές Μ.Μ.Α. διαθέσιμες για τους χρηστές. Στη συνεχεία παρουσιάζονται τα βασικά χαρακτηριστικά και οι λειτουργιές μερικών δημοφιλών Μ.Μ.Α.

4.4.1 Dogpile [59]

Η Dogpile ξεκίνησε την λειτουργία της το 1996 και είναι σήμερα μία από τις πιο δημοφιλής Μ.Μ.Α. Η Dogpile παίρνει αποτελέσματα κυρίως από τις Google, Yahoo!, Bing, Ask και About. Συνολικά διαβιβάζει το ερώτημα ενός χρήστη σε 25 μηχανές αναζήτησης. Αυτό που κάνει είναι απλά να προωθεί το ερώτημα του χρήστη στις πιο πάνω μηχανές αναζήτησης ταυτόχρονα. Η Dogpile δεν ταξινομεί τα επεξεργασμένα αποτελέσματα ερώτησης που συλλέγονται από τις παραπάνω μηχανές

αναζήτησης. Τα μετά-αποτελέσματα είναι μια αλληλουχία από διαφορετικά αποτελέσματα με την ταξινόμηση της εκάστοτε υπηρεσίας που χρησιμοποιείται.

4.4.2 Ithaki [60]

Η Ithaki χρησιμοποιεί ένα μεγάλο αριθμό από μηχανές αναζήτησης για την παραγωγή των μετά-αποτελεσμάτων της και είναι διαθέσιμη σε 14 γλώσσες. Προσφέρει στους χρήστες ειδική αναζήτηση για 35 διαφορετικά θέματα κάνοντας χρήση συγκεκριμένων υπηρεσιών για το κάθε θέμα. Υποστηρίζει όλους τους λογικούς τελεστές, απαλείφει τις διπλοεγγραφές και επιπλέον ο χρήστης έχει την δυνατότητα να αποστείλει μέσω email τα μετά-αποτελέσματα.

4.4.3 MetaCrawler [61]

Η MetaCrawler χρησιμοποιεί τις μεγαλύτερες και καλύτερες μηχανές αναζήτησης όπως τις Google, AltaVista, WebCrawler,Yahoo!, Excite, Lycos και άλλες, καλύπτοντας έτσι ένα τεράστιο εύρος πληροφοριακών πηγών. Όπως και οι περισσότερες M.M.A. διαθέτει 2 βασικά interface, ένα για την απλή και ένα για την προχωρημένη αναζήτηση. Στην προχωρημένη αναζήτηση ο χρήστης μπορεί να επιλέξει τις μηχανές αναζήτησης που θα χρησιμοποιηθούν.

Ο χρήστης μπορεί να αναζητήσει χρησιμοποιώντας τελεστές Boolean, να αναζητήσει ως φράση και έχει την δυνατότητα να επιλέξει το χρονικό διάστημα στο οποίο επιθυμεί να επιστραφούν τα αποτελέσματα. Αφού συγκεντρωθούν τα αποτελέσματα από τις πιο πάνω υπηρεσίες αναζήτησης, διαγράφονται οι διπλοεγγραφές και παρουσιάζονται στο χρήστη τα μετά-αποτελέσματα Η κατάταξη των μετά-αποτελεσμάτων γίνεται συμφώνα με τον αλγόριθμο κατάταξης που χρησιμοποιεί η MetaCrawler και ο οποίος δεν είναι διαθέσιμος.

4.4.4 Intelliseek (Profussion) [62]

Η Intelliseek (Profussion) είναι μια M.M.A. με πολλαπλές λειτουργίες. Η Profussion στέλνει τα ερωτήματα των χρηστών σε πολλαπλές μηχανές αναζήτησης

(AltaVista, LookSmart, Maggellan, WebCrawler, GoTo, AllTheWeb και Yahoo) ανακτά και συγχωνεύει τις διευθύνσεις URL που προκύπτουν. Εντοπίζει και αφαιρει διπλότυπα έγγραφα και δημιουργεί ένα ενιαίο πίνακα κατάταξης. Μπορεί αυτόματα να αναλύσει ένα ερώτημα και βασιζόμενο σε αυτή την ανάλυση να επιλέξει τις κατάλληλες μηχανές αναζήτησης για το ερώτημα.

4.4.5 Ixquick Metasearch [63]

Η Ixquick Metasearch ξεκίνησε την λειτουργία της το 1998 και έχει σαν βάση της την Νέα Υόρκη και την Ολλανδία. Η Ixquick προωθεί τα ερωτήματα των χρηστών σε πολλαπλές μηχανές αναζήτησης (Bing, AltaVista, LookSmart, Euroseek, Excite, FindWhat, AllTheWeb, GoTo, HotBot και Yahoo) ταυτόχρονα, φιλτράροντας τα αποτελέσματα και παρουσιάζοντας στο χρήστη μόνο τα δέκα πρώτα. Αυτά τα δέκα προκύπτουν από τα αποτελέσματα που βρίσκονται στις ψηλότερες θέσεις των μηχανών αναζήτησης που χρησιμοποιήθηκαν (μέθοδος Borda). Δηλαδή μια συγκεκριμένη ιστοσελίδα για να βρίσκεται στα πρώτα δέκα μετά-αποτελέσματα της Ixquick, πρέπει να επιστρέφεται από τις περισσότερες μηχανές αναζήτησης που χρησιμοποιεί και να βρίσκεται στις ψηλότερες θέσεις σε κάθε μια από αυτές.

4.4.6 Copernic [64]

Η Copernic δεν είναι μια διαδικτυακή εφαρμογή ή ένας διαδικτυακός τόπος. Πρόκειται για ένα εμπορικό προϊόν το εγκαθίσταται τοπικά σε έναν υπολογιστή. Συνεπώς, ο χρήστης δεν μπορεί να το χρησιμοποιήσει οπουδήποτε στο κόσμο και σε οποιονδήποτε λειτουργικό σύστημα. Χρησιμοποιεί πολλές υπηρεσίες αναζήτησης όπως: η AltaVista, EuroSeek, Bing, Google, Netscape Netcenter, Fast Search, GoTo, Lycos, HotBot, LookSmart, Magellan, DMOZ, Snap, Web Crawler και Yahoo! Σε αρκετές εκδόσεις υποστηρίζει και κάποιες υπηρεσίες αναζήτησης όπως επιλογή αποτελεσμάτων ανά χώρα. Παρέχει δυνατότητες εξατομίκευσης της ερώτησης ενώ υποστηρίζει καταγραφή ιστορικού και Boolean τελεστές. Επιπλέον, δίνεται η δυνατότητα στον χρήστη να έχει ανά πάσα στιγμή πρόσβαση στις σελίδες της προτίμησης του, οι οποίες ενημερώνονται αυτόματα.

4.4.7 Mamma [65]

Η Mamma χρησιμοποιεί ως πηγές τις υπηρεσίες αναζήτησης AltaVista, Excite, Infoseek, Lycos, WebCrawler και Yahoo! Παρέχει υπηρεσίες αναζήτησης, αρχεία ήχου και εικόνες σε άλλες διαδικτυακές εφαρμογές όπως το Usenet και σε καταλόγους ειδήσεων, χρηματιστηρίων ή εταιριών. Τα μετά-αποτελέσματα παρουσιάζονται στο χρήστη σε μια ενιαία μορφή.

4.4.8 MetaFind [66]

Η MetaFind επιστρέφει ένα προκαθορισμένο αριθμό αποτελεσμάτων από κάθε υπηρεσία αναζήτησης που χρησιμοποιεί. Συγκεκριμένα παίρνει δέκα αποτελέσματα από την AltaVista, δέκα από την Excite, πενήντα από την HotBot, είκοσι πέντε από την HotBot, είκοσι πέντε από την Infoseek, τριάντα από την Planetsearch και πενήντα από την WebCrawler. Τα αποτελέσματα μπορούν να ταξινομηθούν αλφαβητικά ή ανά διεύθυνση του Εξυπηρετητή Ονομάτων Τομέα (Domain Name Server). Ακόμη η MetaFind παρέχει την θέση που κατείχε το κάθε αποτέλεσμα στις υπηρεσίες αναζήτησης που χρησιμοποιήθηκαν.

4.4.9 SavvySearch [67]

Η SavvySearch είναι μια από τις παλαιότερες Μ.Μ.Α. Χρησιμοποιεί πολλές διαφορετικές πηγές για άντληση αποτελεσμάτων. Ανακτά αποτελέσματα από μηχανές αναζήτησης, ομάδες Usenet και διάφορες βάσεις δεδομένων με θεματικές κατηγορίες όπως λογισμικό, διασκέδαση, εμπόριο εκπαίδευση καθώς και αρχεία πολυμέσων. Ο χρήστης έχει την δυνατότητα να καθορίσει το επιστρεφόμενο αριθμό αποτελεσμάτων. Σημαντικό μειονέκτημα της SavvySearch είναι ο σχετικά μεγάλος χρόνος απόκρισης.

4.4.10 Highway61 [68]

Η Highway χρησιμοποιεί τελεστές Boolean, δίνει την δυνατότητα στους χρήστες να επιλέξουν τον όγκο τον επιστρεφόμενων αποτελεσμάτων καθώς και το

χρονικό διάστημα μέσα στο οποίο επιθυμούν να προκύψουν τα αποτελέσματα της αναζήτησης τους. Χρησιμοποιεί γλώσσα με μια πιο ανάλαφρη διάθεση και έναν λιγότερο «επαγγελματικό» τρόπο για να αλληλεπίδραση με το χρήστη. Για παράδειγμα στην επιλογή του χρονικού διαστήματος υπάρχουν οι επιλογές «take your time, I'm going to the bathroom» για τους υπομονετικούς χρήστες, ενώ για τους βιαστικούς υπάρχει η επιλογή «hurry up, you losers!». Στην σελίδα των αποτελεσμάτων παρουσιάζονται τα πρώτα δέκα μετά-αποτελέσματα αναφέροντας και τις πηγές προέλευσης για το κάθε μετά-αποτέλεσμα Ο αριθμός και οι υπηρεσίες αναζήτησης που χρησιμοποιεί η συγκεκριμένη M.M.A. δεν αναφέρονται πουθενά, τουλάχιστον εμφανώς.

4.5 Μέθοδοι συγχώνευσης πληροφορίας από ετερογενείς πηγές πληροφορίας [74].

Η συγχώνευση αποτελεσμάτων από ετερογενείς πηγές πληροφορίας αποτελεί ένα από τα σημαντικότερα μέρη μιας M.M.A. Όταν οι πληροφοριακές ανάγκες ενός χρήστη δεν καλύπτονται από μια μόνο πηγή πληροφοριών τότε ο χειρισμός και η συγχώνευση αποτελεσμάτων από διαφορετικές πηγές είναι ιδιαίτερα χρήσιμη. Ακολούθως μελετάται το πρόβλημα συγχώνευσης και χειρισμού των αποτελεσμάτων από διαφορετικές πηγές αναζήτησης και αναλύονται διάφορες μέθοδοι κατάταξης αποτελεσμάτων από διαφορετικές πηγές.

Οι τεχνικές συγχώνευσης χωρίζονται σε 2 κατηγορίες, στις μεθόδους ενσωμάτωσης και στις μεθόδους απομόνωσης. Οι μέθοδοι ενσωμάτωσης χρειάζονται ειδικές πληροφορίες, για την διαδικασία συγχώνευσης των αποτελεσμάτων, από τις πήγες που αντλούν τα αποτελέσματα τους. Ενώ οι μέθοδοι απομόνωσης μπορούν να εφαρμοστούν χωρίς οποιεσδήποτε εξειδικευμένες πληροφορίες από τις πηγές αυτές.

4.5.1 Μέθοδοι ενσωμάτωσης

Οι μέθοδοι ενσωμάτωσης χρησιμοποιούν ειδικά πρωτόκολλα και ορισμένες λειτουργιές εξυπηρετητών για να συγκρίνουν τις στατιστικές μιας συλλογής από πληροφορίες, επιτρέποντας την παραγωγή συγκρίσιμων βαθμών ομοιότητας στα επιστρεφόμενα αποτελέσματα.

4.5.1.1 Μέθοδος σύγκρισης στατιστικών

Μια προσέγγιση είναι να συγκριθούν οι στατιστικές των συλλογών στους εξυπηρετητές αναζήτησης ή στους πελάτες. Και στις δυο περιπτώσεις, οι εξυπηρετητές χρησιμοποιούν αυτά τα στατιστικά, μαζί με ένα ομοιογενή αλγόριθμο ταξινόμησης, για την παραγωγή συγκρίσιμων αποτελεσμάτων. Ακολούθως το σύστημα ταξινομεί τα αποτελέσματα. Πλεονέκτημα αυτής της προσέγγισης είναι ότι επιτρέπεται στους εξυπηρετητές η παραγωγή συγκρίσιμων αποτελεσμάτων. Μειονέκτημα της μεθόδου αυτής είναι το γεγονός ότι οι εξυπηρετητές πρέπει να λειτουργούν με βάση κάποιο πρωτόκολλο επεξεργασίας των χαρακτηριστικών στατιστικής, ενώ παράλληλα πρέπει να υπάρχει επικοινωνία των συστημάτων που εμπλέκονται πριν από την υποβολή της ερώτησης.

4.5.1.2 Μέθοδος παράλληλης παροχής πληροφοριών και αποτελεσμάτων

Μια διαφορετική προσέγγιση απαιτεί από κάθε εξυπηρετητή να παρέχει πληροφορίες συλλογής παράλληλα με τα αποτελέσματα αναζήτησης. Οι πελάτες έχουν την δυνατότητα να συνδυάσουν τις πληροφορίες της συλλογής από τους μεμονωμένους εξυπηρετητές σε γενικευμένες πληροφορίες και να χρησιμοποιήσουν κάποιο αλγόριθμο ταξινόμησης για την παραγωγή των αποτελεσμάτων. Αυτή η προσέγγιση δεν απαιτεί κάποιο συγχρονισμό πριν την υποβολή της ερώτησης, αλλά μπορεί να εφαρμοστεί μόνο όταν οι εξυπηρετητές υποστηρίζουν το απαραίτητο πρωτόκολλο επικοινωνίας.

4.5.2 Μέθοδοι απομόνωσης

Οι μέθοδοι απομόνωσης χρησιμοποιούν τις πληροφορίες που παρέχουν οι εξυπηρετητές χωρίς την απαίτηση οποιασδήποτε ειδικής λειτουργιάς. Υπάρχουν 4 τέσσερα είδη συγχώνευσης που ανήκουν στην μέθοδο απομόνωσης. Συγχώνευση βάσει ανατιθέμενου βαθμού στάθμισης, βάσει δείκτη βαρύτητας εξυπηρετητή, βάσει ακολουθίας κατάταξης και βάσει περιεχομένου.

4.5.2.1 Συγχώνευση βάσει ανατιθέμενου βαθμού στάθμισης

Σε αυτές τις μεθόδους, τα έγγραφα ταξινομούνται βάσει των βαθμών στάθμισης που παρέχονται από έναν εξυπηρετητή. Αποτελέσματα που παράγονται από ταξινομητές που χρησιμοποιούν ετερογενής αλγόριθμους ταξινόμησης και μη-κοινές συλλογές στατιστικής δεν μπορούν να συγκριθούν.

4.5.2.2 Συγχώνευση βάσει δείκτη βαρύτητας εξυπηρετητή

Σε αυτή τη μέθοδο, χρησιμοποιείται ένας δείκτης βαρύτητας, που δίνετε εκ τον προτέρων, σε κάθε εξυπηρετητή. Συνεπώς ο πιο σημαντικός εξυπηρετητής ή αλλιώς η καλύτερη μηχανή αναζήτησης που χρησιμοποιείται παίρνει το μεγαλύτερο δείκτη βαρύτητας. Με αυτό το τρόπο τα αποτελέσματα που προέρχονται από την μεγαλύτερη σε βαρύτητα μηχανή αναζήτησης ταξινομούνται ψηλότερα στην λίστα κατάταξης των μετά-αποτελεσμάτων.

4.5.2.3 Συγχώνευση βάσει ακολουθίας κατάταξης

Οι μέθοδοι συγχώνευσης βάσει ακολουθίας κατάταξης, χρησιμοποιούν την ταξινόμηση της κάθε μηχανής αναζήτησης (εξυπηρετητή) για την παραγωγή των αποτελεσμάτων. Ποιο συγκεκριμένα, οι θέσεις των αποτελεσμάτων στους εξυπηρετητές είναι αυτή που καθορίζει τη ταξινόμηση των μετά-αποτελεσμάτων. Αυτή η τεχνική μπορεί να συνδυαστεί με την τεχνική συγχώνευσης με βάσει δείκτη βαρύτητας εξυπηρετητή, με αποτέλεσμα ο δείκτης βαρύτητας να παίζει σημαντικό ρόλο στη ταξινόμηση των μετά-αποτελεσμάτων.

4.5.2.4 Συγχώνευση βάσει περιεχομένου

Στις μεθόδους αυτές, το σύστημα του πελάτη «κατεβάζει» (download) το σύνολο των εγγράφων από τους εξυπηρετητές, τα αναλύει με βάσει κάποιο αλγόριθμο και με βάσει τον αλγόριθμο αυτό παράγονται τα μετά-αποτελέσματα. Πλεονέκτημα της μεθόδου αυτής είναι ότι η συγχώνευση των αποτελεσμάτων, θα βασιστεί στο

τρέχον περιεχόμενο των εγγράφων που πιθανόν να έχει αλλαγές σε σχέση με το περιεχόμενο που περιείχε όταν κατηγοριοποιήθηκε από τον εξυπηρετητή. Μειονέκτημα της μεθόδου είναι ότι τα έγγραφα πρέπει να μεταφορτωθούν (download) με συνέπεια να εμπλέκονται οι παράμετροι της χρονικής καθυστέρησης και του εύρους ζώνης για κάθε αναζήτηση.

ΚΕΦΑΛΑΙΟ 5 – ΥΛΟΠΟΙΗΣΗ ΜΙΑΣ ΜΗΧΑΝΗΣ ΜΕΤΑ-ΑΝΑΖΗΤΗΣΗΣ ΚΑΙ ΤΕΚΜΗΡΙΩΣΗ ΚΩΔΙΚΑ.

5.1 Εισαγωγή

Η M.M.A. υλοποιήθηκε στη γλωσσά προγραμματισμού της Java και χρησιμοποιήθηκαν οι πλατφόρμες:

- Eclipse [69] (για το κώδικα) και
- Netbeans [70] (για το σχεδιασμό του γραφικού περιβάλλοντος).

Ακόμη, έγινε χρήση της βιβλιοθήκης Jsoup [3] μιας διεπαφής προγραμματισμού (API) ανοικτού κώδικα (Open Source). Το πλαίσιο του API αυτού παρέχει τα κατάλληλα εργαλεία για εξαγωγή και χειρισμό δεδομένων από ένα HTML κώδικα. Η εφαρμογή εκτελείτε τοπικά στον υπολογιστή, με την προϋπόθεση ότι υπάρχει εγκατεστημένη κάποια έκδοση της java. (Διατίθεται δωρεάν στο http://java.com/en/)

Η M.M.A. που υλοποιήθηκε, χρησιμοποιεί τα αποτελέσματα των τριών πιο διαδεδομένων μηχανών αναζήτησης:

- Yahoo
- Google
- Bing

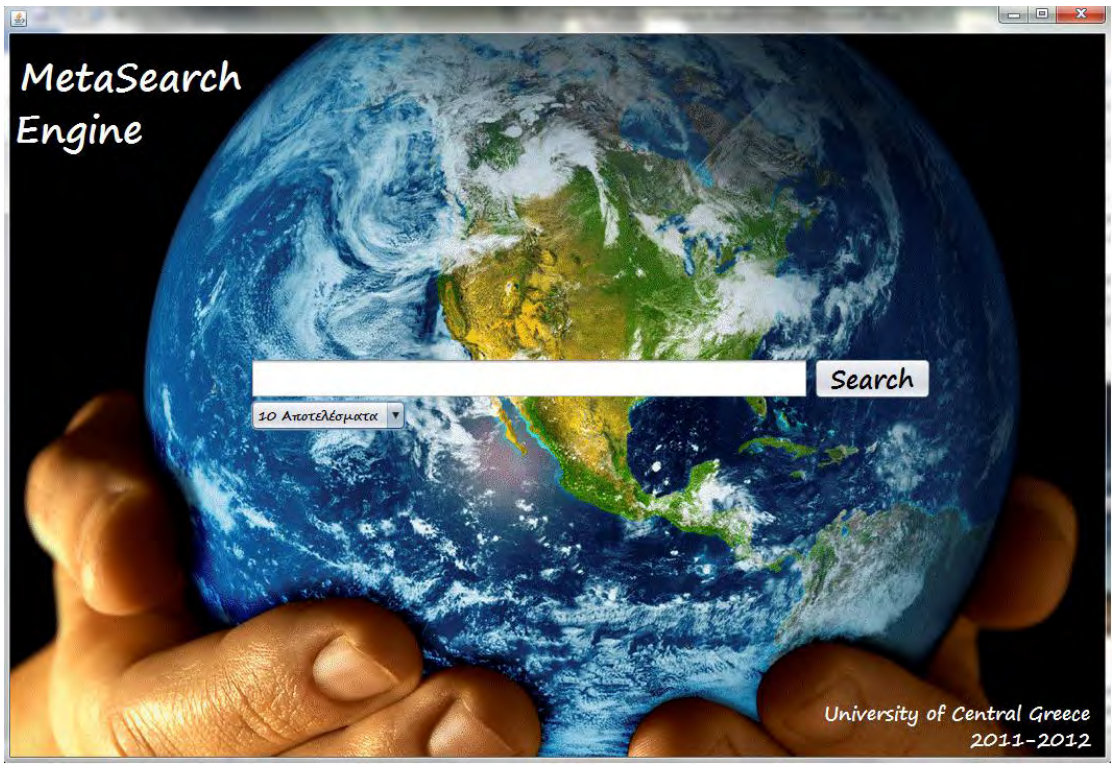
Για την συγχώνευση των αποτελεσμάτων από τις πιο πάνω μηχανές αναζήτησης χρησιμοποιήθηκε η τεχνική του Borda [2].

Σε αυτό το κεφάλαιο, παρουσιάζονται βήμα προς βήμα τα σταδία υλοποίησης της M.M.A καθώς και η τεκμηρίωση του κώδικα που χρησιμοποιήθηκε.

5.2 Διασύνδεση με τον χρήστη

Όταν ο χρήστης «τρέξει» την εφαρμογή, θα εμφανιστεί σε αυτόν η εικόνα 6.1 που είναι η αρχική εικόνα της εφαρμογής μετά-αναζήτησης που υλοποιήθηκε. Αυτή αποτελείται από:

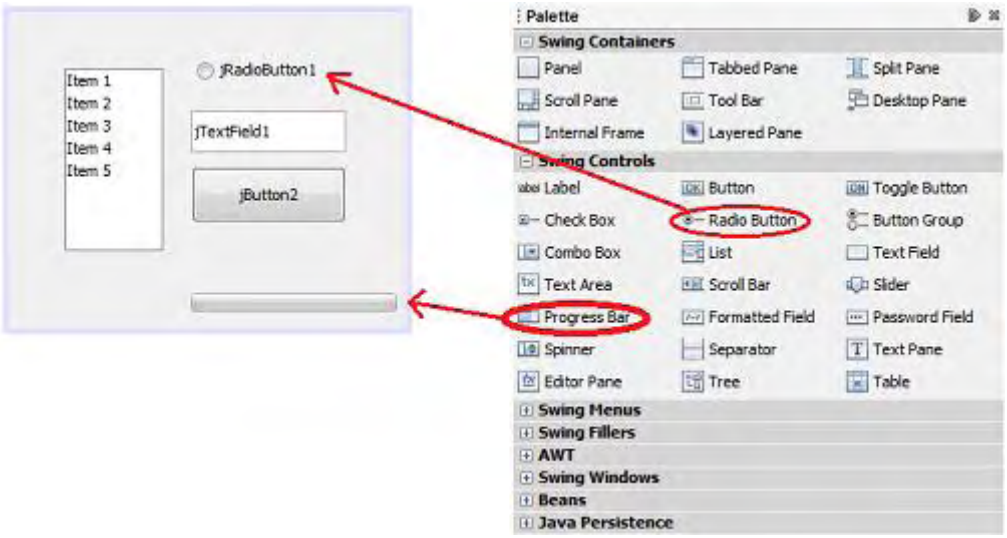
- Ένα πεδίο κειμένου, όπου ο χρήστης εισάγει το ερώτημα του.
- Ένα σύνθετο πλαίσιο όπου ο χρήστης μπορεί να επιλέξει τον αριθμό των επιστρεφόμενων μετά-αποτελεσμάτων που επιθυμεί, με επιλογές 10, 20, 30, 50 και 100.
- Ένα κουμπί αίτησης με όνομα «Search» έτσι ώστε το πρόγραμμα να αρχίσει την αναζήτηση με βάση την λέξη - πρόταση που έδωσε ο χρήστης. Επίσης το πρόγραμμα μπορεί να αρχίσει την αναζήτηση πατώντας το πλήκτρο «Enter» από το πληκτρολόγιο.



Εικόνα 5.1: Αρχική εικόνα εφαρμογής.

Η διεπαφή χρήστη υλοποιήθηκε εξ ολοκλήρου στη πλατφόρμα Netbeans, η οποία παρέχει όλα τα κατάλληλα εργαλεία για εύκολη ανάπτυξη ενός Java GUI (graphical user interface). Πιο συγκεκριμένα, με τη χρήση της «παλέτας» και με την

μέθοδο «drag and drop» [71], μπορείς να επιλέξεις έτοιμα εργαλεία (πινάκες, κουμπιά αιτήσεων κτλ) για την δημιουργία της διεπαφής. Παράλληλα, ο κώδικας παράγεται αυτόματα από τη πλατφόρμα. (εικόνα 5.2)



Εικόνα 5.2: Η παλέτα της πλατφόρμας Netbeans. Ο προγραμματιστής απλά «σέρνει» το εργαλείο που θέλει από την παλέτα και το τοποθετεί στο γραφικό περιβάλλον που δημιουργεί. Το πρόγραμμα αυτόματα παράγει τον κώδικα σε Java.

5.3 Πραγματοποιώντας μια αναζήτηση

Όταν ο χρήστης υποβάλει το ερώτημα του στη Μ.Μ.Α. η εφαρμογή περνά μέσα από διάφορα σταδία μέχρι να επιστραφούν τα μετά-αποτελέσματα πίσω στο χρήστη. Ακολούθως παρουσιάζονται τα διάφορα στάδια μιας αναζήτησης από την χρονική στιγμή που ο χρήστης υποβάλλει το ερώτημα του μέχρι την επιστροφή των αποτελεσμάτων.

5.3.1 Συλλογή παραμέτρων και προετοιμασία για αναζήτηση.

Κατά την πραγματοποίηση μιας αναζήτησης, ενεργοποιείται η κλάση *"Gui.java"* Σε αυτή την κλάση, αρχικά συλλέγεται και αποθηκεύεται σε μια μεταβλητή η βασική παράμετρος της αναζήτησης, δηλαδή το ερώτημα του χρήστη. Στη συνεχεία, δημιουργείται αντικείμενο της κλάσης *"GetResults.java"*.

5.3.2 Ορισμός των URL που περιέχουν τα αποτελέσματα.

Η κλάση *"GetResults.java"* είναι μια από τις κυριότερες κλάσεις της εφαρμογής M.M.A. διότι είναι αυτή που αναζητά τα αποτελέσματα από κάθε μηχανή αναζήτησης.

Σε ένα αλφαριθμητικό με όνομα *"url"*, αποθηκεύουμε το URL, του οποίου ο πηγαίος κώδικας περιέχει τα αποτελέσματα μια μηχανής αναζήτησης.

Για παράδειγμα ο κώδικας:

```
String url = "http://www.google.com/search?num=100&q="+query;
```

,αποθηκεύει σε μια μεταβλητή με όνομα "url", το URL του οποίου ο πηγαίος κώδικας περιέχει τα αποτελέσματα της μηχανής αναζήτησης της Google. Ο διακόπτης "num" χρησιμοποιείται για τον καθορισμό του αριθμού των επιστρεφόμενων αποτελεσμάτων. Προεπιλεγμένα, τα επιστρεφόμενα αποτελέσματα για κάθε μηχανή αναζήτησης οριστήκαν ως τα πρώτα 100. Ο διακόπτης *"q="* χρησιμοποιείται για τον καθορισμό του ερωτήματος του χρήστη. Η μεταβλητή "*query*" είναι ένα αλφαριθμητικό που περιέχει το ερώτημα του χρήστη.

Για την **Google**, όπως προαναφέραμε, το URL αυτό είναι:

url= *"http://www.google.com/search?num=100&q="+query;*

Αντίστοιχα, για την **Yahoo,** το URL αυτό είναι:

url = "http://gr.search.yahoo.com/search?n=100&p="+*query*;

Ομοίως, ο διακόπτης "n" χρησιμοποιείται για τον καθορισμό του αριθμού των επιστρεφόμενων αποτελεσμάτων και ο διακόπτης "p=" χρησιμοποιείται για τον καθορισμό του ερωτήματος του χρήστη.

Τέλος, για την **Bing**, το URL αυτό είναι:

url = "http://www.bing.com/search?q="+*query*+"&count=100";

Ομοίως, ο διακόπτης "count" χρησιμοποιείται για τον καθορισμό του αριθμού των επιστρεφόμενων αποτελεσμάτων και ο διακόπτης "q=" χρησιμοποιείται για τον καθορισμό του ερωτήματος του χρήστη.

5.3.3 Εύρεση αποτελεσμάτων μέσα από τον πηγαίο κώδικα.

Στη συνέχεια καλείτε η μέθοδος της βιβλιοθήκης Jsoup, *"Jsoup.connect(String url)"*, η οποία δέχεται σαν όρισμα το url που δημιουργήσαμε στο προηγούμενο βήμα και δημιουργεί ένα τύπο δεδομένων "document" που περιέχει αποθηκευμένο ολόκληρο τον HTML κώδικα του URL που δόθηκε στη μέθοδο.

Ακόλουθος, χρησιμοποιείται η κλάση της βιβλιοθήκης Jsoup *"selector"*. Η συγκεκριμένη κλάση περιέχει πολλές μεθόδους που βοηθούν σημαντικά στην εύρεση των αποτελεσμάτων. Τα αποτελέσματα αυτά βρίσκονται μέσα σε ετικέτες (tags) που υπάρχουν στον πηγαίο κώδικα της σελίδας με τα αποτελέσματα κάθε μηχανής αναζήτησης. Γι αυτόν τον λόγο υπάρχουν συναρτήσεις που απλά δίνουμε την ετικέτα που θέλουμε και μας επιστρέφει ένα αλφαριθμητικό που περιέχει τα δεδομένα αυτής της ετικέτας.

5.3.3.1 Αναζήτηση πληροφορίας

5.3.3.1.1 Εύρεση αποτελεσμάτων από την Google.

Αφού αποθηκευτεί ο πηγαίος κώδικας της σελίδας των αποτελεσμάτων για την συγκεκριμένη μηχανή αναζήτησης, αρχίζει η αναζήτηση των αποτελεσμάτων. Για κάθε αποτέλεσμα η εφαρμογή αναζητά τον σύνδεσμο προς το αποτέλεσμα.

Στην συνέχεια πειραματιζόμαστε πραγματοποιώντας μια αναζήτηση με ερώτημα «java». Σκοπός του πειράματος είναι να βρούμε σε ποιες ετικέτες βρίσκονται οι σύνδεσμοι των επιστρεφόμενων αποτελεσμάτων. Η μηχανή της Google επιστρέφει, ως προεπιλογή, τα 10 πρώτα αποτελέσματα. Ενδεικτικά η μηχανή της Google επέστρεψε σαν πρώτα 5 αποτελέσματα τους εξής συνδέσμους:

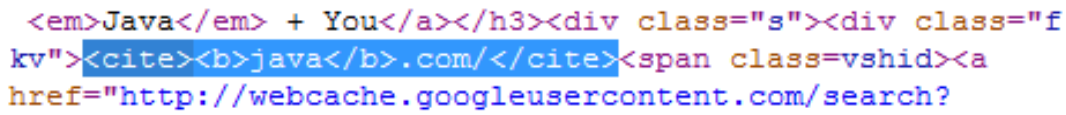
1. java.com/
2. en.wikipedia.org/wiki/Java_(programming_language)
3. en.wikipedia.org/wiki/Java
4. www.oracle.com/technetwork/java/index.html
5. download.oracle.com/javase/tutorial/

Στην εικόνα 5.3 βλέπουμε στιγμιότυπα του αποθηκευμένου πηγαίου κώδικα της Google για μια αναζήτηση με ερώτημα «java».

Εικόνα 5.3: Στιγμιότυπα του πηγαίου κώδικα της Google για αναζήτηση με ερώτημα «java».

Όπως μπορούμε να παρατηρήσουμε, υπάρχει μια ομοιομορφία στο πως είναι χωρισμένα τα αποτελέσματα μεταξύ τους. Πιο συγκριμένα το πρώτο αποτελέσματα είναι αποθηκευμένο σε μια ετικέτα με όνομα *" "*. Μέσα στην ετικέτα αυτή υπάρχει μια άλλη ετικέτα με όνομα *" "* η οποία περιέχει την ετικέτα *<cite>*, η οποία όπως παρατηρούμε περιέχει το σύνδεσμο προς το πρώτο αποτέλεσμα, δηλ. το java.com/. Το συμπέρασμα από το πείραμα είναι ότι στο πηγαίο κώδικα υπάρχουν μόνο 10 ετικέτες * <cite>* και ότι οι σύνδεσμοι των πρώτων 10 αποτελεσμάτων βρίσκονται στις ετικέτες *<cite>* που περιέχουν. Οπόταν πλέον μπορούμε να ξεχωρίσουμε το κάθε ένα αποτέλεσμα.

Στην εικόνα 5.4 βλέπουμε που ακριβώς βρίσκεται ο σύνδεσμος για το πρώτο αποτέλεσμα.



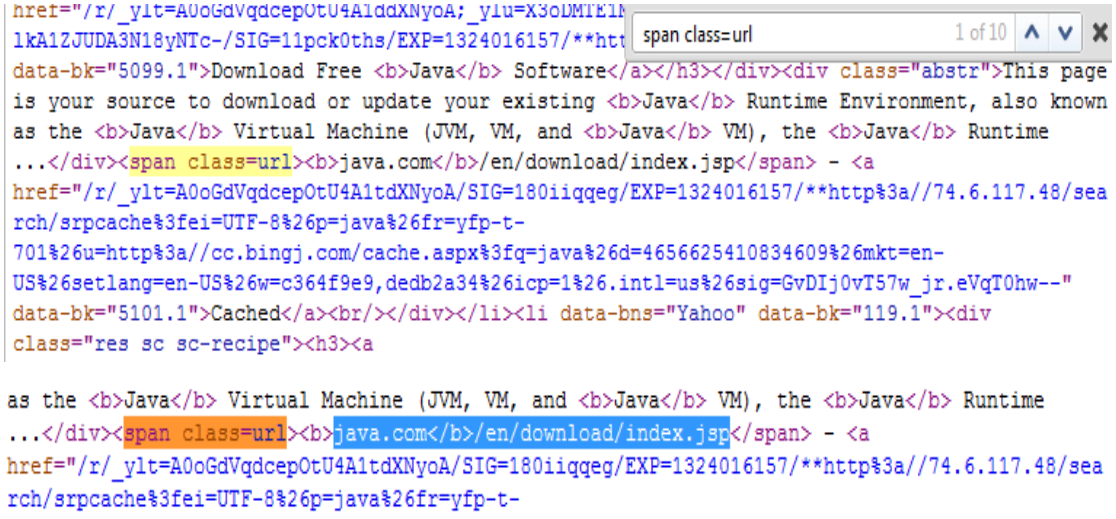
Εικόνα 5.4: Στιγμιότυπο του συνδέσμου προς το πρώτο αποτέλεσμα της Google.

5.3.3.1.2 Εύρεση αποτελεσμάτων από την Yahoo!

Όπως και με την Google, έτσι και με την Yahoo, αποθηκεύουμε αρχικά τον πηγαίο κώδικα των αποτελεσμάτων. Στη συνέχεια εκτελούμε το ίδιο πείραμα για να βρούμε σε ποιες ετικέτες του πηγαίου κώδικα βρίσκονται οι σύνδεσμοι των επιστρεφόμενων αποτελεσμάτων. Αναζητούμε εκ νέου με το ερώτημα «java», αυτή τη φορά στην μηχανή αναζήτησης της Yahoo. Η μηχανή της Yahoo επιστρέφει, ως προεπιλογή, τα 10 πρώτα αποτελέσματα. Ενδεικτικά η μηχανή της Yahoo επέστρεψε σαν πρώτα 5 αποτελέσματα τους εξής συνδέσμους:

1. www.java.com
2. java.com/en/download/index.jsp
3. en.wikipedia.org/wiki/Java_(programming_language)
4. en.wikipedia.org/wiki/Java
5. www.oracle.com/technetwork/java

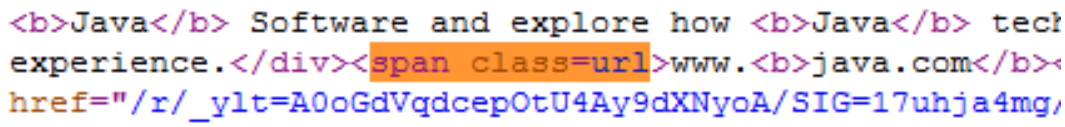
Στην εικόνα 5.5 βλέπουμε στιγμιότυπα του αποθηκευμένου πηγαίου κώδικα της Yahoo για μια αναζήτηση με ερώτημα «java».



Εικόνα 5.5: Στιγμιότυπα του πηγαίου κώδικα της Yahoo για αναζήτηση με ερώτημα «java».

Παρατηρώντας τον πηγαίο κώδικα των αποτελεσμάτων βλέπουμε ότι το κάθε αποτέλεσμα βρίσκεται σε μια ετικέτα με όνομα < *span class=url* >. Στο πηγαίο κώδικα υπάρχουν ακριβώς 10 ετικέτες με αυτό το όνομα, που αντιστοιχούν στα 10 αποτελέσματα.

Στην εικόνα 5.6 βλέπουμε που ακριβώς βρίσκεται ο σύνδεσμος για το πρώτο αποτέλεσμα.



Εικόνα 5.6: Στιγμιότυπο του συνδέσμου προς το πρώτο αποτέλεσμα της Yahoo.

5.3.3.1.3 Εύρεση αποτελεσμάτων από την Bing

Πλέον, έχει απομείνει η συλλογή στοιχείων από τα αποτελέσματα της Bing. Η διαδικασία είναι η ιδία που ακολουθήσαμε στις δυο προηγούμενες μηχανές αναζήτησης. Αρχικά αποθηκεύουμε τον πηγαίο κώδικα των αποτελεσμάτων και στη συνέχεια μελετούμε το κώδικα, εκτελώντας το ίδιο πείραμα, για να βρούμε τους συνδέσμους των επιστρεφόμενων αποτελεσμάτων.

Η μηχανή της Bing επιστρέφει, ως προεπιλογή, τα 10 πρώτα αποτελέσματα. Ενδεικτικά η μηχανή της Bing επέστρεψε σαν πρώτα 5 αποτελέσματα τους εξής συνδέσμους:

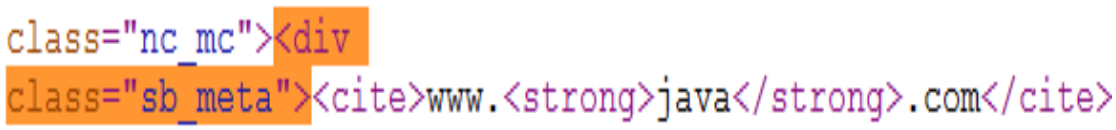
1. www.java.com
2. java.com/en/download/index.jsp

3. en.wikipedia.org/wiki/Java_(programming_language)
4. en.wikipedia.org/wiki/Java
5. www.oracle.com/technetwork/java

Στην εικόνα 5.7 βλέπουμε στιγμιότυπα του αποθηκευμένου πηγαίου κώδικα της Bing για μια αναζήτηση με ερώτημα «java».

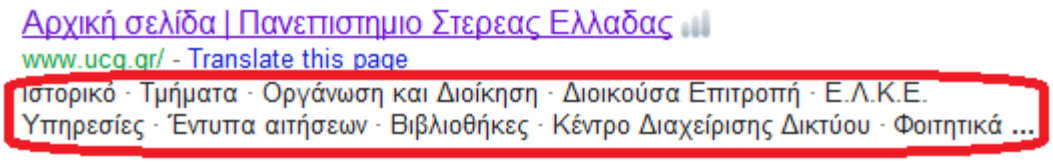
Εικόνα 5.7: Στιγμιότυπα του πηγαίου κώδικα της Bing για αναζήτηση με ερώτημα «java».

Παρατηρώντας τον πηγαίο κώδικα των αποτελεσμάτων βλέπουμε ότι το κάθε αποτέλεσμα βρίσκεται σε μια ετικέτα με όνομα * *, η οποία περιέχει μια άλλη ετικέτα με όνομα *< cite >* που περιέχει το σύνδεσμο προς το πρώτο αποτέλεσμα, δηλ. το "*www.java.com*" , όπως φαίνεται στην εικόνα 5.8.



Εικόνα 5.8: Στιγμιότυπο του συνδέσμου προς το πρώτο αποτέλεσμα της Bing.

Με την ίδια διαδικασία παίρνουμε από κάθε μηχανή αναζήτησης, την περιγραφή που δίνει για το κάθε αποτέλεσμα (εικόνα 5.9).



Εικόνα 5.9: Η περιγραφή που δίνει η Google για την ιστοσελίδα www.ucg.gr.

Τέλος, αφού βρήκαμε σε ποιες ετικέτες βρίσκονται τα αποτελέσματα μέσα από το πηγαίο κώδικα, χρησιμοποιείται η κλάση της βιβλιοθήκης Jsoup *"selector"*, η οποία περιέχει συναρτήσεις που απλά δίνεις την ετικέτα που θέλεις και επιστρέφει ένα αλφαριθμητικό που περιέχει τα δεδομένα αυτής της ετικέτας.

Για παράδειγμα για την επιστροφή των αποτελεσμάτων που περιέχονται στις ετικέτες * <cite>* του πηγαίου κώδικα της Google, χρησιμοποιείται ο ακόλουθος κώδικας:

```
1. Document doc = Jsoup.connect(url).get();
2. Elements links = doc.select(".s cite");
```

Στη πρώτη γραμμή του κώδικα αποθηκεύεται ολόκληρος ο πηγαίος κώδικας, του συνδέσμου (url) που δόθηκε, με όνομα "doc". Στη δεύτερη γραμμή καλείται η συνάρτηση "select", της κλάσης "selector", η οποία επιλέγει τα αποτελέσματα που βρίσκονται στην κλάση "s" και μέσα στην ετικέτα "cite" του "doc" που δημιουργήσαμε στη προηγούμενη γραμμή (*".s cite"*). Τέλος δημιουργείται ένα αλφαριθμητικό με όνομα "links" το όποιο περιέχει μόνο τα περιεχόμενα των ετικετών που θέλουμε. Με αυτό το τρόπο απομονώνουμε από το πηγαίο κώδικα μόνο τις πληροφορίες που χρειαζόμαστε.

Μέχρι αυτό το σημείο, έχουμε αποστείλει το ερώτημα του χρήστη στις τρεις μηχανές αναζήτησης, συλλέξαμε τα αποτελέσματα και τα αποθηκεύσαμε σε τρεις πίνακες, ένας για κάθε μηχανή αναζήτησης. Κάθε πίνακας περιέχει τα πρώτα 100 επιστρεφόμενα αποτελέσματα για την κάθε μηχανή αναζήτησης. Επόμενο βήμα είναι η επεξεργασία των αποτελεσμάτων και η δημιουργία των μετά-αποτελεσμάτων

5.3.4 Συγχώνευση Αποτελεσμάτων

Τελειώνοντας με την συλλογή των αποτελεσμάτων από τις τρεις μηχανές αναζήτησης είμαστε έτοιμοι να περάσουμε στο πιο ουσιαστικό μέρος της εφαρμογής που είναι συγχώνευση αυτών των αποτελεσμάτων.

Σκοπός της συγχώνευσης είναι η παραγωγή ενός μόνο πίνακα αποτελεσμάτων που θα έχει απαλείψει τις διπλότυπες εγγραφές και που θα έχει δώσει σε κάθε αποτέλεσμα μια τιμή θέσης. Στην ουσία, αυτή η τιμή θέσης θα περιγράφει το πόσο καλό είναι ένα αποτέλεσμα έτσι ώστε να το εμφανίσουμε στις πρώτες θέσεις.

Η συγχώνευση των αποτελεσμάτων γίνεται στην κλάση "*MergeResults.java*". Στον κατασκευαστή αυτής της κλάσης περνάμε σαν ορίσματα τους τρεις πίνακες με τα αποτελέσματα από κάθε μηχανή αναζήτησης.

5.3.4.1 Συγχώνευση των τριών πινάκων

Η κλάση "*MergeResults.java*" δέχεται 3 πινάκες που περιέχουν τα αποτελέσματα που δίνει κάθε μια από τις τρεις μηχανές αναζήτησης (πίνακας 5.1). Σε πρώτο στάδιο θα γίνει η συγχώνευση των αποτελεσμάτων. Δηλαδή από το σύνολο των αποτελεσμάτων και των τριών μηχανών αναζήτησης, θα δημιουργήσουμε ένα σύνολο στο οποίο θα έχουν απαλειφθεί οι διπλές εγγραφές του ίδιου αποτελέσματος για τις περιπτώσεις που ένα αποτέλεσμα βρεθεί σε περισσότερες από μια μηχανές αναζήτησης.

Google	Yahoo	Bing
URL 1	**URL 1**	**URL 1**
URL 2	**URL 2**	**URL 2**
.	.	.
.	.	.
.	.	.
URL 100	**URL 100**	**URL 100**

Πίνακας 5.1: Τα πρώτα 100 αποτελέσματα από τις τρεις μηχανές αναζήτησης αρχικά είναι αποθηκευμένα σε 3 διαφορετικούς πίνακες.

Με χρήση της μεθόδου "*System.arraycopy"* της κλάσης Arrays συγχωνεύουμε τους 3 πιο πάνω πίνακες σε ένα. Με το πέρας αυτό του βήματος, έχουμε ένα πίνακα που περιέχει 300 (3*100, πίνακας 5.1) αποτελέσματα όπως φαίνεται και στο πιο κάτω πίνακα (πίνακας 5.2).

Merge Array
URL 1
URL 2
.   .   .
URL 300

Πίνακας 5.2: Πίνακας που περιέχει τα αποτελέσματα και των τριών μηχανών αναζήτησης.

5.3.4.2 Απαλοιφή των διπλότυπων εγγράφων

Επόμενο βήμα είναι η απαλοιφή των διπλότυπων εγγράφων από το πίνακα που δημιουργήσαμε στο προηγούμενο βήμα. Αν δηλαδή δυο αποτελέσματα (για παράδειγμα ένα της Google και ένα της Yahoo) έχουν την ιδία μεταβλητή *"url"*, τότε κρατείται μόνο το ένα αποτέλεσμα.

Για την απαλοιφή των διπλότυπων εγγράφων χρησιμοποιήθηκε μια παραλλαγή του αλγόριθμου ταξινόμησης επιλογής (selection sort) [72]. Ο αλγόριθμος ελέγχει το πρώτο αποτέλεσμα του πίνακα με όλα τα υπόλοιπα, εάν βρει ίδιο αποτέλεσμα τότε το διαγράφει, έπειτα ελέγχεται το δεύτερο αποτέλεσμα του πίνακα με όλα τα υπόλοιπα, η ιδία διαδικασία ακολουθείτε μέχρι το τέλος του πίνακα. Με την ολοκλήρωση του αλγορίθμου ο πίνακας περιέχει N διαφορετικά αποτελέσματα.

N = 300 – M , όπου

- **N**= Αριθμός διαφορετικών αποτελεσμάτων.
- **M**= Διπλότυπα έγραφα.
- **300 =** Τρεις μηχανές αναζήτησης επί 100 αποτελέσματα η κάθε μια.

Τέλος, δημιουργείται ένας πίνακας δυο διαστάσεων με όνομα *«merge»* (πίνακας 5.3) ο οποίος περιέχει Ν γραμμές, όσες δηλαδή και τα διαφορετικά αποτελέσματα και 2 στήλες. Στην πρώτη στήλη είναι αποθηκευμένα τα αποτελέσματα, ενώ η δεύτερη στήλη θα χρησιμοποιηθεί στο επόμενο στάδιο, της βαθμολόγησης των αποτελεσμάτων.

Αποτέλεσμα	**Βαθμολόγια Αποτελέσματος**
URL 1	Score=null
URL 2	Score=null
.	.
.	.
.	.
URL N	Score=null

Πίνακας 5.3: Ο τελικός πίνακας περιέχει Ν διαφορετικά αποτελέσματα και για κάθε ένα από αυτά αντιστοιχεί ένας βαθμός που προς το παρόν παίρνει μηδενική τιμή (null).

5.3.4.3 Βαθμολόγηση και κατάταξη αποτελεσμάτων

Επόμενο στάδιο είναι η βαθμολόγηση των αποτελεσμάτων. Γι αυτό το σκοπό χρησιμοποιήθηκε η τεχνική του "Borda". Η τεχνική αυτή βαθμολογεί τα αποτελέσματα με βάσει τις αποστάσεις που έχουν μεταξύ τους.

Για παράδειγμα ας υποθέσουμε ότι ένα αποτέλεσμα βρέθηκε στη θέση 3 της Google, στην θέση 1 της Yahoo και στη θέση 15 της Bing. Οπόταν σε αυτό το παράδειγμα η μεταβλητή Score = 3+1+15= 19.

Θυμίζουμε ότι για κάθε μηχανή αναζήτησης περνούμε τα 100 πρώτα αποτελέσματα, οπόταν σε περίπτωση που ένα αποτέλεσμα για παράδειγμα βρεθεί στη θέση 3 της Google, στην θέση 1 της Yahoo αλλά δεν υπάρχει στα αποτελέσματα της Bing, τότε η θέση του αποτελέσματος θα δεχθεί ένα είδος «ποινής». Η ποινή αυτή καθορίστηκε στον αριθμό 101, 100 βαθμούς παίρνει το αποτέλεσμα στην θέση 100 μιας μηχανής αναζήτησης, οπόταν 100+1=101 για κάθε αποτέλεσμα που δεν βρίσκεται στα πρώτα 100. Έτσι το Score = (3+1+101)/3=105. Όπως είναι κατανοητό

το αποτέλεσμα με το χαμηλότερο score, θα πάρει την ψηλότερη θέση στη κατάταξη των μετά-αποτελεσμάτων.

Γενικά, η θέση ενός αποτελέσματος σε μια μηχανή αναζήτησης είναι:

$$Score_i^{se} = \begin{cases} r_i^{se}, \text{ αν } 1 \le i \le 100 \\ \\ 100+1=101, \text{ αν } i > 100 \end{cases}$$

όπου:

- r_i^{se} είναι η θέση που βρέθηκε το i-οστό αποτέλεσμα στην μηχανή αναζήτησης,
- 100 ο αριθμός των αποτελεσμάτων που ζητήθηκε από κάθε μηχανή αναζήτησης. 101 για κάθε αποτέλεσμα που δεν βρίσκεται στα πρώτα 100.

Επιπρόσθετα, για τον καθορισμό του score ενός αποτελέσματος δώσαμε ειδικά βάρη σε δυο περιπτώσεις αποτελεσμάτων:

- Α΄ περίπτωση: Εάν ένα συγκεκριμένο αποτέλεσμα βρίσκεται και στις 3 μηχανές αναζήτησης στις θέσεις ≤10, τότε δίνουμε στο αποτέλεσμα αυτό μεγαλύτερη βαρύτητα. Αυτό γίνεται πολλαπλασιάζοντας το score με ένα σταθερό βάρος που το ορίσαμε σαν 0,9. Για παράδειγμα, ένα αποτέλεσμα που βρίσκεται στις θέσεις, 10 της Google, 10 της Yahoo και 10 της Bing, δηλαδή το score=10+10+10=30, θέλουμε να παίρνει μεγαλύτερη βαρύτητα από ένα αποτέλεσμα με το ίδιο score το όποιο όμως βρίσκεται στις θέσεις, 20 της Google, 5 της Yahoo και 5 της Bing score=20+5+5=30. Έτσι πολλαπλασιάζοντας το score με το σταθερό βάρος, δηλαδή score=30×0.9=27, εξασφαλίζουμε ότι το αποτελέσματα που βρίσκεται και στις 3 μηχανές αναζήτησης στις θέσεις ≤10 θα βρίσκεται σε ψιλότερη θέση στην κατάταξη των μετά-αποτελεσμάτων.
- Β΄ περίπτωση – Αντιμετώπιση ελλιπών τιμών: Εάν ένα συγκεκριμένο αποτέλεσμα βρίσκεται στις πρώτες 10 θέσεις δυο εκ των τριών μηχανών αναζήτησης, ενώ δεν υπάρχει στις 100 πρώτες θέσεις της τρίτης μηχανής αναζήτησης, τότε το θεωρούμε σαν ελλιπή τιμή. Αυτό πιθανόν να συμβαίνει για δύο λόγους, είτε η τρίτη μηχανή αναζήτησης δεν κατάφερε να ανακτήσει το συγκεκριμένο αποτέλεσμα είτε το θεωρεί σαν μη σχετικό. Θεωρούμε όμως σε περιπτώσεις όπου το αποτέλεσμα βρέθηκε σε θέσεις ≤10 στις δυο εκ των

τριών, ότι η τρίτη μηχανή αναζήτησης δεν κατάφερε να ανάκτηση το συγκεκριμένο αποτέλεσμα και γι αυτό το λαμβάνουμε σαν ελλιπή τιμή. Σε μια τέτοια περίπτωση για να αντιμετωπίσουμε το πρόβλημα της ελλιπής τιμής, παίρνουμε σαν score τον μέσο όρο των δυο μηχανών αναζήτησης που βρέθηκε το αποτέλεσμα στις θέσεις ≤10. Για παράδειγμα ένα αποτέλεσμα που βρίσκεται στην θέση 3 της Google, 5 της Yahoo και δεν βρίσκεται καθόλου στην Bing, δηλαδή τιμή 101, τότε δίνουμε στο αποτέλεσμα της Bing τον μέσο όρο των Google και Yahoo, δηλαδή (3+5)/2=4, οπόταν θεωρούμε ότι το αποτέλεσμα «έπρεπε» να βρισκόταν στη θέση 4 της Bing.

Σε αυτό το σημείο ο πίνακας 5.3 διαμορφώνεται ως εξής:

Αποτέλεσμα	**Βαθμολόγια Αποτελέσματος**
URL 1	Score=(A+B+Γ)/3
URL 2	Score=(A+B+Γ)/3
. . .	. . .
URL N	Score=(A+B+Γ)/3

Πίνακας 5.4: Ο πίνακας περιέχει N διαφορετικά αποτελέσματα και για κάθε ένα από αυτά αντιστοιχεί ένας βαθμός. Όπου Α η θέση του αποτελέσματος στην Google, Β η θέση του αποτελέσματος στην Yahoo και Γ η θέση του αποτελέσματος στην Bing.

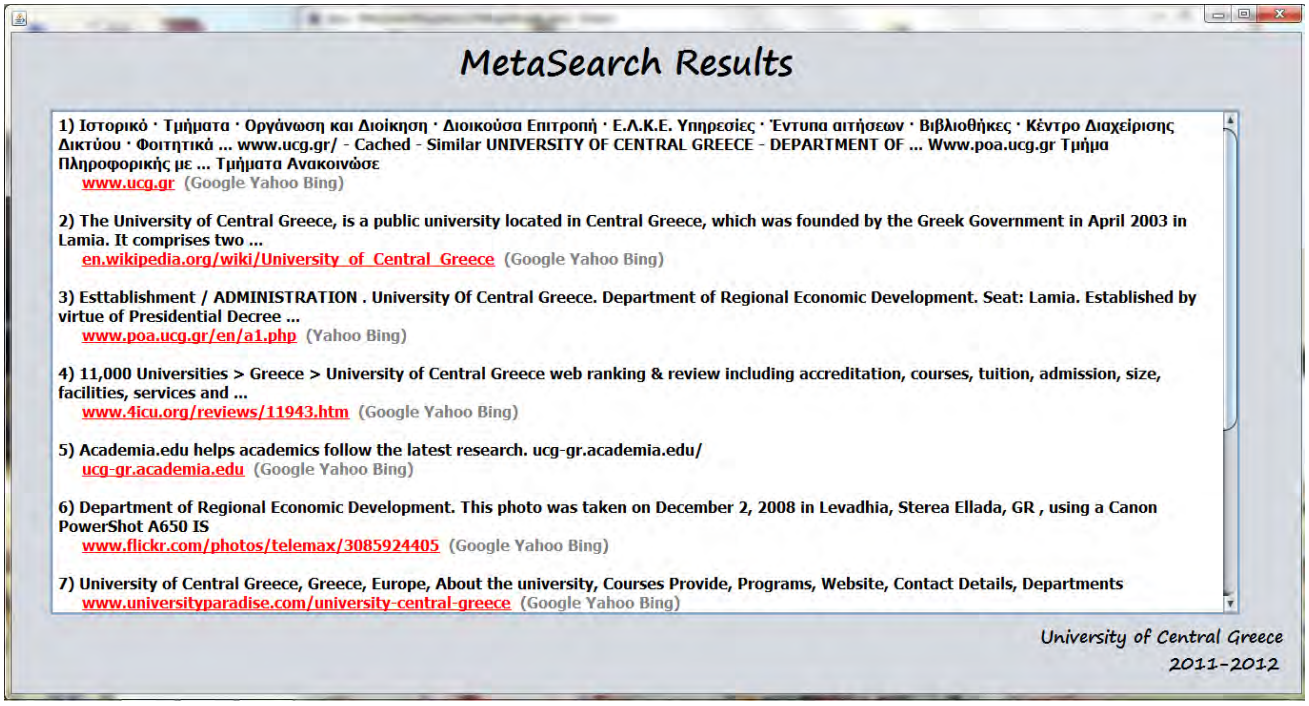
Τέλος, ταξινομούμε τον πιο πάνω πινάκα (πίνακας 5.4) με βάση την δεύτερη στήλη (score). Η ταξινόμηση γίνεται από το μικρότερο score στο μεγαλύτερο με χρήση του αλγόριθμου ταξινόμησης “Bubble sort” [73]. Αφού ολοκληρωθεί η ταξινόμηση, ο πίνακας 5.4 παίρνει την τελική του μορφή (πίνακας 5.5) και απομένει μόνο η παρουσίαση των μετά-αποτελεσμάτων στον χρήστη.

Μετά-αποτελέσματα	**Βαθμολόγια Μετά-αποτελέσματος**
URL 1	Score
URL 2	Score
.	.
.	.
.	.
URL N	Score

Πίνακας 5.5: Ο πίνακας περιέχει N μετά-αποτελέσματα, ταξινομημένα με βάση το score τους από το μικρότερο στο μεγαλύτερο.

5.3.5 Παρουσίαση Μετά-Αποτελεσμάτων

Αφού ολοκληρωθεί η ταξινόμηση απομένει μόνο η παρουσίαση των μετά-αποτελεσμάτων στον χρήστη. Καλείται η κλάση «*GuiResults.java*», η οποία δέχεται σαν παράμετρο τον ταξινομημένο με βάση το score πινάκα merge. Τέλος τα αποτελέσματα εμφανίζονται στο χρήστη (εικόνα 5.6) μέσω μιας διεπαφής χρήστη που σχεδιάστηκε επίσης στη πλατφόρμα «Netbeans».



Εικόνα 5.10: Η διεπαφή χρήστη για την παρουσίαση των μετά-αποτελεσμάτων για αναζήτηση με όρους «university of central Greece».

ΚΕΦΑΛΑΙΟ 6 – ΑΞΙΟΛΟΓΗΣΗ ΑΛΓΟΡΙΘΜΟΥ

6.1 Σύστημα επεξεργασίας της πληροφορίας.

Η παρούσα μηχανή μετά-αναζήτησης αποτελεί ένα Σύστημα Επεξεργασίας της Πληροφορίας (ΣΕΠ). Αυτοματοποιεί τη διαδικασία της επιλογής και της διαχείρισης της πληροφορίας με στόχο την μείωση της υπερφόρτωσης πληροφοριών και την παρουσίαση στον χρήστη ενός απλουστευμένου αλλά πλήρους συνόλου πληροφορίας. Το σύστημα δέχεται ποσά πληροφορίας από τρεις μηχανές αναζήτησης (Google, Yahoo και Bing) και τα επεξεργάζεται υπολογίζοντας παράλληλα ένα βαθμό σχετικότητας αυτών όσον αφορά δυο παραμέτρους:

- Σε πόσες μηχανές αναζήτησης βρέθηκε το κάθε αποτέλεσμα
- Τη θέση των αποτελεσμάτων σε κάθε μια από τις τρεις μηχανές αναζήτησης.

Αυτό που τελικώς παρέχει το σύστημα στο χρήστη είναι ένα σύνολο από σχετικά και μη σχετικά ποσά πληροφορίας. Βασικός σκοπός του συστήματος μας είναι η παρουσίαση στο χρήστη όσο το δυνατόν περισσότερων σχετικών ποσών πληροφορίας

Υπάρχουν τρία βασικά σημεία τα οποία θα συγκεκριμενοποιήσουμε για το σύστημα μας. Αυτά είναι:

1. Η είσοδος.
2. Η έξοδος.
3. Η έννοια της σχετικότητας.

Η είσοδος του συστήματος μας είναι σελίδες του διαδικτύου σε μορφή κειμένου. Πιο συγκεκριμένα, η είσοδος είναι ένα κείμενο που περιέχει τα αποτελέσματα μιας αναζήτησης που πραγματοποιεί κάποιος χρήστης σε κάθε μια από τις τρεις μηχανές αναζήτησης που χρησιμοποιεί η εφαρμογή.

Η έξοδος του συστήματος είναι ένα κείμενο που περιέχει τα μετά-αποτελέσματα. Για την παράγωγη αυτού του κειμένου, το σύστημα επεξεργάζεται την είσοδο κατά μονάδες πληροφορίας, υπολογίζοντας τον βαθμό σχετικότητας τους συμφώνα με τις παραμέτρους που αναφέραμε παραπάνω. Με βάση αυτό το βαθμό σχετικότητας, γίνεται η ταξινόμηση και η παρουσίαση των πληροφοριών στον χρήστη.

Η έννοια της σχετικότητας για το σύστημα μας σχετίζεται στενά με την χρησιμότητα των επιμέρους ποσών πληροφορίας και διαφέρει από χρήστη σε χρήστη. Ένα ποσό πληροφορίας μπορεί να είναι ενδιαφέρον για κάποιο χρήστη αλλά αδιάφορο για κάποιον άλλο.

6.2 Αξιολόγηση απόδοσης και ταξινόμηση

Η αξιολόγηση ενός Σ.Ε.Π. μετράει τη δυνατότητα του συστήματος να ικανοποιεί το χρήστη. Περιλαμβάνει μετρήσεις σχετικά με την αποδοτικότητα και την αποτελεσματικότητα του συστήματος. Η αποτελεσματικότητα είναι ένα μέτρο της δυνατότητας του συστήματος να εντοπίζει την σχετική πληροφορία και να απομονώνει την άσχετη. Όσο πιο αποτελεσματικό είναι ένα σύστημα τόσο περισσότερο ικανοποιεί το χρήστη. Όμως ένα αποτελεσματικό σύστημα πρέπει να είναι και φιλικό προς τον χρήστη. Εδώ εισέρχεται η έννοια της αποδοτικότητας του συστήματος. Η αποδοτικότητα είναι ένα μέτρο απόδοσης σε σχέση με τους πόρους που καταναλώνονται για να παραχθεί η έξοδος.

Η απόφαση ταξινόμησης της εξόδου έχει να κάνει με το αν ένα έγγραφο είναι σχετικό ή όχι για τον χρήστη συμφώνα με μια ορισμένη ανάγκη πληροφοριών. Όπως αναφέραμε παραπάνω, η σχετικότητα είναι μια υποκειμενική έννοια διότι οι κρίσεις διαφορετικών χρηστών για τη σχετικότητα συγκεκριμένης πληροφορίας μπορούν να διαφέρουν σημαντικά. Ως εκ τούτου, η αξιολόγηση ενός Σ.Ε.Π. δεν αποτελεί μια απλή υπόθεση. Εξαρτάται από την ανθρώπινη κρίση και δεν είναι πάντοτε αντικειμενική. Το σύστημα μας αυτοματοποιεί την διαδικασία της ταξινόμησης με βάση τον βαθμό της σχετικότητας που υπολογίζεται για κάθε αποτέλεσμα.

6.3 Μεγέθη αξιολόγησης: Ανάκληση – Ακρίβεια

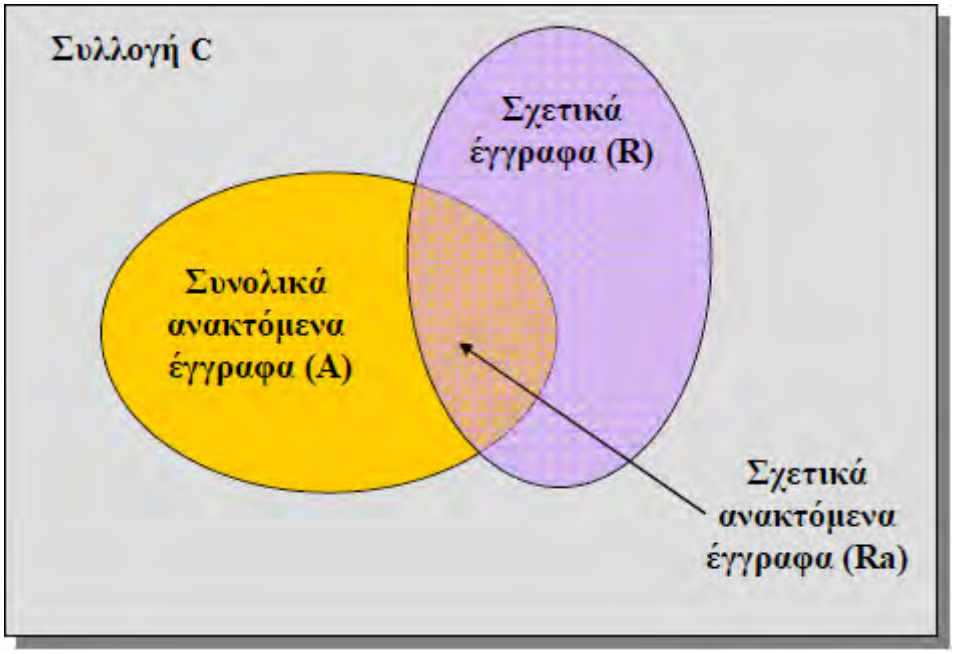
Δείκτες για την μέτρηση της απόδοσης ενός Σ.Ε.Π. είναι η ανάκληση (recall) και η ακρίβεια (precision). Αυτοί οι δείκτες δεν σχετίζονται με τον συνολικό αριθμό των μονάδων πληροφορίας και μπορούν έτσι να προσδιορίσουν την ακρίβεια και την αποτελεσματικότητα του συστήματος.

Ο όρος ανάκληση σε προβλήματα επεξεργασίας σχετίζεται με το αν η πληροφορία είναι σχετική ή όχι όσον αφορά της απαιτήσεις του χρήστη. Είναι μια εκτίμηση της πιθανότητας που προσδίδει το σύστημα μέσω των σχετικών μονάδων πληροφορίας στον χρήστη. Ακρίβεια είναι μια εκτίμηση της πιθανότητας ότι μια μονάδα που παρουσιάζεται στο χρήστη είναι πράγματι σχετική. Οι μετρήσεις των παραπάνω δεικτών χρησιμοποιούνται κυρίως για την εκτίμηση της απόδοσης αλγορίθμων ανάκτησης.

Για τον υπολογισμό αυτών των δεικτών, έστω C μια συλλογή από έγγραφα και q η ερώτηση του χρήστη προς τη συλλογή αυτή. Έστω επίσης A τα συνολικά ανακτώμενα έγγραφα σε σχέση με την ερώτηση q και R το σύνολο όλων τον σχετικών εγγράφων ως προς την ερώτηση q. Τότε το μέγεθος της Ανάκλησης ορίζεται ως ο λόγος του αριθμού των σχετικών ανακτώμενων εγγράφων Ra (το σύνολο που προκύπτει από την τομή των συνόλων A και R) προς το σύνολο των σχετικών εγγράφων R. Το μέγεθος της Ακριβείας ορίζεται ως ο λόγος του αριθμού των σχετικών ανακτώμενων εγγράφων Ra προς τον αριθμό των ανακτώμενων εγγράφων A, σε σχέση με την ερώτηση q. Δηλαδή:

- **Ανάκληση = Ra / R**
- **Ακρίβεια = Ra / A**

Στο σύστημα μας, η συλλογή από έγγραφα C είναι το σύνολο των αποτελεσμάτων για οποιοδήποτε ερώτημα q. Το σύνολο A αποτελείται από τα ανακτώμενα αποτελέσματα για ένα συγκεκριμένο ερώτημα q και R είναι το σύνολο των σχετικών αποτελεσμάτων, δηλαδή των αποτελεσμάτων που ενδιαφέρουν τον χρήστη. Από τα παραπάνω προκύπτει ότι το Ra είναι το σύνολο των σχετικών αποτελεσμάτων από το σύνολο των ανακτώμενων για ένα ερώτημα q και είναι υποσύνολο του R και του A. (Εικόνα 6.1)



Εικόνα 6.1: Ανάκληση και Ακρίβεια σε επίπεδα συνόλων

6.4 Αξιολόγηση συστημάτων

Για την αξιολόγηση του αλγορίθμου που χρησιμοποιεί το σύστημα μας, πραγματοποιήσαμε τέσσερεις αναζητήσεις σε κάθε μια από τις τρεις μηχανές αναζήτησης που χρησιμοποιούνται από το σύστημα μας, με διαφορετικό κάθε φορά ερώτημα, ζητώντας είκοσι αποτελέσματα από κάθε ερώτημα. Τα ερωτήματα που ζητάμε είναι:

- university of central Greece
- platres restaurants
- hyundai ix35 vs kia sportage
- ael football club Cyprus

Στη συνέχεια για κάθε αναζήτηση που πραγματοποιούμε υπολογίζουμε την ανάκληση και την ακρίβεια. Για τον υπολογισμό αυτών των δεικτών αρχικά βλέπουμε ποια από τα είκοσι εμφανιζόμενα αποτελέσματα είναι σχετικά για εμάς, δηλαδή ποια μας ικανοποιούν. Με βάση τον αριθμό των σχετικών αποτελεσμάτων αλλά και την θέση αυτών υπολογίζουμε για κάθε αποτέλεσμα την ακρίβεια και την ανάκληση.

Αφού τελειώσουμε με τον υπολογισμό αυτών των δεικτών για τις τρεις μηχανές αναζήτησης, πραγματοποιούμε την ίδια διαδικασία για την μηχανή μετά-

αναζήτησης που υλοποιήσαμε. Σκοπός της παραπάνω διαδικασίας είναι η αξιολόγηση της εφαρμογής μετά-αναζήτησης σε σύγκριση με τις τρεις μηχανές αναζήτησης που χρησιμοποιούνται.

6.4.1 Αξιολόγηση Google

Αρχικά, θα αξιολογήσουμε τα αποτελέσματα της Google για κάθε μια από τις τέσσερεις αναζητήσεις. Στον Πινάκα 6.1 βλέπουμε τον υπολογισμό της ανάκλησης και της ακρίβειας για μια αναζήτηση με ερώτημα «university of central Greece».

Ερώτημα: university of central Greece					
Σχετικά Αποτελέσματα: {r1, r2, r3, r5, r6, r7, r8, r10, r11, r12, r13, r14, r16, r17,r19, r20}					
Σύνολο Σχετικών: 16					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	ΝΑΙ	1	6,25	100,00
2	r2	ΝΑΙ	2	12,50	100,00
3	r3	ΝΑΙ	3	18,75	100,00
4	r4	ΟΧΙ	3	18,75	75,00
5	r5	ΝΑΙ	4	25,00	80,00
6	r6	ΝΑΙ	5	31,25	83,33
7	r7	ΝΑΙ	6	37,50	85,71
8	r8	ΝΑΙ	7	43,75	87,50
9	r9	ΟΧΙ	7	43,75	77,77
10	r10	ΝΑΙ	8	50,00	80,00
11	r11	ΝΑΙ	9	56,25	81,81
12	r12	ΝΑΙ	10	62,50	83,33
13	r13	ΝΑΙ	11	68,75	84,61
14	r14	ΝΑΙ	12	75,00	85,71
15	r15	ΟΧΙ	12	75,00	80,00
16	r16	ΝΑΙ	13	81,25	81,25
17	r17	ΝΑΙ	14	87,50	82,35
18	r18	ΟΧΙ	14	87,50	77,77
19	r19	ΝΑΙ	15	93,75	78,94
20	r20	ΝΑΙ	16	100,00	80,00

Πινάκας 6.1: Ανάκληση και ακρίβεια ερωτήματος «university of central Greece» για την Google.

Έστω S μια ομάδα που περιέχει ανακτώμενα αποτελέσματα. Πρώτα θεωρούμε ότι η ομάδα S δε περιέχει κανένα ανακτώμενο αποτέλεσμα, δηλαδή S={}. Σε αυτή την ομάδα φυσικά δεν περιέχεται κάποιο σχετικό αποτέλεσμα οπότε το ποσοστό της ανάκλησης και της ακρίβειας είναι 0%. Στη συνεχεία προσθέτουμε στην ομάδα S το πρώτο ανακτώμενο αποτέλεσμα, δηλαδή S={r1}. Αυτό το αποτέλεσμα είναι σχετικό για εμάς, οπότε τώρα η ομάδα S περιέχει 1 σχετικό αποτέλεσμα. Όπως αναφέραμε παραπάνω, η ανάκληση είναι ο λόγος του αριθμού των σχετικών αποτελεσμάτων Ra που περιέχει η ομάδα S ως προς το σύνολο των σχετικών για τον χρήστη αποτελεσμάτων R. Ο αριθμός των σχετικών αποτελεσμάτων της ομάδας S είναι 1, δηλαδή Ra=1 και τα σχετικά αποτελέσματα είναι 16, δηλαδή R=16.
Επομένως, το ποσοστό της ανάκλησης της ομάδας S είναι:

- Ανάκληση % = (Ra/R) * 100 = (1/16) * 100 = 6,25%

Αντίστοιχα, η ακρίβεια είναι ο λόγος του αριθμού των σχετικών αποτελεσμάτων Ra που περιέχει η ομάδα S ως προς το σύνολο των ανακτώμενων αποτελεσμάτων της ομάδας S είναι 1, δηλαδή A=1. Άρα:

- Ακρίβεια % = (Ra/A) * 100 = (1/1) * 100 = 100%

Συνεχίζοντας, προσθέτουμε στην ομάδα S το δεύτερο ανακτώμενο αποτέλεσμα, δηλαδή πλέον S={r1,r2}. Το αποτέλεσμα r2 είναι και αυτό σχετικό για εμάς, οπότε τώρα η ομάδα S περιέχει 2 σχετικά αποτελέσματα (Ra=2). Επίσης, πλέον η ομάδα S περιέχει 2 ανακτώμενα αποτελέσματα, δηλαδή A=2. Από τα παραπάνω προκύπτει ότι:

- Ανάκληση % = (Ra/R) * 100 = (2/16) * 100 = 12,50%
- Ακρίβεια % = (Ra/A) * 100 = (2/2) * 100 = 100%

Προσθέτοντας το τρίτο στη σειρά ανακτώμενο αποτέλεσμα δηλαδή πλέον S={r1,r2,r3}. Το αποτέλεσμα r3 είναι και αυτό σχετικό για εμάς, οπότε τώρα η ομάδα S περιέχει 3 σχετικά αποτελέσματα (Ra=3). Επίσης, πλέον η ομάδα S περιέχει 3 ανακτώμενα αποτελέσματα, δηλαδή A=3. Από τα παραπάνω προκύπτει ότι:

- Ανάκληση % = (Ra/R) * 100 = (3/16) * 100 = 18,75%
- Ακρίβεια % = (Ra/A) * 100 = (3/3) * 100 = 100%

Προσθέτοντας το τέταρτο στη σειρά ανακτώμενο αποτέλεσμα η ομάδα S γίνεται S={r1,r2,r3,r4}. Τώρα ο αριθμός των ανακτώμενων αποτελεσμάτων της ομάδας S είναι 4, (A=4). Το νέο ανακτώμενο αποτέλεσμα r4 δεν είναι σχετικό για

εμάς, οπότε ο αριθμός των σχετικών αποτελεσμάτων της ομάδας S παραμένει 3 (Ra=3). Συμφώνα με τα παραπάνω για την ομάδα S προκύπτει:

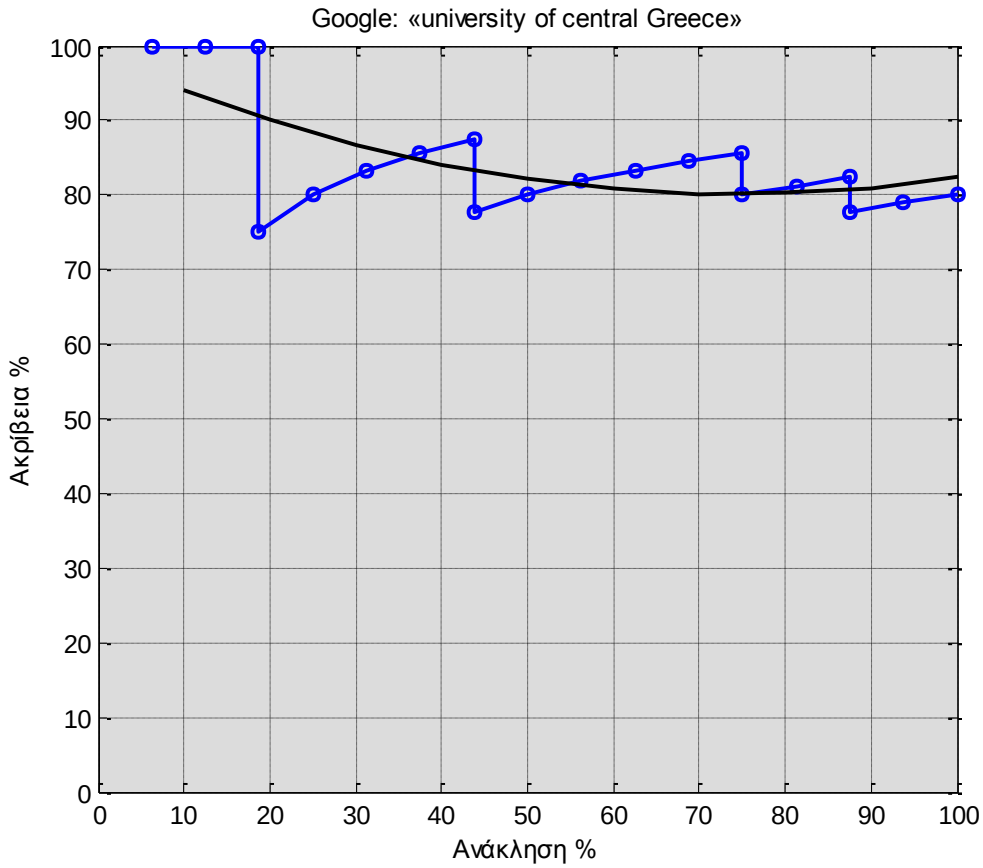
- Ανάκληση % = (Ra/R) * 100 = (3/16) * 100 = 18,75%
- Ακρίβεια % = (Ra/A) * 100 = (3/4) * 100 = 75%

Συνεχίζοντας με τον ίδιο τρόπο, προκύπτουν τα αποτελέσματα που βλέπουμε στο πίνακα 6.1. Παρατηρούμε ότι προσθέτοντας τα ανακτώμενα αποτελέσματα στο σύνολο S, η ανάκληση αυξάνεται συνεχώς και φτάνει το 100% όταν προσθέτουμε στην ομάδα S και το τελευταίο ανακτώμενο αποτέλεσμα. Αυτό είναι αναμενόμενο διότι στο τέλος το σύνολο S περιέχει όλα τα σχετικά για εμάς αποτελέσματα, δηλαδή Ra=R=16. Επίσης, παρατηρούμε ότι η ακρίβεια αυξάνεται όταν προσθέτουμε στην ομάδα S σχετικά ανακτώμενα αποτελέσματα και μειώνεται όταν προσθέτουμε μη σχετικά αποτελέσματα.

Για την καλύτερη κατανόηση των παραπάνω δημιουργούμε την πλήρη καμπύλη ανάκλησης – ακρίβειας. Αρχικά αναπαριστάνουμε το Πίνακα 6.1 σε διάγραμμα. Για κάθε ανακτώμενο έγγραφο, οι τιμές της ανάκλησης και της ακρίβειας αναπαριστούν ένα σημείο (άξονας X η ανάκληση, άξονας Y η ακρίβεια). Συνολικά έχουμε 20 σημεία τα οποία όταν ενωθούν σχηματίζουν το διάγραμμα. Με βάση αυτό το διάγραμμα, κατασκευάζουμε την γραμμή τάσης πολυωνυμικού τύπου διάταξης 2 (2^{nd} order of polynomial type). Από αυτή τη καμπύλη μπορούμε να βγάλουμε χρήσιμα συμπεράσματα για την αποτελεσματικότητα και την απόδοση της μηχανής αναζήτησης για τον συγκεκριμένο χρήστη.

Στο Διάγραμμα 6.1 παρατηρούμε ότι καθώς ανεβαίνει ο δείκτης της ανάκλησης, ο οποίος στην ουσία είναι το ποσοστό εύρεσης σχετικών αποτελεσμάτων από το σύνολο των σχετικών αποτελεσμάτων, μειώνεται ο δείκτης της ακρίβειας, δηλαδή η πιθανότητα το ανακτώμενο αποτέλεσμα να είναι πράγματι σχετικό. Η καμπύλη που δημιουργείται έχει σχετικά μικρή κλίση και ο δείκτης της ακρίβειας κυμαίνεται μεταξύ 80% και 95%.

Από τα παραπάνω συμπεραίνουμε ότι το μεγαλύτερο ποσοστό των σχετικών αποτελεσμάτων βρίσκεται στα αρχικά και μεσαία ανακτώμενα αποτελέσματα και η πιθανότητα ένα ανακτώμενο αποτέλεσμα να είναι σχετικό, αν και μειώνεται, διατηρείται σε υψηλές τιμές.

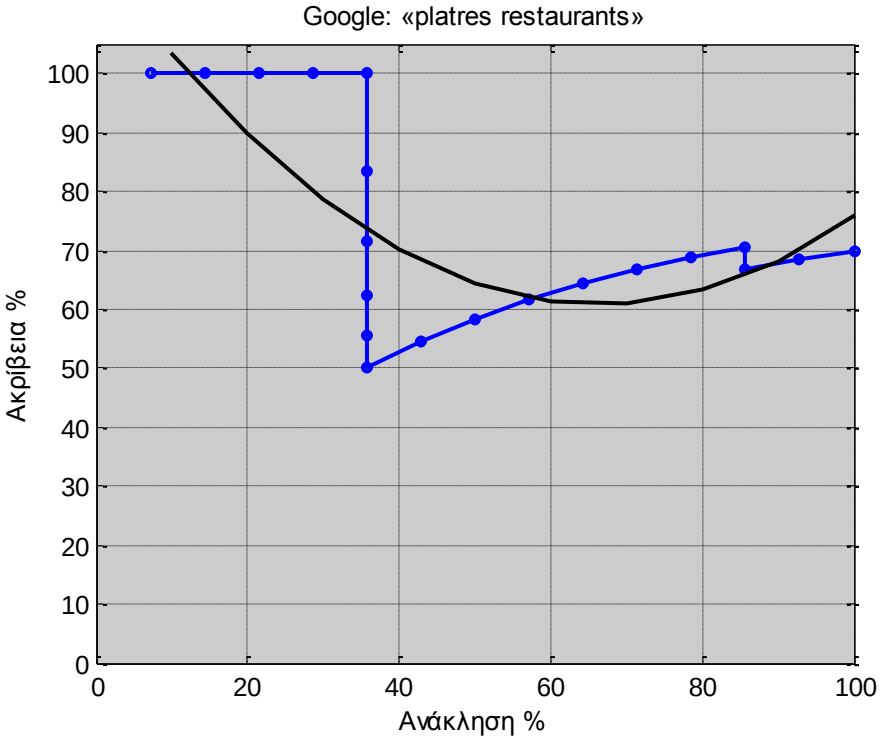


Διάγραμμα 6.1: Διάγραμμα ανάκλησης - ακρίβειας ερωτήματος «university of central Greece» για την Google.

Συνεχίζουμε την αξιολόγηση των αποτελεσμάτων της Google, πραγματοποιώντας την ίδια διαδικασία για το ερώτημα «Limassol restaurants». Στο πίνακα 6.2 βλέπουμε τις τιμές της ανάκλησης και της ακρίβειας για το συγκεκριμένο ερώτημα. Από τα 20 συνολικά ανακτώμενα αποτελέσματα, 14 είναι σχετικά για εμάς.

Ερώτημα: platres restaurants					
Σχετικά Αποτελέσματα: {r1, r2, r3, r4, r5, r11, r12, r13, r14, r15, r16,r17,r19, r20}					
Σύνολο Σχετικών: 14					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	7,14	100,00
2	r2	NAI	2	14,28	100,00
3	r3	NAI	3	21,42	100,00
4	r4	NAI	4	28,57	100,00
5	r5	NAI	5	35,71	100,00
6	r6	OXI	5	35,71	83,33
7	r7	OXI	5	35,71	71,42
8	r8	OXI	5	35,71	62,50
9	r9	OXI	5	35,71	55,55
10	r10	OXI	5	35,71	50.00
11	r11	NAI	6	42,85	54,54
12	r12	NAI	7	49,99	58,33
13	r13	NAI	8	57,14	61,53
14	r14	NAI	9	64,28	64,28
15	r15	NAI	10	71,42	66,66
16	r16	NAI	11	78,57	68,75
17	r17	NAI	12	85,71	70,58
18	r18	OXI	12	85,71	66,66
19	r19	NAI	13	92,85	68,42
20	r20	NAI	14	100,00	70,00

Πινάκας 6.2: Ανάκληση και ακρίβεια ερωτήματος «platres restaurants» για την Google.



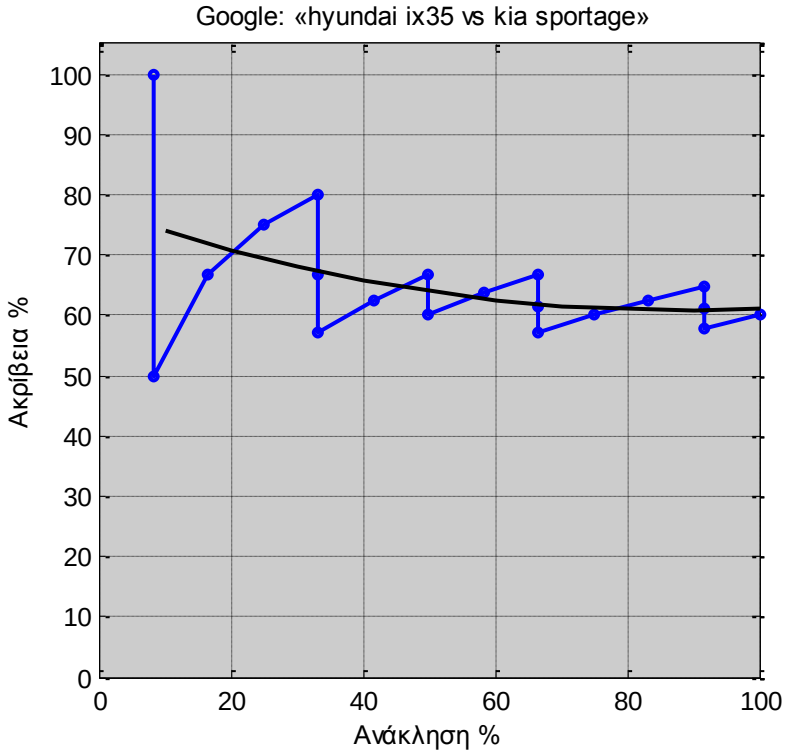
Διάγραμμα 6.2: Διάγραμμα ανάκλησης - ακρίβειας ερωτήματος «platres restaurants» για την Google.

Στο Διάγραμμα 6.2, βλέπουμε την πλήρη καμπύλη ανάκλησης – ακρίβειας του παραπάνω πίνακα. Παρατηρούμε ότι ενώ αυξάνεται το ποσοστό της ανάκλησης, η ακρίβεια μειώνεται συνεχώς, φτάνει σε ένα κατώτατο επίπεδο 60% και στη συνεχεία αυξάνεται γραμμικά και φτάνει μέχρι το 80%.

Συνεχίζουμε με τον ίδιο τρόπο για τα υπόλοιπα 2 ερωτήματα. Στους πίνακες 6.3 και 6.4 βλέπουμε τις τιμές της ανάκλησης και της ακρίβειας για κάθε ένα ερώτημα. Για κάθε πίνακα έχουμε το αντίστοιχο διάγραμμα της πλήρους καμπύλης – ανάκλησης – ακρίβειας. (Διαγράμματα 6.3 και 6.4)

Ερώτημα: hyundai ix35 vs kia sportage					
Σχετικά Αποτελέσματα: {r1,r3, r4, r5, r8, r9, r11, r12, r15, r16,r17, r20}					
Σύνολο Σχετικών: 12					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	8,33	100,00
2	r2	OXI	1	8,33	50,00
3	r3	NAI	2	16,66	66,66
4	r4	NAI	3	24,99	75,00
5	r5	NAI	4	33,33	80,00
6	r6	OXI	4	33,33	66,66
7	r7	OXI	4	33,33	57,14
8	r8	NAI	5	41,66	62,50
9	r9	NAI	6	49,99	66,66
10	r10	OXI	6	49,99	60,00
11	r11	NAI	7	58,33	63,63
12	r12	NAI	8	66,66	66,66
13	r13	OXI	8	66,66	61,53
14	r14	OXI	8	66,66	57,14
15	r15	NAI	9	74,99	60,00
16	r16	NAI	10	83,33	62,50
17	r17	NAI	11	91,66	64,70
18	r18	OXI	11	91,66	61,11
19	r19	OXI	11	91,66	57,89
20	r20	NAI	12	100,00	60,00

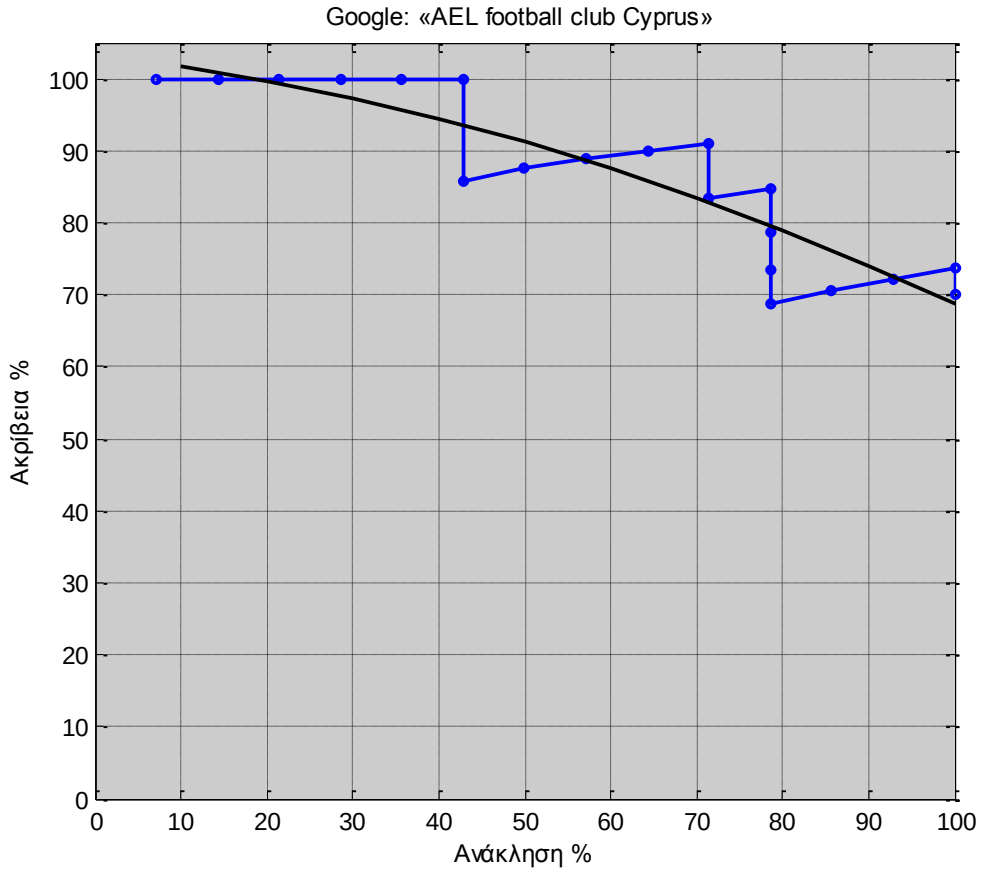
Πινάκας 6.3: Ανάκληση και ακρίβεια ερωτήματος «hyundai ix35 vs kia sportage» για την Google.



Διάγραμμα 6.3: Διάγραμμα ανάκλησης - ακρίβειας ερωτήματος «hyundai ix35 vs kia sportage» για την Google.

Ερώτημα: AEL football club Cyprus					
Σχετικά Αποτελέσματα: {r1,r2,r3,r4,r5,r6,r8,r9,r10,r11,r13,r17,r18,r19**}**					
Σύνολο Σχετικών: 14					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	7,14	100,00
2	r2	NAI	2	14,28	100,00
3	r3	NAI	3	21,42	100,00
4	r4	NAI	4	28,56	100,00
5	r5	NAI	5	35,70	100,00
6	r6	NAI	6	42,84	100,00
7	r7	OXI	6	42,84	85,71
8	r8	NAI	7	49,98	87,50
9	r9	NAI	8	57,12	88,88
10	r10	NAI	9	64,26	90,00
11	r11	NAI	10	71,40	90,90
12	r12	OXI	10	71,40	83,33
13	r13	NAI	11	78,54	84,61
14	r14	OXI	11	78,54	78,57
15	r15	OXI	11	78,54	73,33
16	r16	OXI	11	78,54	68,75
17	r17	NAI	12	85,68	70,58
18	r18	NAI	13	92,82	72,22
19	r19	NAI	14	100,00	73,68
20	r20	OXI	14	100,00	70,00

Πινάκας 6.4: Ανάκληση και ακρίβεια ερωτήματος «AEL football club Cyprus» για την Google.

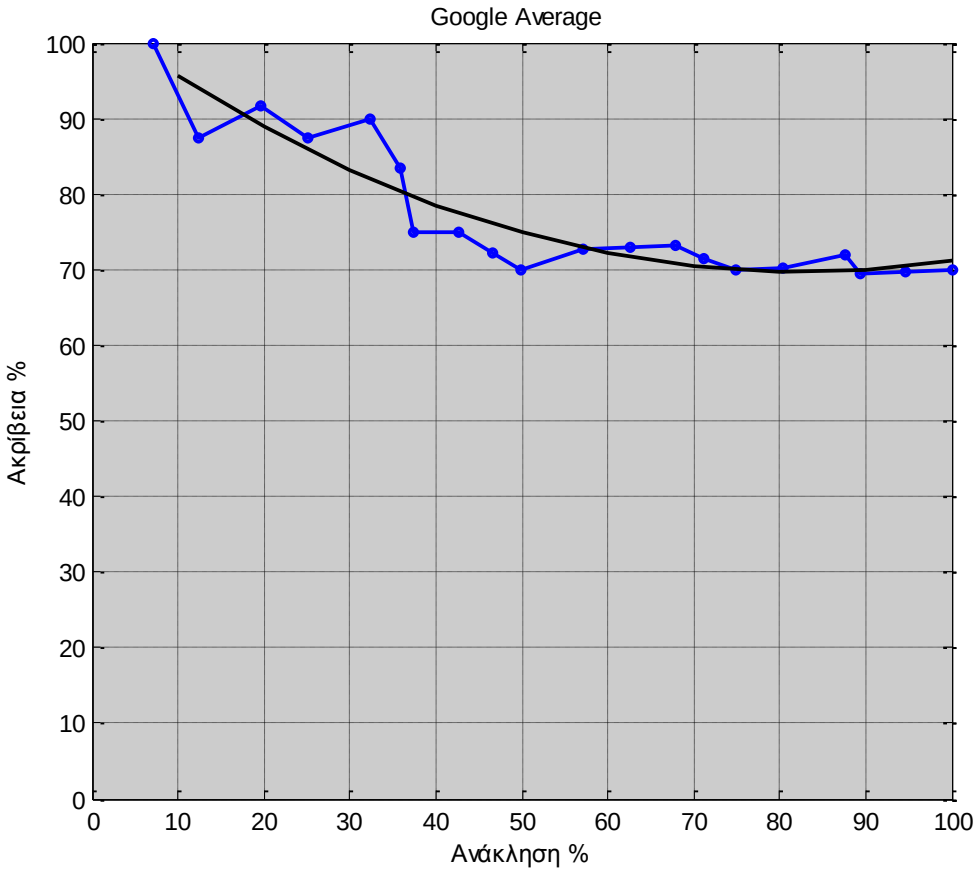


Διάγραμμα 6.4: Διάγραμμα ανάκλησης - ακρίβειας ερωτήματος «AEL football club Cyprus» για την Google.

Παρατηρούμε ότι για κάθε διαφορετικό ερώτημα, οι αντίστοιχοι πίνακες με τους δείκτες της ανάκλησης και της ακρίβειας παρουσιάζουν σημαντικές διαφορές. Γι' αυτό το λόγο δημιουργούμε έναν πίνακα του οποίου κάθε γραμμή είναι ο μέσος όρος των αντίστοιχων γραμμών των τεσσάρων ερωτημάτων. Δηλαδή, για κάθε ένα από το σύνολο των 20 ανακτώμενων αποτελεσμάτων βρίσκουμε το μέσο όρο, ανά επίπεδο, των τιμών της ανάκλησης και της ακρίβειας των τεσσάρων διαφορετικών ερωτημάτων (Πίνακας 8.5).

Μέσος Όρος		
Αριθμός Ανακτώμενων	**Ανάκληση %**	**Ακρίβεια %**
0	0,00	0,00
1	7,21	100,00
2	12,34	87,50
3	19,56	91,68
4	25,21	87,50
5	32,43	90,00
6	35,78	83,33
7	37,34	74,99
8	42,75	75,00
9	46,56	72,21
10	49,99	70,00
11	57,20	72,72
12	62,63	72,91
13	67,77	73,07
14	71,12	71,42
15	74,98	69,99
16	80,42	70,31
17	87,63	72,05
18	89,42	69,44
19	94,56	69,73
20	100,00	70,00

Πίνακας 6.5: Μέσος όρος Ανάκλησης και Ακρίβειας για την Google.



Διάγραμμα 6.5: Διάγραμμα μέσου όρου Ανάκλησης και Ακρίβειας για την Google.

Από την πλήρη καμπύλη ανάκλησης – ακρίβειας του Διαγράμματος 6.1, παρατηρούμε ότι ενώ αυξάνεται το ποσοστό της ανάκλησης, η ακρίβεια κυμαίνεται σε υψηλές τιμές με ένα σταθερό μικρό ρυθμό μείωσης, ξεκινώντας από το 95% και φτάνοντας περίπου στο 70%. Αυτό σημαίνει ότι έχουμε ένα υψηλό ποσοστό σχετικών αποτελεσμάτων. Πράγματι, για τα 20 ανακτώμενα αποτελέσματα κάθε αναζήτησης, έχουμε μέσο όρο 14 σχετικά αποτελέσματα, δηλαδή το 70%. Επίσης, λόγω της κλίσης που έχει η πλήρης καμπύλη, συμπεραίνουμε ότι τα σχετικά αποτελέσματα είναι περισσότερα σε τιμές ανάκλησης 10% - 40%, ενώ στη συνέχεια κατανέμονται ομοιόμορφα.

Σε αυτό το σημείο να υπενθυμίσουμε ότι όλα τα παραπάνω συμπεράσματα αφορούν τον συγκεκριμένο χρήστη. Για κάποιον άλλο χρήστη, οι τιμές της ανάκλησης και της ακρίβειας για κάθε ερώτημα πιθανόν να ήταν αρκετά

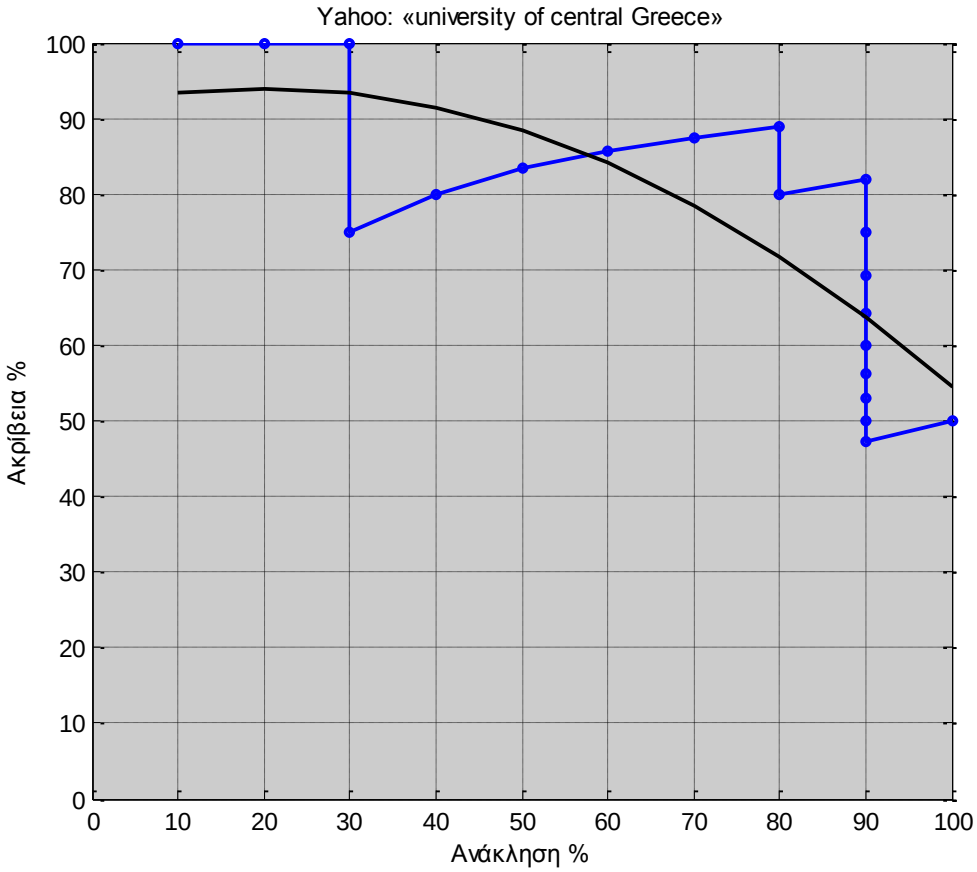
διαφορετικές και επομένως τα συμπεράσματα που θα πρόεκυπταν θα ήταν διαφορετικά.

6.4.2 Αξιολόγηση Yahoo

Συνεχίζουμε, πραγματοποιώντας την ίδια διαδικασία για την αξιολόγηση της Yahoo. Τα ερωτήματα που θα υποβάλουμε είναι τα ίδια και φυσικά ο χρήστης που θα πραγματοποιήσει τις αναζητήσεις είναι ο ίδιος.

Ερώτημα: university of central Greece					
Σχετικά Αποτελέσματα: {r1,r2,r3,r5,r6,r7,r8,r9,r11,r20}					
Σύνολο Σχετικών: 10					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	10,00	100,00
2	r2	NAI	2	20,00	100,00
3	r3	NAI	3	30,00	100,00
4	r4	OXI	3	30,00	75,00
5	r5	NAI	4	40,00	80,00
6	r6	NAI	5	50,00	83,33
7	r7	NAI	6	60,00	85,71
8	r8	NAI	7	70,00	87,50
9	r9	NAI	8	80,00	88,88
10	r10	OXI	8	80,00	80,00
11	r11	NAI	9	90,00	81,81
12	r12	OXI	9	90,00	75,00
13	r13	OXI	9	90,00	69,23
14	r14	OXI	9	90,00	64,28
15	r15	OXI	9	90,00	60,00
16	r16	OXI	9	90,00	56,25
17	r17	OXI	9	90,00	52,94
18	r18	OXI	9	90,00	50,00
19	r19	OXI	9	90,00	47,36
20	r20	NAI	10	100,00	50,00

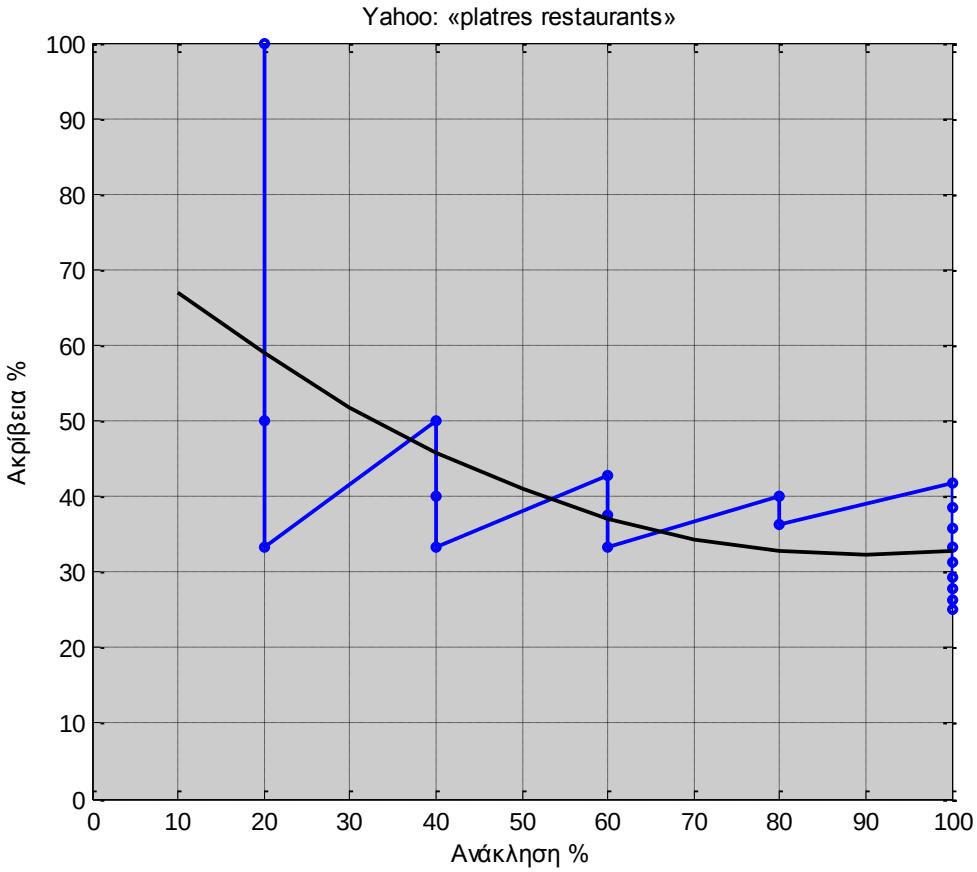
Πινάκας 6.6: Ανάκληση και ακρίβεια ερωτήματος «university of central Greece» για την Yahoo.



Διάγραμμα 6.6: Διάγραμμα ανάκλησης - ακρίβειας ερωτήματος «university of central Greece» για την Yahoo.

Ερώτημα: platres restaurants					
Σχετικά Αποτελέσματα: {r1,r4,r7,r10,r12}					
Σύνολο Σχετικών: 5					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	20,00	100,00
2	r2	OXI	1	20,00	50,00
3	r3	OXI	1	20,00	33,33
4	r4	NAI	2	40,00	50,00
5	r5	OXI	2	40,00	40,00
6	r6	OXI	2	40,00	33,33
7	r7	NAI	3	60,00	42,85
8	r8	OXI	3	60,00	37,50
9	r9	OXI	3	60,00	33,33
10	r10	NAI	4	80,00	40,00
11	r11	OXI	4	80,00	36,36
12	r12	NAI	5	100,00	41,66
13	r13	OXI	5	100,00	38,46
14	r14	OXI	5	100,00	35,71
15	r15	OXI	5	100,00	33,33
16	r16	OXI	5	100,00	31,25
17	r17	OXI	5	100,00	29,41
18	r18	OXI	5	100,00	27,77
19	r19	OXI	5	100,00	26,31
20	r20	OXI	5	100,00	25,00

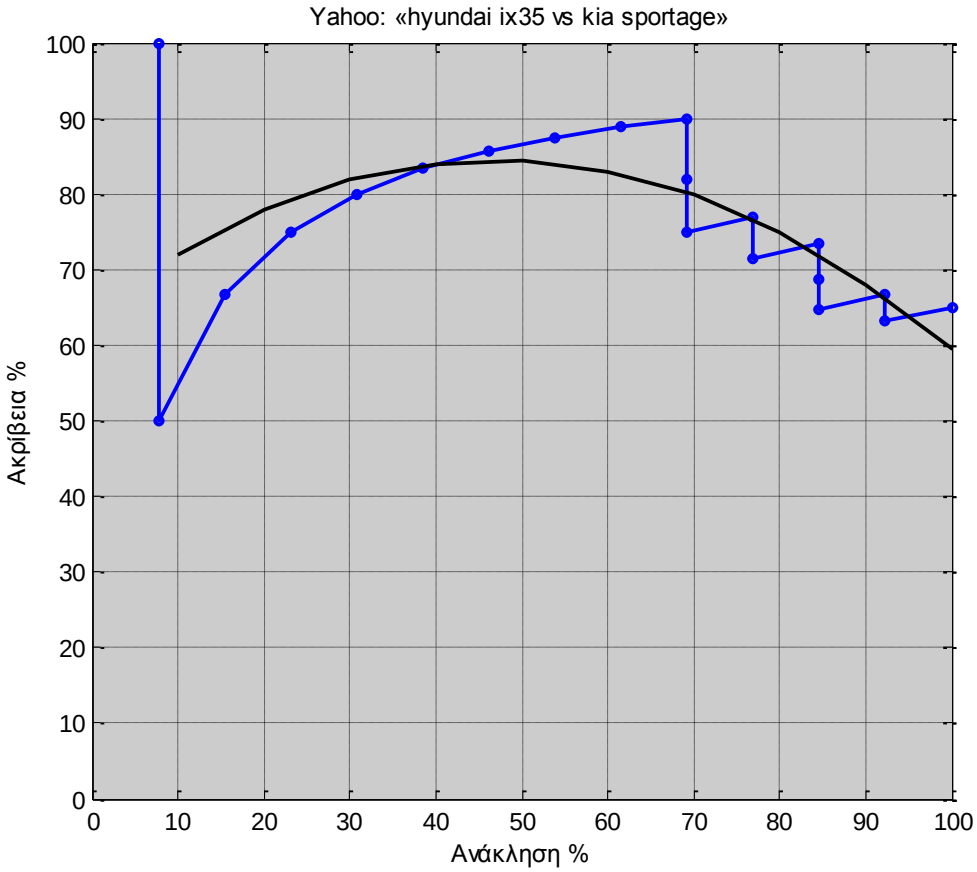
Πινάκας 6.7: Ανάκληση και ακρίβεια ερωτήματος «platres restaurants» για την Yahoo.



Διάγραμμα 6.7: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «platres restaurants» για την Yahoo.

Ερώτημα: hyundai ix35 vs kia sportage					
Σχετικά Αποτελέσματα: {r1,r3,r4,r5,r6,r7,r8,r9,r10,r13,r15,r18,r20}					
Σύνολο Σχετικών: 13					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	7,69	100,00
2	r2	OXI	1	7,69	50,00
3	r3	NAI	2	15,38	66,66
4	r4	NAI	3	23,07	75,00
5	r5	NAI	4	30,76	80,00
6	r6	NAI	5	38,46	83,33
7	r7	NAI	6	46,15	85,71
8	r8	NAI	7	53,84	87,50
9	r9	NAI	8	61,53	88,88
10	r10	NAI	9	69,23	90,00
11	r11	OXI	9	69,23	81,81
12	r12	OXI	9	69,23	75,00
13	r13	NAI	10	76,92	76,92
14	r14	OXI	10	76,92	71,42
15	r15	NAI	11	84,61	73,33
16	r16	OXI	11	84,61	68,75
17	r17	OXI	11	84,61	64,70
18	r18	NAI	12	92,30	66,66
19	r19	OXI	12	92,30	63,15
20	r20	NAI	13	100,00	65,00

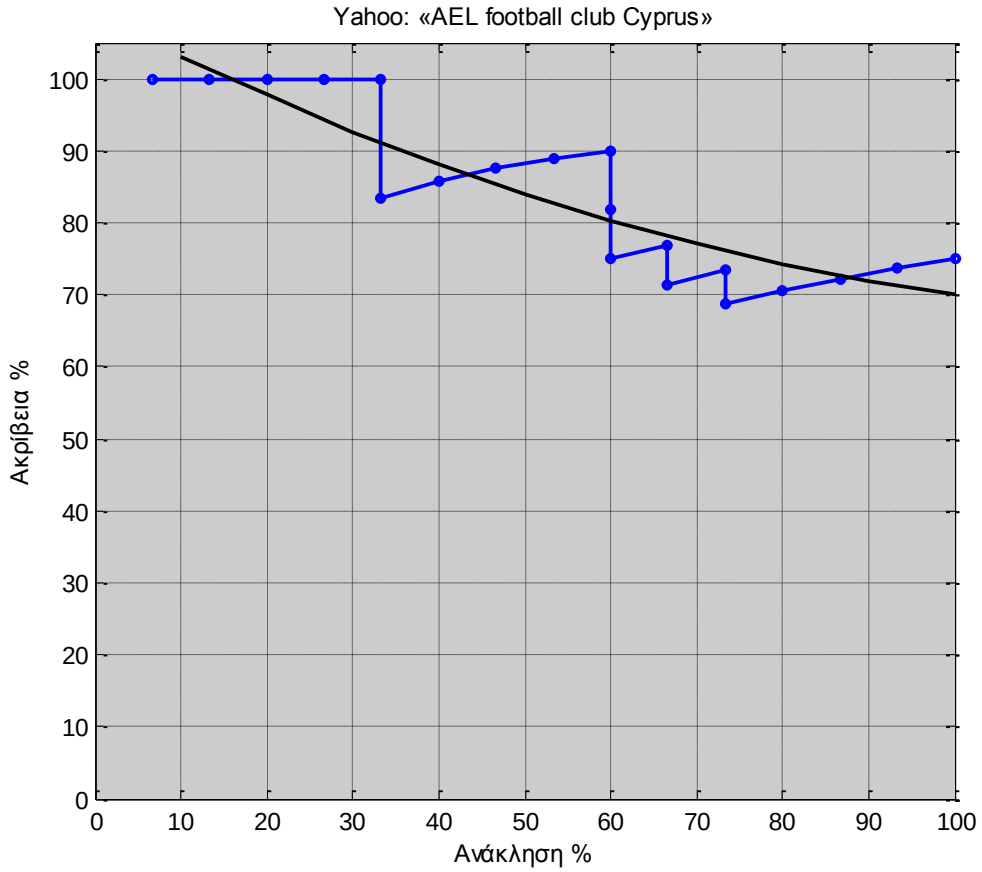
Πινάκας 6.8: Ανάκληση και ακρίβεια ερωτήματος «hyundai ix35 vs kia sportage» για την Yahoo.



Διάγραμμα 6.8: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «hyundai ix35 vs kia sportage» για την Yahoo.

Ερώτημα: AEL football club Cyprus					
Σχετικά Αποτελέσματα: {r1,r2,r3,r4,r5,r7,r8,r9,r10,r13,r15,r17,r18,r19,r20}					
Σύνολο Σχετικών: 15					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	6,66	100,00
2	r2	NAI	2	13,33	100,00
3	r3	NAI	3	19,99	100,00
4	r4	NAI	4	26,66	100,00
5	r5	NAI	5	33,33	100,00
6	r6	OXI	5	33,33	83,33
7	r7	NAI	6	39,99	85,71
8	r8	NAI	7	46,66	87,50
9	r9	NAI	8	53,33	88,88
10	r10	NAI	9	59,99	90,00
11	r11	OXI	9	59,99	81,81
12	r12	OXI	9	59,99	75,00
13	r13	NAI	10	66,66	76,92
14	r14	OXI	10	66,66	71,42
15	r15	NAI	11	73,33	73,33
16	r16	OXI	11	73,33	68,75
17	r17	NAI	12	79,99	70,58
18	r18	NAI	13	86,66	72,22
19	r19	NAI	14	93,33	73,68
20	r20	NAI	15	100,00	75,00

Πινάκας 6.9: Ανάκληση και ακρίβεια ερωτήματος «AEL football club Cyprus» για την Yahoo.

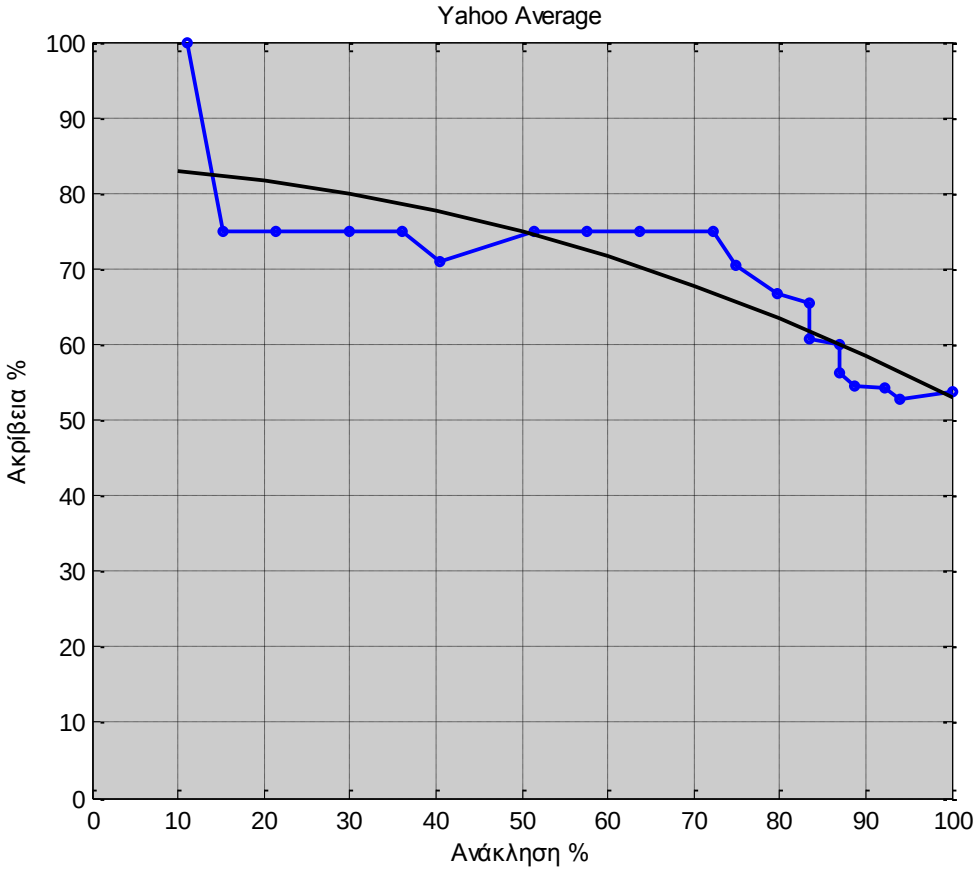


Διάγραμμα 6.9: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «AEL football club Cyprus» για την Yahoo.

Μέσος Όρος		
Αριθμός Ανακτώμενων	**Ανάκληση %**	**Ακρίβεια %**
0	0,00	0,00
1	11,08	100,00
2	15,25	75,00
3	21,34	74,99
4	29,93	75,00
5	36,02	75,00
6	40,44	70,83
7	51,53	74,99
8	57,62	75,00

9	63,71	74,99
10	72,30	75,00
11	74,80	70,44
12	79,80	66,66
13	83,39	65,38
14	83,39	60,70
15	86,98	59,99
16	86,98	56,25
17	88,65	54,40
18	92,24	54,16
19	93,90	52,65
20	100,00	53,75

Πίνακας 6.10: Μέσος όρος Ανάκλησης και Ακρίβειας για την Yahoo.



Διάγραμμα 6.10: Διάγραμμα μέσου όρου Ανάκλησης και Ακρίβειας για την Yahoo.

Από την πλήρη καμπύλη ανάκληση – ακριβείας του Διαγράμματος 6.10, παρατηρούμε ότι καθώς αυξάνεται η ανάκληση, η ακρίβεια έχει ένα σταθερό ρυθμό μείωσης, ξεκινώντας από το 80% και φτάνοντας λίγο κάτω από το 55%. Αυτό σημαίνει ότι και το ποσοστό των σχετικών για εμάς αποτελεσμάτων είναι λίγο κάτω από το 55%. Πράγματι, για τα 20 ανακτώμενα αποτελέσματα κάθε αναζήτησης, έχουμε μέσο όρο 10,75 σχετικά αποτελέσματα, δηλαδή το 53,75%. Επίσης, λόγω της σταθερής μείωσης της ακρίβειας συμπεραίνουμε ότι καθώς αυξάνεται το ποσοστό εύρεσης σχετικών αποτελεσμάτων, μειώνεται η πιθανότητα το ανακτώμενο αποτέλεσμα να είναι σχετικό. Οπότε τα περισσότερα σχετικά (για εμάς) αποτελέσματα της Yahoo βρίσκονται στις αρχικές θέσεις των αποτελεσμάτων.

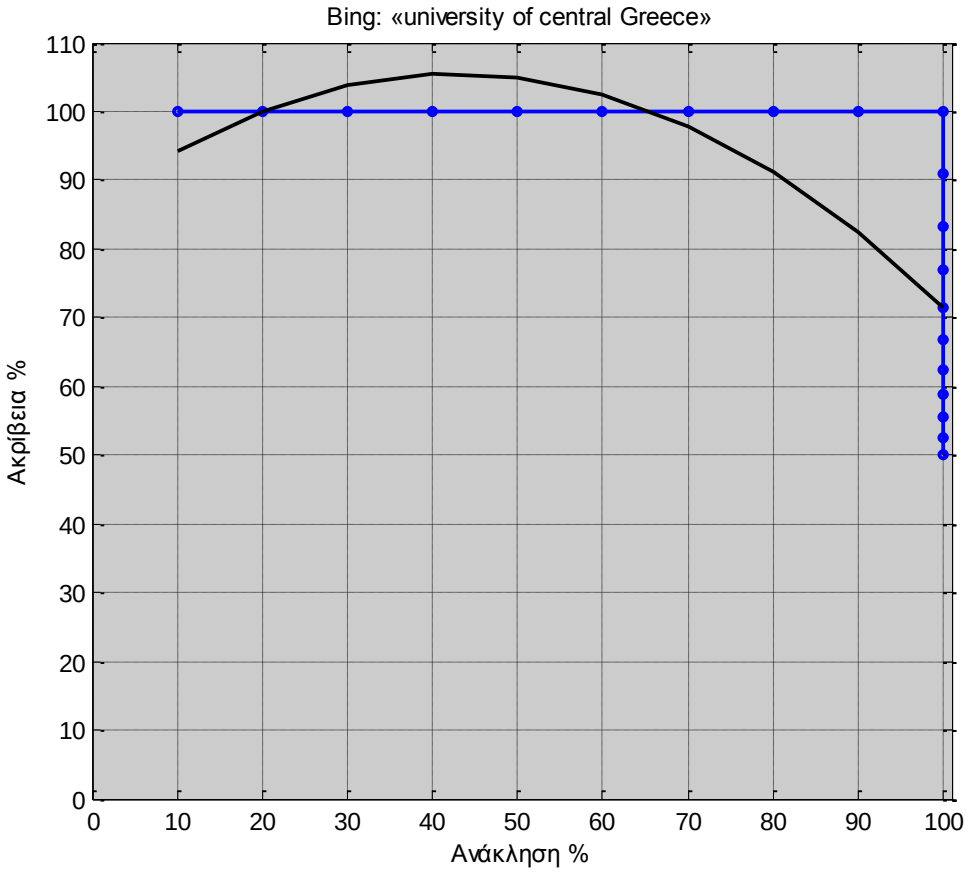
Συγκρίνοντας τα παραπάνω αποτελέσματα με τα αντίστοιχα της Google, προκύπτει το συμπέρασμα ότι για τον χρήστη που πραγματοποίησε τις αναζητήσεις, η μηχανή αναζήτησης Google είναι αποδοτικότερη από τη Yahoo. Το ποσοστό των σχετικών αποτελεσμάτων (από τα συνολικά 20 ανακτώμενα) είναι αρκετά μεγαλύτερο στη Google (70%) από αυτό της Yahoo (53,75%). Θετικό στοιχείο για την Yahoo είναι ότι τα περισσότερα σχετικά αποτελέσματα τα εμφανίζει στις αρχικές θέσεις. Η Google τα εμφανίζει σχεδόν ομοιόμορφα, με λίγα περισσότερα στις αρχικές θέσεις.

6.4.3 Αξιολόγηση Bing.

Συνεχίζουμε πραγματοποιώντας την ίδια διαδικασία για την αξιολόγησης της μηχανής αναζήτησης Bing.

Ερώτημα: university of central Greece					
Σχετικά Αποτελέσματα: {r1,r2,r3,r4,r5,r6,r7,r8,r9,r10}					
Σύνολο Σχετικών: 10					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	ΝΑΙ	1	10,00	100,00
2	r2	ΝΑΙ	2	20,00	100,00
3	r3	ΝΑΙ	3	30,00	100,00
4	r4	ΝΑΙ	4	40,00	100,00
5	r5	ΝΑΙ	5	50,00	100,00
6	r6	ΝΑΙ	6	60,00	100,00
7	r7	ΝΑΙ	7	70,00	100,00
8	r8	ΝΑΙ	8	80,00	100,00
9	r9	ΝΑΙ	9	90,00	100,00
10	r10	ΝΑΙ	10	100,00	100,00
11	r11	ΟΧΙ	10	100,00	90,90
12	r12	ΟΧΙ	10	100,00	83,33
13	r13	ΟΧΙ	10	100,00	76,92
14	r14	ΟΧΙ	10	100,00	71,42
15	r15	ΟΧΙ	10	100,00	66,66
16	r16	ΟΧΙ	10	100,00	62,50
17	r17	ΟΧΙ	10	100,00	58,82
18	r18	ΟΧΙ	10	100,00	55,55
19	r19	ΟΧΙ	10	100,00	52,63
20	r20	ΟΧΙ	10	100,00	50,00

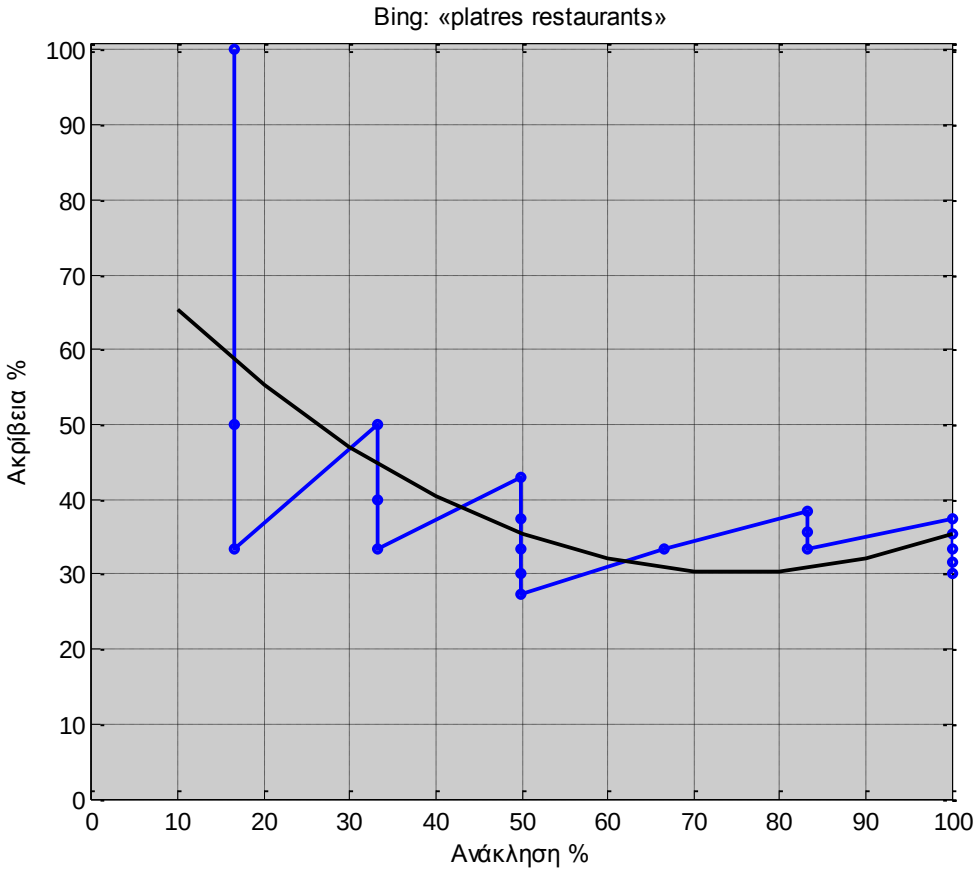
Πινάκας 6.11: Ανάκληση και ακρίβεια ερωτήματος «university of central Greece» για την Bing.



Διάγραμμα 6.11: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «university of central Greece» για την Bing.

Ερώτημα: platres restaurants					
Σχετικά Αποτελέσματα: {r1,r4,r7,r12,r13,r16}					
Σύνολο Σχετικών: 6					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	16,66	100,00
2	r2	OXI	1	16,66	50,00
3	r3	OXI	1	16,66	33,33
4	r4	NAI	2	33,32	50,00
5	r5	OXI	2	33,32	40,00
6	r6	OXI	2	33,32	33,33
7	r7	NAI	3	49,98	42,85
8	r8	OXI	3	49,98	37,50
9	r9	OXI	3	49,98	33,33
10	r10	OXI	3	49,98	30,00
11	r11	OXI	3	49,98	27,27
12	r12	NAI	4	66,64	33,33
13	r13	NAI	5	83,30	38,46
14	r14	OXI	5	83,30	35,71
15	r15	OXI	5	83,30	33,33
16	r16	NAI	6	100,00	37,50
17	r17	OXI	6	100,00	35,29
18	r18	OXI	6	100,00	33,33
19	r19	OXI	6	100,00	31,57
20	r20	OXI	6	100,00	30,00

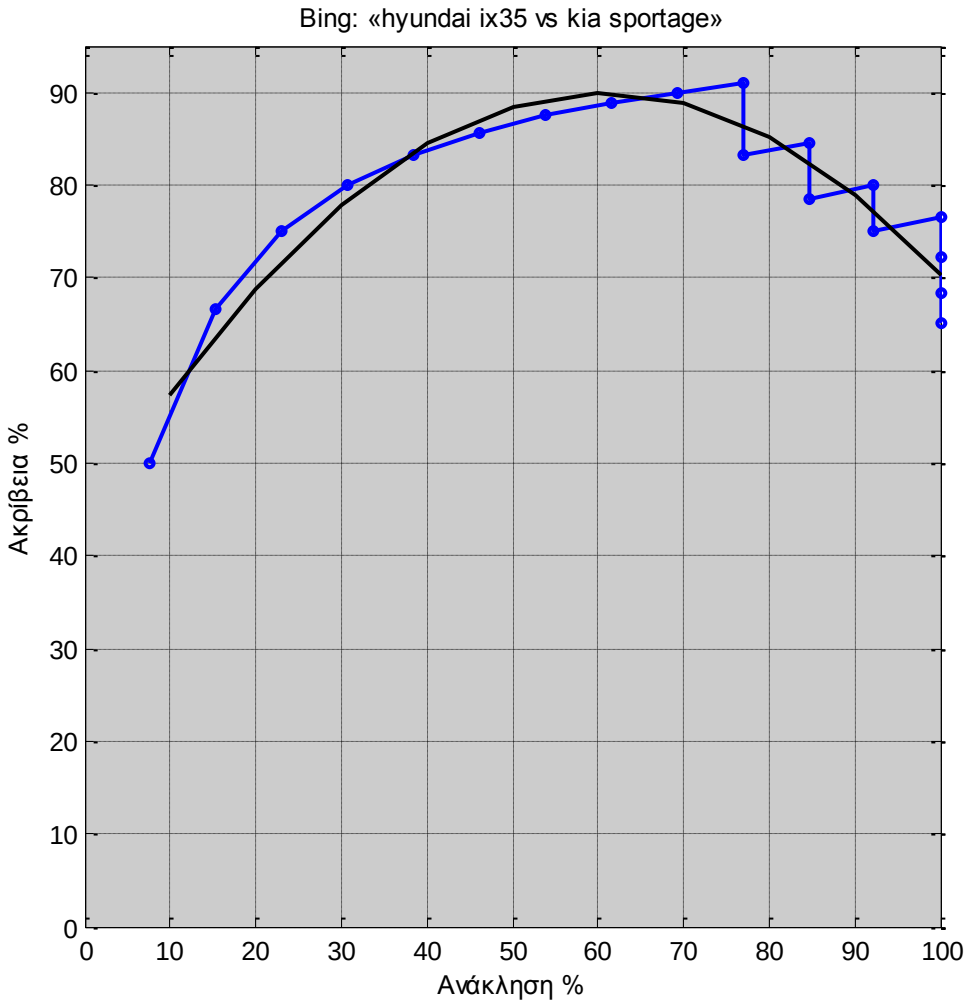
Πινάκας 6.12: Ανάκληση και ακρίβεια ερωτήματος «platres restaurants» για την Bing.



Διάγραμμα 6.12: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «platres restaurants» για την Bing.

Ερώτημα: hyundai ix35 vs kia sportage					
Σχετικά Αποτελέσματα: {r2,r3,r4,r5,r6,r7,r8,r9,r10,r11,r13,r15,r17}					
Σύνολο Σχετικών: 13					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	OXI	0	0,00	0,00
2	r2	NAI	1	7,69	50,00
3	r3	NAI	2	15,38	66,66
4	r4	NAI	3	23,07	75,00
5	r5	NAI	4	30,76	80,00
6	r6	NAI	5	38,45	83,33
7	r7	NAI	6	46,14	85,71
8	r8	NAI	7	53,83	87,50
9	r9	NAI	8	61,52	88,88
10	r10	NAI	9	69,21	90,00
11	r11	NAI	10	76,90	90,90
12	r12	OXI	10	76,90	83,33
13	r13	NAI	11	84,59	84,61
14	r14	OXI	11	84,59	78,57
15	r15	NAI	12	92,28	80,00
16	r16	OXI	12	92,28	75,00
17	r17	NAI	13	100,00	76,47
18	r18	OXI	13	100,00	72,22
19	r19	OXI	13	100,00	68,42
20	r20	OXI	13	100,00	65,00

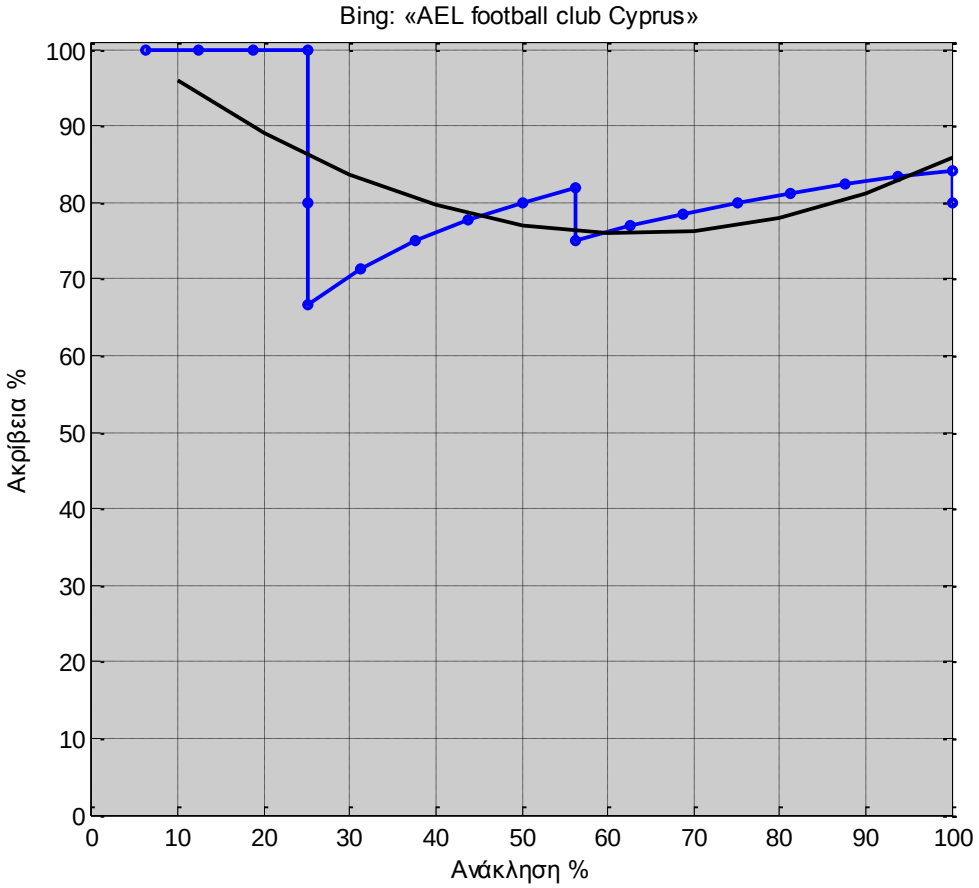
Πινάκας 6.13: Ανάκληση και ακρίβεια ερωτήματος «hyundai ix35 vs kia sportage» για την Bing.



Διάγραμμα 6.13: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «hyundai ix35 vs kia sportage» για την Bing.

Ερώτημα: AEL football club Cyprus					
Σχετικά Αποτελέσματα: {r1,r2,r3,r4,r7,r8,r9,r10,r11,r13,r14,r15,r16,r17,r18,r19}					
Σύνολο Σχετικών: 16					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	6,25	100,00
2	r2	NAI	2	12,50	100,00
3	r3	NAI	3	18,75	100,00
4	r4	NAI	4	25,00	100,00
5	r5	OXI	4	25,00	80,00
6	r6	OXI	4	25,00	66,66
7	r7	NAI	5	31,25	71,42
8	r8	NAI	6	37,50	75,00
9	r9	NAI	7	43,75	77,77
10	r10	NAI	8	50,00	80,00
11	r11	NAI	9	56,25	81,81
12	r12	OXI	9	56,25	75,00
13	r13	NAI	10	62,50	76,92
14	r14	NAI	11	68,75	78,57
15	r15	NAI	12	75,00	80,00
16	r16	NAI	13	81,25	81,25
17	r17	NAI	14	87,50	82,35
18	r18	NAI	15	93,75	83,33
19	r19	NAI	16	100,00	84,21
20	r20	OXI	16	100,00	80,00

Πινάκας 6.14: Ανάκληση και ακρίβεια ερωτήματος «AEL football club Cyprus» για την Bing.

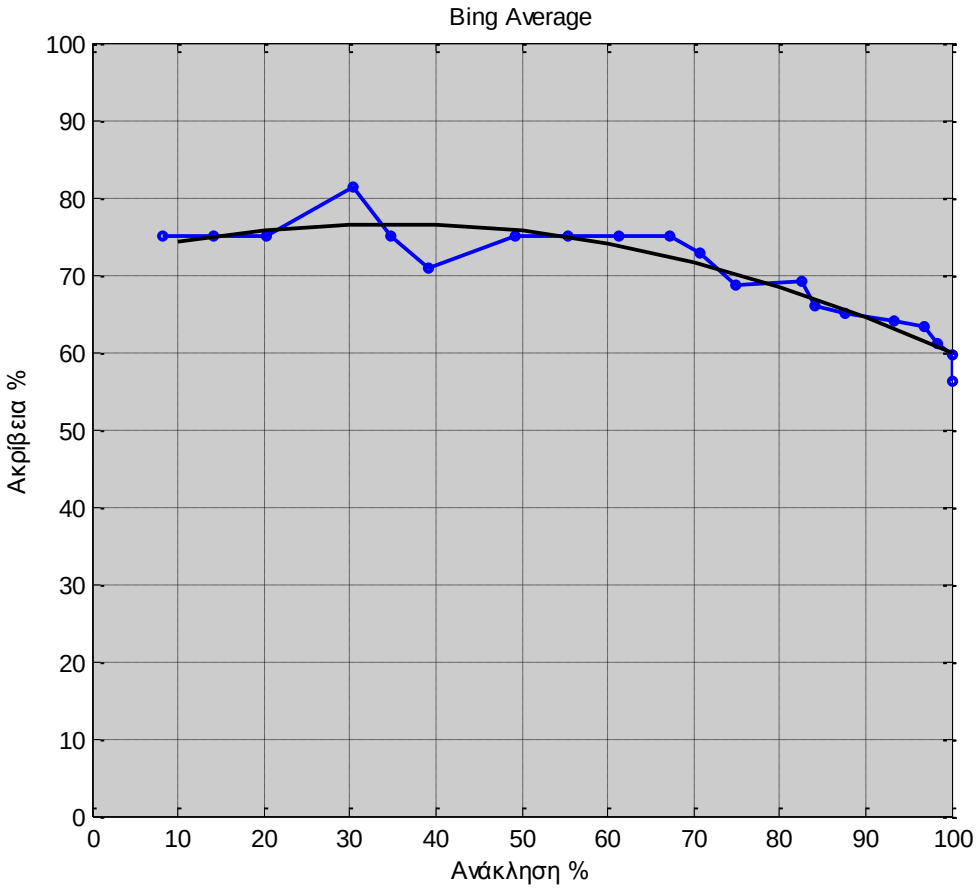


Διάγραμμα 6.14: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «AEL football club Cyprus» για την Bing.

Μέσος Όρος		
Αριθμός Ανακτώμενων	**Ανάκληση %**	**Ακρίβεια %**
0	0,00	0,00
1	8,22	75,00
2	14,21	75,00
3	20,19	74,99
4	30,34	81,25
5	34,77	75,00
6	39,19	70,83
7	49,34	74,99

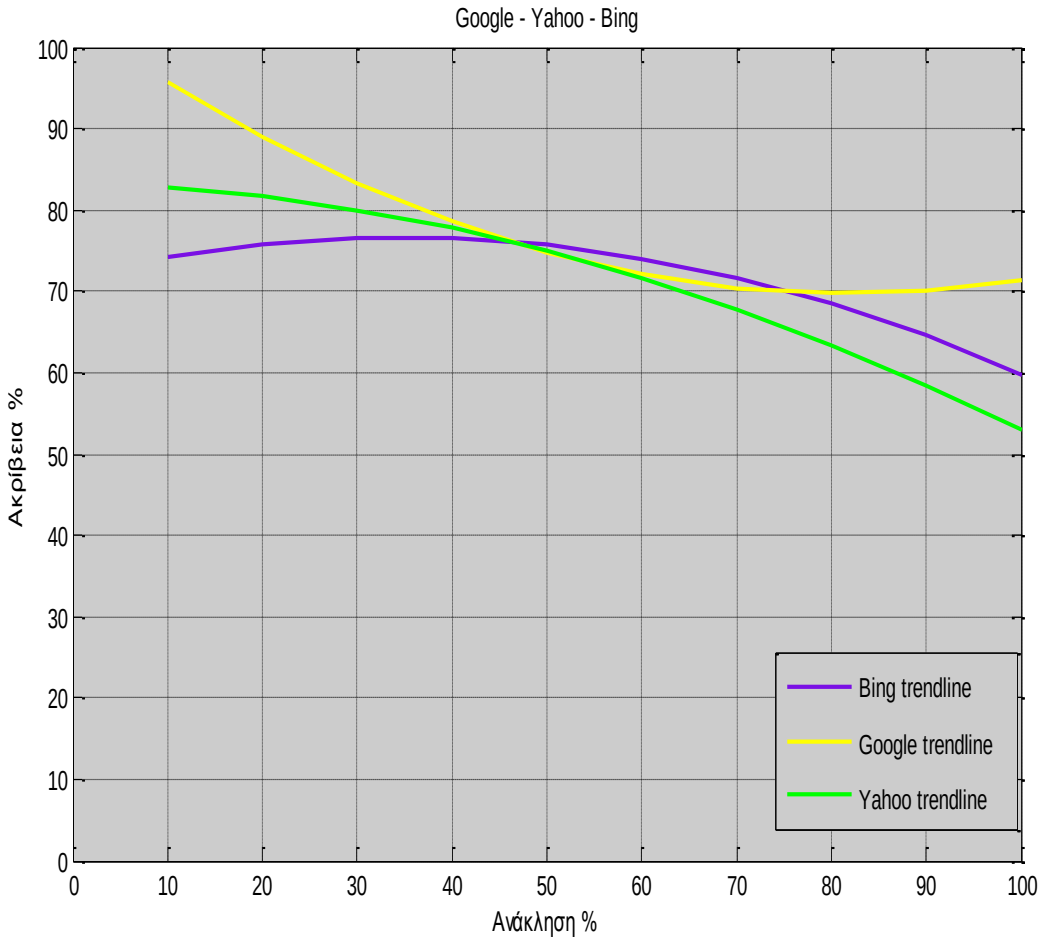
8	55,32	75,00
9	61,31	74,99
10	67,29	75,00
11	70,78	72,72
12	74,94	68,74
13	82,59	69,21
14	84,16	66,06
15	87,64	64,99
16	93,38	64,06
17	96,87	63,23
18	98,43	61,10
19	100,00	59,64
20	100,00	56,25

Πίνακας 6.15: Μέσος όρος Ανάκλησης και Ακρίβειας για την Bing.



Διάγραμμα 6.15: Διάγραμμα μέσου όρου Ανάκλησης και Ακρίβειας για την Bing.

Από την πλήρη καμπύλη ανάκλησης – ακρίβειας του Διαγράμματος 6.15, παρατηρούμε ότι καθώς αυξάνεται ανάκληση από το 0 μέχρι το 50%, η ακρίβεια παραμένει σταθερή σε ποσοστό κοντά στο 75%. Αυτό σημαίνει ότι και το ποσοστό των σχετικών αποτελεσμάτων είναι κοντά στο 75%. Ακολούθως, καθώς αυξάνεται η ανάκληση ξεπερνώντας το 50%, η ακρίβεια αρχίζει να μειώνεται και σταθεροποιείται λίγο κάτω από το 60%. Δηλαδή, το ποσοστό των σχετικών αποτελεσμάτων για εμάς είναι κοντά στο 60%. Πράγματι, για τα 20 ανακτώμενα αποτελέσματα κάθε αναζήτησης, έχουμε μέσο όρο 11,25 σχετικά αποτελέσματα, δηλαδή το 56,25%. Επίσης λόγω της μείωσης της ακρίβειας συμπεραίνουμε ότι καθώς αυξάνεται το ποσοστό εύρεσης σχετικών αποτελεσμάτων, μειώνεται η πιθανότητα το ανακτώμενο αποτέλεσμα να είναι σχετικό. Οπότε τα περισσότερα σχετικά (για εμάς) αποτελέσματα της Bing βρίσκονται στις αρχικές θέσεις των αποτελεσμάτων.



Διάγραμμα 6.16: Σύγκριση πλήρους καμπύλης ανάκλησης – ακρίβειας των 3 μηχανών αναζήτησης (Google, Yahoo, Bing)

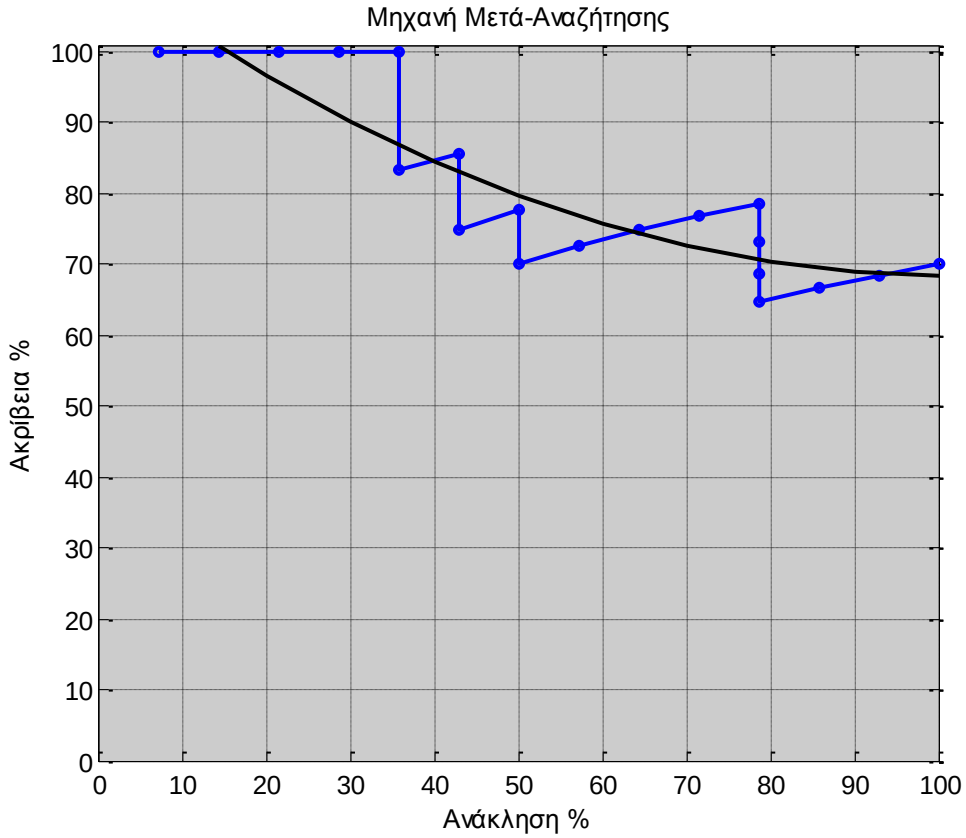
Στο Διάγραμμα 6.16, βλέπουμε συγκεντρωτικά τις πλήρες καμπύλες ανάκλησης ακρίβειας των τριών μηχανών αναζήτησης που μελετάμε. Για να βγάλουμε συμπεράσματα για το ποια μηχανή αναζήτησης είναι η αποδοτικότερη, χρησιμοποιούμε το σημείο ισορροπίας της κάθε καμπύλης. Σημείο ισορροπίας είναι εκεί όπου η ανάκληση ισούται με την ακρίβεια. Όσο υψηλότερο είναι το σημείο ισορροπίας, τόσο πιο αποδοτική θεωρείται μια μηχανή αναζήτησης. Συγκρίνοντας τις τρεις μηχανές αναζήτησης το υψηλότερο σημείο ισορροπίας έχει η Bing με ποσοστό κοντά στο 72%, ακολουθεί η Google με ποσοστό 70% και τέλος η Yahoo με ποσοστό κοντά στο 68%. Βέβαια, η Google έχει πολύ υψηλότερο ποσοστό ακρίβειας στα αρχικά επίπεδα ανάκλησης και περισσότερα συνολικά σχετικά αποτελέσματα από την Bing. Επίσης, η Yahoo έχει μεγαλύτερο ποσοστό ακρίβειας στα αρχικά στάδια ανάκλησης από την Bing αλλά λιγότερα συνολικά σχετικά αποτελέσματα.

6.4.4 Αξιολόγηση Μηχανής Μετά-Αναζήτησης

Για την αξιολόγηση της μηχανής μετά-αναζήτησης πραγματοποιούμε την ίδια διαδικασία όπως παραπάνω. Πιο συγκεκριμένα, ο χρήστης υποβάλει στην Μ.Μ.Α. τα τέσσερα ερωτήματα που υποβλήθηκαν στις τρεις μηχανές αναζήτησης που χρησιμοποιεί η εφαρμογή (Google, Yahoo, Bing). Σκοπός μας είναι η παραγωγή συγκρίσιμων αποτελεσμάτων για την διαπίστωση της ποιότητας των αποτελεσμάτων που προσφέρει η εφαρμογή σε σχέση με τις τρεις μηχανές αναζήτησης.

Ερώτημα: university of central Greece					
Σχετικά Αποτελέσματα: {r1,r2,r3,r4,r5,r7,r9,r11,r12,r13,r14,r18,r19,r20}					
Σύνολο Σχετικών: 14					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	7,14	100,00
2	r2	NAI	2	14,28	100,00
3	r3	NAI	3	21,42	100,00
4	r4	NAI	4	28,57	100,00
5	r5	NAI	5	35,71	100,00
6	r6	OXI	5	35,71	83,33
7	r7	NAI	6	42,85	85,71
8	r8	OXI	6	42,85	75,00
9	r9	NAI	7	49,99	77,77
10	r10	OXI	7	49,99	70,00
11	r11	NAI	8	57,14	72,72
12	r12	NAI	9	64,28	75,00
13	r13	NAI	10	71,42	76,92
14	r14	NAI	11	78,57	78,57
15	r15	OXI	11	78,57	73,33
16	r16	OXI	11	78,57	68,75
17	r17	OXI	11	78,57	64,70
18	r18	NAI	12	85,71	66,66
19	r19	NAI	13	92,85	68,42
20	r20	NAI	14	100,00	70,00

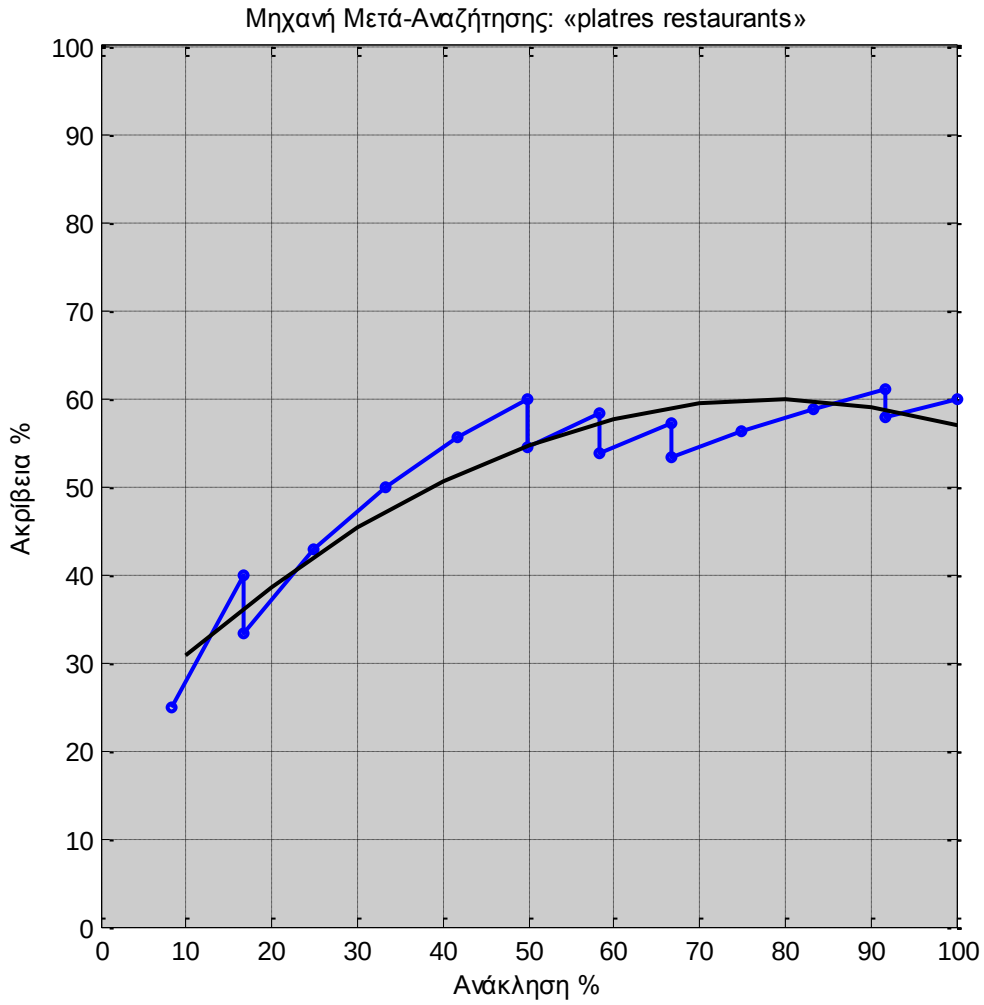
Πινάκας 6.16: Ανάκληση και ακρίβεια ερωτήματος «university of central Greece» για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.



Διάγραμμα 6.17: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «university of central Greece» για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.

Ερώτημα: platres restaurants					
Σχετικά Αποτελέσματα: {r4,r5,r7,r8,r9,r10,r12,r14,r16,r17,r18,r20}					
Σύνολο Σχετικών: 12					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	OXI	0	0,00	0,00
2	r2	OXI	0	0,00	0,00
3	r3	OXI	0	0,00	0,00
4	r4	NAI	1	8,33	25,00
5	r5	NAI	2	16,66	40,00
6	r6	OXI	2	16,66	33,33
7	r7	NAI	3	24,99	42,85
8	r8	NAI	4	33,32	50,00
9	r9	NAI	5	41,65	55,55
10	r10	NAI	6	49,98	60,00
11	r11	OXI	6	49,98	54,54
12	r12	NAI	7	58,31	58,33
13	r13	OXI	7	58,31	53,84
14	r14	NAI	8	66,64	57,14
15	r15	OXI	8	66,64	53,33
16	r16	NAI	9	74,97	56,25
17	r17	NAI	10	83,30	58,82
18	r18	NAI	11	91,63	61,11
19	r19	OXI	11	91,63	57,89
20	r20	NAI	12	100,00	60,00

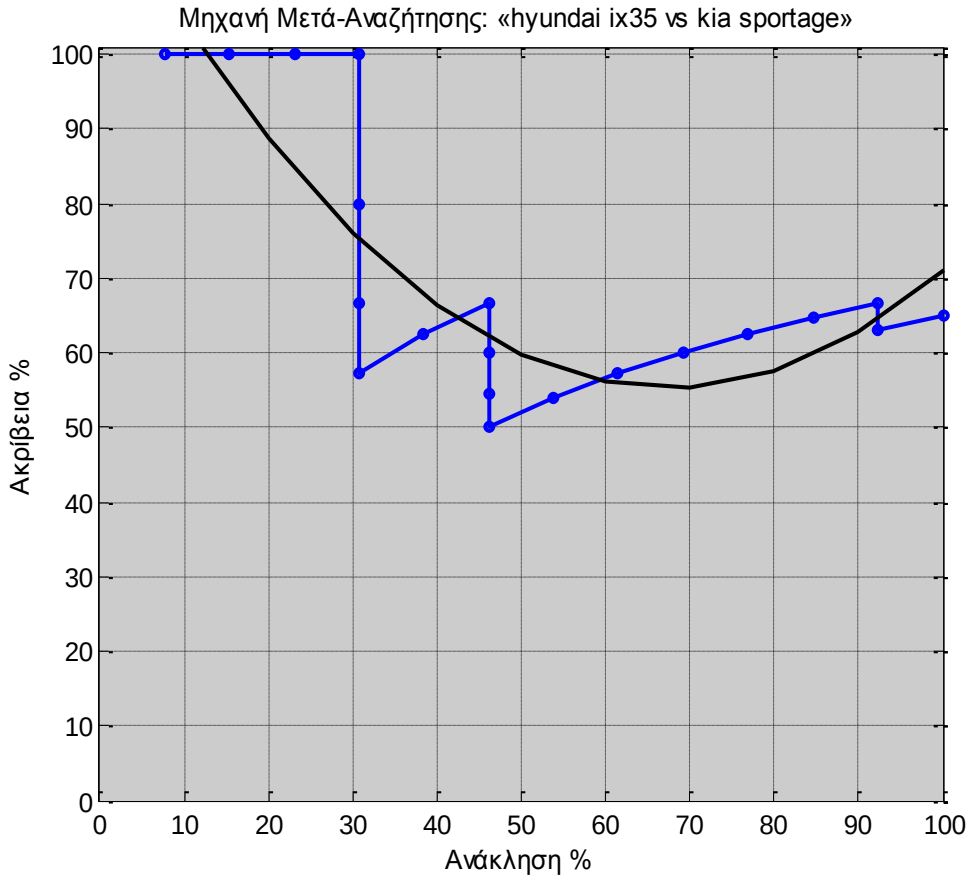
Πινάκας 6.17: Ανάκληση και ακρίβεια ερωτήματος «platres restaurants» για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.



Διάγραμμα 6.18: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «platres restaurants» για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.

Ερώτημα: hyundai ix35 vs kia sportage					
Σχετικά Αποτελέσματα: {r1,r2,r3,r4,r8,r9,r13,r14,r15,r16,r17,r18,r20}					
Σύνολο Σχετικών: 13					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	7,69	100,00
2	r2	NAI	2	15,38	100,00
3	r3	NAI	3	23,07	100,00
4	r4	NAI	4	30,76	100,00
5	r5	OXI	4	30,76	80,00
6	r6	OXI	4	30,76	66,66
7	r7	OXI	4	30,76	57,14
8	r8	NAI	5	38,45	62,50
9	r9	NAI	6	46,14	66,66
10	r10	OXI	6	46,14	60,00
11	r11	OXI	6	46,14	54,54
12	r12	OXI	6	46,14	50,00
13	r13	NAI	7	53,83	53,84
14	r14	NAI	8	61,52	57,14
15	r15	NAI	9	69,21	60,00
16	r16	NAI	10	76,90	62,50
17	r17	NAI	11	84,59	64,70
18	r18	NAI	12	92,28	66,66
19	r19	OXI	12	92,28	63,15
20	r20	NAI	13	100,00	65,00

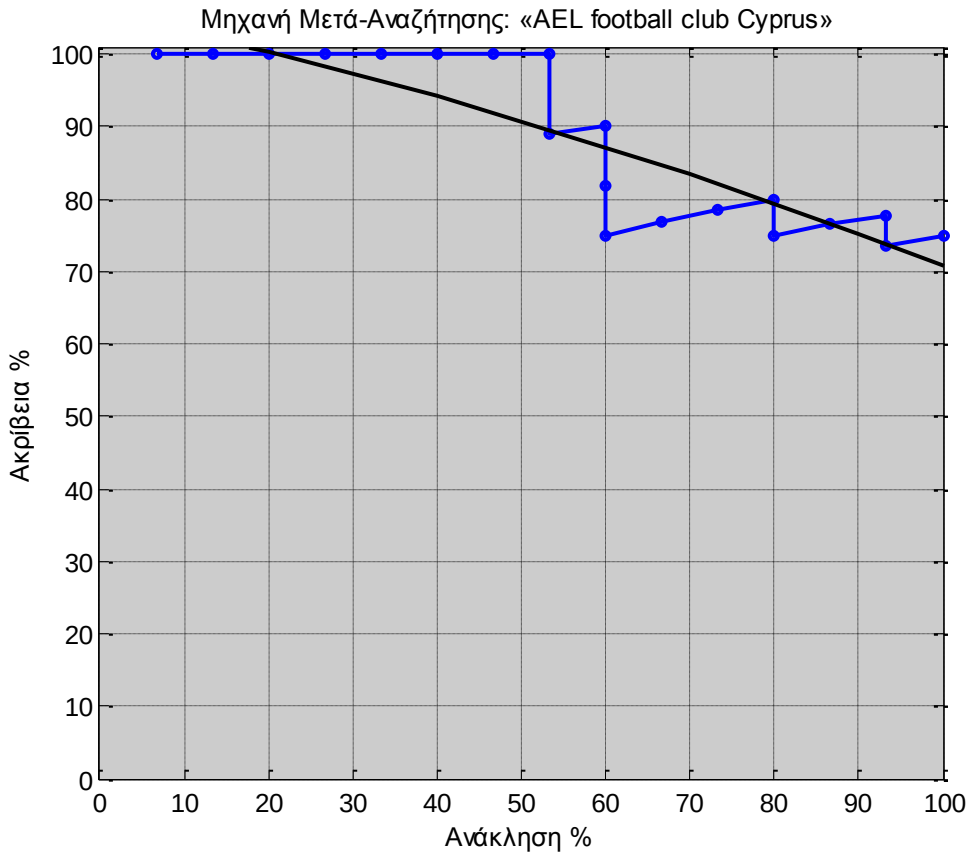
Πινάκας 6.18: Ανάκληση και ακρίβεια ερωτήματος «hyundai ix35 vs kia sportage» για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.



Διάγραμμα 6.19: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «hyundai ix35 vs kia sportage» για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.

Ερώτημα: AEL football club Cyprus					
Σχετικά Αποτελέσματα: {r1,r2,r3,r4,r5,r6,r7,r8,r10,r13,r14,r15,r17,r18,r20}					
Σύνολο Σχετικών: 15					
Αριθμός Ανακτώμενων	**Νέο Ανακτώμενο**	**Είναι σχετικό;**	**Αριθμός Σχετικών**	**Ανάκληση %**	**Ακρίβεια %**
0	-	-	0	0,00	0,00
1	r1	NAI	1	6,66	100,00
2	r2	NAI	2	13,32	100,00
3	r3	NAI	3	19,98	100,00
4	r4	NAI	4	26,64	100,00
5	r5	NAI	5	33,30	100,00
6	r6	NAI	6	39,96	100,00
7	r7	NAI	7	46,62	100,00
8	r8	NAI	8	53,28	100,00
9	r9	OXI	8	53,28	88,88
10	r10	NAI	9	59,94	90,00
11	r11	OXI	9	59,94	81,81
12	r12	OXI	9	59,94	75,00
13	r13	NAI	10	66,60	76,92
14	r14	NAI	11	73,26	78,57
15	r15	NAI	12	79,92	80,00
16	r16	OXI	12	79,92	75,00
17	r17	NAI	13	86,58	76,47
18	r18	NAI	14	93,24	77,77
19	r19	OXI	14	93,24	73,68
20	r20	NAI	15	100,00	75,00

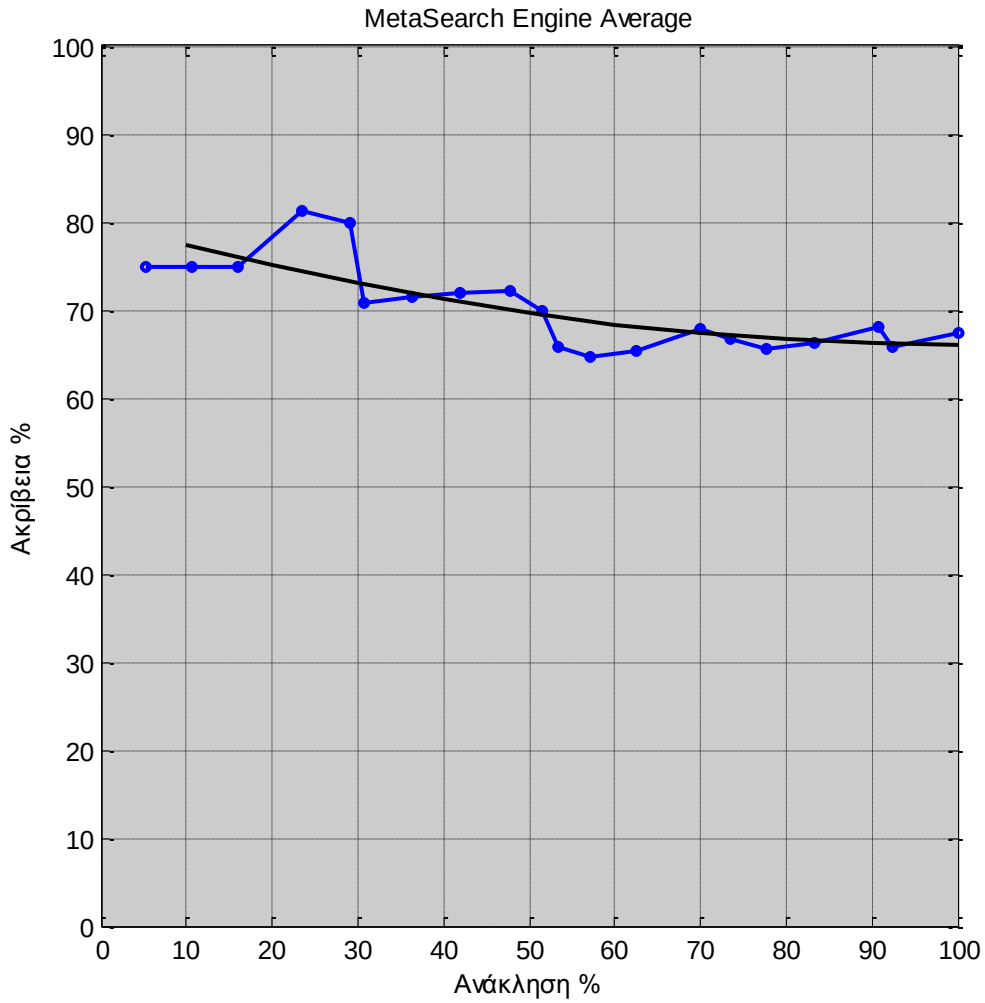
Πινάκας 6.19: Ανάκληση και ακρίβεια ερωτήματος «AEL football club Cyprus» για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.



Διάγραμμα 6.20: Διάγραμμα ανάκλησης – ακρίβειας ερωτήματος «AEL football club Cyprus» για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.

Μέσος Όρος		
Αριθμός Ανακτώμενων	**Ανάκληση %**	**Ακρίβεια %**
0	0,00	0,00
1	5,37	75,00
2	10,74	75,00
3	16,11	75,00
4	23,57	81,25
5	29,10	80,00
6	30,77	70,83
7	36,30	71,42
8	41,97	71,87
9	47,76	72,21
10	51,51	70,00
11	53,30	65,90
12	57,16	64,58
13	62,54	65,38
14	69,99	67,85
15	73,58	66,66
16	77,59	65,62
17	83,26	66,17
18	90,71	68,05
19	92,50	65,78
20	100,00	67,50

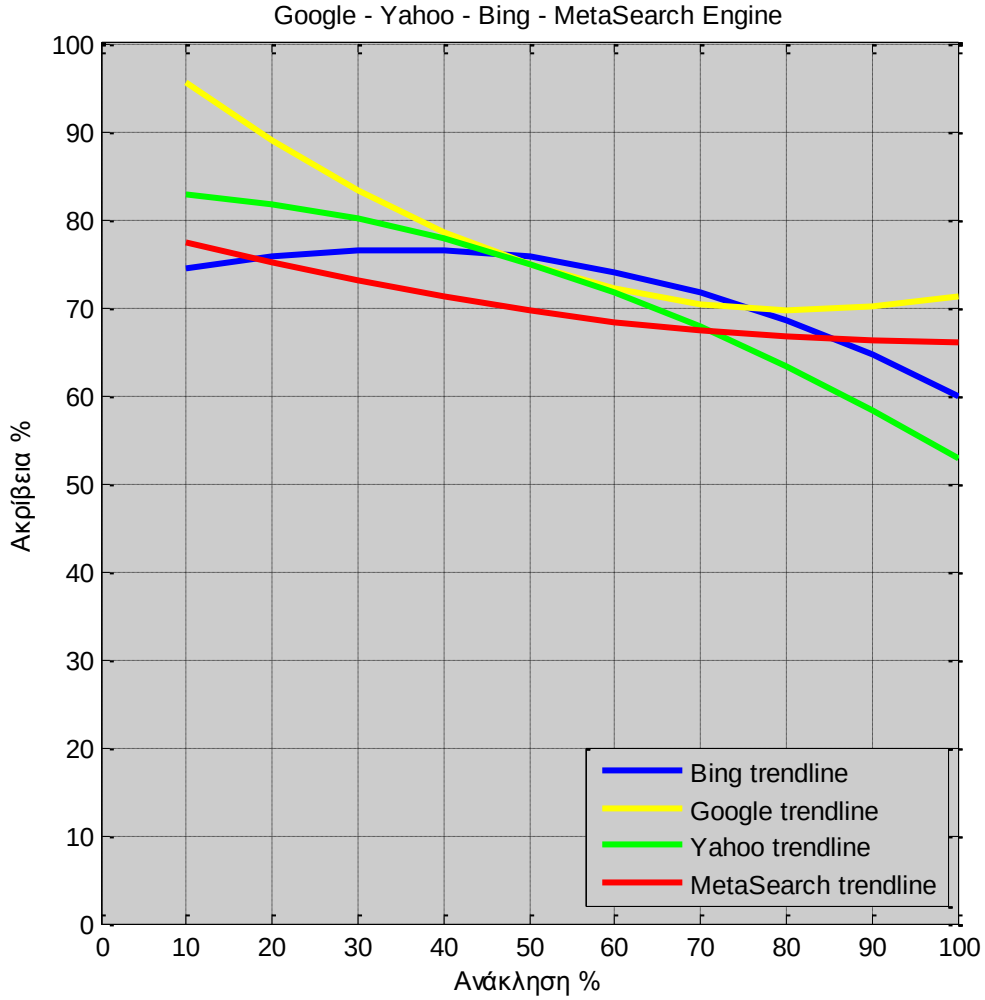
Πίνακας 6.20: Μέσος όρος Ανάκλησης και Ακρίβειας για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.



Διάγραμμα 6.21: Διάγραμμα μέσου όρου Ανάκλησης και Ακρίβειας για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε.

Από την πλήρη καμπύλη ανάκλησης – ακρίβειας του Διαγράμματος 6.21, παρατηρούμε ότι καθώς αυξάνεται η ανάκληση, η ακρίβεια διατηρείται σε σχετικά υψηλές τιμές έχοντας όμως μικρή μείωση, ξεκινώντας από περίπου 75% και καταλήγοντας σταθεροποιείται κοντά στο 67%. Αυτό σημαίνει ότι και το ποσοστό των σχετικών για εμάς αποτελεσμάτων είναι κοντά στο 67%. Πράγματι, για τα 20 ανακτώμενα αποτελέσματα κάθε αναζήτησης, έχουμε μέσο όρο 13,5 σχετικά αποτελέσματα, δηλαδή 67,50%. Επίσης, λόγω της μικρής μείωσης της ακρίβειας, συμπεραίνουμε ότι καθώς αυξάνεται το ποσοστό εύρεσης σχετικών αποτελεσμάτων, μειώνεται, η πιθανότητα το ανακτώμενο αποτέλεσμα να είναι σχετικό. Οπότε τα περισσότερα σχετικά (για εμάς) αποτελέσματα βρίσκονται στις πρώτες δέκα θέσεις

των αποτελεσμάτων, ενώ ακολούθως η πιθανότητα να είναι σχετικό ένα αποτέλεσμα σταθεροποιείται σε ποσοστό 67%.

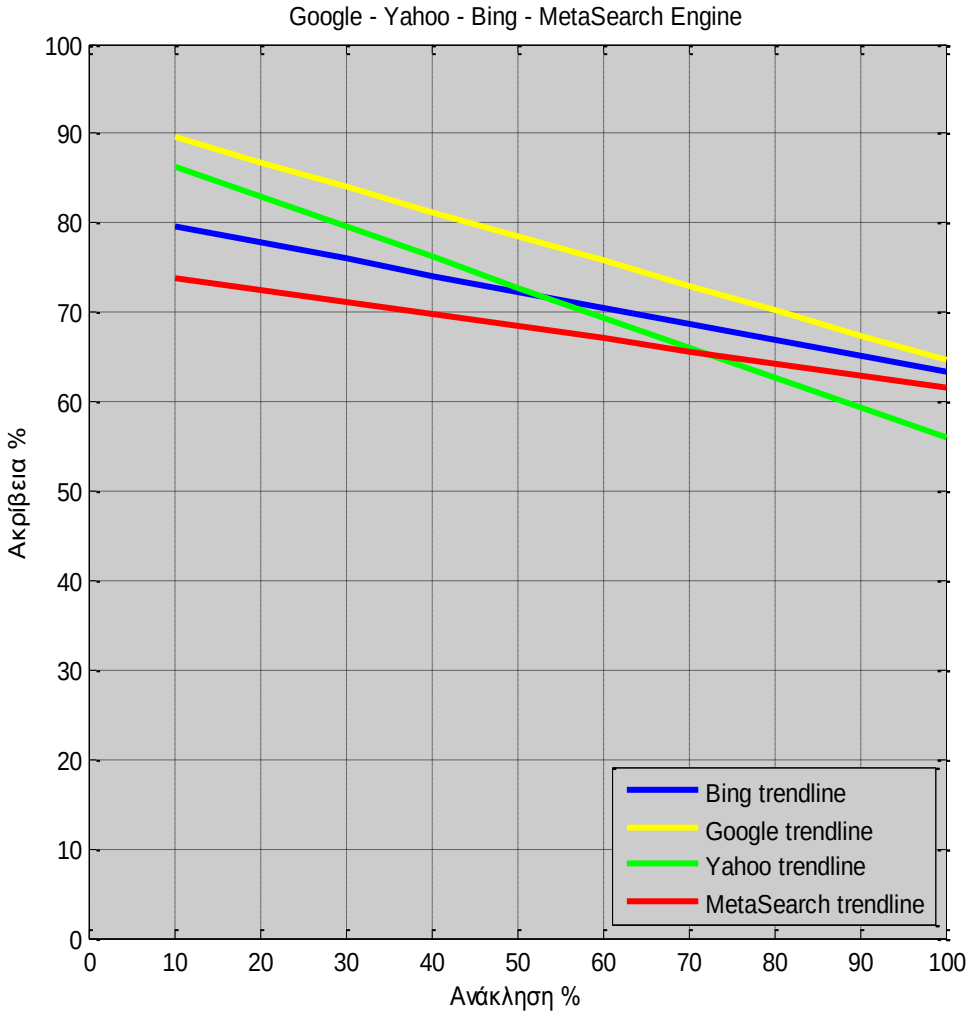


Διάγραμμα 6.22: Σύγκριση πλήρους καμπύλης ανάκλησης – ακρίβειας (γραμμή τάσης ανάκλησης - ακρίβειας πολυωνυμικού τύπου διάταξης 2) των 3 μηχανών αναζήτησης (Google, Yahoo, Bing) και της μηχανής μετά-αναζήτησης που υλοποιήθηκε.

Στο διάγραμμα 6.22, βλέπουμε συγκεντρωτικά την πλήρη καμπύλη ανάκλησης – ακρίβειας της μηχανής μετά-αναζήτησης μαζί με τις αντίστοιχες καμπύλες των τριών μηχανών αναζήτησης που μελετάμε. Συγκρίνοντας αυτές τις καμπύλες προκύπτει το συμπέρασμα ότι για τον χρήστη που πραγματοποίησε τις αναζητήσεις, η μηχανή μετά-αναζήτησης έχει το χαμηλότερο σημείο ισορροπίας, περίπου 67%, σε σχέση με Bing 72%, Google 70% και Yahoo 68%. Βέβαια, η M.M.A έχει αρκετά υψηλότερο ποσοστό σχετικών αποτελεσμάτων με μέσο όρο

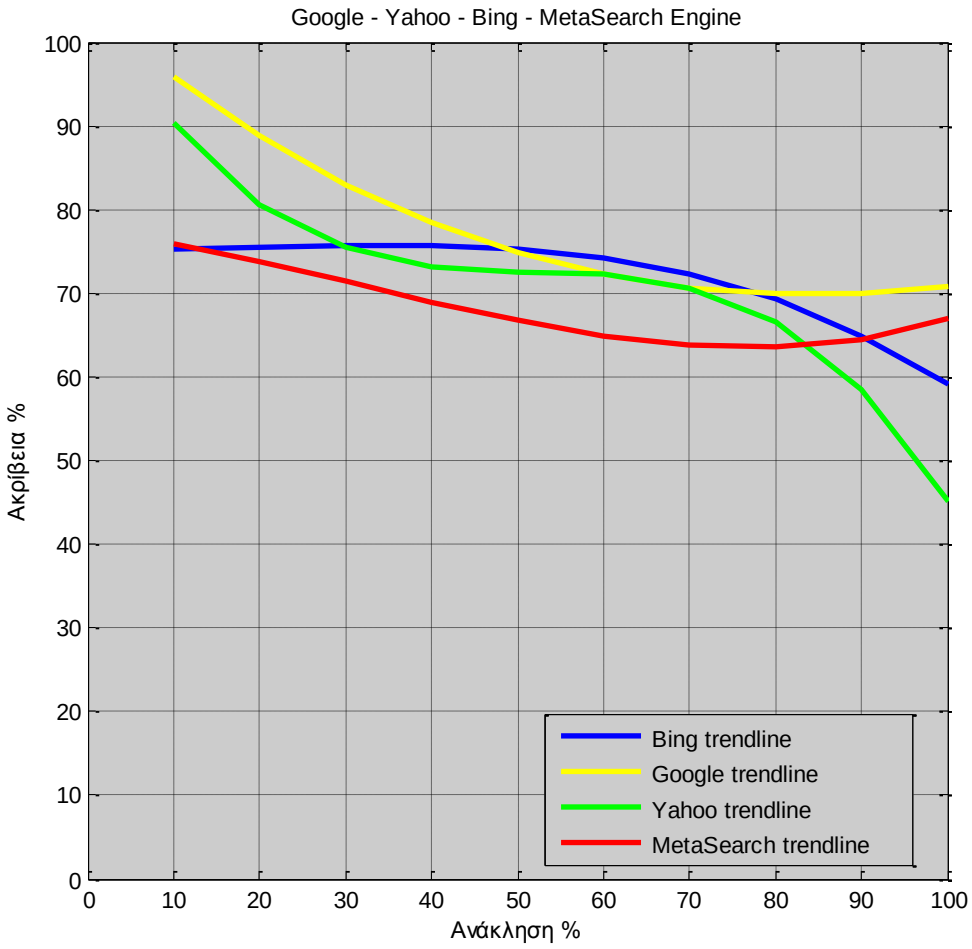
67,5%, σε σχέση με αυτό της Bing (56%) και αυτό της Yahoo (54%), αλλά χαμηλότερο από αυτό της Google (70%).

Σε αυτό το σημείο πρέπει να τονίσουμε ότι για την μηχανή μετά-αναζήτησης αξιολογούμε τα 20 πρώτα αποτελέσματα και όχι όλο το σύνολο των μετά αποτελεσμάτων που προκύπτουν από την ένωση των αποτελεσμάτων των Google, Yahoo και Bing. Επίσης, η σειρά με την οποία εμφανίζονται επηρεάζεται από τον αριθμό των μηχανών αναζήτησης στις οποίες βρέθηκε το αποτέλεσμα (μπορεί να βρέθηκε μόνο σε μια, σε δυο ή και στις τρεις μηχανές αναζήτησης).



Διάγραμμα 6.23: Σύγκριση πλήρους καμπύλης ανάκλησης – ακρίβειας (γραμμή τάσης ανάκλησης - ακρίβειας πολυωνυμικού τύπου διάταξης 1) των 3 μηχανών αναζήτησης (Google, Yahoo, Bing) και της μηχανής μετά-αναζήτησης που υλοποιήθηκε.

Στο διάγραμμα 6.23, βλέπουμε συγκεντρωτικά την πλήρη καμπύλη ανάκλησης – ακρίβειας (διάταξης 1) της μηχανής μετά-αναζήτησης μαζί με τις αντίστοιχες καμπύλες των τριών μηχανών αναζήτησης που μελετάμε. Παρατηρούμε ότι όταν χρησιμοποιήσουμε την γραμμή τάσης ανάκλησης - ακρίβειας πολυωνυμικού τύπου διάταξης 1 προκύπτουν διαφορετικά συμπεράσματα όσο αφορά την αποδοτικότητα των μηχανών αναζήτησης. Συγκεκριμένα, η Google έχει πλέον το υψηλότερο σημείο ισορροπίας με περίπου 72%, ακολουθεί η Bing με 69%, η Yahoo με 67% και η M.M.A. με 66%.



Διάγραμμα 6.24: Σύγκριση πλήρους καμπύλης ανάκλησης – ακρίβειας (γραμμή τάσης ανάκλησης - ακρίβειας πολυωνυμικού τύπου διάταξης 3) των 3 μηχανών αναζήτησης (Google, Yahoo, Bing) και της μηχανής μετά-αναζήτησης που υλοποιήθηκε.

Στο διάγραμμα 6.24, βλέπουμε συγκεντρωτικά την πλήρη καμπύλη ανάκλησης – ακρίβειας (διάταξης 3) της μηχανής μετά-αναζήτησης μαζί με τις

αντίστοιχες καμπύλες των τριών μηχανών αναζήτησης που μελετάμε. Παρατηρούμε ότι όταν χρησιμοποιήσουμε την γραμμή τάσης ανάκλησης - ακρίβειας πολυωνυμικού τύπου διάταξης 3 προκύπτει το συμπέρασμα ότι για τον χρήστη που πραγματοποίησε τις αναζητήσεις, η μηχανή μετά-αναζήτησης έχει το χαμηλότερο σημείο ισορροπίας, περίπου 65%, σε σχέση με Bing 72%, Google 71% και Yahoo 70%.

6.5 ΑΞΙΟΛΟΓΗΣΗ ΠΡΟΤΕΙΝΟΜΕΝΗΣ ΜΗΧΑΝΗΣ ΜΕΤΑ–ΑΝΑΖΗΤΗΣΗΣ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Για την αξιολόγηση της M.M.A., ο χρήστης υποβάλει στην M.M.A. τα τέσσερα ερωτήματα («university of central Greece», «platres restaurants», «hyundai ix35 vs kia sportage» και «ael football club Cyprus») που υποβλήθηκαν στις τρεις μηχανές αναζήτησης (Google, Yahoo, Bing) που χρησιμοποιεί η εφαρμογή. Σκοπός μας είναι η παραγωγή συγκρίσιμων αποτελεσμάτων για την διαπίστωση της ποιότητας των αποτελεσμάτων που προσφέρει η εφαρμογή σε σχέση με τις τρεις μηχανές αναζήτησης.

Η μελέτη και αξιολόγηση που πραγματοποιήσαμε, μας έδωσε χρήσιμα συμπεράσματα, τόσο για τις χρησιμοποιούμενες Μηχανές Αναζήτησης όσο και για την εφαρμογή που δημιουργήσαμε.

Αρχικά, διαπιστώσαμε ότι για τον χρήστη που πραγματοποίησε την αξιολόγηση, η μηχανή αναζήτησης με τα περισσότερα επιθυμητά – σχετικά αποτελέσματα είναι η Google και ακολουθούν, με αρκετή διαφορά η Bing και η Yahoo, οι οποίες είναι σχεδόν ισοδύναμες, με λίγο καλύτερη την Bing. Επίσης, διαπιστώσαμε ότι η μηχανή της Bing έχει το υψηλότερο σημείο ισορροπίας 72% σε σχέση με τις Google 70%, Yahoo 68%, και τη M.M.A. 67%, άρα μπορεί να θεωρηθεί ως η αποδοτικότερη υπηρεσία αναζήτησης.

Για την Μηχανή Μετά-Αναζήτησης που υλοποιήθηκε τα συμπεράσματα που προέκυψαν είναι αρκετά καλά. Συγκρίνοντας των αριθμό των σχετικών αποτελεσμάτων, στα πρώτα 20 αποτελέσματα, με αυτά των τριών μηχανών αναζήτησης παρατηρήσαμε ότι μόνο η Google έχει περισσότερα σχετικά αποτελέσματα. Παρόλα αυτά, έχει το χαμηλότερο σημείο ισορροπίας σε σχέση με τις τρεις μηχανές αναζήτησης, οπόταν μπορεί να θεωρηθεί ως η λιγότερη αποδοτική.

Σε αυτό το σημείο πρέπει να τονίσουμε ότι τα παραπάνω αποτελέσματα και συμπεράσματα αφορούν ένα συγκεκριμένο χρήστη. Για κάποιον άλλο χρήστη, ίσως να είχαμε διαφορετικά αποτελέσματα και άρα διαφορετικά συμπεράσματα.

Η αξιολόγηση μπορεί να γίνει περισσότερο αξιόπιστη και έγκυρη με την εκτέλεση της από πολλούς χρήστες. Με αυτό τον τρόπο θα μπορούσαμε να βγάλουμε ένα πιο γενικό συμπέρασμα.

ΑΝΑΦΟΡΕΣ

[1] «…systems that process links on the web to find new content hit a milestone: 1 trillion (as in 1,000,000,000,000) unique URLs on the web at once…»,
http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

[2] Borda count method
http://en.wikipedia.org/wiki/Borda_count

[3] Jsoup
http://jsoup.org/

[4] Vannevar Bush, "As We May Think",
http://en.wikipedia.org/wiki/As_We_May_Think

[5] Text Retrieval Conference(TREC),
http://trec.nist.gov

[6] Gopher
http://www.codeghost.com/gopher_history.html

[7] Mathew Gray - Wandex
http://en.wikipedia.org/wiki/World_Wide_Web_Wanderer

[8] C. Silverstein, M. Henzinger, H. Marais, M. Moricz,
"Analysis of a Very Large AltaVista"

[9] Spamdexing
http://en.wikipedia.org/wiki/Spamdexing

[10] Content spam
http://en.wikipedia.org/wiki/Spamdexing#Content_spam

[11] Page Rank
http://en.wikipedia.org/wiki/PageRank

[12] Hits algorithm
http://en.wikipedia.org/wiki/HITS_algorithm

[13] cookie stuffing
http://en.wikipedia.org/wiki/Cookie_stuffing

[14] Web Crawler
http://en.wikipedia.org/wiki/Web_crawler

[15] Urls

http://en.wikipedia.org/wiki/Urls

[16] Duplicate hosts and search engines

http://www.seobythesea.com/2006/06/duplicate-content-issues-and-search-engines/

[17] Meta Data

http://en.wikipedia.org/wiki/Metadata

[18] HTML

http://el.wikipedia.org/wiki/HTML

[19] XML

http://www.w3.org/XML/ και σημειώσεις κ. Αναγνωστοπούλου του μαθήματος Διασυνδεδεμένα συστήματα υπολογιστών. (XML_technology.pdf)

[20] Ορισμός Εξατομίκευσης.

M. Eirinaki, M. Vazirgiannis, «**Web Mining for Web Personalization**», ACM Transactions on Internet Technology (TOIT), Pages: 1 – 27, February 2003.

[21] Mobasher, B., Cooley, R., and Srivastava, J., «**Automatic personalization based on web usage mining**», Communications of the ACM, Pages: 142–151, 2000

[22] Nasraoui O., Krishnapuram R., Joshi A., and Kamdar T., «**Automatic Web User Profiling and Personalization using Robust Fuzzy Relational Clustering**», in «**ECommerce and Intelligent Methods**» in the series "Studies in Fuzziness and Soft Computing", J. Segovia, P. Szczepaniak, and M. Niedzwiedzinski, Ed, SpringerVerlag, 2002

[23] Μηχανή αναζήτησης

http://en.wikipedia.org/wiki/Web_search_engine

[24] Alan Emtage – Archie

http://en.wikipedia.org/wiki/Alan_Emtage

[25] File transfer protocol

http://en.wikipedia.org/wiki/File_Transfer_Protocol

[26] The internet Gopher

http://en.wikipedia.org/wiki/Mark_P._McCahill

[27] Very Easy Rodent Oriented Net-wide Index to Computerized Archives (Veronica)

http://linux.about.com/cs/linux101/g/veronicalparver.htm

[28] Jonzy's Universal Gopher Hierarchy Excavation And Display (Jughead),

http://dictionary.babylon.com/JUGHEAD

[29] Wide Web wanderer, Matthew Gray

http://en.wikipedia.org/wiki/World_Wide_Web_Wanderer

[30] Wandex

http://tech-fact.blogspot.com/2006/01/wandex.html

[31] Archie-Like Indexing of the Web (ALIWEB), Martijn Koster,

http://www.aliweb.com

[32] JumpStation

http://en.wikipedia.org/wiki/JumpStation

[33] Lycos

http://en.wikipedia.org/wiki/Lycos

[34] Magellan

http://en.wikipedia.org/wiki/Magellan_(search_engine)

[35] Excite

http://en.wikipedia.org/wiki/Excite

[36] Infoseek

http://en.wikipedia.org/wiki/Infoseek

[37] Inktomi

http://en.wikipedia.org/wiki/Inktomi

[38] Northern Light Group

http://en.wikipedia.org/wiki/Northern_Light_Group

[39] AltaVista

http://en.wikipedia.org/wiki/AltaVista

[40] Yahoo!

http://en.wikipedia.org/wiki/Yahoo!

[41] Netscape

http://en.wikipedia.org/wiki/Netscape

[42] Inbound links

http://en.wikipedia.org/wiki/Inbound_link

[43] Msn search

http://en.wikipedia.org/wiki/Msn_search

[44] Open Directory

http://en.wikipedia.org/wiki/Open_Directory_Project

[45] Nielsen Online

http://www.nielsen-online.com

[46] AOL

http://www.aol.com/

[47] ASK

http://www.ask.com/

[48] Experian Hitwise

http://www.hitwise.com

[49] Merriam-Webster

http://www.merriam-webster.com/

[50] Meta element

http://en.wikipedia.org/wiki/Meta_element

[51] Κύκλος ζωής μια αναζήτησης στη Google

http://www.google.com.gr/intl/el/corporate/tech.html

[52] Υπηρεσίες αναζήτησης Google

http://www.google.com/support/websearch/bin/static.py?page=guide.cs&guide=1221265&answer=142143

[53] Επιλογές αναζήτησης

http://www.google.com/support/websearch/bin/static.py?page=guide.cs&guide=1221265&answer=136861

[54] Yahoo! Slurp

http://en.wikipedia.org/wiki/Yahoo!_Slurp

[55] LookSmart

http://en.wikipedia.org/wiki/LookSmart

[56] Live Search

http://en.wikipedia.org/wiki/Bing#Live_Search

[57] Αλγοριθμοι Clustering

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/

[58] Μέγεθος ευρετηρίων των μηχανών αναζήτησης

http://www.worldwidewebsize.com/

[59] Dogpile

http://en.wikipedia.org/wiki/Dogpile

http://www.dogpile.com

[60] Ithaki

http://www.ithaki.net/indexu.htm

[61] Metacrowler

http://en.wikipedia.org/wiki/Metacrawler

http://www.metacrawler.com/

[62] Intelliseek (Profussion)

http://www.profusion.com/

[63] Ixquick Metasearch

http://en.wikipedia.org/wiki/Ixquick

http://www.ixquick.com/

[64] Copernic

http://www.copernic.com/

[65] Mamma

http://en.wikipedia.org/wiki/Mamma.com

http://www.mamma.com/

[66] MetaFind

http://www.softsea.com/review/MetaFind.html

[67] SavvySearch

http://www.search.com/

[68] Highway61

http://www.highway61.com/

[69] Eclipse

http://www.eclipse.org/

[70] Netbeans

http://netbeans.org/

[71] Drag and drop

http://en.wikipedia.org/wiki/Drag-and-drop

[72] Selection Sort

http://en.wikipedia.org/wiki/Selection_sort

[73] Bubble Sort

http://en.wikipedia.org/wiki/Bubble_sort

[74] Μέθοδοι απομόνωσης

SELVADURAI, SANTTHOSH BABU. **Implementing a Metasearch Framework with Content-Directed Result Merging.** (Under the direction of Dr. Gregory T. Byrd.)

ΠΑΡΑΡΤΗΜΑ – ΠΗΓΑΙΟΣ ΚΩΔΙΚΑΣ

Class Gui.java

```
import java.io.IOException;
import javax.swing.text.BadLocationException;

public class Gui extends javax.swing.JFrame {

   /**
            *
            */
           private static final long serialVersionUID = 1L;
           /** Creates new form ContactEditorUI */

           public Gui() {
       initComponents();
   }

           private void initComponents() {

       jLabel2 = new javax.swing.JLabel();
       jTextField1 = new javax.swing.JTextField();
       jButton1 = new javax.swing.JButton();
       jLabel3 = new javax.swing.JLabel();
       jLabel4 = new javax.swing.JLabel();
       jLabel5 = new javax.swing.JLabel();
       jComboBox1 = new javax.swing.JComboBox();
       jLabel1 = new javax.swing.JLabel();

       setDefaultCloseOperation(javax.swing.WindowConstants.EXIT_ON_CLOSE);
       setIconImages(null);
       setMinimumSize(new java.awt.Dimension(1065, 730));
       getContentPane().setLayout(null);

       jLabel2.setFont(new java.awt.Font("Segoe Print", 1, 18)); // NOI18N
       jLabel2.setForeground(new java.awt.Color(255, 255, 255));
       jLabel2.setText("University of Central Greece");
       getContentPane().add(jLabel2);
       jLabel2.setBounds(780, 630, 260, 40);

       jTextField1.setFont(new java.awt.Font("Tahoma", 1, 14)); // NOI18N
       jTextField1.addActionListener(new java.awt.event.ActionListener() {
          public void actionPerformed(java.awt.event.ActionEvent evt) {
             try {
                                                 jTextField1ActionPerformed(evt);
                                       } catch (BadLocationException e) {

                                                 e.printStackTrace();
```

```
                                    }
        }
    });
    getContentPane().add(jTextField1);
    jTextField1.setBounds(230, 310, 535, 39);

    jButton1.setFont(new java.awt.Font("Segoe Print", 1, 24));
    jButton1.setText("Search");
    jButton1.addActionListener(new java.awt.event.ActionListener() {

    public void actionPerformed(java.awt.event.ActionEvent evt) {
        try {
                                        jButton1ActionPerformed(evt);
                                    } catch (BadLocationException e) {

                                        e.printStackTrace();
                                    }
        }
    });

    getContentPane().add(jButton1);
    jButton1.setBounds(770, 310, 113, 40);

    jLabel3.setFont(new java.awt.Font("Segoe Print", 1, 18));
    jLabel3.setForeground(new java.awt.Color(255, 255, 255));
    jLabel3.setText("2011-2012");
    getContentPane().add(jLabel3);
    jLabel3.setBounds(920, 660, 120, 30);

    jLabel4.setFont(new java.awt.Font("Segoe Print", 1, 36));
    jLabel4.setForeground(new java.awt.Color(255, 255, 255));
    jLabel4.setHorizontalAlignment(javax.swing.SwingConstants.CENTER);
    jLabel4.setText("Engine");
    getContentPane().add(jLabel4);
    jLabel4.setBounds(0, 50, 130, 80);

    jLabel5.setFont(new java.awt.Font("Segoe Print", 1, 36));
    jLabel5.setForeground(new java.awt.Color(255, 255, 255));
    jLabel5.setHorizontalAlignment(javax.swing.SwingConstants.CENTER);
    jLabel5.setText("MetaSearch");
    getContentPane().add(jLabel5);
    jLabel5.setBounds(0, 0, 230, 80);

    jComboBox1.setFont(new java.awt.Font("Segoe Print", 1, 12));
    jComboBox1.setModel(new javax.swing.DefaultComboBoxModel(new String[] { "10
Αποτελέσματα", "20 Αποτελέσματα", "30 Αποτελέσματα", "50 Αποτελέσματα", "100
Αποτελέσματα" }));
    getContentPane().add(jComboBox1);
    jComboBox1.setBounds(230, 350, 150, 30);

    jLabel1.setIcon(new javax.swing.ImageIcon(getClass().getResource("earth.gif")));
```

```
        jLabel1.setMaximumSize(new java.awt.Dimension(1100, 1050));
        jLabel1.setMinimumSize(new java.awt.Dimension(1100, 1100));
        jLabel1.setPreferredSize(new java.awt.Dimension(1100, 1050));
        getContentPane().add(jLabel1);
        jLabel1.setBounds(0, 0, 1050, 790);

        pack();
    }// </editor-fold>

private void jTextField1ActionPerformed(java.awt.event.ActionEvent evt) throws
BadLocationException  {
            String query = jTextField1.getText();

            int count=10;
            String str = (String)jComboBox1.getSelectedItem();
                        if(str.equals("10 Αποτελέσματα"))count=10;
                        if(str.equals("20 Αποτελέσματα"))count=20;
                        if(str.equals("30 Αποτελέσματα"))count=30;
                         if(str.equals("50 Αποτελέσματα"))count=50;
                        if(str.equals("100 Αποτελέσματα"))count=100;

             setVisible(false);

    try {
                        new GetResults(query,count);
            } catch (IOException e) {

                        e.printStackTrace();
            }

}

private void jButton1ActionPerformed(java.awt.event.ActionEvent evt) throws
BadLocationException  {
            String query = jTextField1.getText();

            int count=10;
            String str = (String)jComboBox1.getSelectedItem();
                        if(str.equals("10 Αποτελέσματα"))count=10;
                        if(str.equals("20 Αποτελέσματα"))count=20;
                         if(str.equals("30 Αποτελέσματα"))count=30;
                         if(str.equals("50 Αποτελέσματα"))count=50;
                         if(str.equals("100 Αποτελέσματα"))count=100;

             setVisible(false);
    try {
                        new GetResults(query,count);
            } catch (IOException e) {

                        e.printStackTrace();
            }
}
```

```
    /**
     * @param args the command line arguments
     */
    public static void main(String args[]) {
        /* Set the Nimbus look and feel */
        //<editor-fold defaultstate="collapsed" desc=" Look and feel setting code (optional) ">
        /* If Nimbus (introduced in Java SE 6) is not available, stay with the default look and
feel.
         * For details see
http://download.oracle.com/javase/tutorial/uiswing/lookandfeel/plaf.html
         */
        try {
            for (javax.swing.UIManager.LookAndFeelInfo info :
javax.swing.UIManager.getInstalledLookAndFeels()) {
                if ("Nimbus".equals(info.getName())) {
                    javax.swing.UIManager.setLookAndFeel(info.getClassName());
                    break;
                }
            }
        } catch (ClassNotFoundException ex) {

java.util.logging.Logger.getLogger(Gui.class.getName()).log(java.util.logging.Level.SEVERE
, null, ex);
        } catch (InstantiationException ex) {

java.util.logging.Logger.getLogger(Gui.class.getName()).log(java.util.logging.Level.SEVERE
, null, ex);
        } catch (IllegalAccessException ex) {

java.util.logging.Logger.getLogger(Gui.class.getName()).log(java.util.logging.Level.SEVERE
, null, ex);
        } catch (javax.swing.UnsupportedLookAndFeelException ex) {

java.util.logging.Logger.getLogger(Gui.class.getName()).log(java.util.logging.Level.SEVERE
, null, ex);
        }
        //</editor-fold>

        /* Create and display the form */
        java.awt.EventQueue.invokeLater(new Runnable() {

            public void run() {
                new Gui().setVisible(true);
            }
        });
    }
    private javax.swing.JButton jButton1;
    private javax.swing.JLabel jLabel1;
    private javax.swing.JLabel jLabel2;
    private javax.swing.JLabel jLabel3;
    private javax.swing.JLabel jLabel4;
```

```
    private javax.swing.JLabel jLabel5;
    private javax.swing.JComboBox<?> jComboBox1;
    private javax.swing.JTextField jTextField1;
    // End of variables declaration
}
```

Class GetResults.java

```
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;
import java.io.IOException;
import java.lang.System;
import javax.swing.text.BadLocationException;

public class GetResults {

        public GetResults(String key, int count) throws IOException,
BadLocationException {

        char[] cArray = key.toCharArray();
        for (int i=0; i<cArray.length; i++ ){
            if (cArray[i]==' ')
                    cArray[i]='+';
                }
        String key2 = new String(cArray);
        String [] BingResults=BingResults(key2);
        String [] YahooResults=YahooResults(key2);
        String [] GoogleResults=GoogleResults(key2);
        String [] GoogleText=GoogleText(key2);
        String [] BingText=BingText(key2);
        String [] YahooText=YahooText(key2);

        newMergeResults(BingResults,YahooResults,GoogleResults,count,GoogleText,Bing
Text,YahooText);
        }

        private String[] YahooText(String key2) throws IOException {
         int i=0;
         int j=0;
         int k=0;
        String []pinakas1= new String[100];
        String []pinakas2= new String[40];
        String []pinakas3= new String[20];

        String url1 = "http://search.yahoo.com/search?n=40&p="+key2;
        String url2 = "http://search.yahoo.com/search?n=40&p="+key2+"&b=41";
        String url3 = "http://search.yahoo.com/search?n=20&p="+key2+"&b=81";
```

```
            Document doc1 = Jsoup.connect(url1).get();
            Document doc2 = Jsoup.connect(url2).get();
            Document doc3 = Jsoup.connect(url3).get();

            Elements links1 = doc1.select(".abstr");  //(".body") finds  elements anywhere under
class "body"
            Elements links2 = doc2.select(".abstr");
            Elements links3 = doc3.select(".abstr");

            for (Element link : links1) {
          pinakas1[i]=link.text(); //text:Gets the combined text of this element and all its children
          i++;
        }
            for (Element link : links2) {
           pinakas2[j]=link.text();
           j++;
         }
            for (Element link : links3) {
           pinakas3[k]=link.text();
           k++;
         }
            System.arraycopy(pinakas2, 0, pinakas1, 40, 40);
            System.arraycopy(pinakas3, 0, pinakas1, 80, 20);

            return pinakas1;
            }

            private String[] BingText(String key2) throws IOException {
             int i=0;
             int j=0;
            String []pinakas1= new String[100];
            String []pinakas2= new String[55];
            String url1 = "http://www.bing.com/search?q="+key2+"&count=50";
            String url2 = "http://www.bing.com/search?q="+key2+"&count=50&first=51";

            Document doc1 = Jsoup.connect(url1).get();
            Document doc2 = Jsoup.connect(url2).get();

            Elements links1 = doc1.select(".sa_cc p");  //(".body p") finds p elements anywhere
under a block with class "body"
            Elements links2 = doc2.select(".sa_cc p");

            for (Element link : links1) {
          pinakas1[i]=link.text(); //text:Gets the combined text of this element and all its children
          i++;
        }
            for (Element link : links2) {
           pinakas2[j]=link.text();
           j++;
         }
            System.arraycopy(pinakas2, 0, pinakas1, 50, 50);
            return pinakas1 ;
```

```
        }

        private String[] GoogleText(String key2) throws IOException {
        int i=0;
        int j=0;
        String []pinakas1= new String[100];
        String []pinakas2= new String[55];
        String url1 = "http://www.google.com/search?num=50&q="+key2;
        String url2 = "http://www.google.com/search?num=50&q="+key2+"&start=50";

        Document doc1 = Jsoup.connect(url1).userAgent("Internet Explorer").get();
        Document doc2 = Jsoup.connect(url2).userAgent("Internet Explorer").get();

        Elements links1 = doc1.select(".s");  //(".body p") finds p elements anywhere under a
block with class "body"
        Elements links2 = doc2.select(".s");

        for (Element link : links1) {
       pinakas1[i]=link.text(); //text:Gets the combined text of this element and all its children
       i++;
     }
        for (Element link : links2) {
       pinakas2[j]=link.text(); //text:Gets the combined text of this element and all its
children
       j++;
     }
        System.arraycopy(pinakas2, 0, pinakas1, 50, 50);
        for (int y=0; y<pinakas1.length; y++){
                if(pinakas1[y]==null) pinakas1[y]="xx";
        }


        for (int y=0; y<pinakas1.length; y++){
                char[] cArray = pinakas1[y].toCharArray();
                int count=0;
                for (int z=0; z<cArray.length; z++){count++;}
                if (pinakas1[y]!="xx")
                pinakas1[y]=pinakas1[y].substring(0,count-19);
  }
                return pinakas1;
        }

        private String[] GoogleResults(String key2)throws IOException {

         int i=0;
         int j=0;
        String []pinakas1= new String[100];
        String []pinakas2= new String[55];
        String url1 = "http://www.google.com/search?num=50&q="+key2;
        String url2 = "http://www.google.com/search?num=50&q="+key2+"&start=50";

        Document doc1 = Jsoup.connect(url1).userAgent("Internet Explorer").get();
```

```
        Document doc2 = Jsoup.connect(url2).userAgent("Internet Explorer").get();

        Elements links1 = doc1.select(".s cite");  //(".body p") finds p elements anywhere
under a block with class "body"
        Elements links2 = doc2.select(".s cite");

        for (Element link : links1) {
       pinakas1[i]=link.text(); //text:Gets the combined text of this element and all its children
       i++;
     }
        for (Element link : links2) {
       pinakas2[j]=link.text();
       j++;
     }
        System.arraycopy(pinakas2, 0, pinakas1, 50, 50);

        for (int y=0; y<pinakas1.length; y++){
                if(pinakas1[y]==null) pinakas1[y]="xx";
        }


        int count;
        for (int y=0; y<pinakas1.length; y++){
                char[] cArray = pinakas1[y].toCharArray();
                count=0;
        if (pinakas1[y].contains("›")){
        for (int k=0; k<cArray.length; k++ ){
                count++;
                        if (cArray[k]=='›'){
                        pinakas1[y]=pinakas1[y].substring(0,count-2);
                        break;}
                }
        }
        for (int k=0; k<cArray.length; k++) count++;
        if (cArray[count-1]=='/') pinakas1[y]=pinakas1[y].substring(0,count-1);
        }
        return pinakas1;
        }

        private String[] YahooResults(String key2) throws IOException {
                int i=0;
        int j=0;
         int k=0;
        String []pinakas1= new String[100];
        String []pinakas2= new String[45];
        String []pinakas3= new String[25];

        String url1 = "http://search.yahoo.com/search?n=40&p="+key2;
        String url2 = "http://search.yahoo.com/search?n=40&p="+key2+"&b=41";
        String url3 = "http://search.yahoo.com/search?n=20&p="+key2+"&b=81";

        Document doc1 = Jsoup.connect(url1).get();
```

```
            Document doc2 = Jsoup.connect(url2).get();
            Document doc3 = Jsoup.connect(url3).get();

            Elements links1 = doc1.select(".url");  //(".body p") finds p elements anywhere under
a block with class "body"
            Elements links2 = doc2.select(".url");
            Elements links3 = doc3.select(".url");


            for (Element link : links1) {
          pinakas1[i]=link.text(); //text:Gets the combined text of this element and all its children
          i++;
        }
            for (Element link : links2) {
           pinakas2[j]=link.text();
           j++;
         }
            for (Element link : links3) {
           pinakas3[k]=link.text();
           k++;
         }
            System.arraycopy(pinakas2, 0, pinakas1, 40, 40);
            System.arraycopy(pinakas3, 0, pinakas1, 80, 20);
            for (int y=0; y<pinakas1.length; y++){
                        if(pinakas1[y]==null) pinakas1[y]="";}

                        return pinakas1;
            }

            private String[] BingResults(String key2) throws IOException {
             int i=0;
             int j=0;
            String []pinakas1= new String[100];
            String []pinakas2= new String[55];
                        String url1 = "http://www.bing.com/search?q="+key2+"&count=50";
            String url2 = "http://www.bing.com/search?q="+key2+"&count=50&first=51";

            Document doc1 = Jsoup.connect(url1).get();
            Document doc2 = Jsoup.connect(url2).get();

            Elements links1 = doc1.select(".sb_meta cite");  //(".body p") finds p elements
anywhere under a block with class "body"
            Elements links2 = doc2.select(".sb_meta cite");

            for (Element link : links1) {
          pinakas1[i]=link.text(); //text:Gets the combined text of this element and all its children
          i++;
        }
            for (Element link : links2) {
           pinakas2[j]=link.text();
           j++;
         }
```

```
        System.arraycopy(pinakas2, 0, pinakas1, 50, 50);
        for (int y=0; y<pinakas1.length; y++){
                if(pinakas1[y]==null) pinakas1[y]="";}

        return pinakas1;
        }
}
```

Class MergeResults.java

```
import java.util.Arrays;
import java.lang.System;
import java.lang.String;
import javax.swing.text.BadLocationException;

public class MergeResults  {

        public MergeResults(String[] bingResults, String[] yahooResults,
         String[] googleResults, int count2, String[] googleText, String[] bingText, String[]
yahooText) throws BadLocationException {

        String[] sum=new String[300];
        System.arraycopy(bingResults, 0, sum, 0, 100);
        System.arraycopy(yahooResults, 0, sum, 100, 100);
        System.arraycopy(googleResults, 0, sum, 200, 100);


        int count=0;
         for (int x=0; x<sum.length-1; x++){
                for (int y=x+1; y<sum.length; y++){
                        if (sum[x].equals(sum[y]))
                          sum[y]="zzz";
                }
          }

        Arrays.sort(sum);

        for (int x=0; x<sum.length; x++){
                if (sum[x]!="zzz")count++;
                }

        int a,b,c,sums;

        String [ ][ ] merge= new String [count][2];
        for (int x=0; x<count; x++){
                a=b=c=0;
                merge[x][0]=sum[x];
                for (int y=0; y<100; y++){
                        if (merge[x][0].equals(bingResults[y]))  a=y+1;
                        if (merge[x][0].equals(yahooResults[y])  b=y+1;
```

```
                    if (merge[x][0].equals(googleResults[y])) c=y+1;
                    }
                      if (a==0) a=101;
                      if (b==0) b=101;
                      if (c==0) c=101;

 if(a==101 && b<10 && c<10){a=(int)(b+c)/2;}
 if(a<10 && b==101 && c<10){b=(int)(a+c)/2;}
 if(a<10 && b<10 && c==101){c=(int)(a+b)/2;}

 sums=(int)(a+b+c);

 if (a<10 && b<10 && c<10) {
 sums=(int) (sums * 0.9);}
 merge[x][1]=Integer.toString(sums);
}

String temp1,temp2;
int x1,x2;

        for(int i = 0; i < count; i++){
                for(int j = count-1; j>i; j--){

                  x1=Integer.parseInt(merge[j-1][1]);
                  x2=Integer.parseInt(merge[j][1]);

                        if(x1>x2){

                         temp1 = merge[j-1][1];
                         temp2 = merge[j-1][0];
                         merge[j-1][1]=merge[j][1];
                         merge[j-1][0]=merge[j][0];
                         merge[j][1]=temp1;
                         merge[j][0]=tem
                        }
                }
          }

String [ ] text= new String [count];
for (int x=0; x<count; x++){
     for (int y=0; y<100; y++){
     if (merge[x][0].equals(googleResults[y])){ text[x]=googleText[y];break;}
else if (merge[x][0].equals(googleResults[y])){ text[x]=bingText[y];break;}
else if (merge[x][0].equals(yahooResults[y])){ text[x]=yahooText[y];break;}
               }
                    }
                for(int x=0; x<text.length; x++){
                      if(text[x]==null)
                      text[x]="No description available";
                }
                new GuiResults(merge,count2,text);
```

```
        String[] engine= new String [count];

                        for (int x=0; x<count; x++){
                                String google="";
                                   String yahoo="";
                                   String bing="";

        for (int y=0; y<100; y++){
           if(merge[x][0].equals(googleResults[y])) google="Google ";
           if(merge[x][0].equals(yahooResults[y])) yahoo="Yahoo ";
           if(merge[x][0].equals(bingResults[y]))  bing="Bing";
                                   }

           engine[x]=(google+yahoo+bing);

                                                           }
           new GuiResults(merge,count2,text,engine);
}
}
```

Class GuiResults.java

```
import java.awt.Color;
import java.net.MalformedURLException;
import javax.swing.text.BadLocationException;
import javax.swing.text.SimpleAttributeSet;
import javax.swing.text.StyleConstants;
import javax.swing.text.StyledDocument;

public class GuiResults extends javax.swing.JFrame {

   public GuiResults(String[][] merge, int count2, String[] text, String[] engine) throws
BadLocationException {

        setVisible(true);
        try {
                        initComponents(merge,count2,text,engine);
                } catch (MalformedURLException e) {
                        // TODO Auto-generated catch block
                        e.printStackTrace();
                }
   }


   private void initComponents(String[][] merge, int count2, String[] text, String[] engine)
throws BadLocationException, MalformedURLException {

      jScrollPane1 = new javax.swing.JScrollPane();
```

```
    jTextPane1 = new javax.swing.JTextPane();
    jLabel2 = new javax.swing.JLabel();
    jLabel3 = new javax.swing.JLabel();
    jLabel1 = new javax.swing.JLabel();

    setDefaultCloseOperation(javax.swing.WindowConstants.EXIT_ON_CLOSE);
    setBackground(new java.awt.Color(204, 204, 204));
    setForeground(java.awt.Color.white);
    setIconImages(null);
    setMinimumSize(new java.awt.Dimension(1370, 730));
    getContentPane().setLayout(null);

    jTextPane1.setFont(new java.awt.Font("Tahoma", 1, 16));
    jTextPane1.setEditable(false);
    jScrollPane1.setViewportView(jTextPane1);

    getContentPane().add(jScrollPane1);
    jScrollPane1.setBounds(40, 80, 1250, 530);

    jLabel2.setFont(new java.awt.Font("Segoe Print", 1, 18));
    jLabel2.setText("University of Central Greece");
    getContentPane().add(jLabel2);
    jLabel2.setBounds(1080, 615, 260, 35);

    jLabel3.setFont(new java.awt.Font("Segoe Print", 1, 18));
    jLabel3.setText("2011-2012");
    getContentPane().add(jLabel3);
    jLabel3.setBounds(1215, 650, 120, 20);

    jLabel1.setFont(new java.awt.Font("Segoe Print", 1, 36));
    jLabel1.setText("MetaSearch Results");
    getContentPane().add(jLabel1);
    jLabel1.setBounds(470, 10, 370, 40);

    StyledDocument doc = jTextPane1.getStyledDocument();
    SimpleAttributeSet style1 = new SimpleAttributeSet();
    SimpleAttributeSet style2 = new SimpleAttributeSet();
    StyleConstants.setForeground(style1, Color.red);
    StyleConstants.setForeground(style2, Color.gray);
    StyleConstants.setBold(style1, true);
    StyleConstants.setUnderline(style1, true);

    for (int x=0; x<count2; x++){
        doc.insertString(doc.getLength(), (x+1)+") " +text[x]+"\n"+"     ", null);
        doc.insertString(doc.getLength(),"" + merge[x][0], style1);
        doc.insertString(doc.getLength(),"  (" + engine[x] + ")\n\n", style2);

       }
    pack();
}
private javax.swing.JLabel jLabel1;
private javax.swing.JLabel jLabel2;
```

```
    private javax.swing.JLabel jLabel3;
    private javax.swing.JScrollPane jScrollPane1;
    private javax.swing.JTextPane jTextPane1;
    // End of variables declaration
}
```