ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΣΠΟΥΔΩΝ ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΔΗΜΗΤΡΙΟΣ Ν. ΖΑΧΟΣ

ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΒΕΛΤΙΩΣΗ ΤΗΣ ΑΠΟΔΟΣΗΣ ΚΑΤΑΝΕΜΗΜΕΝΩΝ ΣΥΣΤΗΜΑΤΩΝ

Θεσσαλονίκη 2003

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΣΠΟΥΔΩΝ ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΔΗΜΗΤΡΙΟΣ Ν. ΖΑΧΟΣ

ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΒΕΛΤΙΩΣΗ ΤΗΣ ΑΠΟΔΟΣΗΣ ΚΑΤΑΝΕΜΗΜΕΝΩΝ ΣΥΣΤΗΜΑΤΩΝ

Επιβλέπων Καθηγητής: Μανωλόπουλος Ιωάννης
Εξεταστής: Μαργαρίτης Κωνσταντίνος
Ημερομηνία Παρουσίασης: 14 Φεβρουαρίου 2003

Θεσσαλονίκη 2003

ΖΑΧΟΣ Ν. ΔΗΜΗΤΡΙΟΣ

"Δεν ψάχνω, βρίσκω..."

Pablo Picasso

Περίληψη

Το θέμα της παρούσας διπλωματικής εργασίας είναι η βελτίωση της απόδοσης σε Κατανεμημένα Συστήματα, όπως για παράδειγμα ο Παγκόσμιος Ιστός (World Wide Web) και τα Κινητά Περιβάλλοντα Υπολογισμού (Mobile Environments), χρησιμοποιώντας τεχνικές που αναπτύχθηκαν από τον ερευνητικό χώρο της Εξόρυξης Δεδομένων (Data Mining).

Τα κίνητρα της εργασίας μας είναι η αξιοποίηση του τεράστιου όγκου των δεδομένων που συσσωρεύονται σε αυτά τα κατανεμημένα συστήματα με σκοπό την καλύτερη κατανομή των πόρων του συστήματος.

Γενικός στόχος της παρούσας εργασίας είναι να μελετήσει το κατά πόσο είναι εφικτό να υιοθετηθούν οι προαναφερθείσες τεχνικές σε πραγματικά περιβάλλοντα. Για το λόγο αυτό θα αναζητηθεί η ύπαρξη “κανονικότητας” (patterns) στην συμπεριφορά των χρηστών των συγκεκριμένων συστημάτων. Επιπλέον, θα εξερευνηθεί εάν η κανονικότητα αυτή μπορεί να αξιοποιηθεί εύκολα από το σύστημα.

Πιο συγκεκριμένα, η παρούσα εργασία θα μελετήσει την ύπαρξη κανονικότητας στις επισκέψεις χρηστών σε πραγματικούς Δικτυακούς τόπους (Web sites). Αυτή η κανονικότητα μπορεί να αξιοποιηθεί για τους σκοπούς της πρόβλεψης της συμπεριφοράς των επισκεπτών.

Στη μελέτη θα χρησιμοποιηθούν τα πραγματικά δεδομένα των επισκέψεων στον δικτυακό τόπο του Μακεδονικού Πρακτορείου Ειδήσεων (http://www.mpa.gr).

Η εργασία απαντά καταφατικά στο ερώτημα της ύπαρξης κανονικότητας. Επιπλέον, ένα από τα ευρήματά της είναι ότι οι κανόνες που περιγράφουν αυτή την κανονικότητα είναι σχετικά λίγοι τον αριθμό και μπορούν να βρεθούν πολύ γρήγορα, χωρίς επιβάρυνση του συστήματος. Το γενικό συμπέρασμα της μελέτης είναι ότι η εξόρυξη γνώσης από μεγάλο όγκο πρωτογενών δεδομένων σε κατανεμημένα συστήματα, μπορεί να χρησιμοποιηθεί επιτυχώς, έχοντας απώτερο στόχο τη βελτίωση της επίδοσης των συστημάτων αυτών.

Abstract

The present thesis focuses on the improvement of the performance of Distributed Systems, such as the World Wide Web (WWW) and the Mobile Environments, using modern techniques developed by the research field of Data Mining.

The motivation of our work is to process and exploit the great amount of data gathered by these distributed systems, and improve the systems' performance through better resource management.

A general objective of the current work is to study the feasibility of all the formerly mentioned techniques facing problems and manipulating data taken from the real world. For this reason, we examine the existence of regularities (patterns) regarding the users' behavior in the former distributed systems. Moreover, we will examine if these patterns can be used easily to build on the systems' performance. More specifically, the present thesis will study the existence of regularities in the users' visits of real web sites of the world wide web. This regularity can be used in order to make predictions on the visitors' behavior. In our study we will use and process real web user traces taken by the web site of the Macedonian Press Agency (http://www.mpa.gr/).

The results of the present work prove the existence of regularity in the users' trace. What is more, one of its conclusions is that the rules describing this regularity are not so many and can be quickly found without requiring many system resources. The general conclusion of the present study is that the implementation of Knowledge Discovery procedures on a great amount of data found in modern distributed systems, can prove very successful, especially for the improvement of the overall performance in these systems.

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Δ' εξάμηνου σπουδών στο Διατμηματικό Μεταπτυχιακό στα Πληροφοριακά Συστήματα (Master in Information Systems – MIS) του Πανεπιστημίου Μακεδονίας κατά τη χρονική περίοδο Σεπτεμβρίου 2002 – Φεβρουαρίου 2003.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Ι. Μανωλόπουλο (Τμήμα Πληροφορικής-Α.Π.Θ.) και τον εξεταστή καθηγητή κ. Κ. Μαργαρίτη (Τμήμα Εφαρμοσμένης Πληροφορικής-ΠΑ.ΜΑΚ.) για την επιστημονική καθοδήγηση που μου παρείχαν σε όλη τη διάρκεια της εκπόνησης της διπλωματικής εργασίας. Ιδιαίτερα θα ήθελα να ευχαριστήσω τον κ. Ι. Μανωλόπουλο για τη δυνατότητα που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα, στα πλαίσια του οποίου είχαμε μια άψογη συνεργασία. Επίσης, ευχαριστώ τον κ. Δ. Κατσαρό υποψήφιο Διδάκτορα στο τμήμα Πληροφορικής του Α.Π.Θ., για την πολύτιμη βοήθειά του στο ερευνητικό και τεχνικό μέρος της εργασίας. Η συμβολή του στην τελική μορφή της διπλωματικής εργασίας, ήταν ουσιαστική και απολύτως απαραίτητη.

Επίσης, ευχαριστώ θερμά το Μακεδονικό Πρακτορείο Ειδήσεων και ιδιαίτερα τον υπεύθυνο για την ανάπτυξη του ηλεκτρονικού κόμβου, κ. Λ. Μακρή για την άψογη συνεργασία και την παραχώρηση των χρήσιμων και εμπιστευτικών δεδομένων για ερευνητικούς σκοπούς. Εύχομαι η προσπάθεια για αξιοποίηση του μεγάλου όγκου των δεδομένων να καρποφορήσει στο μέλλον ακόμα περισσότερο.

Θα ήταν παράλειψη από μέρους μου να μην ευχαριστήσω το Ινστιτούτο Πληροφορικής και Τηλεματικής (Ι.Π.ΤΗΛ.) του Εθνικού Κέντρου Έρευνας και Τεχνολογικής Ανάπτυξης (Ε.Κ.Ε.Τ.Α.) και τον διευθυντή του κ. Μ. Στρίντζη για την τιμητική υποτροφία που μου παρείχαν για την εκπόνηση των μεταπτυχιακών μου σπουδών.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την πολύπλευρη συμπαράστασή της στο πρόσωπό μου, όλα αυτά τα χρόνια των σπουδών. Δεν μπορώ να μην αναφερθώ στους παλιούς και νέους φίλους για την υπομονή και την πολύτιμη στήριξή τους στις δύσκολες στιγμές. Η διπλωματική αυτή εργασία είναι αφιερωμένη σε όλους αυτούς.

Ζάχος Δημήτρης
Θεσσαλονίκη, Φεβρουάριος 2003

Περιεχόμενα

1 Κεφάλαιο: Ζητήματα επίδοσης σε κατανεμημένα συστήματα

Την τελευταία δεκαετία παρατηρείται έντονη δραστηριότητα σχετικά με την ανάπτυξη κατανεμημένων συστημάτων όπως ο Παγκόσμιος Ιστός και τα Κινητά Περιβάλλοντα Υπολογισμών. Οι σύγχρονες εφαρμογές που αναπτύσσονται για τα συστήματα αυτά και η ευρεία διάδοσή τους, έχουν ως αποτέλεσμα να έχουν γίνει ιδιαίτερα δημοφιλή, προκαλώντας όμως έτσι προβλήματα απόδοσης λόγω των περιορισμένων διαθεσίμων πόρων που δεν αρκούν για την εξυπηρέτηση του αυξημένου όγκου των χρηστών. Ως χαρακτηριστικό παράδειγμα αναφέρουμε τις καθυστερήσεις που παρατηρούνται για την μεταφόρτωση πολλών σελίδων στο διαδίκτυο. Παρόμοια προβλήματα απόδοσης, παρατηρούνται στα κινητά περιβάλλοντα υπολογισμών, όπου το διαθέσιμο εύρος ζώνης δεν επαρκεί για να καλύψει τις αυξανόμενες απαιτήσεις των χρηστών.

Στο εισαγωγικό αυτό κεφάλαιο παραθέτουμε τις θεμελιώδεις αρχές που διέπουν τα δύο αυτά συστήματα με ιδιαίτερη έμφαση στην ανάλυση της περίπτωσης του Παγκόσμιου Ιστού.

1.1 Ο Παγκόσμιος Ιστός

1.1.1 Αρχιτεκτονική του Παγκόσμιου Ιστού

Από τις αρχές της δεκαετίας του '90, ο Παγκόσμιος Ιστός ή World Wide Web έγινε η πιο δημοφιλής υπηρεσία του διαδικτύου αφού μας επιτρέπει να βλέπουμε όλα τα περιεχόμενα του διαδικτύου σε ένα γραφικό και εύκολο στη χρήση περιβάλλον. Ο Παγκόσμιος Ιστός είναι ένα Κατανεμημένο Σύστημα που έχει υλοποιηθεί σύμφωνα με το μοντέλο πελάτη-εξυπηρετητή (Client/Server Model) που επιτρέπει στους χρήστες να ανακαλούν πληροφορίες χωρίς να χρειάζεται να γνωρίζουν το πού είναι αποθηκευμένες αυτές οι πληροφορίες.

Στο μοντέλο client-server, ο πελάτης (client) υποβάλλει μια αίτηση και ο εξυπηρετητής (server) επιστρέφει μια ανταπόκριση ή κάνει μια σειρά από ενέργειες για να ανταποκριθεί. Ο server μπορεί να ενεργοποιείται άμεσα για την αίτηση αυτή ή να προσθέτει την αίτηση σε μια ουρά. Η τοποθέτηση της αίτησης σε μια ουρά, μπορεί να σημαίνει ότι η αίτηση πρέπει να τεθεί σε αναμονή για να εξυπηρετηθεί. Ένα παράδειγμα αφορά στην εκτύπωση ενός κειμένου σε έναν εκτυπωτή δικτύου. Ο server τοποθετεί την αίτηση σε μια ουρά μαζί με αιτήσεις εκτυπώσεων και από άλλους clients. Μετά επεξεργάζεται την αίτηση με βάση την σειρά προτεραιότητας, η οποία σε αυτή την περίπτωση καθορίζεται από τη σειρά με την οποία ο server παρέλαβε την απαίτηση.

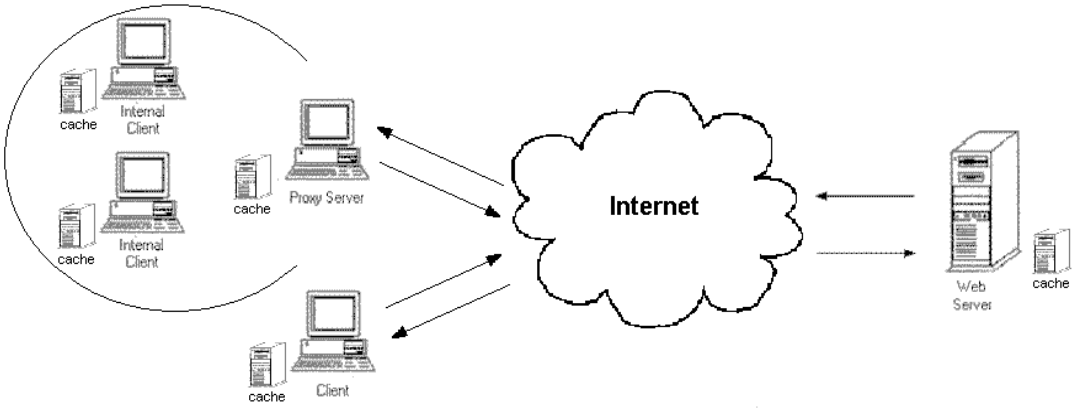
Server:

Ο server βρίσκεται στο επίκεντρο της αρχιτεκτονικής του μοντέλου. Πρόκειται συνήθως για έναν ισχυρό υπολογιστή, όπου εγκαθίσταται το λογισμικό server, με κύρια αποστολή να ρυθμίζει τη διακίνηση πακέτων δεδομένων μεταξύ διαφόρων clients. Είναι ένας κεντρικός υπολογιστής ο οποίος εξυπηρετεί άλλους υπολογιστές (Clients) και προμηθεύει τους χρήστες αυτών των υπολογιστών με τα δεδομένα που του ζητήθηκαν.

Client:

Πρόκειται για το πρόγραμμα «πελάτης» με το οποίο ζητούνται πληροφορίες από έναν εξυπηρετητή (Server). Οι clients είναι υπολογιστές που τρέχουν το κατάλληλο λογισμικό ώστε να μπορούν να επικοινωνούν με τον server. Ζητούν από αυτόν τα αρχεία που χρειάζονται, τα οποία βρίσκονται είτε στη βάση δεδομένων του είτε σε άλλη βάση ή ακόμη σε άλλους clients και ο server αναλαμβάνει να τους εξυπηρετήσει.

Το παρακάτω σχήμα απεικονίζει γραφικά την αρχιτεκτονική του Παγκόσμιου Ιστού.



Σχήμα 1. Αρχιτεκτονική Παγκοσμίου Ιστού.

Η φράση «αρχιτεκτονική client-server» δεν προσδιορίζει μονοσήμαντα τον τρόπο με τον οποίο λειτουργεί το εν λόγω μοντέλο. Υπάρχουν διάφορες παραλλαγές που χρησιμοποιούνται ανάλογα με το είδος και το μέγεθος των αναγκών κατά περίπτωση, αλλά βέβαια ξεφεύγουν από τα πλαίσια της παρούσας αναφοράς. Γενικά, κυρίαρχο χαρακτηριστικό του μοντέλου είναι η ανταπόκρισή του στις απαιτήσεις για κεντρική διαχείριση, αξιόπιστη αποθήκευση δεδομένων και ασφάλεια συναλλαγών.

Η βασική μονάδα πληροφοριών στον Ιστό είναι ένα έγγραφο (που συχνά ονομάζεται σελίδα). Οι πληροφορίες που περιέχονται σε μια σελίδα του Ιστού

μπορούν να πάρουν πολύ διαφορετικές μορφές: εκτός από κείμενο, τα έγγραφα του Ιστού, μπορούν να περιέχουν εικόνες, ηχητικά αποσπάσματα και οδηγίες για την εκτέλεση προγραμμάτων, είτε τοπικά είτε στον υπολογιστή που φιλοξενεί έναν εξυπηρετητή Ιστού.

Οι εξυπηρετητές δικτύου αναμένουν τις ενδεχόμενες εντολές του πρωτοκόλλου HTTP στη θύρα TCP 80 και απαντούν ανάλογα με το περιεχόμενο της εντολής. Οι απαντήσεις στις αιτήσεις των χρηστών είναι συνήθως Υπερκείμενο (hyper-text), δομημένο σύμφωνα με το πρότυπο HTML και που μπορεί να περιλαμβάνει κείμενο, ήχο, στατικές και κινούμενες εικόνες, ακόμα και video.

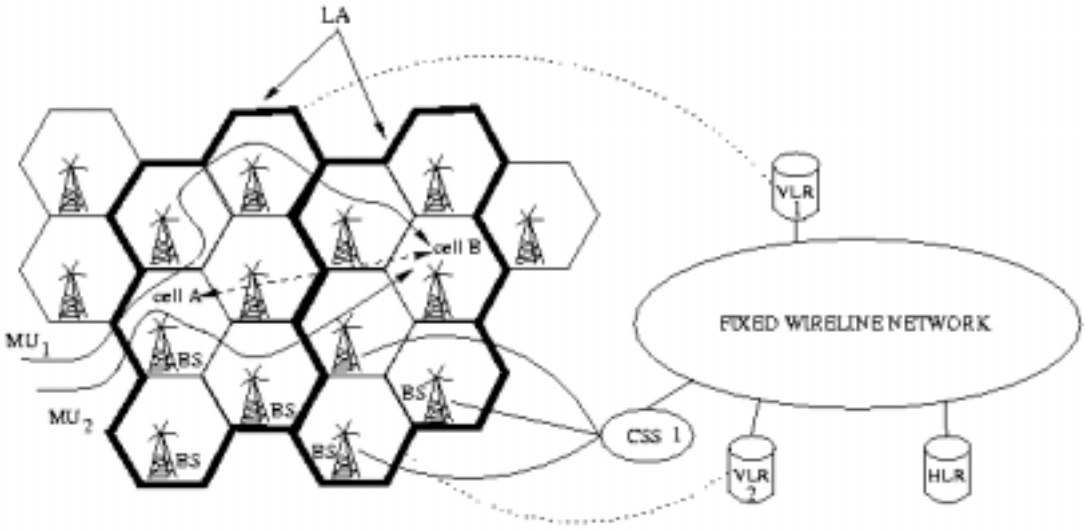
Το πρόγραμμα-πελάτης που χρησιμοποιείται για τις πληροφορίες του Ιστού ονομάζεται φυλλομετρητής (browser). Το πρωτόκολλο που χρησιμοποιείται από τον Παγκόσμιο Ιστό είναι το HTTP (Hypertext Transfer Protocol) και είναι το σύστημα που χρησιμοποιούν οι browsers και οι servers για να επικοινωνούν μεταξύ τους. Η γλώσσα επικοινωνίας που χρησιμοποιείται για την παρουσίαση και τη σύνδεση των εγγράφων είναι η HTML (Hypertext Mark-up Language). Παραπομπές σε άλλες σελίδες ή περιεχόμενο γίνονται με την τυποποιημένη χρήση των Uniform Resource Locators (URL). Με τη χρήση των προγραμμάτων περιήγησης (browsers) οι πληροφορίες παρουσιάζονται σε μορφή ιστοσελίδων (web pages). Ένας Web Browser για παράδειγμα θεωρείται ως πελάτης (client) σε ένα Web Server. Να σημειωθεί όμως, ότι ο Παγκόσμιος Ιστός αποτελεί μόνο μία από τις πολλές δυνατότητες επικοινωνίας που διαθέτει το Internet.

1.1.2 Μεγάλος Χρόνος απόκρισης

Το σημαντικότερο πρόβλημα για το οποίο παραπονιούνται οι χρήστες του διαδικτύου είναι ο μεγάλος χρόνος απόκρισης και οι καθυστερήσεις μεταφοράς από τους εξυπηρετητές των αντικειμένων που ζητούν. Ο μηχανισμός του Caching που έχει προταθεί για τη βελτίωση της κατάστασης, έχει αποδειχθεί ότι δεν μπορεί πάντα να λύσει το πρόβλημα, επειδή υπάρχει ένας μεγάλος αριθμός εγγράφων που ζητιούνται μόνο μία φορά από το χρήστη. Θα βοηθούσε η ύπαρξη ενός μηχανισμού ικανού να προβλέψει τις μελλοντικές αιτήσεις για έγγραφα από τους χρήστες προς τους εξυπηρετητές.

1.2 Κινητά περιβάλλοντα υπολογισμού

Οι ραγδαίες τεχνολογικές εξελίξεις τόσο στην Πληροφορική, όσο και στις Τηλεπικοινωνίες, έδωσαν μεγάλη ώθηση για την ανάπτυξη συστημάτων προσωπικής επικοινωνίας (Personal Communications Systems - PCS). Τα συστήματα PCS, στηρίζονται στη συνεχή διαθεσιμότητα υπηρεσιών που διευκολύνουν την πρόσβαση σε διάφορα είδη πληροφοριών όπως φωνή, κείμενο, εικόνες, βίντεο, ανεξάρτητα από το χρόνο, τον τόπο και τον τρόπο πρόσβασης σε αυτά. Οι εν λόγω υπηρεσίες βασίζονται στις έννοιες της Ασύρματης Πρόσβασης και στα Δίκτυα που επιτρέπουν την επικοινωνία ανάμεσα σε ανθρώπους και τοποθεσίες και όχι μόνο ανάμεσα σε τοποθεσίες.



Σχήμα 2. Τυπική αρχιτεκτονική κινητού περιβάλλοντος υπολογισμών.

Αντίθετα με τα συνηθισμένα στατικά δίκτυα, όπως για παράδειγμα το δημόσιο δίκτυο σταθερής τηλεφωνίας, τα PCS επιτρέπουν τη δυναμική επανατοποθέτηση των κινητών τερματικών, με συνέπεια τα σημεία πρόσβασης των χρηστών στο δίκτυο να αλλάζουν καθώς οι χρήστες μετακινούνται σε διαφορετικές τοποθεσίες. Αυτή η κινητικότητα, προκαλεί μια σειρά από προβλήματα, μεταξύ των οποίων η διαχείριση των επιμέρους περιοχών και η κατανομή του εύρους συχνότητας. Η διαχείριση περιοχής αποτελείται από τις επιμέρους διαδικασίες: (α) της χωρικής καταγραφής (Location) και (β) της αναζήτησης (Paging). Η πρώτη διαδικασία επιτρέπει στο σύστημα να καταγράφει τις πληροφορίες (ακριβείς ή κατά προσέγγιση) σχετικά με τη θέση του κάθε

χρήστη, έτσι ώστε να είναι σε θέση να τον εντοπίζει όποτε αυτό χρειάζεται. Η διαδικασία της αναζήτησης αποτελείται από την αποστολή μηνυμάτων προς όλες τις κατευθύνσεις (κελιά) για τον εντοπισμό του χρήστη.

Τα ζητήματα σχετικά με την κατανομή εύρους ζώνης, προκύπτουν λόγω της φύσης των PCS, καθώς αυτά βασίζονται στην έννοια της κατανομής εύρους ζώνης μετά από ζήτηση. Το σύστημα δηλαδή παραχωρεί πόρους, μόνον αφού προηγουμένως αυτοί έχουν ζητηθεί από κάποιο χρήστη (bandwidth-on-demand).

Η έρευνα για τη διαχείριση περιοχής, έχει εστιαστεί κυρίως σε ζητήματα που αφορούν την ανανέωσης της βάσης δεδομένων και σχετικά λίγα έχουν γίνει προς την κατεύθυνση της πρόβλεψης. Η πρόβλεψη της θέσης του χρήστη αποτελεί μία δυναμική στρατηγική, κατά την οποία το PCS εκτιμά εκ των προτέρων τη θέση του κινητού πελάτη και μπορεί να αξιοποιηθεί με πολλούς τρόπους, για να βελτιωθεί η γενική αποτελεσματικότητα του συστήματος.

1.2.1 Ζητήματα κατανομής εύρους ζώνης

Γενικά, λόγω του περιορισμένου εύρους συχνότητας, υπάρχει κάποιο μέγιστο όριο όσον αφορά τον αριθμό των χρηστών που μπορούν συγχρόνως να εξυπηρετηθούν σε ένα κελί. Επομένως, υπάρχει η πιθανότητα κατά τη μετακίνηση του χρήστη από τη μια περιοχή στην άλλη να διακοπεί η σύνδεσή του ή να απορριφθεί μία νέα σύνδεση, λόγω της έλλειψης του απαιτούμενου εύρους ζώνης. Η πρόβλεψη της μετακίνησης θα μπορούσε να βοηθήσει σημαντικά στη λήψη αποφάσεων σχετικά με την κατανομή του διαθέσιμου εύρους ζώνης στα γειτονικά κελιά. Αντί να παραχωρείται “τυφλά” το εύρος ζώνης σε όλα τα κελιά, θα μπορούσε να παραχωρηθεί μόνο στα κελιά που είναι περισσότερο πιθανό να μετακινηθεί ο χρήστης. Επιπλέον, η σπατάλη πόρων θα μπορούσε να μειωθεί κατά την αναζήτηση του χρήστη από τη διαδικασία εντοπισμού, καθώς το σύστημα θα ήταν σε θέση να αποστέλλει μηνύματα, υπολογίζοντας προηγουμένως την πιθανότητα να βρίσκεται σε καθένα από αυτά.

1.3 Το ζήτημα της πρόβλεψης

Από την παρουσίαση των δυο προηγούμενων παραδειγμάτων κατανεμημένων συστημάτων βλέπουμε ότι κοινός παρονομαστής για την βελτίωση της απόδοσής τους είναι η πρόβλεψη – μελλοντικών αιτήσεων για αντικείμενα (στο Web) και της θέσης ενός κινητού (σε mobile περιβάλλοντα).

2 Κεφάλαιο: Εξόρυξη Δεδομένων

Στην ενότητα αυτή θα προσπαθήσουμε να προσεγγίσουμε το γνωστικό πεδίο της Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων (KDD: Knowledge Discovery in Databases). Αρχικά θα αναφερθούμε στο πλαίσιο και τις γενικές αρχές που διέπουν τον εν λόγω ερευνητικό χώρο και τη σχέση που έχει με την Εξόρυξη Δεδομένων (Data Mining). Θα αναφερθούμε στους στόχους της έρευνας με τη χρήση τεχνικών Εξόρυξης Δεδομένων, καθώς και στα πιθανά πρωτογενή δεδομένα τα οποία μπορούμε να διαχειριστούμε. Κλείνοντας την εισαγωγική αυτή υπό-ενότητα, θα παρουσιάσουμε μία προτεινόμενη αρχιτεκτονική ενός συστήματος εξόρυξης δεδομένων.

Στο δεύτερο μέρος της ενότητας κάνουμε μία εκτενέστερη αναφορά στα στάδια από τα οποία αποτελείται η διαδικασία ανακάλυψης γνώσης καθώς και στις τεχνικές που μπορούμε να υιοθετήσουμε ανάλογα με τα δεδομένα που αναλύονται και τα επιδιωκόμενα αποτελέσματα. Έτσι, αρχικά περιγράφονται τα διαδοχικά στάδια τα οποία ακολουθούνται συνήθως και στα οποία διακρίνουμε την Διαδικασία Ανακάλυψης Γνώσης. Τέλος, αναλύουμε τις διαφορετικές τεχνικές και τους αλγορίθμους που χρησιμοποιεί η καθεμιά για την εξόρυξη δεδομένων. Περισσότερη έμφαση δίδεται στη μέθοδο της Εξόρυξης Κανόνων Συσχέτισης και στην τεχνική του αλγορίθμου Apriori που χρησιμοποιείται σε αυτή τη μέθοδο.

2.1 Γενικά

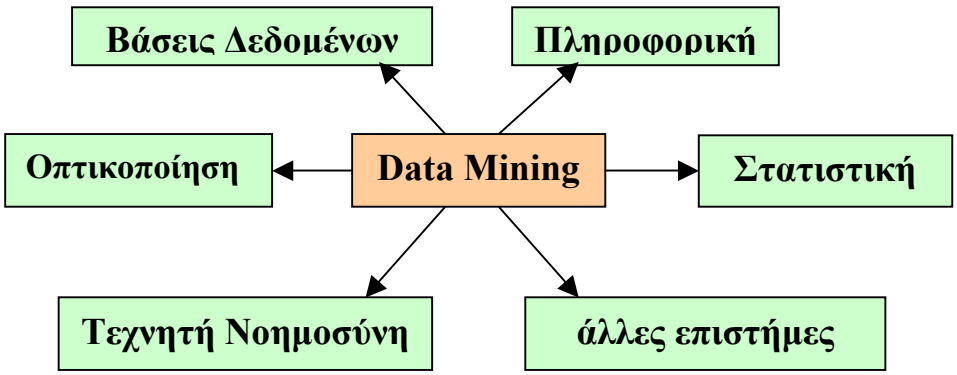
Όλοι μας είμαστε εξοικειωμένοι με το φαινόμενο της **έκρηξης πληροφοριών**. Τα Πρακτικά του Συνεδρίου για την Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Conference in Knowledge Discovery in Databases -KDD) του 1995 ξεκινούν ως εξής:

> « ...εκτιμάται ότι η ποσότητα της πληροφορίας σε όλο τον κόσμο διπλασιάζεται κάθε είκοσι μήνες. Τι θα πρέπει να κάνουμε με αυτόν τον κατακλυσμό από ακατέργαστα δεδομένα; Ασφαλώς, λίγα από αυτά θα τα δει το ανθρώπινο μάτι.»

Τα Πληροφοριακά Συστήματα που έχουν διεισδύσει σε όλες τις πτυχές της ανθρώπινης δραστηριότητας, συλλέγουν τα δεδομένα από τις αναρίθμητες καθημερινές συναλλαγές μας: στα ταμεία των πολυκαταστημάτων και των super-markets, στα αυτόματα μηχανήματα των τραπεζών, στην έκδοση αεροπορικών εισιτηρίων, στις τηλεφωνικές κλήσεις κινητής και σταθερής τηλεφωνίας, κλπ. Η αξιοποίηση της πληροφορίας που προκύπτει από αυτά τα δεδομένα, είναι δυνατό

να αποτελέσει ένα ισχυρό συγκριτικό πλεονέκτημα για κάθε οργανισμό, στο σύγχρονο ανταγωνιστικό περιβάλλον.

Η **Εξόρυξη Δεδομένων** έχει προσελκύσει στις μέρες μας την ιδιαίτερη προσοχή και το ενδιαφέρον της επιστημονικής και της επιχειρηματικής κοινότητας. Ο κυριότερος λόγος για αυτήν την τάση είναι η ύπαρξη τεράστιου όγκου δεδομένων και η ανάγκη για άμεση μετατροπή των δεδομένων αυτών σε χρήσιμη **Πληροφορία** και **Γνώση**. Έτσι, το τρίπτυχο Δεδομένα – Πληροφορία – Γνώση στην «Κοινωνία της Πληροφορίας», στηρίζεται σε μεγάλο βαθμό στην ανάπτυξη αποτελεσματικών και έγκυρων εργαλείων εξόρυξης δεδομένων. Η πληροφορία και γνώση που αποκτάται μπορεί να αξιοποιηθεί σε μια πλειάδα από πεδία εφαρμογής όπως: η διοίκηση επιχειρήσεων και λήψη αποφάσεων, ο έλεγχος παραγωγής, η ανάλυση αγοράς, ο μηχανικός σχεδιασμός (science engineering) και επιστημονικές έρευνες (science exploration).



Σχήμα 3 Data Mining και τα σχετικά πεδία έρευνας.

Η ευρεία χρήση των υπολογιστών στις συναλλαγές μας σε όλους τους τομείς της σύγχρονης κοινωνίας (στο χώρο των επιχειρήσεων, της βιομηχανίας, των επιστημών) καθώς και τα πολλαπλά πλεονεκτήματα που παρέχουν τα σύγχρονα εργαλεία συλλογής δεδομένων (Συστήματα Διαχείρισης Βάσεων Δεδομένων, Αρχείων, κλπ) έχουν οδηγήσει στην συγκέντρωση μεγάλου όγκου πληροφορίας. Για το λόγο αυτό ένα νέο πεδίο έρευνας το οποίο αφορά στην διαδικασία εξόρυξης γνώσης και πληροφορίας από μεγάλα συστήματα βάσεων δεδομένων (KDD and Data Mining) άρχισε να κάνει την εμφάνιση του. Το ερευνητικό αυτό πεδίο καθώς αποτελεί συνισταμένη πολλών επιστημών, έχει επίσης προσελκύσει το έντονο ενδιαφέρον πολλών και διαφορετικών επιστημονικών πεδίων, όπως οι βάσεις δεδομένων, η τεχνητή νοημοσύνη, η στατιστική, η παρουσίαση των δεδομένων (Data Visualization).

2.2 Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (KDD)

Ως Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων(Knowledge Discovery in Databases) μπορούμε να ορίσουμε τη **διαδικασία της εύρεσης νέων, έγκυρων, χρήσιμων και κατανοητών προτύπων από τα δεδομένα**. Θα μπορούσαμε να πούμε ότι η Εξόρυξη Δεδομένων είναι ένα στάδιο της ευρύτερης διαδικασίας ανακάλυψης γνώσης. Παρόλα' αυτά ο όρος Data Mining(εξόρυξη δεδομένων) στην ερευνητική κοινότητα έχει επικρατήσει εξίσου με τον όρο Ανακάλυψη Γνώσης που αποδίδει καλύτερα και πιο ολοκληρωμένα το σύνολο της διαδικασίας. Παραπλήσια ορολογία που χρησιμοποιείται προκειμένου να προσδιοριστεί αυτή η ερευνητική περιοχή είναι : Ανακάλυψη γνώσης σε βάσεις δεδομένων (KDD), Εξαγωγή γνώσης, Ανάλυση Δεδομένων / Προτύπων, Ανάκτηση Δεδομένων κλπ.

Τι είδους δεδομένα ενδείκνυνται για μελέτη και ανάλυση από εργαλεία εξόρυξης δεδομένων;

- Σχεσιακές βάσεις δεδομένων
- Αποθήκες δεδομένων
- Βάσεις δεδομένων συναλλαγών
- Αντικειμενοστραφείς βάσεις δεδομένων
- Βάσεις Χωρικών δεδομένων
- Χρονοσειρές και χρονικά δεδομένα
- Βάσεις δεδομένων κειμένου και πολυμέσων
- Το Διαδίκτυο

2.3 Στόχοι της Εξόρυξης Δεδομένων και Ανακάλυψης Γνώσης

Η Εξόρυξη Δεδομένων (Data Mining) αποτελεί μια ραγδαία αναπτυσσόμενη τεχνολογία και περιλαμβάνει ένα σύνολο από μεθοδολογίες που υποστηρίζουν, μεταξύ των άλλων, την εξαγωγή ανεξερεύνητης γνώσης από μεγάλες βάσεις δεδομένων, αποθήκες δεδομένων, δεξαμενές πληροφορίας, χωρικές βάσεις δεδομένων, βάσεις κειμένου και πολυμέσων κ.ά.. Η Εξόρυξη Δεδομένων χρησιμοποιεί διάφορες τεχνικές, όπως ανάλυση περιοχών, κανόνες συσχέτισης, κατηγοριοποίηση, δένδρα αποφάσεων, εννοιολογική περιγραφή, κλπ που θα αναλύσουμε σε επόμενη ενότητα, οι οποίες βρίσκουν εφαρμογή σε πολλά και διαφορετικά πεδία:

- Εμπορικές εφαρμογές (Μάρκετινγκ, Customer Relation Management-CRM, Ανάλυση Καλαθιού αγοράς)
- Τηλεπικοινωνίες (Πρόβλεψη για την κατανομή πόρων, ανακάλυψη απάτης)
- Τραπεζικά συστήματα (Ανάλυση τάσεων, δανειοδότηση)
- Εφαρμογές στην Υγεία (DNA mining, Βιοτεχνολογία)
- Τεχνολογίες Διαδικτύου (Web mining, Web log analysis, Ηλεκτρονικό Εμπόριο)
- Εξόρυξη δεδομένων από κείμενα (βιβλιοθήκες, βάσεις δεδομένων, άλλα έγγραφα)

Οι κυριότεροι στόχοι της Εξόρυξης Δεδομένων μπορούν να συνοψιστούν στα εξής:

1. Πρόβλεψη: Η εξόρυξη δεδομένων μπορεί να δείξει την συμπεριφορά κάποιων γνωρισμάτων των δεδομένων στο μέλλον.
2. Ταυτοποίηση: Οι μορφές των δεδομένων μπορούν να χρησιμοποιηθούν για να προσδιορισθεί η ύπαρξη ενός προϊόντος, ενός γεγονότος, η μιας δραστηριότητας.
3. Ταξινόμηση: Η εξόρυξη δεδομένων μπορεί να διαμερίσει τα δεδομένα ώστε να μπορούν να προσδιορισθούν διαφορετικές κλάσεις ή κατηγορίες με βάση συνδυασμούς παραμέτρων.
4. Βελτιστοποίηση: Ένας ενδεχόμενος στόχος της εξόρυξης δεδομένων μπορεί να είναι η βελτιστοποίηση της χρήσης μέσων όπως ο χρόνος, ο χώρος, το χρήμα, ή τα υλικά και η μεγιστοποίηση των μεταβλητών εξόδου όπως οι πωλήσεις ή τα κέρδη δοθέντων κάποιων περιορισμών. Σαν τέτοιος, αυτός ο στόχος της εξόρυξης δεδομένων προσομοιάζει την αντικειμενική συνάρτηση που χρησιμοποιείται στα προβλήματα επιχειρησιακής έρευνας που αντιμετωπίζει βελτιστοποιήσεις υπό περιορισμούς.

2.4 Αρχιτεκτονική Συστήματος Ανακάλυψης Γνώσης

Παρακάτω θα περιγράψουμε σύντομα τα επίπεδα της αρχιτεκτονικής ενός τυπικού συστήματος Ανακάλυψης Γνώσης.

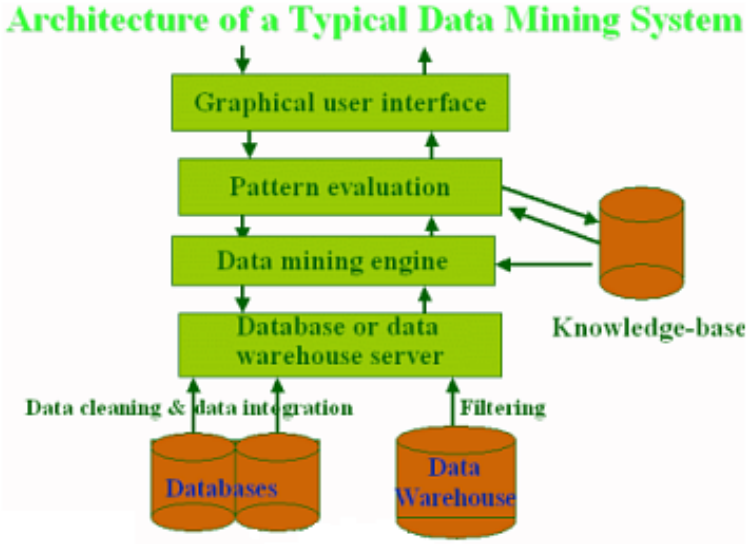
- **Βάσεις ή Αποθήκες Δεδομένων (Databases or Data Warehouses)**:

Εδώ είναι αποθηκευμένα τα δεδομένα που μας ενδιαφέρουν. Οι αποθηκευτικοί χώροι μπορεί να είναι είτε απλές ή κατανεμημένες βάσεις δεδομένων, είτε

αποθήκες δεδομένων. Σε αυτά τα δεδομένα θα εφαρμοσθούν τα αρχικά στάδια της διαδικασίας εξόρυξης δεδομένων (καθαρισμός και ενοποίηση των δεδομένων)

- **Εξυπηρετητής Βάσης ή Αποθήκης Δεδομένων (DB or data warehouse Server)**:

Ο εξυπηρετητής του συστήματος της βάσης ή της αποθήκης δεδομένων είναι επιφορτισμένος με την μεταφορά των κατάλληλων δεδομένων ανάλογα με τις απαιτήσεις των χρηστών του συστήματος.



Σχήμα 4 Αρχιτεκτονική συστήματος Ανακάλυψης Γνώσης.

- **Βάση Γνώσης (Knowledge - base)**:

Αναφέρεται σε όλες τις πληροφορίες και τα δεδομένα της γνωστικής περιοχής που αναλύεται. Μπορεί να περιλαμβάνει πληροφορίες και μετά-δεδομένα που περιγράφουν τις δομές και τις ιεραρχίες των υπό ανάλυση δεδομένων. Επίσης, μπορεί να περιλαμβάνει και τις αντιλήψεις και τις εξειδικευμένες γνώσεις των ειδικών της περιοχής που μελετάται. Αυτές οι πληροφορίες και γνώσεις καθορίζουν κατ' αυτόν τον τρόπο και κάποια σημαντικά μεγέθη της ανάλυσης όπως τα ελάχιστα και μέγιστα επιτρεπτά όρια κάποιων μεταβλητών, τα διαστήματα εμπιστοσύνης κλπ.

- **Μηχανή εξόρυξης δεδομένων (Data mining engine):**

Είναι το απολύτως απαραίτητο στοιχείο στην αρχιτεκτονική του συστήματος εξόρυξης δεδομένων. Περιλαμβάνει τα λειτουργικά τμήματα που σχετίζονται με τις τεχνικές της συσχέτισης, ταξινόμησης, ομαδοποίησης κλπ.

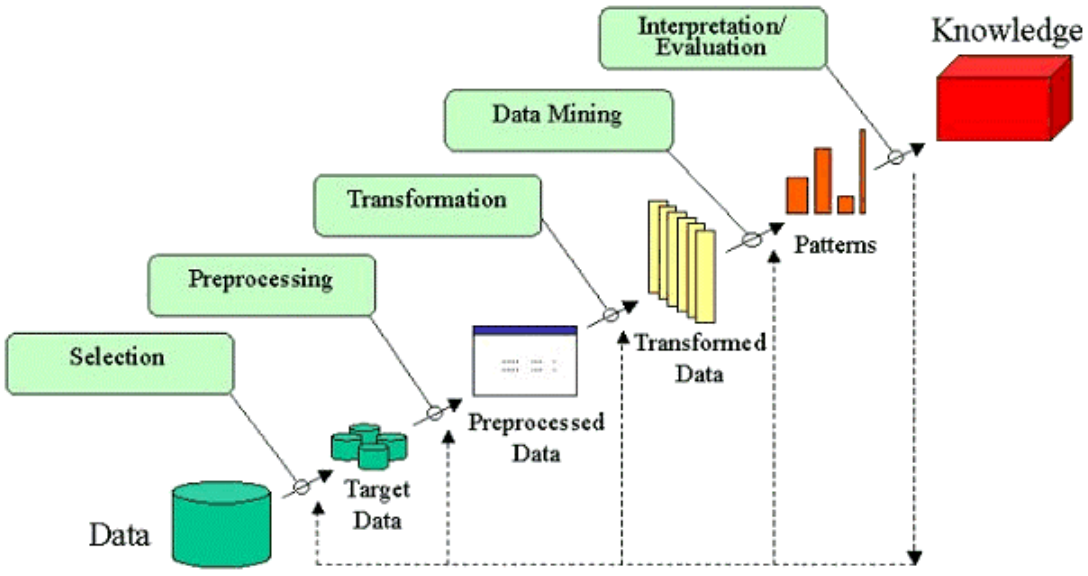
- **Αξιολόγηση των προτύπων που προκύπτουν (Pattern Evaluation):**

Το στοιχείο αυτό της αρχιτεκτονικής του συστήματος εξόρυξης δεδομένων αναφέρεται συνήθως στην αλληλεπίδραση των μεγεθών που προσδιορίζουν το βαθμό του ενδιαφέροντος ενός κανόνα, και βοηθά να εστιάσουμε καλύτερα την προσοχή προς αυτά. Μπορεί να χρησιμοποιηθούν κατώφλια (thresholds) σημαντικότητας των προτύπων και μερικές φορές είναι δυνατό να ενσωματωθεί αυτό το στοιχείο της αξιολόγησης με την εξόρυξη αυτή καθεαυτή.

- **Γραφικό περιβάλλον Διεπαφής με το Χρήστη (Graphical User Interface):**

Είναι το επίπεδο της επικοινωνίας του συστήματος με τον χρήστη. Επιτρέπει στον χρήστη να αλληλεπιδρά με το σύστημα, εισάγοντας τους δικούς του στόχους και τα δικά του ερωτήματα(queries). Επιπλέον, μπορεί ο χρήστης να ελέγξει το σχήμα της βάσης ή της αποθήκης δεδομένων και τα αποτελέσματα της διαδικασίας σε κάθε στάδιό της και να κάνει σε εκείνο το σημείο τις επιλογές του. Τέλος σημαντικό στοιχείο του Γραφικού Περιβάλλοντος διεπαφής είναι η παρουσίαση των αποτελεσμάτων σε διαφορετικές φόρμες.

2.5 Στάδια ανάπτυξης της διαδικασίας Εξόρυξης Δεδομένων



Σχήμα 5 Τα στάδια της διαδικασίας Ανακάλυψης Γνώσης.

Παρακάτω αναλύουμε ξεχωριστά, το καθένα από τα στάδια της διαδικασίας Ανακάλυψης Γνώσης:

2.5.1 Επιλογή δεδομένων

Κατά την έναρξη της διαδικασίας εξαγωγής γνώσης, είναι απαραίτητο να εστιάσουμε την προσοχή μας όχι στο σύνολο των διαθέσιμων στοιχείων, αλλά στις μεταβλητές ή τα δείγματα δεδομένων από τα οποία πρόκειται να εξαχθεί η γνώση. Αυτόματα περιορίζεται η μελέτη στα συγκεκριμένα δεδομένα που μας ενδιαφέρουν, και αφαιρούνται από την ανάλυσή μας όλα τα υπόλοιπα.

Τελικά, θα πρέπει να έχει οριστεί το σύνολο των δεδομένων στα οποία θα εφαρμοστεί η διαδικασία, επιλέγεται δηλαδή το training set, τα κατάλληλα πεδία κλπ. Συνήθως τα δεδομένα είναι οργανωμένα σε σχεσιακές βάσεις δεδομένων (relational databases) που προορίζονται για άλλη χρήση. Επειδή οι αλγόριθμοι που εκτελούν την αναζήτηση γνώσης δεν μπορούν να εκτελεστούν με μεγάλη ευκολία σε πολλαπλούς πίνακες δεδομένων, απαιτείται η εξαγωγή των δεδομένων από αυτές και η οργάνωσή τους σε δομές που είναι καλύτερα προσβάσιμες από τους συγκεκριμένους αλγόριθμους.

2.5.2 Καθαρισμός των δεδομένων

Στο στάδιο αυτό είναι απαραίτητο να γίνει και ο καθορισμός των πρωτογενών δεδομένων. Είναι πιθανό τα δεδομένα να περιλαμβάνουν θόρυβο ή να μην είναι πλήρη και ολοκληρωμένα. Επομένως, είναι σημαντικό να μελετηθούν οι στρατηγικές με τις οποίες θα αντιμετωπίσουμε αυτά τα προβλήματα, όπως για παράδειγμα η προσπάθεια για την αφαίρεση του θορύβου ή αντικατάστασης των ελλιπών στοιχείων.

Ο καθαρισμός των δεδομένων είναι από τα σημαντικότερα στάδια της εξόρυξης δεδομένων και πολλές φορές μπορεί να απαιτεί πολλούς περισσότερους πόρους και χρόνο απ’ ότι αρχικά φαίνεται.

2.5.3 Ενσωμάτωση δεδομένων από διαφορετικές πηγές

Στο στάδιο αυτό συνδυάζονται τα δεδομένα που μπορεί να προέρχονται από διαφορετικές πηγές όπως απομακρυσμένες μεταξύ τους βάσεις δεδομένων, αποθήκες δεδομένων και άλλα μέσα αποθήκευσης πληροφοριών.

2.5.4 Μετασχηματισμός/ Κωδικοποίηση δεδομένων

Στο στάδιο αυτό της διαδικασίας τα δεδομένα μετασχηματίζονται σε δομές που διευκολύνουν την αναζήτηση γνώσης, όπως για παράδειγμα η μείωση του αριθμού των υπό εξέταση μεταβλητών (dimensionality reduction) και η

ομοιόμορφη κωδικοποίηση της ποιοτικά ίδιας πληροφορίας (π.χ. πεδίο salary σε μια βάση δεδομένων και payment σε μια άλλη).

Ο μετασχηματισμός των δεδομένων μπορεί επίσης να επιτευχθεί με διαφορετικούς τρόπους όπως η αφαίρεση των τιμών που έχουν θόρυβο και να περιλαμβάνει την περίληψη και σύνοψη των δεδομένων. Τέλος, είναι δυνατό να δημιουργηθούν νέες μεταβλητές που θα συνθέτουν κάποια επιμέρους πεδία των δεδομένων.

Στο σημείο αυτό ολοκληρώνεται το κεφάλαιο της προ-επεξεργασίας των δεδομένων που στοχεύει στη βελτίωση της αποτελεσματικότητας των διαδικασιών που ακολουθούν.

2.5.5 Εξόρυξη δεδομένων (Data Mining)

Στο στάδιο αυτό επιλέγεται ο αλγόριθμος και γενικότερα η «ευφυής διαδικασία» που θα ακολουθηθεί με σκοπό την εξαγωγή προτύπων από τα δεδομένα. Μπορούμε να διακρίνουμε τις εξής επιμέρους φάσεις:

- *Επιλογή εργασιών εξόρυξης δεδομένων*: Αποφασίζουμε ποιες τεχνικές θα υιοθετηθούν και ποιες εργασίες θα εκτελεστούν αντίστοιχα (π.χ. clustering, classification, κλπ)
- *Επιλογή αλγορίθμου εξόρυξης δεδομένων*: Επιλέγουμε τις μεθόδους που πρόκειται να χρησιμοποιηθούν για την αναζήτηση προτύπων στα δεδομένα. Αυτό περιλαμβάνει απόφαση για το ποια μοντέλα και παράμετροι είναι κατάλληλα να χρησιμοποιηθούν (π.χ. μοντέλα για δεδομένα με λεκτικές τιμές είναι διαφορετικά από τα μοντέλα για δεδομένα με αριθμητικές τιμές), καθώς και αντιστοίχηση μίας δεδομένης μεθόδου data mining με τα συνολικά κριτήρια της διαδικασίας KDD (π.χ. ο τελικός χρήστης μπορεί να ενδιαφέρεται περισσότερο να κατανοήσει το μοντέλο, παρά τις μεθόδους πρόβλεψης).
- *Κυρίως διαδικασία εξόρυξης δεδομένων*: Αναζήτηση των προτύπων που μας ενδιαφέρουν σε μία συγκεκριμένη μορφή αναπαράστασης ή σε ένα σύνολο τέτοιων αναπαραστάσεων, όπως classification rules, decision trees, clustering κλπ.

Επίσης, ένα άλλο είδος διάκρισης που μπορούμε να εισάγουμε, είναι το κατά πόσο τα πρότυπα που θα επιδιώξουμε να εξάγουμε θα έχουν περισσότερο χαρακτηριστικά πληροφόρησης και λιγότερες δυνατότητες πρόβλεψης, ή το

αντίθετο. Κάτι τέτοιο ορίζει την προσέγγιση που επιδιώκουμε και επιτυγχάνεται με την επιλογή διαφορετικών μεθόδων σε κάθε περίπτωση.

2.5.6 Ερμηνεία και αξιολόγηση αποτελεσμάτων

Βασιζόμενοι στις μετρήσεις που κάνουμε πάνω στα αποτελέσματα της ανάλυσης που προηγήθηκε στα προηγούμενα στάδια, μπορούμε να ορίσουμε διάφορα μεγέθη που περιγράφουν το βαθμό αξιοπιστίας και το πόσο ενδιαφέροντα μπορεί να είναι τα πρότυπα που εξάγονται. Επίσης μπορούν να ληφθούν υπόψη οι διαγραμματικές απεικονίσεις των μετρήσεών μας.

Καθώς η διαδικασία εξόρυξης δεδομένων είναι μία δυναμική διαδικασία, τα αποτελέσματα της ερμηνείας και αξιολόγησης των νέων προτύπων μπορούν να τροφοδοτήσουν εκ νέου τα προηγούμενα στάδια της ανάλυσης με νέα δεδομένα και αρχές που θα πρέπει να ληφθούν υπόψη, για την πιο ολοκληρωμένη και επικεντρωμένη στα στοιχεία που μας ενδιαφέρουν ανάλυση.

2.5.7 Παρουσίαση Νέας Γνώσης – Λήψη αποφάσεων

Στο τελικό στάδιο της διαδικασίας εξόρυξης δεδομένων, είναι απαραίτητο όλη η γνώση και τα δεδομένα που συγκεντρώθηκαν, να παρουσιαστούν με τέτοιο τρόπο ώστε να είναι εφικτή η χρησιμοποίησή τους για την περαιτέρω λήψη αποφάσεων. Έτσι σημαντικά ζητήματα που τίθενται είναι αυτά που αφορούν την οπτικοποίηση των αποτελεσμάτων και την αναπαράσταση γνώσης και προς αυτήν την κατεύθυνση αναπτύσσεται σημαντικά η έρευνα στις μέρες μας. Απλούστερες δομές παρουσίασης αποτελούν οι αναφορές με τα δεδομένα που προκύπτουν, γραφικές παραστάσεις και διαγράμματα κατανομών κλπ.

Η νέα γνώση που προκύπτει μπορεί είτε να χρησιμοποιηθεί από τους ειδικούς στη γνωστική περιοχή, είτε να αποθηκευθεί εκ νέου μαζί με το υπόλοιπο γνωστικό υπόβαθρο (knowledge base). Έτσι η διαδικασία της εξόρυξης δεδομένων είναι δυναμική και μπορεί να τροφοδοτήσει νέες διαδικασίες για ανακάλυψης γνώσης.

2.6 Τεχνικές Εξόρυξης Δεδομένων

2.6.1 Εξόρυξη Συσχετίσεων (Association Mining)

Η Ανακάλυψη Γνώσης με τη χρήση τεχνικών εξόρυξης συσχετίσεων αναφέρεται στην ανακάλυψη κανόνων, συχνών προτύπων και γενικότερα

συσχετίσεων ανάμεσα στα δεδομένα. Ένας κανόνας συσχέτισης είναι της μορφής X ή Y όπου τα $X=\{x_1, x_2,\ldots, x_n\}$ και $Y=\{Y_1, Y_2,\ldots, Y_m\}$ είναι σύνολα αντικειμένων, με τα x_i και y_j να είναι διακριτά αντικείμενα για κάθε i και j. Η συσχέτιση αυτή ερμηνεύεται ως εξής: αν για παράδειγμα ένας πελάτης αγοράζει το X προϊόν είναι πιθανό να αγοράσει επίσης και το Y. Γενικά κάθε κανόνας συσχέτισης έχει την μορφή LHS (Left-Hand-Side Αριστερό μέλος) ή RHS (Right-Hand-Side Δεξιό μέλος), όπου τα LHS και RHS είναι σύνολα αντικειμένων.

Δύο μεγέθη πολύ σημαντικά για τον υπολογισμό της αξιοπιστίας ενός κανόνα και στα οποία θα αναφερθούμε εκτενέστερα είναι **η στήριξη** και **η εμπιστοσύνη** του κανόνα. Στήριξη του κανόνα LHS ή RHS είναι το ποσοστό των δοσοληψιών που περιλαμβάνουν όλα τα αντικείμενα της ένωσης LHSURHS. Η εμπιστοσύνη του κανόνα συσχέτισης LHS ή RHS είναι το ποσοστό των δοσοληψιών που περιλαμβάνουν επίσης το RHS. Ένας άλλος όρος για την εμπιστοσύνη είναι η ισχύς του κανόνα

Οι κανόνες συσχέτισης αποτυπώνονται με την εξής μορφή:

- "**Body** →**Head** [Support, Confidence]"

Παράδειγμα ενός κανόνα συσχέτισης:

- αγοράζει (X, “Πάνες”) → αγοράζει (X, “Μπύρες”) [0.5%, 60%]

Πρόκειται για ένα κλασικό παράδειγμα κανόνα συσχέτισης από την ανάλυση του καλαθιού αγοράς (Market basket analysis). Ο κανόνας μας υποδεικνύει ότι για κάποιον καταναλωτή X ισχύει ότι σε μία συναλλαγή του στο σούπερ-μάρκετ με στήριξη 0.5% και εμπιστοσύνη 60%, θα αγοράσει συγχρόνως πάνες και μπύρες.

Σχήμα 6 Παράδειγμα ανάλυσης καλαθιού αγοράς (market basket analysis).

Δημιουργία όλων των αντικειμένων που έχουν στήριξη παραπάνω από το όριο που έχει τεθεί. Αυτά τα σύνολα αντικειμένων ονομάζονται μεγάλα σύνολα

αντικειμένων. Μεγάλο εδώ σημαίνει μεγάλη στήριξη. Για κάθε μεγάλο σύνολο αντικειμένων, δημιουργούνται όλοι οι κανόνες με ελάχιστη εμπιστοσύνη με τον ακόλουθο τρόπο: για ένα μεγάλο σύνολο αντικειμένων X και Y∩X, έστω Z=X-Y τότε αν στήριξη(X)/στήριξη(Z) ή ελάχιστη εμπιστοσύνη, ο κανόνας Z⇒Y (δηλαδή X-Y ή Y) είναι ένας έγκυρος κανόνας.

Μπορούμε να διακρίνουμε τους κανόνες συσχέτισης σε πολυδιάστατους και μονοδιάστατους. Ένα παράδειγμα της διαφοράς είναι οι δύο κανόνες :

(α) ηλικία(X, "20..29") ^ εισόδημα(X, "20..29K") ⇒ αγοράζει(X, "PC") [support = 2%, confidence = 60%]

(β) αγοράζει(T, "Computer") ⇒ αγοράζει(x, "software") [1%, 75%],

όπου στον πρώτο κανόνα (α) υπάρχει συσχέτιση σε δύο διαστάσεις ενώ στον (β) μόνο σε μία.

Ένας κανόνας σε γενικές γραμμές έχει ενδιαφέρον αν είναι:

- εύκολα κατανοητός,
- έγκυρος αν εφαρμοστεί δοκιμαστικά σε νέα δεδομένα
- χρήσιμος στο γνωστικό πεδίο που ερευνάται
- πρωτότυπος

Αντικειμενικά μεγέθη υπολογισμού της αξίας ενός κανόνα είναι αυτά που βασίζονται σε στατιστικά μεγέθη που αναλύουν τα δεδομένα, όπως η στήριξη και η εμπιστοσύνη, ενώ ταυτόχρονα μπορούμε να θεωρήσουμε και κάποια υποκειμενικά μεγέθη που βασίζονται στις αντιλήψεις που έχουν οι χρήστες για τα δεδομένα και μπορεί να είναι ο βαθμός πρωτοτυπίας και κατανοητότητας των κανόνων.

Ο Αλγόριθμος Apriori

Μία από τις περισσότερο γνωστές μεθόδους Εξόρυξης Συσχετίσεων είναι η ανακάλυψη συχνών προτύπων με την υλοποίηση του αλγόριθμου Apriori. Ο αλγόριθμος αυτός δέχεται σαν είσοδό του ένα σύνολο από συναλλαγές π.χ. το καλάθι αγοράς και εξάγει κανόνες συσχέτισης για επαναλαμβανόμενα στοιχεία των συναλλαγών αυτών.

Ο αλγόριθμος Apriori των Agrawal και Srikant(1994) στηρίζεται στην παρατήρηση ότι εάν ένα σύνολο από στοιχεία (items) των συναλλαγών δεν είναι αρκετά συχνά επαναλαμβανόμενο και δεν έχει την κατάλληλη στήριξη, τότε και κάθε υπέρ-σύνολό του δεν θα είναι συχνά επαναλαμβανόμενο και άρα θα πρέπει εκ των προτέρων (A priori) να αποκλειστεί από την διαδικασία εξόρυξης

συσχετίσεων, καθώς κάθε προσπάθεια υπολογισμού της σημασίας του θα είναι περιττή. Για παράδειγμα, αν γνωρίζουμε ότι το σύνολο {Α,Β} δεν επαληθεύει τη στήριξη, τότε συνεπάγεται ότι ούτε το σύνολο {Α, Β, Γ} ούτε το σύνολο {Α, Β, Δ} θα την επαληθεύουν.

Τα βασικά βήματα του αλγορίθμου είναι τα εξής:

1. Καθορισμός της στήριξης για καθένα από τα χαρακτηριστικά του συνόλου των συναλλαγών.
2. Διαγραφή όλων των απλών χαρακτηριστικών που δεν υποστηρίζονται αρκετά από τα δεδομένα.
3. Δημιουργία ζευγών από όλα τα μονά χαρακτηριστικά του συνόλου
4. Αν υπάρχουν ζεύγη στο προηγούμενο βήμα υπολογίζεται η στήριξή τους, διαφορετικά τερματίζεται η διαδικασία.
5. Δημιουργία τριάδων από τα ζεύγη που δημιουργήθηκαν στο προηγούμενο βήμα.
6. Αν έχουν δημιουργηθεί τριάδες υπολογίζεται η στήριξή τους, διαφορετικά τερματίζει η διαδικασία.
7. Η διαδικασία συνεχίζεται μέχρι να μην υπάρχουν άλλα υποψήφια (υπέρ-)σύνολα που να μπορούν να δημιουργηθούν

Μία προσέγγιση στον **ψευδοκώδικα** του αλγορίθμου Apriori είναι η εξής:

C_k: Candidate itemset of size k, L_k : frequent itemset of size k

L_1 = {frequent items};

for (k = 1; L_k != 0; k++) **do begin**

C_{k+1} = candidates generated from L_k;

for each transaction t in database do

increment the count of all candidates in C_{k+1} that are contained in it L_{k+1} = candidates in C_{k+1} with min_support

end

return $\cup_k L_k$;

2.6.2 Ταξινόμηση και Πρόβλεψη (Classification and Prediction)

Η εύρεση μοντέλων που να περιγράφουν τα δεδομένα και να τα διακρίνουν σε κατηγορίες με στόχο τη μελλοντική πρόβλεψη τιμών. Ο διάκριση των τιμών σε κατηγορίες είναι επιβλεπόμενη γνώση και ορίζεται από κάποιον άνθρωπο που χειρίζεται το σύστημα εξόρυξης δεδομένων. Εφόσον μια μεταβλητή ανήκει σε μια

συγκεκριμένη κατηγορία, τότε μπορούμε με κάποιο ποσοστό επιτυχίας να προβλέψουμε το εύρος των τιμών κάποιων σχετιζόμενων παραμέτρων. Η υλοποίηση αυτής της μεθόδου μπορεί να γίνει με διάφορους τρόπους μεταξύ των οποίων η χρήση νευρωνικών δικτύων, δένδρων αποφάσεων, κανόνων κατηγοριοποίησης κλπ. Για παράδειγμα, η κατηγοριοποίηση χωρών με βάση τις κλιματικές τους διαφορές, ή κατηγοριοποίηση αυτοκινήτων με βάση την χρήση καυσίμων.

Η εργασία της ταξινόμησης χαρακτηρίζεται από τον ακριβή καθορισμό των κλάσεων και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ-ταξινομημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορεί να εφαρμοστεί για να ταξινομήσει δεδομένα που δεν έχουν ακόμα ταξινομηθεί. Δεν έχουν τοποθετηθεί δηλαδή σε κάποια από τις υπάρχουσες κλάσεις.

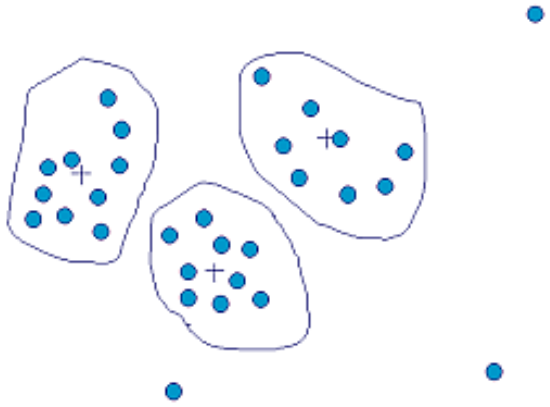
Στις περισσότερες περιπτώσεις, υπάρχει ένας περιορισμένος αριθμός κλάσεων και εμείς θα πρέπει να τοποθετήσουμε κάθε εγγραφή στην κατάλληλη κλάση. Για το σκοπό αυτό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο κατηγορίες. Η πρώτη χρησιμοποιεί ***Δέντρα Αποφάσεων*** *(Decision Trees)* και η δεύτερη ***Νευρωνικά Δίκτυα*** *(Neural Networks).* Και οι δύο στηρίζονται στην ιδέα της “εκπαίδευσης” (training) με τη βοήθεια ενός υποσυνόλου δεδομένων που ονομάζεται *training set*. Το υποσύνολο αυτό επιλέγεται σαν αντιπροσωπευτικό δείγμα του συνολικού όγκου δεδομένων. Με την εφαρμογή της διαδικασίας εκπαίδευσης καθορίζονται κάποια πρότυπα για τις κατηγορίες δεδομένων. Έτσι, όταν προκύψει ένα νέο δεδομένο τότε μπορεί εύκολα να κατηγοριοποιηθεί. Για τη διαδικασία αυτή χρησιμοποιούνται είτε τεχνικές βασισμένες στα νευρωνικά δίκτυα, είτε συμβολικές τεχνικές. Στις πρώτες υπάρχει το φαινόμενο της αμφίδρομης αναμετάδοσης και επεξεργασίας δεδομένων, ενώ στη δεύτερη υπάρχουν μοντέλα δένδρων αποφάσεων ή μοντέλα για IF…THEN…ELSE ανάλυση.

Η πρόβλεψη προτύπων αναφέρεται σε κάποιες αριθμητικές τιμές που πιθανότατα είναι άγνωστες ή λείπουν από τα δεδομένα.

2.6.3 Ομαδοποίηση (Clustering)

Η διαδικασία της ομαδοποίησης φυσικών ή αφηρημένων αντικειμένων σε παρόμοια αντικείμενα ονομάζεται ομαδοποίηση ή μη επιβλεπόμενη ταξινόμηση. Η διαφορά από την ταξινόμηση είναι ότι στην ομαδοποίηση δεν είναι γνωστά εκ των

προτέρων τα σύνολα των δεδομένων που θα προκύψουν. Προσδιορίζονται συστάδες (clusters) ή περιοχές με μεγαλύτερη πυκνότητα τιμών, σύμφωνα με κάποια απόσταση σε ένα σύνολο πολυδιάστατων δεδομένων. Αντίθετα στην ταξινόμηση, ο πληθυσμός διαιρείται σε κλάσεις αναθέτοντας κάθε στοιχείο ή εγγραφή σε μία προκαθορισμένη κλάση με βάση ένα μοντέλο που αναπτύσσεται μέσω της εκπαίδευσης του με παραδείγματα που έχουν κατηγοριοποιηθεί εκ των προτέρων.



Σχήμα 7 Σχηματική απεικόνιση της μεθόδου της Ομαδοποίησης (Clustering).

Στο clustering οι εγγραφές ομαδοποιούνται σε σύνολα με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους. Επαφίεται σε εμάς να καθορίσουμε την σημασία που θα έχει κάθε ένα από τα clusters που προκύπτουν. Για παράδειγμα, τα clusters συμπτωμάτων μπορεί να υποδεικνύουν διαφορετικές ασθένειες, clusters που περιλαμβάνουν τα χαρακτηριστικά που σχετίζονται με τα *φύλλα* και τον *καρπό* φυτών μπορεί να υποδεικνύουν διαφορετικές ποικιλίες ενός φυτού.

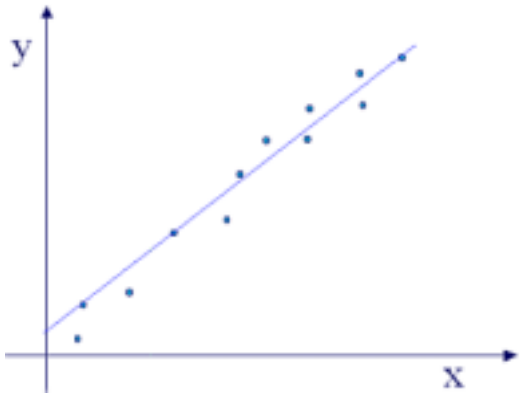
Η μέθοδος clustering μπορεί να χρησιμοποιηθεί και σαν εισαγωγή σε κάποια άλλη μορφή data mining ή μοντελοποίησης. Για παράδειγμα, το clustering μπορεί να χρησιμοποιηθεί σαν πρώτο βήμα στην προσπάθεια μερισμού της αγοράς. Αντί δηλαδή να προσπαθούμε να προσδιορίσουμε τι είδος προώθησης (promotion) θα ταίριαζε καλύτερα σε κάθε πελάτη, μπορούμε να κατηγοριοποιήσουμε τους πελάτες αρχικά σε ομάδες (clusters) ατόμων που παρουσιάζουν τις ίδιες συνήθειες σχετικά με την αγορά προϊόντων και στην συνέχεια να προσδιορίσουμε το είδος του promotion που ταιριάζει σε κάθε ομάδα.

2.6.4 Ανακάλυψη Σειριακών Προτύπων (Serial Analysis)

Η ανακάλυψη των σειριακών προτύπων βασίζεται στην έννοια μιας ακολουθίας συνόλων αντικειμένων. Κατά συνέπεια μπορούμε να πούμε ότι στην ανακάλυψη σειριακών προτύπων σημαντικός παράγοντας είναι ο χρόνος.

Η στήριξη για μια ακολουθία S συνόλων αντικειμένων είναι το ποσοστό των ακολουθιών U στις οποίες η S είναι υπό-ακολουθία. Το πρόβλημα του προσδιορισμού ακολουθιών προτύπων, είναι να βρεθούν όλες οι υπό-ακολουθίες που έχουν ελάχιστη στήριξη που ορίζεται από το χρήστη

2.6.5 Παλινδρόμηση (Regression)



Σχήμα 8 Σχηματική απεικόνιση της Παλινδρόμησης.

Η παλινδρόμηση αναφέρεται στην εκμάθηση μίας συνάρτησης η οποία αντιστοιχεί τα δεδομένα σε μία μεταβλητή πρόβλεψης (prediction variable) πραγματικής τιμής. Οι εφαρμογές του regression είναι πάρα πολλές π.χ. εκτίμηση της πιθανότητας ένας ασθενής να έχει κάποια ασθένεια δεδομένων των αποτελεσμάτων ενός συνόλου διαγνωστικών tests, πρόβλεψη της ζήτησης ενός νέου προϊόντος από τους πελάτες σαν συνάρτηση των εξόδων για διαφήμιση.

2.6.6 Περίληψη (Summarization)

Περιλαμβάνει μεθόδους για την εύρεση μίας περιγραφής για ένα υποσύνολο δεδομένων. Ένα απλό παράδειγμα θα μπορούσε να είναι η εκτίμηση της μέσης και της τυπικής απόκλισης για όλα τα πεδία. Πιο εξεζητημένες λειτουργίες περιλαμβάνουν την παραγωγή συνοπτικών κανόνων, τεχνικές παρουσίασης πολλαπλών μεταβλητών και την ανακάλυψη λειτουργικών σχέσεων μεταξύ των μεταβλητών. Οι εργασίες του *summarization* χρησιμοποιούνται συχνά

στην αλληλεπιδραστική ανάλυση δεδομένων και στην αυτοματοποιημένη παραγωγή αναφορών.

2.6.7 Χαρακτηρισμός και διάκριση (Characterization and discrimination)

Είναι δυνατό να επιτευχθεί η γενίκευση, περίληψη και αντιπαραβολή των δεδομένων. Π.χ. σύνολο υψηλών και χαμηλών τιμών μιας μεταβλητής

2.6.8 Ανάλυση εύρεσης ακραίων τιμών (Outlier analysis)

Αν θεωρήσουμε ότι την περίπτωση που μία μεταβλητή ή ένα σύνολο από αντικείμενα της ανάλυσής μας, παρουσιάζουν τιμές εκτός του αποδεκτού συνόλου τιμών (βλ. Σχήμα Clustering), τότε υπάρχει μεγάλη πιθανότητα να πρόκειται για θόρυβο στα δεδομένα μας, που θα πρέπει να αφαιρεθεί και να συνεχισθεί η διαδικασία της ανακάλυψης γνώσης. Πολύ σημαντική είναι όμως και η περίπτωση ανακάλυψης λάθους, που πολλές φορές μπορεί να υποκρύπτει την ύπαρξη απάτης που θα πρέπει να αντιμετωπιστεί με ανάλογο τρόπο.

3 Κεφάλαιο: Web Prefetching

Στα προηγούμενα κεφάλαια αναπτύχθηκαν θέματα που αφορούσαν κυρίως στην εξόρυξη δεδομένων, χωρίς όμως να γίνεται λόγος για την πρακτική εφαρμογή της ανάλυσης αυτής. Η περίπτωση με την οποία θα ασχοληθούμε αναλυτικότερα, αφορά στον Παγκόσμιο Ιστό και την αναζήτηση προβλέψεων σχετικά με την πλοήγηση των επισκεπτών σε δικτυακούς κόμβους. Η διαδικασία στην οποία θα αναφερθούμε είναι η προ-φόρτωση (*Prefetching*) σελίδων στον υπολογιστή του χρήστη, χωρίς να μας απασχολήσει η περαιτέρω τεχνική υλοποίηση της εν λόγω μεθόδου.

Τα τελευταία χρόνια ο Παγκόσμιος Ιστός έχει καταστεί το πρωταρχικό μέσο για τη διακίνηση πληροφοριών. Η ευρεία διάδοσή του και η υιοθέτησή του ως πρωτεύοντος μέσου επικοινωνίας, έχει επιφέρει μεγάλο φόρτο στους εξυπηρετητές δικτύου του παγκόσμιου ιστού. Μία σημαντική παρατήρηση, που η αξιοποίησή της μπορεί να οδηγήσει στην καλύτερη χρήση και εξοικονόμηση των περιορισμένων διαθέσιμων πόρων είναι η ύπαρξη ισχυρής τοπικότητας ανάμεσα στα δεδομένα που ζητούν οι χρήστες. Έχει παρατηρηθεί ότι υπάρχει μεγάλη συσχέτιση, τόσο χωρικά όσο και χρονικά στα δεδομένα που ζητούνται στο διαδίκτυο. Μία εφαρμογή που βασίζεται στις προηγούμενες διαπιστώσεις είναι η διαδικασία του Prefetching που θα περιγράψουμε στην ενότητα που ακολουθεί.

Το Web prefetching είναι άρρηκτα συνδεδεμένο με το μηχανισμό πρόβλεψης των επόμενων επισκέψεων ενός χρήστη κατά την πλοήγησή του σε ένα δικτυακό κόμβο. Συνοπτικά μπορούμε να πούμε ότι οι προβλέψεις που εξάγονται θα πρέπει να βασίζονται στις προηγούμενες επισκέψεις του συγκεκριμένου χρήστη, σε συνδυασμό με τη μελέτη του αρχείου πρόσβασης στον κόμβο. Σε αυτό το κεφάλαιο, προσδιορίζονται οι παράγοντες που επηρεάζουν την απόδοση των αλγορίθμων prefetching.

Στην πρώτη υπό-ενότητα θα αναφερθούμε πιο αναλυτικά στα διάφορα εισαγωγικά ζητήματα στην έννοια του Prefetching. Στη συνέχεια, θα προχωρήσουμε σε μια πιο επισταμένη αναφορά στον μηχανισμό με τον οποίο υλοποιείται η διαδικασία του Prefetching. Τέλος, θα κλείσουμε με ζητήματα που τίθενται ειδικότερα στο χώρο του Διαδικτύου, συνδέοντας το Prefetching με την εξόρυξη δεδομένων από αρχεία καταγραφής επισκέψεων από web servers και προσεγγίζοντας το ευρύτατο θέμα του προσδιορισμού του συνόλου των αιτήσεων ενός χρήστη που αποτελούν μία κοινή σύνοδο. Το τελευταίο θέμα, είναι παρόμοιο

με τον καθορισμό των συναλλαγών των χρηστών κατά την ανάλυση του καλαθιού αγοράς, με μόνη διαφορά τα ιδιαίτερα ζητήματα που προκύπτουν για το διαδίκτυο.

3.1 Η έννοια του Prefetching

Με τον όρο “Prefetching” - *Προ-φόρτωση σελίδων* αναφερόμαστε στη διαδικασία προσδιορισμού των μελλοντικών αιτήσεων του χρήστη για τα διάφορα αντικείμενα του Παγκόσμιου Ιστού και στην τοποθέτηση αυτών των αντικείμενων στην cache, στο υπόβαθρο της λειτουργίας σύνδεσης του χρήστη, πριν από την ρητή αίτηση του χρήστη προς τον κόμβο. Το Prefetching βασίζεται στη τοπικότητα των αιτήσεων του χρήστη, δηλαδή στο ότι υπάρχουν συσχετιζόμενες μεταξύ τους αναφορές σε έγγραφα, που παρουσιάζονται σε διαφορετικά σύνολα αιτήσεων και εκμεταλλεύεται το (νεκρό) χρόνο μη απασχόλησης του χρήστη, δηλαδή το χρόνο μεταξύ των διαδοχικών αιτημάτων. Το βασικότερο πλεονέκτημα είναι ότι αποτρέπεται η υπό-εκμετάλλευση του εύρους ζώνης (Bandwidth underutilization) καλύπτοντας το τμήμα της λανθάνουσας αυτής κατάστασης. Αντίθετα, μια υπερβολικά εντατική διαδικασία Prefetching, μπορεί να προκαλέσει αντίστοιχα την υπερφόρτωση του δικτύου. Χωρίς την προσεκτικά σχεδιασμένη διαδικασία prefetching, πολλά από τα έγγραφα που έχουν μεταφερθεί δεν θα είναι δυνατό να χρησιμοποιηθούν από το χρήστη, σπαταλώντας κατά συνέπεια και πάλι πολύτιμο εύρος ζώνης.

Συμπερασματικά, μια αποτελεσματική διαδικασία Prefetching, σε συνδυασμό με έναν μηχανισμό ελέγχου του ποσοστού των εγγράφων που μεταφέρονται, μπορεί να δράσει καθοριστικά για τη λειτουργία του δικτύου και να κάνει περισσότερο ομαλή την «κυκλοφορία» μειώνοντας σημαντικά τις καθυστερήσεις και βελτιώνοντας έτσι γενικότερα την απόδοσή του.

Διακρίνουμε δύο κύριες προσεγγίσεις στη διαδικασία του prefetching. Είτε ο χρήστης θα ενημερώσει το σύστημα για τις μελλοντικές απαιτήσεις του ή με έναν αυτοματοποιημένο τρόπο και χωρίς την επίγνωση του χρήστη, το σύστημα θα προσδιορίσει τις προβλέψεις βασισμένες στην ακολουθία των προηγούμενων αναφορών του χρήστη. Η δεύτερη περίπτωση της μεθόδου των προβλέψεων, είναι περισσότερο βιώσιμη, ειδικά κάτω από την προϋπόθεση ότι υπάρχει αρκετή τοπικότητα στις αιτήσεις των χρηστών, λόγω του γεγονότος ότι μια τέτοια prefetching μέθοδος χρησιμοποιεί το ιστορικό των αιτημάτων προκειμένου να γίνουν οι προβλέψεις. Όπως καταλαβαίνει κανείς, σε έναν καλοσχεδιασμένο κόμβο

με πολλούς συνδέσμους σε κάθε σελίδα του, η εν λόγω μέθοδος αναμένεται να παρουσιάσει καλύτερα αποτελέσματα.

Επίσης, οι υπάρχοντες προβλεπτικοί αλγόριθμοι Prefetching που εξετάζονται στα επιστημονικά πεδία των Βάσεων Δεδομένων, Συστημάτων Αρχείων και πρόσφατα στον Παγκόσμιο Ιστό, μπορούν να ταξινομηθούν σε δύο οικογένειες:

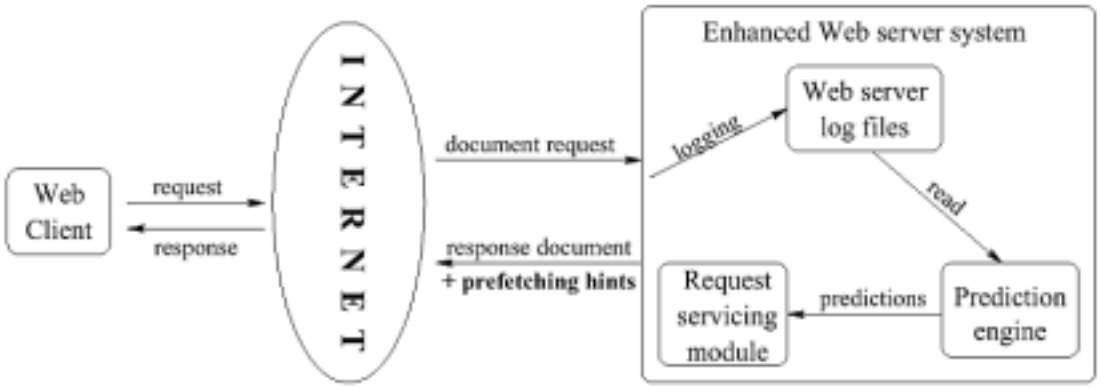
(α) σε εκείνους που χρησιμοποιούν ένα γράφημα, που ονομάζεται Γράφημα Εξαρτήσεων (Dependency Graph - DG) για να ορίζουν τα πρότυπα των επισκέψεων. Η συγκεκριμένη μέθοδος υπολογίζει μόνο τις εξαρτήσεις ανάμεσα σε ζεύγη αιτήσεων και οι αιτήσεις αυτές δεν απαιτείται να είναι διαδοχικές.

(β) σε εκείνους που χρησιμοποιούν μία μέθοδο που υιοθετείται από τον χώρο της Συμπίεσης Κειμένων, που ονομάζεται Πρόβλεψη με μερική αντιστοιχία (Prediction by Partial Match - PPM). Στην περίπτωση αυτή οι αναφορές θα πρέπει να είναι διαδοχικές.

3.2 Ο μηχανισμός του Prefetching

Ο προσδιορισμός των μελλοντικών αιτήσεων με βάση την διαδικασία προβλεπτικού Prefetching μπορεί να εφαρμοστεί με την υλοποίηση ενός μηχανισμού ο οποίος αφού επεξεργαστεί τις προηγούμενες αναφορές παράγει την πιθανότητα μελλοντικής πρόσβασης. Ο μηχανισμός πρόβλεψης μπορεί να βρίσκεται είτε στην πλευρά του χρήστη είτε στην πλευρά του εξυπηρετητή(Server/Client Architecture). Στην δεύτερη περίπτωση, χρησιμοποιεί το σύνολο των προηγούμενων αναφορών για να βρει τους συσχετισμούς και ξεκινάει τη διαδικασία του Prefetching. Δεν είναι απαραίτητο να γίνει οποιαδήποτε τροποποίηση στην υπάρχουσα υποδομή του κόμβου (π.χ. HTTP πρωτόκολλο, εξυπηρετητές κλπ) ούτε στους φυλλομετρητές (browser) δικτύου, στην περίπτωση που το εργαλείο “Prefetcher” τρέχει ως proxy στον φυλλομετρητή.

Ο βασικός περιορισμός αυτής της προσέγγισης είναι ότι οι χρήστες τις περισσότερες φορές δεν έχουν πρόσβαση στην πληροφορία σχετικά με τη διασύνδεση των εγγράφων καθώς πλοηγούνται σε ένα μεγάλο σύνολο από έγγραφα σε διαφορετικούς εξυπηρετητές. Αφετέρου, οι εξυπηρετητές δικτύου βρίσκονται σε καλύτερη θέση αναφορικά με τις προβλέψεις για τις μελλοντικές αναφορές, δεδομένου ότι καταγράφουν ένα σημαντικό μέρος των αιτημάτων από όλους τους χρήστες Διαδικτύου για τα στοιχεία τα οποία είναι ελέγχουν.



Σχήμα 9 Ο μηχανισμός του Prefetching

3.3 Κανονικότητα στις επισκέψεις των χρηστών

Ο τρόπος με τον οποίο πλοηγούνται οι χρήστες σε έναν δικτυακό κόμβο δεν εξαρτάται αποκλειστικά από τα ενδιαφέροντά τους, αλλά και από τη δομή που έχουν οι δικτυακοί αυτοί τόποι. Πιο συγκεκριμένα, αν ένας χρήστης σε μία χρονική στιγμή έχει επιλέξει για παράδειγμα ένα έγγραφο Δ, το επόμενο έγγραφο που θα επιλέξει να επισκεφθεί θα είναι μέσα από το σύνολο των συνδέσμων που περιλαμβάνονται στη σελίδα Δ. Αυτή η επιλογή, σε γενικές γραμμές, βασίζεται στα έγγραφα που ο χρήστης έχει ήδη επισκεφθεί (δηλ., στο Δ και σε οποιαδήποτε άλλα έγγραφα πριν από Δ). Διαφορετικά, ο χρήστης πλοηγείται τυχαία στην περιοχή, χωρίς να αναζητά κάποιες ιδιαίτερες πληροφορίες. Η προηγούμενη περίπτωση αθροίζεται στις εξαρτήσεις που προκύπτουν για τα έγγραφα του κόμβου που επισκέπτονται οι χρήστες. Αυτό το μοντέλο της πλοήγησης των χρηστών περιγράφεται με μια ακολουθία Markov σε ένα γράφημα του οποίου οι κόμβοι είναι τα έγγραφα της περιοχής και τα τόξα είναι οι συνδέσεις μεταξύ των εγγράφων.

Οι χρήστες που δεν εξερευνούν τυχαία μια περιοχή Ιστού επισκέπτονται συνήθως τις σελίδες σύμφωνα με ένα πρότυπο. Επομένως, μια ακολουθία πρόσβασης χρηστών περιέχει τις σελίδες που ανήκουν σε ένα από διάφορα πρότυπα. Όμως, ένας χρήστης μπορεί επίσης να επισκεφτεί και άλλες σελίδες οι οποίες δεν ανήκουν σε κάποιο πρότυπο. Συνεπώς, μια ακολουθία πρόσβασης χρηστών μπορεί να περιέχει σελίδες που ανήκουν σε ένα πρότυπο και μεταξύ αυτών διάφορες άλλες σελίδες που να μην ανήκουν σε αυτό.

Το μήκος των ακολουθιών πρόσβασης χρηστών, οι αλληλεξαρτήσεις μεταξύ των προσβάσεων και η ύπαρξη σελίδων (μέσα στις ακολουθίες) που δεν

ανήκουν στα πρότυπα, είναι παράμετροι που εξαρτώνται από το είδος του δικτυακού κόμβου. Στην περίπτωση των μικρών websites, ο αντίκτυπος αυτών των παραμέτρων μπορεί να είναι μικρός λόγω της περιορισμένης πλοήγησης σε εναλλακτικές λύσεις. Αντίθετα, σε μεγαλύτερες δικτυακές περιοχές, με μεγάλο αριθμό εγγράφων και αρκετά υψηλή συνδετικότητα (που μας θυμίζει τις παραδοσιακές βάσεις δεδομένων υπερκειμένων), όπου παρουσιάζονται εναλλακτικές λύσεις για την πλοήγηση των χρηστών, ο αντίκτυπος αυτών των παραμέτρων είναι σημαντικός. Αυτό το είδος δικτυακών χώρων αναμένεται να γίνει περισσότερο δημοφιλές στο προσεχές μέλλον, όταν η διαδικασία δημιουργίας και συντήρησης των κόμβων του διαδικτύου θα είναι περισσότερο αυτοματοποιημένη, χρησιμοποιώντας τα εργαλεία διαχείρισης περιεχομένου για το διαδίκτυο σε συνεργασία με συστήματα βάσεων δεδομένων.

3.4 Η ιδιαιτερότητα του Prefetching στον Παγκόσμιο Ιστό

Οι υπάρχοντες αλγόριθμοι prefetching για τον Παγκόσμιο Ιστό δεν αναγνωρίζουν τα εξειδικευμένα χαρακτηριστικά του διαδικτύου. Πιο συγκεκριμένα, δύο σημαντικοί παράγοντες που πρέπει να λαμβάνονται υπόψη είναι:

- Η σειρά των εξαρτήσεων μεταξύ των εγγράφων στα πρότυπα που προκύπτουν.
- Η παρεμβολή των εγγράφων στα πρότυπα, και η επισήμανση ότι αυτές μπορεί να αποτελούν τυχαίες επισκέψεις μέσα στις συνόδους των χρηστών. Κάτι τέτοιο αποτελεί ουσιαστικά θόρυβο και θα πρέπει να απομακρυνθεί από τα δεδομένα και τα αντικείμενα από τα οποία αποτελούνται τα πρότυπα.

Αυτοί οι παράγοντες προκύπτουν τόσο από το περιεχόμενο των εγγράφων όσο και από τη δομή του δικτυακού κόμβου που μας ενδιαφέρει (τους συνδέσμους της κάθε σελίδας κλπ).

Η επιλογή των προσεχών σελίδων από τον χρήστη εξαρτάται στην πλειονότητα των περιπτώσεων από τις διάφορες σελίδες που επισκέφθηκε προηγουμένως.

Ένας χρήστης του Παγκόσμιου Ιστού μπορεί να ακολουθήσει κατά τη διάρκεια μιας συνόδου, συνδέσμους με σελίδες που ανήκουν σε ένα από τα διάφορα πρότυπα που έχουν προκύψει. Εντούτοις, κατά τη διάρκεια της ίδιας συνόδου μπορεί ο χρήστης να πλοηγηθεί σε άλλες σελίδες που δεν ανήκουν σε

αυτό το πρότυπο (ή που μπορεί να μην ανήκει σε κανένα από όλα τα ορισμένα πρότυπα). Κατά συνέπεια, μια σύνοδος χρήστη μπορεί να περιέχει τόσο έγγραφα που ανήκουν στα προσδιορισμένα πρότυπα όσο και άλλα που δεν ανήκουν σε αυτά, παρά μόνο παρεμβάλλονται σε αυτά.

Με βάση τα όσα προαναφέρθηκαν, γίνεται φανερό ότι οι τεχνικές ανακάλυψης κανόνων συσχέτισης στα αρχεία καταγραφής των επισκέψεων μπορούν να αποτελέσουν τη λύση για το συγκεκριμένο πρόβλημα του Prefetching. Οι συγκεκριμένες τεχνικές μπορούν να βοηθήσουν:

- Στην αξιοποίηση του αρχείου των προσβάσεων των χρηστών, για τον υπολογισμό των πιθανοτήτων ζήτησης ενός αντικειμένου (ή ενός συνόλου αντικειμένων), έχοντας ως δεδομένο ότι έχει ζητηθεί κάποιο άλλο αντικείμενο (ή άλλο σύνολο αντικειμένων).
- Στον προσδιορισμό των κανόνων της μορφής $d_1,d_2, \ldots, d_k \Rightarrow dm, d_{m+1}, \ldots, d_n$ που περιγράφουν τις στατιστικά ισχυρές αλληλεξαρτήσεις μεταξύ των αντικειμένων του εξυπηρετητή.
- Στην ενεργοποίηση των κανόνων για τα οποία το head τους περιέχει τις προσβάσεις που έχει κάνει ένας χρήστης μέχρι ενός χρονικού σημείου και την φόρτωση των σελίδων που συμφωνούν με το body των κανόνων.

Είναι προφανές ότι μία διαδικασία prefetching θα πρέπει να μπορεί να αντιμετωπίσει συσχετίσεις υψηλής κατάταξης. Επομένως, θα πρέπει να επιτρέπονται κανόνες με περισσότερα από ένα έγγραφα στο head του κανόνα και ενδεχομένως για τους κανόνες με περισσότερα από ένα έγγραφα στο body. Ως εκ τούτου, η διαδικασία prefetching πρέπει να είναι σε θέση να επιλέξει προσαρμοστικά την κατάλληλη μέγιστη ακολουθία.

3.5 Προσδιορισμός των συνόδων (Transactions) των επισκεπτών

Η έρευνα για το Prefetching με χρήση πρόβλεψης στο διαδίκτυο, έχει συμπεριλάβει το σημαντικότατο ζήτημα της επεξεργασίας των αρχείων καταγραφής των επισκέψεων στον εξυπηρετητή (log files) και τον καθορισμό των συναλλαγών (transactions) των χρηστών. Δεδομένου ότι είναι ένα απαραίτητο βήμα για κάθε prefetching μέθοδο στο διαδίκτυο, έχουν προταθεί παρόμοιες προσεγγίσεις στο σχηματισμό του συνόλου των συναλλαγών από τα log files.

Οι επισκέπτες του διαδικτυακού κόμβου, είναι συνδεδεμένοι σε αυτόν κατά ένα χρονικό διάστημα που δεν είναι σαφές χρονικά. Ο επακριβής ορισμός των

επισκέψεων του κάθε χρήστη ή αλλιώς ο προσδιορισμός των συναλλαγών των χρηστών απαιτεί αρχικά τον προσδιορισμό των περιόδων επικοινωνίας του χρήστη από τα log files. Οι προσβάσεις κάθε χρήστη ομαδοποιούνται στις περιόδους επικοινωνίας σύμφωνα με τη χρονική απόστασή τους. Στη συνέχεια, μπορούν να υποβληθούν σε περαιτέρω επεξεργασία με διάφορες μεθόδους. Έτσι, οι περίοδοι επικοινωνίας των χρηστών υποδιαιρούνται ανάλογα με τις μέγιστες δυνατές αναφορές τους. Αυτό το χαρακτηριστικό φιλτράρει το στοιχείο των αναφορών προς τα πίσω (που γίνονται συνήθως με την χρήση του πλήκτρου “Back” σε έναν φυλλομετρητή δικτύου), οι οποίες εξυπηρετούν μόνο σκοπούς πλοήγησης στον κόμβο. Οι καθαρισμένες περίοδοι επικοινωνίας χρηστών αποτελούν τις συναλλαγές των χρηστών. Ο καθορισμός των περιόδων επικοινωνίας των χρηστών στοχεύει στις ανεξάρτητες προσβάσεις που γίνονται από τους διαφορετικούς χρήστες ή από τον ίδιο χρήστη σε απομακρυσμένες μεταξύ τους χρονικές στιγμές.

3.6 Εξόρυξη σε Web δεδομένα

Είδαμε στα παραπάνω πώς η εξόρυξη δεδομένων στα αρχεία επισκέψεων μπορεί να χρησιμοποιηθεί για να προβλέψουμε τις αιτήσεις των χρηστών προς τον εξυπηρετητή δικτύου, σε συνδυασμό με την διαδικασία Prefetching σελίδων. Βέβαια οι τεχνικές εξόρυξης δεδομένων μπορούν να βοηθήσουν σε μια πληθώρα άλλων επεξεργασιών, όπως:

- στην εξαγωγή στατιστικής πληροφορίας
- στην ανακάλυψη προτύπων (patterns) στα δεδομένα σχετικά με την συμπεριφορά του χρήστη όσον αφορά την πλοήγηση
- ενδεχόμενες θεματικές ενότητες των ενδιαφερόντων χρηστών
- ομάδες χρηστών με κοινά ενδιαφέροντα
- συσχετισμούς ανάμεσα σε σελίδες του κόμβου.

Είναι φανερό ότι υπάρχει μεγάλη ποικιλία τεχνικών που μπορούν να εφαρμοστούν με διαφορετικά αποτελέσματα και στόχους η κάθε μία. Αξίζει να σημειωθεί ότι οι τεχνικές αυτές πέρα από τα log files στα οποία κυρίως εστιάζουμε την προσοχή μας στην παρούσα εργασία, είναι δυνατό να εφαρμοσθούν ανάλογα με τα αποτελέσματα που μας ενδιαφέρει να μελετήσουμε, και σε άλλες πολύτιμες βάσεις δεδομένων του διαδικτύου όπως είναι:

- τα αρχεία καταγραφής προγραμμάτων-πρακτόρων,

- τα μετά-δεδομένα
- το περιεχόμενο των σελίδων
- ο τρόπος δόμησης του κόμβου
- τα αποθηκευμένα στον Η/Υ του χρήστη cookies κ.ά.

Στην παρούσα εργασία, επικεντρώνουμε στην σύνδεση της ανακάλυψης κανόνων συσχέτισης με το Prefetching. Βέβαια, δεν τέθηκαν κάποια σχετικά ζητήματα που αφορούν στο πότε θα πρέπει να εφαρμοσθεί η εν λόγω διαδικασία και το πού θα πρέπει να αποθηκευτούν τα έγγραφα αυτά. Αν το Prefetching εφαρμοσθεί αρκετά νωρίς, τότε είναι πολύ πιθανό να αντικαταστήσει από την κρυφή μνήμη έγγραφα που χρειάζονται, ενώ αν εφαρμοσθεί πολύ αργά μέσα στην εξέλιξη της σύνδεσης του χρήστη στον κόμβο, κατά πάσα πιθανότητα δεν θα ωφελήσει στην βελτίωση της πλοήγησης. Απαιτείται λοιπόν κατάλληλη επιλογή του χρόνου για τη διαδικασία του Prefetching. Επίσης, ο χώρος αποθήκευσης είναι πολύ σημαντικός και έχουν προταθεί για το σκοπό αυτό, πολλές εναλλακτικές αποθήκευσης στη διάρκεια της ροής δεδομένων.

Στην παρούσα εργασία δεν θα προχωρήσουμε σε μεγαλύτερη εμβάθυνση όσον αφορά στην πρακτική εφαρμογή της διαδικασίας Prefetching γιατί κάτι τέτοιο απαιτεί τον έλεγχο μιας πλειάδας παραγόντων που αλληλεπιδρούν μεταξύ τους. Η περαιτέρω ανάλυση εμπλέκει μια σειρά από τεχνικά θέματα όπως η επιλογή της πολιτικής αντικατάστασης των εγγράφων στην κρυφή μνήμα τόσο του χρήστη όσο και του εξυπηρετητή, ζητήματα που αφορούν τους Proxy εξυπηρετητές κ.ά. Όλα αυτά τα ζητήματα είναι ορθογώνια με την παρούσα μελέτη.

4 Κεφάλαιο: Ανάλυση των δεδομένων εισόδου

Στο σημείο αυτό θα πρέπει να αναφερθούμε και να περιγράψουμε τα πραγματικά δεδομένα που χρησιμοποιήθηκαν για την έρευνα σχετικά με τη διαδικασία Prefetching για τη βελτίωση της απόδοσης στον παγκόσμιο. Ως δεδομένα εισόδου για την ανάλυσή μας επιλέχθηκαν τα αρχεία καταγραφής των επισκέψεων στον διαδικτυακό κόμβο του Μακεδονικού Πρακτορείου Ειδήσεων για το χρονικό διάστημα πέντε εβδομάδων (Αύγουστος/ Σεπτέμβριος 2002). Να σημειωθεί ότι ο ηλεκτρονικός κόμβος του Μ.Π.Ε. αποτελεί ένα web site με πολλές ημερήσιες επισκέψεις (έχει κατά μέσο όρο 25.000 με 30.000 μοναδικούς επισκέπτες καθημερινά), γεγονός που καθιστά το χρονικό διάστημα που επιλέχθηκε αρκετά μεγάλο για να μπορέσουμε να καταλήξουμε σε ικανοποιητικά συμπεράσματα σχετικά με την εξόρυξη προτύπων και κανόνων συσχέτισης στα δεδομένα που θα χρησιμοποιηθούν για τη διαδικασία Prefetching.

4.1 **Στατιστική Ανάλυση των log files του κόμβου** http://www.mpa.gr

Τα αρχεία καταγραφής των επισκέψεων που μας παραχωρήθηκαν από το Μ.Π.Ε. αφορούν το διάστημα πέντε εβδομάδων, από την 01/08/2002 ως τις 05/09/2002 και είναι χωρισμένα σε πέντε αντίστοιχα αρχεία, ένα για την κάθε εβδομάδα. Για λόγους πληρότητας να αναφέρουμε εδώ ότι το συνολικό μέγεθος των log files (που είναι όπως έχει αναφερθεί σε άλλο σημείο απλά αρχεία κειμένου) που επεξεργαστήκαμε φτάνει το 1,2 Gigabyte.

Για τη στατιστική ανάλυση των δεδομένων των log files μετά από έρευνα στον χώρο των εργαλείων που χρησιμοποιούνται ευρέως, προτιμήθηκε το πρόγραμμα “Analog” (http://www.analog.cx) που ανήκει στην κατηγορία του Ελεύθερου Λογισμικού (freeware) / Ανοικτού Κώδικα (Open Source). Για τη δημιουργία των γραφικών παραστάσεων που παραθέτουμε χρησιμοποιήθηκε το λογισμικό “Report Magic” (http://www.reportmagic.com) που ανήκει στην ίδια κατηγορία λογισμικού.

Η παρουσίαση των διαφόρων στατιστικών στοιχείων που ακολουθεί, έχει την ίδια λογική διαχωρισμού σε πέντε διαφορετικά αρχεία ανάλογα με την εβδομάδα που μελετάται. Οι αναφορές που παρουσιάζονται για τις διαδοχικές εβδομάδες είναι οι εξής:

- Αναφορά Ημερήσιων Επισκέψεων (Daily Report)
- Αναφορά Ωριαίων επισκέψεων (Hourly Report)
- Αναφορά Λέξεων Αναζήτησης (Word)

- Αναφορά Καταλόγων (Directory Report)
- Αναφορά Σελίδων (Search word Report)
- Αναφορά Μεγέθους αρχείων (File size Report)
- Αναφορά Σελίδων (Requested pages Report)

Στη συνέχεια παρουσιάζονται οι αναφορές αυτές για το κάθε αρχείο, και έπειτα συνοψίζονται τα συμπεράσματα της ανάλυσης για το σύνολο της υπό εξέταση περιόδου.

4.1.1 Αρχείο Καταγραφής 1: Εβδομάδα: 1η – 8η Αυγούστου 2003

Στον παρακάτω πίνακα παρουσιάζονται τα βασικότερα μεγέθη που αφορούν την ανάλυση του 1ου αρχείου επισκέψεων (log file mpa1, χρονικό διάστημα 1-8/8/2003).

Γενική Αναφορά	
Χρόνος Πρώτης Επίσκεψης	**Aug 1, 2002 03:00**
Χρόνος Τελευταίας Επίσκεψης	**Aug 8, 2002 02:59**
Επιτυχημένες αιτήσεις	**513555 Requests**
Επιτυχημένες αιτήσεις σελίδων	**173873 Requests for pages**
Αποτυχημένες αιτήσεις	**2628 Requests**
Αναδρομολογημένες αιτήσεις	**1460 Requests**
Μοναδικοί επισκέπτες	**42493 Hosts**
Κατεστραμμένες γραμμές αρχείου log file	**208 Lines**
Μη χρήσιμες καταχωρήσεις αρχείου log file	**761595 Lines**
Σύνολο δεδομένων που μεταφέρθηκαν	**3.940 GB**

Βλέπουμε ότι το σύνολο των σελίδων που επισκέφθηκαν οι χρήστες είναι **173873**, αριθμός ιδιαίτερα μεγάλος για την χρονική περίοδο της μίας εβδομάδας, γεγονός που υποδηλώνει αφενός την ύπαρξη μεγάλου πλήθους σελίδων που στην προκειμένη περίπτωση είναι οι δημοσιευμένες στον κόμβο ειδήσεις, και αφετέρου την ύπαρξη μεγάλου αριθμού χρηστών. Και οι δύο αυτές παρατηρήσεις καθιστούν το Prefetching σελίδων με πρόβλεψη, μέσα από την μελέτη των προτύπων

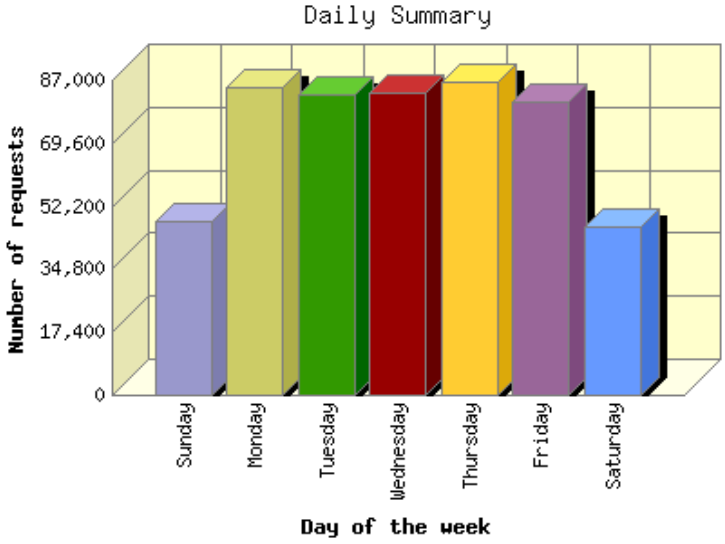
πρόσβασης στον κόμβο, μία διαδικασία που μπορεί να αποβεί πολύ σημαντική για την ομαλή και αποδοτική προσφορά υπηρεσιών στους τελικούς χρήστες, δηλαδή τους συνδρομητές και απλούς επισκέπτες του κόμβου.

Οι συνολικές επιτυχημένες αιτήσεις στον κόμβο είναι **513555**, ενώ οι αντίστοιχες αποτυχημένες αιτήσεις είναι συνολικά **2628**.

Ένα ακόμα στατιστικό στοιχείο άξιο σχολιασμού είναι οι γραμμές από το αρχικό αρχείο log file που επεξεργαστήκαμε και οι οποίες δεν συμπεριλαμβάνονται στην περαιτέρω μελέτη. Αυτές οι γραμμές αφορούν αιτήσεις για διάφορα αρχεία εικόνων που χρησιμοποιούνται για τη μορφοποίηση των σελίδων του κόμβου, «κρυφά» αρχεία στα οποία δεν έχει πρόσβαση ο επισκέπτης όπως ο φάκελος εισαγωγής των άρθρων που δημοσιεύονται ή ο φάκελος διαχείρισης του κόμβου και κάποιοι δευτερεύοντες φάκελοι με αρχεία τα οποία περιλαμβάνονται στις σελίδες του κόμβου, αλλά στις οποίες οι επισκέπτες δεν έχουν ποτέ άμεση πρόσβαση μέσω κάποιου συνδέσμου. Τέλος, ο συνολικός όγκος των δεδομένων που μετακινήθηκαν ανέρχεται στα **3.940 GB**.

4.1.1.1 Αναφορά ημερήσιων επισκέψεων

Στο παρακάτω διάγραμμα παρουσιάζουμε σχηματικά την ημερήσια κατανομή των επισκέψεων στον κόμβο. Το συμπέρασμα στο οποίο καταλήγει αβίαστα κανείς, είναι ότι η κύρια δραστηριότητα στον κόμβο κατανέμεται σχεδόν ομοιόμορφα στις εργάσιμες μέρες. Βλέπουμε ότι κατά την εβδομάδα 1-8/8/2003 οι επισκέψεις στον κόμβο που έγιναν κατά τη διάρκεια του σαββατοκύριακου ήταν σχεδόν υπό-διπλάσιες απ' ότι τις καθημερινές.

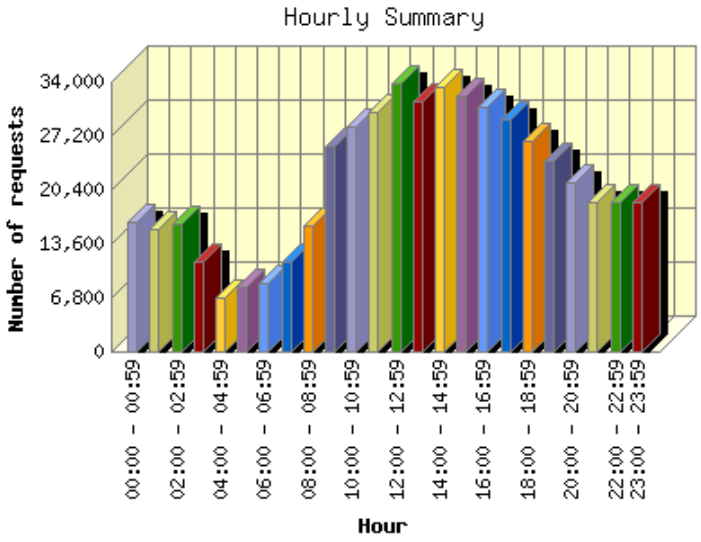


Στον πίνακα που ακολουθεί μπορούμε να δούμε και συγκεκριμένα στοιχεία σχετικά με τις συνολικές ημερήσιες επισκέψεις στον κόμβο.

Ημέρα		Αρ. Αιτήσεων	Αρ. Αιτήσεων Σελίδων
1.	Κυριακή	47,880	17,813
2.	Δευτέρα	85,041	28,043
3.	Τρίτη	82,900	26,532
4.	Τετάρτη	83,554	28,216
5.	Πέμπτη	86,256	28,065
6.	Παρασκευή	81,205	27,266
7.	Σάββατο	46,719	17,938

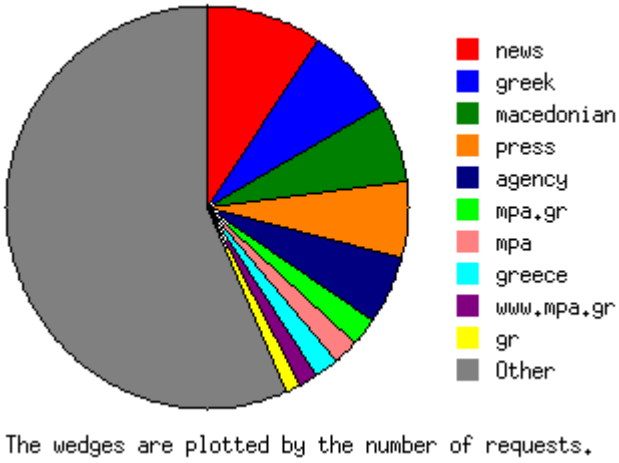
4.1.1.2 Αναφορά ωριαίων επισκέψεων

Η αναφορά των ωριαίων επισκέψεων στον κόμβο, που παρουσιάζεται παρακάτω, είναι πολύ σημαντική όσον αφορά την κατανομή πόρων για τη διαδικασία Prefetching που θα εφαρμοστεί. Παρατηρούμε ότι οι ώρες με πολύ μεγάλη δραστηριότητα και πολλές επισκέψεις στον κόμβο είναι οι εργάσιμες ώρες της μέρας με αιχμή το διάστημα μεταξύ των ωρών 13:00 με 16:00.



4.1.1.3 Αναφορά λέξεων αναζήτησης

Στη συγκεκριμένη αναφορά παρουσιάζεται μία λίστα με τις λέξεις κλειδιά που χρησιμοποίησαν οι τελικοί επισκέπτες, προκειμένου να βρουν την ηλεκτρονική διεύθυνση του δικτυακού χώρου.

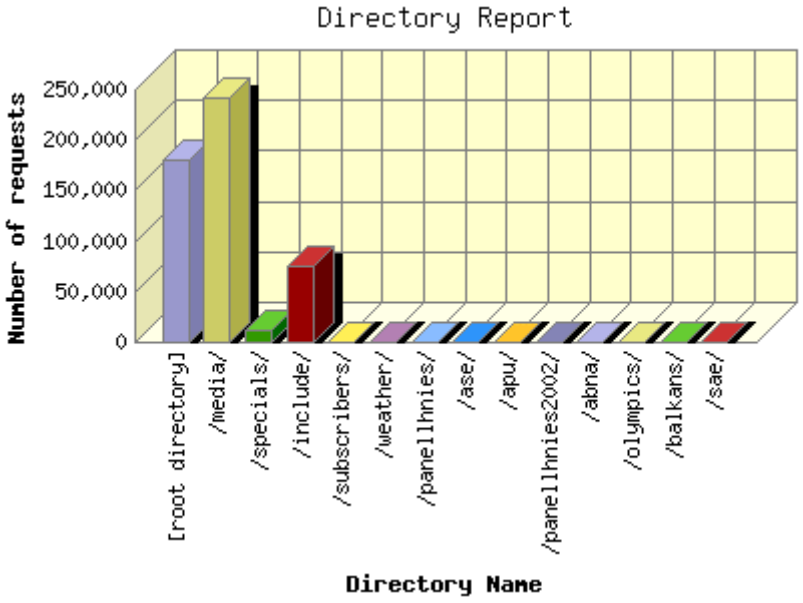


Search Word		Αρ. Αιτήσεων
1.	News	76
2.	Greek	59
3.	macedonian	51
4.	Press	49
5.	agency	43
6.	mpa.gr	19
7.	Mpa	16
8.	greece	16
9.	www.mpa.gr	12
10.	Gr	10

4.1.1.4 Αναφορά καταλόγων

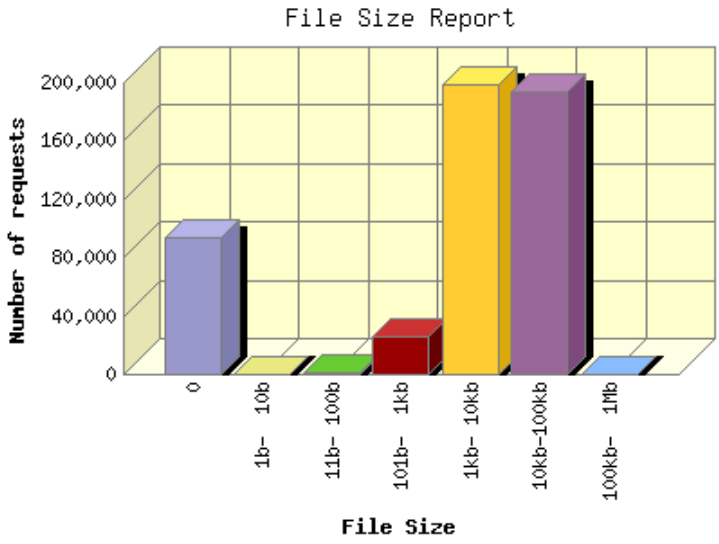
Στο διάγραμμα που ακολουθεί βλέπουμε την κατανομή των αιτήσεων των επισκεπτών του Δικτυακού Κόμβου του Μακεδονικού Πρακτορείου Ειδήσεων, σε κάθε κατάλογο του κόμβου. Όπως παρατηρεί κανείς, τις περισσότερες αιτήσεις συγκεντρώνει για την εβδομάδα 01/08/2003 με 08/08/2003 ο κατάλογος /media, ενώ στη δεύτερη θέση με μεγάλη διαφορά από τους υπολοίπους καταλόγους είναι ο κεντρικός κατάλογος του site. Παρατηρώντας τη δομή του κόμβου, αποδεικνύεται ότι η μεγάλη πλειοψηφία των προσπελάσιμων από το χρήστη αρχείων βρίσκεται αποθηκευμένη στον κεντρικό κατάλογο (root directory). Τα

αρχεία αυτά είναι κατά κύριο λόγο τα αρχεία με τα άρθρα που δημοσιεύονται από τον εν λόγω ειδησεογραφικό οργανισμό. Αντίθετα στον φάκελο /media βρίσκεται αποθηκευμένο κυρίως το φωτογραφικό υλικό του κόμβου.



4.1.1.5 Αναφορά μεγέθους αρχείων

Στην επόμενη αναφορά βλέπουμε την κατανομή των μεγεθών των αρχείων που ζητούν οι χρήστες από τον διαδικτυακό κόμβο. Το μέγεθος των αρχείων είναι πολύ σημαντικό για την ανάλυσή μας, καθώς επηρεάζει τη διαδικασία της πρόβλεψης και του Prefetching των σελίδων. Παρατηρούμε ότι τα αρχεία έχουν στην συντριπτική τους πλειοψηφία κατά μέσο όρο πολύ μικρό σχετικά μέγεθος (1-100Kbytes). Το στοιχείο αυτό μπορεί στην περίπτωση του Μακεδονικού Πρακτορείου Ειδήσεων να αποδειχθεί πολύ σημαντικό για να καταστεί αποτελεσματική η διαδικασία του Prefetching και να βελτιωθεί η απόδοση του κόμβου.



4.1.1.6 Αναφορά σελίδων

Η παρακάτω αναφορά παρουσιάζει τις σελίδες που είχαν τις περισσότερες αιτήσεις κατά το χρονικό διάστημα το οποίο εξετάζουμε (01/08/2003 με 08/08/2003). Παρατηρούμε ποια από τα αρχεία αυτά συμπεριλαμβάνονται στους κανόνες στους οποίους καταλήγουμε με τη βοήθεια του αλγορίθμου Apriori.

reqs: %bytes: last time: file
-----: ------: ----------: ----
97127: 45.34%: 8/Aug/02 02:59: /article.html
657: 0.28%: 8/Aug/02 02:24: /article.html?doc_id=285727
626: 0.22%: 8/Aug/02 02:52: /article.html?doc_id=285278
622: 0.24%: 8/Aug/02 02:24: /article.html?doc_id=285535
554: 0.27%: 8/Aug/02 02:52: /article.html?doc_id=285574
554: 0.24%: 8/Aug/02 02:51: /article.html?doc_id=285962
530: 0.24%: 8/Aug/02 02:45: /article.html?doc_id=286062
…
63724: 47.58%: 8/Aug/02 02:59: /
8754: 7.74%: 8/Aug/02 02:59: /?page=greece
4266: 4.44%: 8/Aug/02 02:59: /?page=balkans
4047: 2.76%: 8/Aug/02 02:59: /?page=english
3635: 2.59%: 8/Aug/02 02:59: /?page=economy
3447: 2.57%: 8/Aug/02 02:59: /?page=home
3351: 2.12%: 8/Aug/02 02:59: /?page=world
2907: 1.30%: 8/Aug/02 02:59: /?page=sports
2296: 1.19%: 8/Aug/02 02:59: /?page=culture
625: 0.53%: 8/Aug/02 02:59: /?page=russian
2795: 1.61%: 8/Aug/02 01:35: /search.html
368: 0.23%: 8/Aug/02 01:26: /search.html?lang=el
711: 1.10%: 7/Aug/02 23:59: /titles.html
473: 0.27%: 8/Aug/02 02:36: /weather/map.html

401: 0.20%: 8/Aug/02 02:51: /ase/
372: 0.10%: 7/Aug/02 14:25: /media/thumb_show.html
317: 0.10%: 8/Aug/02 02:40: /weather/
265: 0.01%: 8/Aug/02 02:37: /specials/oreinosaugust/index.htm
231: 0.37%: 8/Aug/02 00:54: /subscribers/
214: 0.12%: 8/Aug/02 02:14: /specials/patriarchate/

4.2 Παρατηρήσεις

Χωρίς να εμβαθύνουμε ιδιαίτερα στην ανάλυση της δομής και του τρόπου λειτουργίας του δικτυακού κόμβου του Μ.Π.Ε., θα πρέπει αρχικά να σχολιάσουμε κάποια βασικά χαρακτηριστικά του, που αναμένεται να επαληθευθούν ή ενδεχομένως να καταρριφθούν από την διαδικασία Ανακάλυψης Γνώσης.

Πρόκειται για έναν κόμβο πολύ καλά οργανωμένο νοηματικά. Ως ειδησεογραφικός οργανισμός, το προϊόν που προσφέρει είναι πρωτίστως η μετάδοση της ειδησεογραφικής πληροφορίας άμεσα και έγκαιρα, γεγονός που πετυχαίνει με την παρουσία του στο διαδίκτυο. Μετά από προσωπική συνέντευξη με τους επικεφαλής της ομάδας σύνταξης και τεχνικής υποστήριξης στον κόμβο, ενημερωθήκαμε για το Σύστημα Διαχείρισης Περιεχομένου που χρησιμοποιεί το Μ.Π.Ε.. Συγκεκριμένα, χρησιμοποιεί την πλατφόρμα Databound για τη δυναμική παραγωγή και διαχείριση των ιστοσελίδων.

Η κύρια ενημερωτική οντότητα στην οποία επιθυμούν να έχουν πρόσβαση οι επισκέπτες του κόμβου, είναι τα άρθρα τα οποία είναι χωρισμένα σε θεματικές ενότητες και η πρόσβαση σε αυτά γίνεται δυναμικά, καλώντας από τον κύριο κατάλογο τη σελίδα article.html δίνοντας ως παράμετρο τον κωδικό του άρθρου, για παράδειγμα /article.html?doc_id=288868 . Έτσι, όλα τα άρθρα ανεξάρτητα από το μέγεθός τους, την ημερομηνία έκδοσής τους ή την θεματική ενότητα στην οποία ανήκουν, χαρακτηρίζονται αποκλειστικά από έναν κωδικό με τον οποίο είναι αποθηκευμένα στην αντίστοιχη βάση δεδομένων. Φυσικά, οι χρήστες δεν είναι υποχρεωμένοι να γνωρίζουν όλους τους κωδικούς όλων των άρθρων και γι' αυτό το λόγο υπάρχουν οι πολυάριθμοι σύνδεσμοι προς τα άρθρα και η δυνατότητα αναζήτησης με βάση λέξεις-κλειδιά και άλλες παραμέτρους.

Ένα άλλο σημαντικό χαρακτηριστικό, είναι η ύπαρξη πολλών συνδέσμων που υποβοηθούν την πλοήγηση, σε κάθε σελίδα του κόμβου και επιπλέον η επιλεκτική εμφάνιση συνδέσμων προς άλλα σχετικά άρθρα. Οι σύνδεσμοι αυτοί είτε επιλέγονται από κάποιον διαχειριστή του κόμβου, είτε παρατίθενται

οργανωμένα σύμφωνα με τη θεματική ενότητα στην οποία ανήκουν και τη χρονική σειρά εισαγωγής τους στον κόμβο.

Παρατηρούμε ότι σε γενικές γραμμές υπάρχει μία σταθερότητα και ομαλότητα όσον αφορά στις αιτήσεις προς τον κόμβο σε όλη τη διάρκεια της χρονικής περιόδου που εξετάζουμε, και αυτό προκύπτει από την εβδομαδιαία εξέταση των αρχείων καταγραφής των επισκέψεων στον κόμβο. Η κύρια δραστηριότητα των επισκεπτών, σύμφωνα με τα στατιστικά, φαίνεται να προσελκύεται γύρω από τα άρθρα που δημοσιεύονται στον κόμβο με αρκετά μεγάλη συχνότητα. Δίχως αμφιβολία, σημαντικό στοιχείο της ανάλυσης αποτελούν και τα περισσότερο σταθερά στοιχεία στον κόμβο, όπως τα διάφορα αφιερώματα Ασφαλώς, σημαντική αναμένεται να είναι η εφαρμογή της Διαδικασίας Εξόρυξης Δεδομένων για να επαληθεύσει ή να διαφοροποιήσει τις μέχρι τώρα διαπιστώσεις.

5 Κεφάλαιο: Αξιολόγηση της τεχνικής

Στο προηγούμενο κεφάλαιο της εργασίας, αποκτήσαμε μία συνολική εικόνα για το είδος των πραγματικών δεδομένων που χρησιμοποιήθηκαν για την μελέτη της περίπτωσης που μας απασχολεί. Γνωρίζουμε δηλαδή τα ποιοτικά χαρακτηριστικά των δεδομένων μας και είμαστε σε θέση να αναπτύξουμε με βάση αυτά, τη διαδικασία Ανακάλυψης Γνώσης.

Στην ενότητα που ακολουθεί, αρχικά θα περιγράψουμε την εφαρμογή της διαδικασίας εξόρυξης δεδομένων που αναπτύξαμε και τα διάφορα στάδια που ακολουθήσαμε με σκοπό την Ανακάλυψη Γνώσης. Όλα τα στάδια της KDD διαδικασίας, που αναπτύχθηκαν διεξοδικά σε προηγούμενο κεφάλαιο, ακολουθήθηκαν για την επεξεργασία ενός αρχείου καταγραφής των επισκέψεων και την εξαγωγή συμπερασμάτων. Πιο συγκεκριμένα εκτελέσαμε τα εξής στάδια:

(α) Προετοιμασία και καθαρισμός των δεδομένων

(β) Ανακάλυψη προτύπων στα δεδομένα

(γ) Ανάλυση των αποτελεσμάτων

Εκτός από την πρακτική εφαρμογή όμως των σταδίων εξόρυξης, η παρούσα εργασία καλείται να απαντήσει στο ερώτημα κατά πόσο μπορεί να έχει σημαντικά αποτελέσματα η συγκεκριμένη μέθοδος για πραγματικά δεδομένα από τον κόσμο του Διαδικτύου. Έτσι, παραθέτουμε την αξιολόγηση των αποτελεσμάτων που προέκυψαν από την ανάπτυξη της διαδικασίας Ανακάλυψης Γνώσης. Η αξιολόγηση αυτή, αφορά στις μετρήσεις των βασικότερων μεγεθών που μας ενδιαφέρουν όπως ο αριθμός των κανόνων που εξάγονται, το μέγεθός τους και ο χρόνος εκτέλεσης.

5.1 Εφαρμογή των Σταδίων της διαδικασίας Ανακάλυψης Γνώσης – Πορεία Ανάπτυξης

Αρχικά πρώτη προτεραιότητα μας, σύμφωνα με τη ανάλυση των επιμέρους σταδίων όπως αυτά ορίστηκαν στο 2^{ο} Κεφάλαιο της παρούσας εργασίας, είναι απαραίτητο να ορισθεί το **είδος της μεθόδου** που θα ακολουθήσουμε, και **οι στόχοι μας** γιατί έτσι καθορίζονται ακολούθως και τα υπόλοιπα στάδια της επεξεργασίας. Στόχος μας λοιπόν σε αφηρημένο επίπεδο, είναι αρχικά να καθορίσουμε τις «**συναλλαγές**» των επισκεπτών σε κάποιο ηλεκτρονικό κόμβο με σκοπό να τις επεξεργαστούμε και να εξαγάγουνε **κανόνες συσχέτισης** για τη συμπεριφορά των επισκεπτών. Οι κανόνες αυτοί θα τροφοδοτήσουν σε τελικό επίπεδο το μηχανισμό **Prefetching**, όπως αυτός έχει ορισθεί στο 4^{ο} Κεφάλαιο, με απώτερο στόχο τη βελτίωση της απόδοσης του ηλεκτρονικού Κόμβου.

Στα πλαίσια της παρούσας εργασίας αναπτύχθηκε κατάλληλο λογισμικό, το οποίο αφορά την προ-επεξεργασία και το μετασχηματισμό των δεδομένων και την εφαρμογή του αλγορίθμου εξόρυξης δεδομένων. Διακρίνουμε το λογισμικό αυτό σε δύο επιμέρους μονάδες, τα προγράμματα **log-processor** και **Apriori** που επιτελούν την επεξεργασία των δεδομένων και την εφαρμογή του αλγορίθμου εξόρυξης Apriori αντίστοιχα και τα οποία έχουν αναπτυχθεί στη γλώσσα προγραμματισμού C (περισσότερα στο Παράρτημα).

Αφού ορίσαμε τον στόχο της ανάλυσης που θα επιχειρηθεί, θα πρέπει να ασχοληθούμε με τα πρωτογενή δεδομένα. Το πρώτο στάδιο της επεξεργασίας αφορά τη συλλογή των δεδομένων και τον την προ-επεξεργασία τους με υπό-διαδικασίες όπως : ο καθαρισμός των δεδομένων, η ενοποίηση από διάφορες πηγές, η αφαίρεση των άχρηστων για την επεξεργασία δεδομένων και του θορύβου. Επιλέξαμε τα δεδομένα που μας ενδιαφέρει να μελετήσουμε και αυτά αφορούν όπως αναφέραμε και στο προηγούμενο κεφάλαιο, τα αρχεία καταγραφής των επισκέψεων στον ηλεκτρονικό κόμβο του Μακεδονικού Πρακτορείου Ειδήσεων (Μ.Π.Ε.) κατά το χρονικό διάστημα 01/08/2003 έως 05/09/2003.

Τα αρχεία καταγραφής των επισκέψεων ή αλλιώς (web server) log files όπως αποκαλούνται, ακολουθούν την Τυπική Μορφή των αρχείων καταγραφής (Common log file format) όπως ορίζεται από την διεθνή κοινοπραξία προτύπων για τον Παγκόσμιο Ιστό W3C (World Wide Web Consortium). Κάθε γραμμή του αρχείου καταγραφής αντιπροσωπεύει μία αίτηση για κάποιο έγγραφο του εξυπηρετητή. Τα δεδομένα τα οποία καταγράφονται στο αρχείο αφορούν στην διεύθυνση του επισκέπτη (client), την ώρα που γίνεται η αίτηση, τη μέθοδο και έχουν την εξής δομή:

- remotehost rfc931 authuser [date] "method URL" status bytes

- Remotehost : Η διεύθυνση του χρήστη (IP ή DNS).
- rfc931/Authuser : Χαρακτηριστικά του χρήστη.
- [date] : Ημέρα και ώρα της αίτησης..
- "request" : Η αίτηση όπως ακριβώς πληκτρολογήθηκε από το χρήστη.
- Status : Ο κωδικός κατάστασης του HTTPπου επιστρέφεται στο χρήστη.
- Bytes : Το μέγεθος σε bytes της απάντησης του εξυπηρετητή.

Για παράδειγμα:

129.128.29.28 – [14/Feb/2003:08:25:32 -0200] “GET /specials/?greece.html HTTP/1.0” 200 8817

Το επόμενο στάδιο αφορά στον καθορισμό του training set σύμφωνα με την ορολογία της εξόρυξης δεδομένων, δηλαδή των δεδομένων εκείνων που αφορούν τη μελέτη και την αφαίρεση όλων των υπολοίπων. Πρώτα απ' όλα, επιλέγουμε ποιους τύπους αρχείων θα συμπεριλάβουμε στην ανάλυσή μας και ποιους θα αφαιρέσουμε. Στη φάση αυτή οι τύποι αρχείων που επιλέγουμε να μην συμπεριλάβουμε είναι στην περαιτέρω επεξεργασία είναι τα εξής:

- Αρχεία τύπου "PHP"
- Αρχεία τύπου "Cgi-bin"
- Εικόνες τύπου ".jpg" ή ".gif"

Στη συνέχεια κάνουμε κάποιους μετασχηματισμούς που θα διευκολύνουν τους υπολογισμούς και την επεξεργασία των δεδομένων στο επόμενο στάδιο. Αυτοί οι μετασχηματισμοί έχουν να κάνουν με την αντιστοίχηση των μοναδικών IP διευθύνσεων των επισκεπτών καθώς και των εγγράφων του κόμβου σε ακέραιους αριθμούς, έτσι ώστε να απλοποιηθούν οι υπολογισμοί στη συνέχεια. Ακόμα αφαιρούνται οι αιτήσεις και τα έγγραφα που έχουν μηδενικό μέγεθος.

Στη φάση αυτή της προ-επεξεργασίας, γίνεται ο καθαρισμός του αρχείου δεδομένων από τα αρχεία εκείνα που ζητούνται πολύ λίγες φορές συνολικά και τα οποία δεν θα επηρεάσουν την ανάλυση. Υπολογίζουμε ένα ποσοστό ελάχιστης εμφάνισης (0,001%) για το σύνολο των αιτήσεων στο αρχείο καταγραφής και στη συνέχεια όλες οι αιτήσεις που έχουν απομείνει αποτελούν το «καθαρισμένο» log file, το οποίο είναι έτοιμο για το επόμενο στάδιο της επεξεργασίας, που είναι ο ορισμός των transactions του κάθε χρήστη.

Τα στοιχεία που λαμβάνουμε υπόψη και που καθορίζουν το σχηματισμό των συναλλαγών των επισκεπτών του ηλεκτρονικού κόμβου, είναι ο ακέραιος αριθμός που αντιπροσωπεύει την IP διεύθυνση του επισκέπτη και το χρονικό διάστημα (stride) που μεσολαβεί ανάμεσα σε δύο διαδοχικές αιτήσεις από την ίδιο επισκέπτη.

Έτσι, ορίζουμε το χρονικό αυτό διάστημα σε **2 ώρες** και το μέγιστο αριθμό αιτήσεων που μπορούν να περιλαμβάνονται σε μία επίσκεψη στο όριο των **15** εγγράφων.

5.2 Αποτελέσματα

Ο δεύτερος κύριος στόχος της εργασίας είναι να εξερευνήσει τη δυνατότητα αξιοποίησης της γνώσης που παρήγαγε η διαδικασίας εξόρυξης στην τεχνική του Web Prefetching. Η γνώση που τελικά παρήγαγε η διαδικασία εξόρυξης είναι ένα

σύνολο κανόνων. Επιθυμούμε να εξετάσουμε ένα σύνολο ζητημάτων που σχετίζονται με αυτούς τους κανόνες.

Πολύ συχνά αναφέρεται ότι το σύνολο των κανόνων που παράγονται με τη διαδικασία εύρεσης συσχετίσεων (association rule mining) είναι πολύ μεγαλύτερο από την βάση δεδομένων που εξορύχτηκε και συνεπώς η παραγόμενη γνώση δεν μπορεί να αξιοποιηθεί. Πρώτος στόχος μας λοιπόν είναι να απαντήσουμε στο ερώτημα εάν παράγονται κάποιοι κανόνες και πόσοι είναι αυτοί. Επιπλέον επιθυμούμε να μελετήσουμε τα χαρακτηριστικά αυτών των κανόνων, γιατί παρόλο που οι μέχρι τώρα μελέτες επικεντρώνονται μόνο στην εύρεση των μεγάλων συνόλων (large itemsets), για την εφαρμογή του Prefetching μας ενδιαφέρει και η μορφή των κανόνων.

Δεύτερος στόχος είναι να εξερευνήσουμε το ζήτημα του κατά πόσο η διαδικασίας εύρεσης κανόνων μπορεί να επαναλαμβάνεται αρκετά συχνά. Αυτό μας ενδιαφέρει επειδή οι περισσότεροι σημερινοί διαδικτυακοί τόποι (Web sites) δέχονται μεγάλο αριθμό επισκέψεων από χρήστες και αυτό συνεπάγεται ότι οι “κανονικότητες” στις επισκέψεις μεταβάλλονται αρκετά συχνά. Θα θέλαμε να μπορούμε να εκτελούμε τη διαδικασίας εύρεσης κανόνων αρκετά συχνά και συνεπώς μας ενδιαφέρει ο χρόνος εκτέλεσης του αλγορίθμου.

Στις επόμενες παραγράφους θα μελετήσουμε τα προαναφερθέντα ζητήματα, δηλαδή τη δομή και το πλήθος των κανόνων και το χρόνο εκτέλεσης.

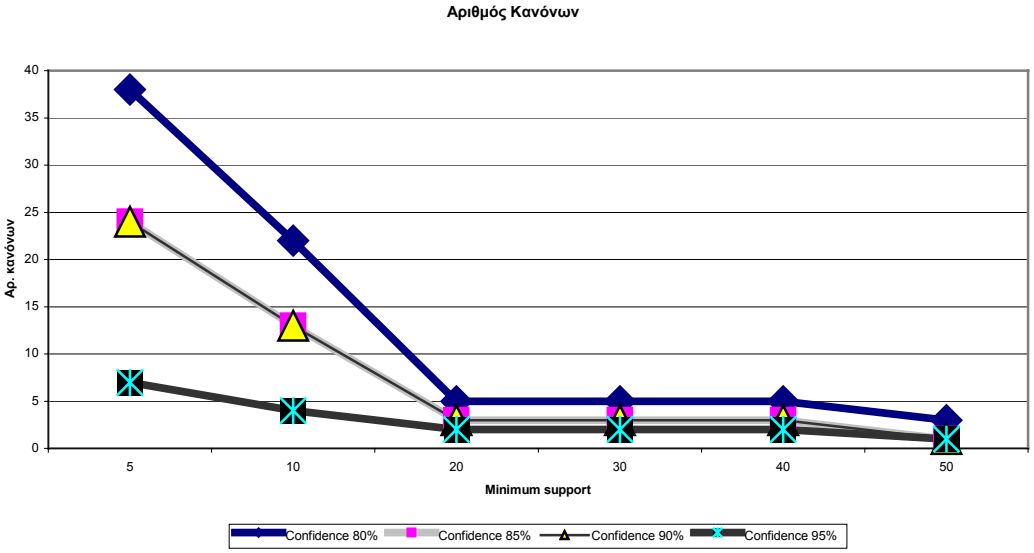
5.2.1 Αριθμός κανόνων

Ο αριθμός των κανόνων που παράγονται σε συνάρτηση με τα διάφορα επίπεδα για τα Support και Confidence, είναι πολύ σημαντικά στοιχεία της ανάλυσής μας, επειδή καθορίζει το πόσο μεγάλη εφαρμογή μπορεί να έχουν οι κανόνες αυτοί ή όχι.

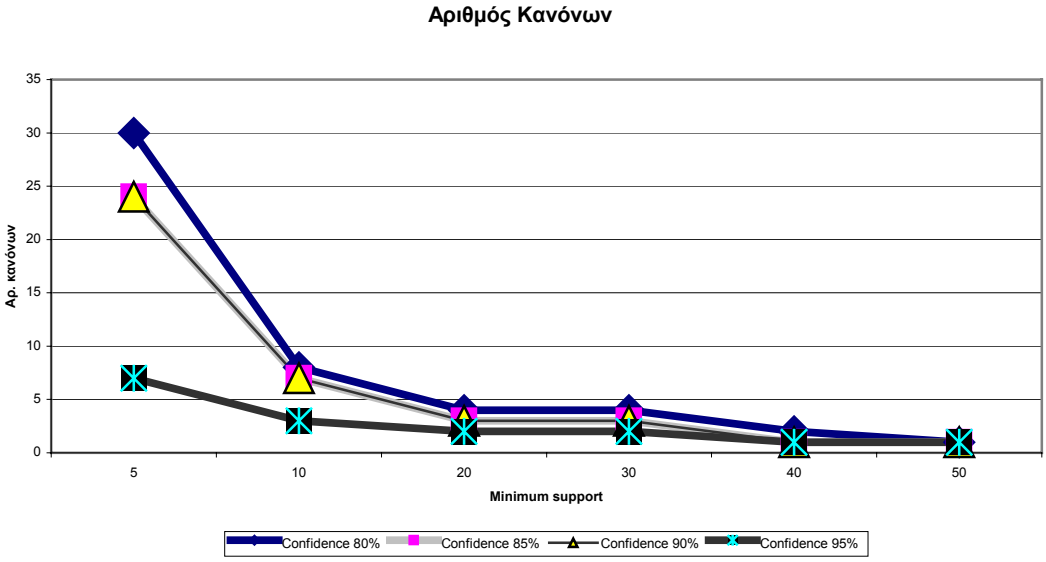
Ένα μεγάλο πλήθος κανόνων σημαίνει πρακτικά ότι είναι πολύ δύσκολο να επιλέξουμε τον κατάλληλο κανόνα τον οποίο θα ενεργοποιήσουμε για να κάνουμε Prefetching. Εάν πυροδοτήσουμε (κάνουμε firing) όλους τους σχετικούς κανόνες, τότε θα προκαλέσουμε μεγάλο network traffic και αντί έχουμε οφέλη από το Prefetching θα προκαλέσουμε συμφόρηση (congestion) στο δίκτυο. Από την άλλη πλευρά, ένας μικρός αριθμός κανόνων, θα σημαίνει ότι το όφελος από το Prefetching (δηλαδή η αύξηση του cache hit ratio) θα είναι πολύ μικρό.

Παρουσιάζουμε στα διαγράμματα που ακολουθούν τις μετρήσεις που κάναμε για το πλήθος των κανόνων που παράγονται (αρχεία mpa1 και mpa2).

Αξίζει να σημειωθεί, ότι και για τα υπόλοιπα αρχεία που δεν φαίνεται εδώ, ο αριθμός των κανόνων που παρουσιάζονται, έχουμε σχεδόν πανομοιότυπα αποτελέσματα:



Σχήμα 10. Αριθμός Κανόνων Συσχέτισης (αρχείο mpa1).



Σχήμα 11. Αριθμός Κανόνων Συσχέτισης (αρχείο mpa2).

Από τα παραπάνω γραφήματα εξάγουμε τα ακόλουθα ποιοτικά συμπεράσματα. Το πρώτο είναι ότι πράγματι ο αριθμός των παραγόμενων κανόνων είναι σχετικά μικρός. Αυτό σημαίνει ότι μπορούν εύκολα να αποθηκευτούν στην κύρια μνήμη. Το δεύτερο συμπέρασμα είναι ότι ο αριθμός των κανόνων που εξάγονται από την επεξεργασία των δεδομένων, είναι συνεχώς μειούμενος καθώς αυξάνεται η ελάχιστη στήριξή τους και αυτό επαληθεύεται για

όλα τα επίπεδα του βαθμού εμπιστοσύνης. Κάτι τέτοιο είναι απολύτως φυσιολογικό, αν σκεφτεί κανείς ότι είναι απόρροια του ότι αυξάνονται συνεχώς οι ελάχιστες προϋποθέσεις που θέτουμε.

Από την επεξεργασία του αρχείου mpa1, βλέπουμε ότι για το ελάχιστο επίπεδο στήριξης που μελετάμε που είναι το 5% και για τον ελάχιστο βαθμό εμπιστοσύνης 80%, παράγονται συνολικά 38 κανόνες. Σε κάθε αύξηση της ελάχιστης εμπιστοσύνης κατά πέντε ποσοστιαίες μονάδες, παρατηρείται μείωση του αριθμού των κανόνων που παράγονται σχεδόν κατά το ήμισυ. Τελικά, όταν αυξάνεται το επίπεδο ελάχιστης στήριξης στο 50% φτάνουμε να βρίσκουμε τρεις μόνο κανόνες συσχέτισης. Κάτι ανάλογο συμβαίνει σε γενικές γραμμές και για τα υπόλοιπα αρχεία που επεξεργαζόμαστε, όπου παρατηρούμε ότι ο αριθμός των κανόνων που παράγονται σταδιακά μειώνεται και τείνει στο μηδέν όταν αυξηθούν τα επίπεδα της ελάχιστης στήριξης και του βαθμού εμπιστοσύνης.

Θα πρέπει εδώ να σημειώσουμε ότι δεν εξετάζουμε χαμηλότερα επίπεδα εμπιστοσύνης, επειδή επιθυμούμε να έχουμε ικανοποιητική ακρίβεια στην πρόβλεψη των μελλοντικών αιτήσεων των χρηστών.

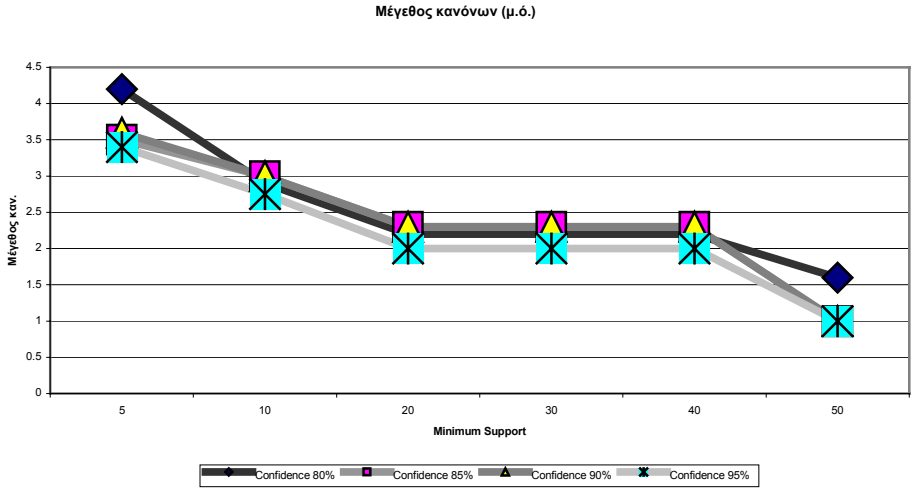
Συνεπώς, από τις μετρήσεις για το σύνολο των κανόνων συσχέτισης που παράγονται από την επεξεργασία των δεδομένων, φτάνουμε στο συμπέρασμα ότι είναι ικανοποιητικό το πλήθος των κανόνων. Έχουμε ένα σύνολο από αρκετά ισχυρούς κανόνες και σε επίπεδα support και confidence που μας επιτρέπουν να βασίσουμε σε αυτούς την πρόβλεψή μας για τη διαδρομή του χρήστη στο δικτυακό κόμβο και να τροφοδοτήσουμε έτσι το μηχανισμό της διαδικασίας προ-φόρτωσης σελίδων (Prefetching).

5.2.2 Μέγεθος κανόνων

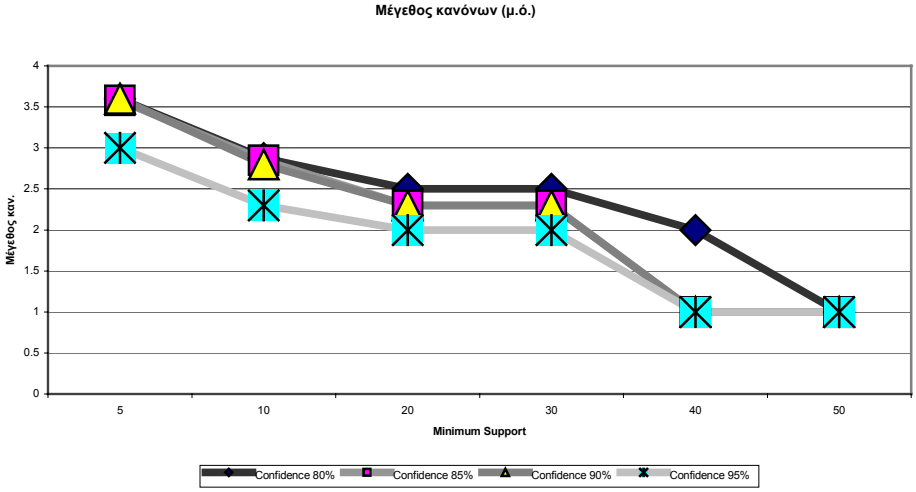
Το μέγεθος ενός κανόνα συσχέτισης αναφέρεται στο πλήθος των στοιχείων (items) που αποτελούν τον κανόνα, και που παρουσιάζονται στο body και στο head. Το μέγεθος αυτό είναι επίσης ένα πολύ σημαντικό στοιχείο για την αξιολόγηση της έρευνάς μας, καθώς προσδιορίζει ποιοτικά τους κανόνες που εξάγονται. Ένας κανόνας που αποτελείται από πολύ λίγα στοιχεία, αλλά και ένας κανόνας με πολύ μεγάλο μέγεθος, στην πραγματικότητα θα είναι δύσκολο να χρησιμοποιηθούν προκειμένου να κάνουμε προβλέψεις και να δώσουν έτσι βοήθεια στον μηχανισμό του Prefetching. Άρα εκείνο που θα επιδιώκαμε θα ήταν η εξαγωγή ενός μέσου αριθμού κανόνων που να μπορεί να χρησιμοποιηθεί με

σκοπό την πρόβλεψη της επόμενης επιλογής του χρήστη, άλλα και με ένα μικρό αριθμό στοιχείων που να αποτελούν τον κανόνα.

Στα διαγράμματα που ακολουθούν υπολογίζεται το μέγεθος των κανόνων (χωρίς να υπολογίζεται το head του κανόνα) που εξάγονται για τα διαφορετικά επίπεδα της ελάχιστης στήριξης και του βαθμού εμπιστοσύνης των κανόνων. Να σημειωθεί ότι επιλέξαμε να εξαγάγουμε μόνο τους κανόνες οι οποίοι έχουν μόνο ένα στοιχείο στο head τους. Ο λόγος είναι ότι για το Prefetching μας ενδιαφέρει να μπορούμε να κάνουμε μια πρόβλεψη σχετικά γρήγορα χωρίς να περιμένουμε από τον χρήστη να εκτελέσει πολλές επισκέψεις μέσα στον δικτυακό τόπο.



Σχήμα 12 Μέγεθος Κανόνων Συσχέτισης (αρχείο mpa1).



Σχήμα 13 Μέγεθος Κανόνων Συσχέτισης (αρχείο mpa1).

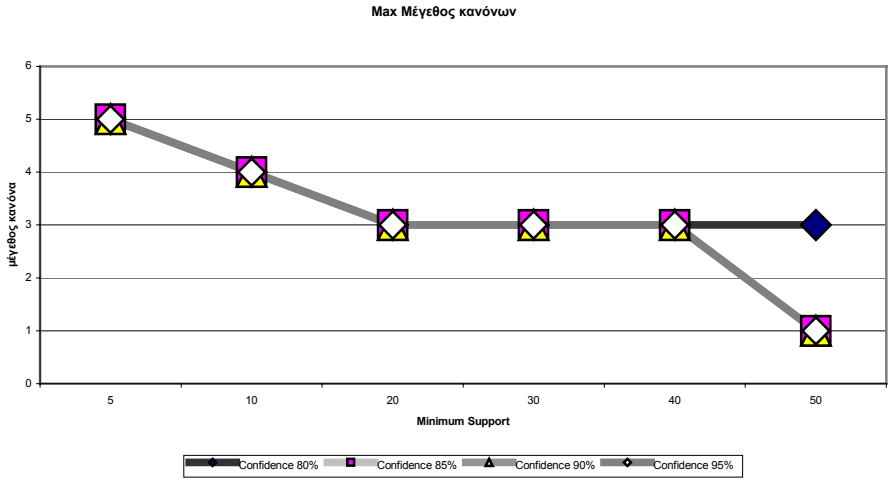
Από τα παραπάνω σχεδιαγράμματα συμπεραίνουμε το μέσο μήκος ενός κανόνα είναι σχετικά μικρό, το οποίο είναι ενθαρρυντικό γιατί η εκτέλεση του κανόνα δεν αναμένεται να επιφέρει μεγάλο network traffic.

Επίσης, παρατηρούμε ότι το μέσο μέγεθος των κανόνων που εξάγονται μειώνεται σταθερά καθώς αυξάνεται το επίπεδο της ελάχιστης στήριξης των κανόνων, για όλα τα επίπεδα του βαθμού εμπιστοσύνης. Παρατηρούμε ότι το μέσο μέγεθος των κανόνων ξεκινάει από το ύψος των 3,6 στοιχείων ανά κανόνα και μειώνεται όσο αυξάνονται τα ποιοτικά μεγέθη του support και confidence. Βλέπουμε δηλαδή, ότι έχουμε τη δημιουργία σχετικά μικρού μεγέθους κανόνων, γεγονός που επίσης είναι πολύ θετικό για την εξαγωγή προβλέψεων.

5.2.3 Μέγιστο μέγεθος κανόνων

Ο μέγιστος αριθμός στοιχείων που συμμετέχει σε κάθε κανόνα σχετίζεται άμεσα με το προηγούμενο σημείο όπου υπολογίσαμε το μέσο μέγεθος του κάθε κανόνα. Όπως είδαμε εκεί, το σχετικά μικρό πλήθος των κανόνων που προκύπτουν ευνοεί την εξαγωγή προβλέψεων. Το δεδομένο αυτό σε συνδυασμό με το μέγιστο πλήθος σε κάθε επίπεδο της διαδικασίας, μπορούμε να πούμε ότι περιγράφει καλύτερα την εικόνα των αποτελεσμάτων.

Στο διάγραμμα που ακολουθεί παρουσιάζεται η διακύμανση του κανόνα με το μέγιστο πλήθος στοιχείων που το αποτελούν για κάθε επίπεδο για την ελάχιστη στήριξη και το βαθμό εμπιστοσύνης, για το αρχείο mpa1. Και για τα υπόλοιπα αρχεία καταγραφής επισκέψεων της χρονικής περιόδου που μας απασχολεί, παρουσιάζονται με μεγάλη ακρίβεια τα ίδια αποτελέσματα:

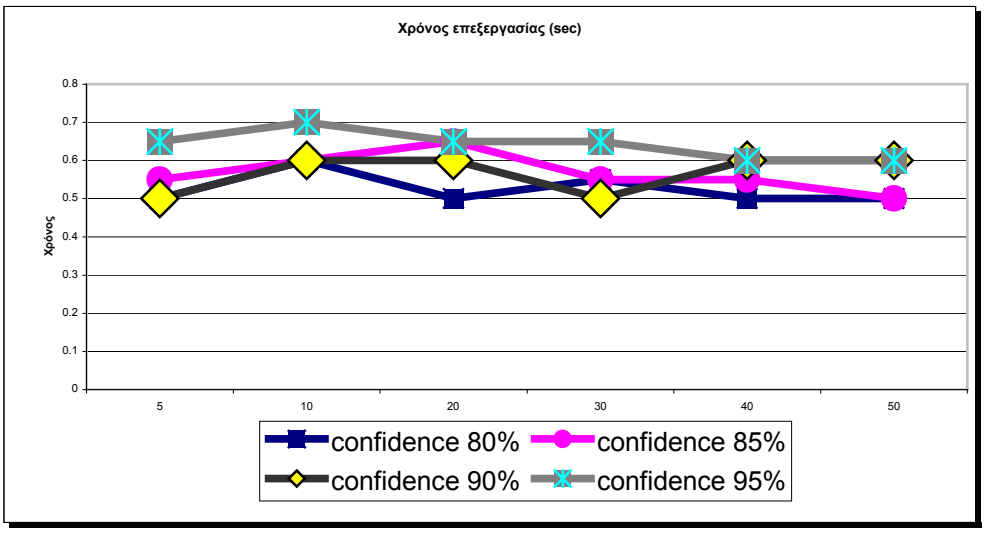


Σχήμα 14 Μέγιστο μέγεθος κανόνα συσχέτισης (αρχείο mpa1)

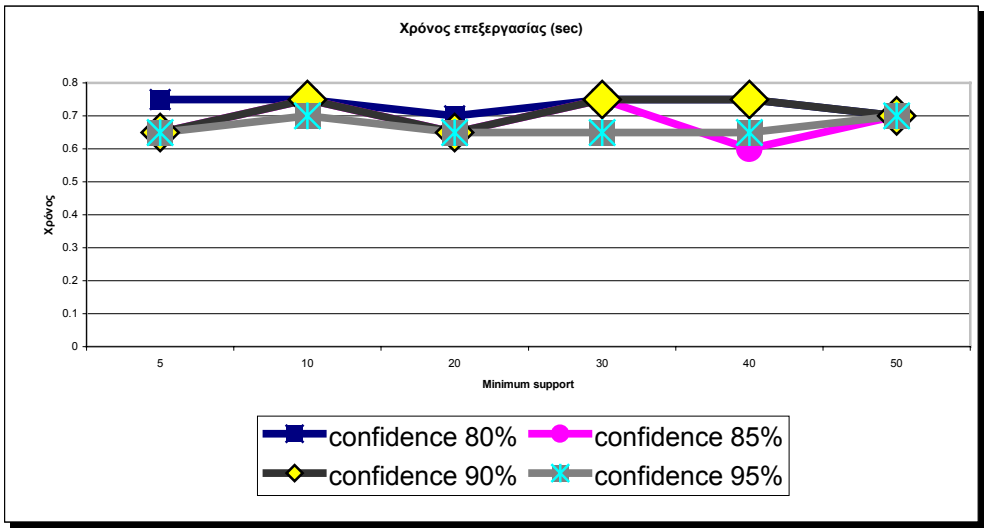
Βλέπουμε από το διάγραμμα, ότι ο κανόνας με το μέγιστο μέγεθος ανεξάρτητα με το επίπεδο του βαθμού εμπιστοσύνης, επηρεάζεται από την ελάχιστη στήριξη και παίρνει διακριτές τιμές μεταξύ ενός και πέντε στοιχείων, έχοντας κατά μέσο όρο τρία στοιχεία. Κάτι τέτοιο είναι πολύ θετικό, καθώς αποδεικνύεται ότι το μέγιστο μέγεθος των κανόνων που παράγονται δεν ξεπερνάει κατά πολύ το αντίστοιχο μέσο μέγεθος. Έτσι, διατηρούνται οι κανόνες σε ένα πολύ καλό επίπεδο για τον σκοπό για τον οποίο τους χρειαζόμαστε.

5.2.4 Χρόνος επεξεργασίας

Πολύ σημαντικό στοιχείο για την αξιολόγηση των αποτελεσμάτων που προκύπτουν είναι ο χρόνος επεξεργασίας ο οποίος απαιτήθηκε. Συγκεκριμένες μετρήσεις έχουμε καταγράψει για τη μονάδα του προγράμματος που εξάγει τους κανόνες συσχέτισης με εφαρμογή τους αλγορίθμου Apriori. Τα αποτελέσματα των μετρήσεων για τα πρώτα δύο από τα πέντε εβδομαδιαία αρχεία log files που επεξεργαστήκαμε, παρουσιάζονται στα δύο διαγράμματα που ακολουθούν. Τα διαγράμματα για τα υπόλοιπα αρχεία καταγραφής ακολουθούν παρόμοια δομή και γι' αυτό δεν τα παραθέτουμε εδώ (βλ. Παράρτημα).



Σχήμα 15. Χρόνος επεξεργασίας (αρχείο mpa1).



Σχήμα 16. Χρόνος επεξεργασίας mpa2.

Παρατηρούμε ότι ο χρόνος επεξεργασίας του προγράμματος εξαγωγής κανόνων συσχέτισης επηρεάζεται ελάχιστα από τα μεγέθη της εμπιστοσύνης (Support) και της στήριξης (Confidence). Μπορούμε να πούμε ότι ο χρόνος επεξεργασίας διατηρείται σταθερός με μικρές διακυμάνσεις. Γενικά, φαίνεται ότι όσο αυξάνεται το ποσοστό της εμπιστοσύνης και της στήριξης των κανόνων, τόσο παρατηρείται μια τάση για μείωση του χρόνου επεξεργασίας, καθώς μειώνεται εξαρχής το σύνολο των πιθανών κανόνων συσχέτισης και ο αλγόριθμος Apriori είναι έτσι περισσότερο αποτελεσματικός.

Συνολικά, αθροίζοντας τον χρόνο επεξεργασίας σε ολόκληρη τη διάρκεια της διαδικασίας από το στάδιο της προ-επεξεργασίας των δεδομένων και του προσδιορισμού των ομάδων επισκέψεων των χρηστών, μπορούμε να πούμε ότι κυμαίνεται σε μεγέθη μικρότερα της τάξεως του ενός λεπτού (40-60 δευτερόλεπτα) για το καθένα από τα αρχεία καταγραφής των επισκέψεων. Συμπεραίνουμε έτσι, ότι η όλη διαδικασία είναι αρκετά σύντομη και ενδείκνυται ακόμα και για συχνή εκτέλεση σε περιπτώσεις ηλεκτρονικών κόμβων με υψηλή συχνότητα ανανέωσης των περιεχομένων τους.

5.3 Συμπεράσματα

Συμπερασματικά, μπορούμε να πούμε ότι από τις παραπάνω μετρήσεις αποδεικνύεται ότι η εξαγωγή των κανόνων συσχέτισης με την μέθοδο την οποία ακολουθήσαμε είναι ιδιαίτερα επιτυχής. Κατ' αρχήν αποδεικνύεται ότι η διαδικασία

εξόρυξης δεδομένων και ο αλγόριθμος που επιλέχθηκε παράγει κανόνες συσχέτισης ανάμεσα στα πραγματικά δεδομένα που επεξεργαστήκαμε και πολύ περισσότερο, τους παράγει με πολύ καλούς χρόνους επεξεργασίας των δεδομένων και εκτέλεσης του λογισμικού. Σε δεύτερο επίπεδο, οι κανόνες αυτοί είναι πολύ ικανοποιητικοί ποιοτικά, λόγω του πλήθους τους, δεν είναι πολλοί για τα επίπεδα support και confidence που μας ενδιαφέρουν, και του μεγέθους τους, δεν είναι ούτε υπερβολικά μεγάλοι ούτε πολύ μικροί. Είναι δηλαδή μέσα στα αποδεκτά όρια που απαιτούνται για το σκοπό που τα χρησιμοποιούμε.

6 Κεφάλαιο: Κατανομή εύρους ζώνης σε κινητά περιβάλλοντα

Όπως έχουμε αναφέρει στο εισαγωγικό κεφάλαιο, στα κινητά περιβάλλοντα υπολογισμών αντίθετα με τα συνηθισμένα στατικά δίκτυα υποστηρίζεται η δυναμική επανατοποθέτηση των κινητών τερματικών, με συνέπεια τα σημεία πρόσβασης των χρηστών στο δίκτυο να αλλάζουν καθώς οι χρήστες μετακινούνται σε διαφορετικές τοποθεσίες. Αυτή η κινητικότητα, προκαλεί μια σειρά από προβλήματα, μεταξύ των οποίων η διαχείριση των επιμέρους περιοχών και η κατανομή του εύρους συχνότητας. Τα ζητήματα που τίθενται αφορούν στις επιμέρους διαδικασίες: (α) της χωρικής καταγραφής (Location) και (β) της αναζήτησης (Paging). Η πρώτη διαδικασία επιτρέπει στο σύστημα να καταγράφει τις πληροφορίες (ακριβείς ή κατά προσέγγιση) σχετικά με τη θέση του κάθε χρήστη, έτσι ώστε να είναι σε θέση να τον εντοπίζει όποτε αυτό χρειάζεται. Η διαδικασία της αναζήτησης αποτελείται από την αποστολή μηνυμάτων προς όλες τις κατευθύνσεις (κελιά) για τον εντοπισμό του χρήστη.

Στο κεφάλαιο αυτό θα επιχειρήσουμε να περιγράψουμε μία μέθοδο για μελέτη και επεξεργασία των διαδρομών των χρηστών σε ένα γεωγραφικό χώρο – πλέγμα με σκοπό την πρόβλεψη με βάση τα πρότυπα της κίνησης των χρηστών που θα προκύψουν. Η διαδικασία είναι απλή: οργανώνουμε τις τροχιές των χρηστών σε ομάδες ανάλογα με την ομοιότητά τους και έπειτα ορίζουμε τις περισσότερο αντιπροσωπευτικές τροχιές και με αυτές προσπαθούμε να ταυτίσουμε την τρέχουσα τροχιά του χρήστη και να συμβουλέψουμε τη διαδικασία εκχώρησης εύρους ζώνης για την μελλοντική κίνησή του.

6.1 Δυναμική ομαδοποίηση των διαδρομών του χρήστη

Η τεχνική που θα χρησιμοποιηθεί για την οργάνωση των διαδρομών των χρηστών σε παρόμοιες ομάδες είναι η ομαδοποίηση (Clustering). Παρόλο που στη βιβλιογραφία έχουν προταθεί πολλοί αποτελεσματικοί αλγόριθμοι για την μέθοδο clustering, θα πρέπει να εστιάσουμε το ενδιαφέρον μας σε εκείνους τους αλγορίθμους που δεν περιορίζονται σε επεξεργασία αντικειμένων στο Ευκλείδειο χώρο. Τέτοιου είδους είναι οι ιεραρχικοί συσσωρευτικοί αλγόριθμοι.

Θα πρέπει να οριστεί ένα μέτρο της απόστασης ανάμεσα στις διαδρομές καθώς αυτές δεν είναι απλά σημεία, αλλά ακολουθίες από κελιά που έχουν επισκεφθεί οι χρήστες. Η έννοια της απόστασης ανάμεσα σε μία τροχιά $A=<\alpha_1,...,\alpha_m>$ και μία τροχιά $B=<b_1,\ldots,b_m>$ θα πρέπει να αναφέρεται τόσο στον αριθμό των κελιών που είναι κοινά και στις δύο διαδρομές, αλλά και στην σειρά με την οποία επισκέφθηκε ο χρήστης τα κελιά αυτά. Οι προηγούμενες διαπιστώσεις

μας οδηγούν στην επιλογή του γνωστού μεγέθους της edit distance για την σύγκριση πεπερασμένων αλφαριθμητικών. Για να συμπεριλάβουμε στην ανάλυσή μας τον γεωμετρικό μη – Ευκλείδειο χώρο, θα προσθέσουμε την έννοια του βάρους στις αποστάσεις και τα μεγέθη που μας ενδιαφέρει να προσδιορίσουμε θα είναι weighted edit distance.

Επιπλέον εφόσον καθοριστούν οι ομάδες (clusters) από παρόμοιες τροχιές, μας ενδιαφέρει να υπολογίσουμε τις αποστάσεις ανάμεσα στις τροχιές που δημιουργούνται. Σε αυτήν την περίπτωση οι συσσωρευτικοί αλγόριθμοι σε γενικές γραμμές χρησιμοποιούν τις αποστάσεις ανάμεσα στα κέντρα των clusters για να μετρήσουν τις αποστάσεις ανάμεσα στα διαφορετικά clusters. Όταν πρόκειται για μη ευκλείδειους χώρους όπως στην περίπτωσή του κινητού περιβάλλοντος με τις τροχιές των χρηστών, μπορούμε να ορίζουμε ως την πιο αντιπροσωπευτική τροχιά εκείνη με το ελάχιστο άθροισμα αποστάσεων από τις υπόλοιπες τροχιές.

6.2 On-line πρόβλεψη της κίνησης

Το αποτέλεσμα της επεξεργασίας των διαδρομών των χρηστών στο πλέγμα του κινητού περιβάλλοντος που περιγράψαμε προηγουμένως, είναι ο καθορισμός ενός συνόλου από clusters, τα οποία αποτελούνται το καθένα από ένα σύνολο από αντιπροσωπευτικές τροχιές. Επομένως, αυτές οι τροχιές ορίζονται ως τα πρότυπα με βάση τα οποία κινούνται οι χρήστες. Έτσι, όταν εξετάζεται η διαδρομή ενός χρήστη, εκείνο που μας ενδιαφέρει είναι να βρούμε κατ' αρχήν με ποια από όλες τις αντιπροσωπευτικές διαδρομές ταιριάζει η τρέχουσα διαδρομή. Στη συνέχεια, σύμφωνα με την πρότυπη αυτή διαδρομή, προσδιορίζονται τα πιο πιθανά να μετακινηθεί ο χρήστης σημεία, και σύμφωνα με αυτά υλοποιείται η κατανομή πόρων.

Έστω ότι η τρέχουσα διαδρομή ενός χρήστη είναι $A = <\alpha_1,...,\alpha_\kappa>$ όπου τα α_1 και α_κ είναι τα κελιά που επισκέφθηκε πρόσφατα ο χρήστης. Σε αυτό το στάδιο η διαδρομή A θα πρέπει να συγκριθεί με όλες τις αντιπροσωπευτικές διαδρομές της μορφής $B = <b_1,\ldots,b_n>$ έτσι ώστε να βρεθούν οι διαδρομές αυτές που θα έχουν απόσταση d_m μικρότερη από οποιοδήποτε ελάχιστο κατώφλι t_m . Επίσης, σε αυτόν τον υπολογισμό θα πρέπει να λάβουμε υπόψη ότι η διαδρομή A ενός χρήστη, θα μπορούσε να είναι τμήμα (υπό-ακολουθία) της αντιπροσωπευτικής διαδρομής B.

Η λύση στο πρόβλημα του υπολογισμού της d_m απόστασης έχει δοθεί και βασίζεται στον δυναμικό προγραμματισμό.

7 Κεφάλαιο: Συμπεράσματα – Προτάσεις

Στην προηγούμενη ενότητα εξετάσαμε το βαθμό αξιοπιστίας της μεθόδου που αναπτύχθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας. Το μέτρο της αξιολόγησης της μεθόδου, είναι συνάρτηση των τεχνικών εξόρυξης δεδομένων που χρησιμοποιούμε. Οι τεχνικές εξόρυξης που υλοποιήθηκαν αφορούν στην εξόρυξη Κανόνων Συσχέτισης (Association Rules), και εφαρμόζεται εφόσον προηγουμένως έχουμε επεξεργαστεί τα δεδομένα και έχουμε προσδιορίσει τα σύνολα των επισκέψεων(transactions) στον υπό μελέτη ηλεκτρονικό κόμβο.

Τα σημαντικότερα μεγέθη που λάβαμε υπόψη μας για αξιολογήσουμε τους κανόνες που έχουν παράγει, ήταν η εξέταση διαφόρων μεγεθών για τα διαφορετικά επίπεδα ελάχιστης στήριξης (Minimum support) και εμπιστοσύνης (Confidence) του κάθε κανόνα. Οι μετρήσεις που κάναμε και μας βοήθησαν προς αυτήν την κατεύθυνση, αλλά και για την γενικότερη αξιολόγηση της ανάλυσής μας ήταν:

- Ο αριθμός των κανόνων που παράγονται
- Το μέγεθος των κανόνων
- Το μέγιστο μέγεθος κανόνων
- Ο χρόνος επεξεργασίας της διαδικασίας εξόρυξης.

Στην ενότητα αυτή καταλήγουμε στα τελικά συμπεράσματα της εργασίας. Θα προσπαθήσουμε να απαντήσουμε στα ερωτήματα που τέθηκαν αρχικά και αφορούν στο κατά πόσο μπορεί να εφαρμοστεί η διαδικασία εξόρυξης δεδομένων χρησιμοποιώντας πραγματικά δεδομένα του Παγκοσμίου Ιστού. Επίσης, πόσο χρήσιμα ήταν τα πρότυπα που προέκυψαν από τη διαδικασία εξόρυξης που χρησιμοποιήσαμε και κατά πόσο μπορούμε να τα αξιοποιήσουμε για να τροφοδοτήσουμε το μηχανισμό του web Prefetching;

Τέλος, θα αναφερθούμε στα γενικότερα συμπεράσματα της ανάλυσής μας και στις προτάσεις μας για μελλοντική εργασία στο θέμα.

7.1 Συμπεράσματα

Ο στόχος της διαδικασίας εξόρυξης δεδομένων που αναπτύχθηκε ήταν να παραχθούν κανόνες συσχέτισης, με βάση τους οποίους να κάνουμε προβλέψεις για τη διαδικασία Prefetching. Οι επιμέρους στόχοι αφορούσαν τη διερεύνηση της

ποιότητας των κανόνων αυτών που στη συνέχεια θα χρησιμοποιηθούν για να τροφοδοτήσουν την διαδικασία Prefetching.

Σύμφωνα με τα δεδομένα του προηγούμενου κεφαλαίου, σχετικά με την αξιολόγηση των αποτελεσμάτων, προκύπτουν θετικά συμπεράσματα για την διαδικασία που ακολουθήσαμε για την επεξεργασία των δεδομένων των επισκέψεων στον ηλεκτρονικό κόμβο του Μακεδονικού Πρακτορείου Ειδήσεων (Μ.Π.Ε.). Το πρώτο γενικό συμπέρασμα στο οποίο καταλήγουμε είναι ότι στα πραγματικά δεδομένα εισόδου που διαπραγματευτήκαμε συναντάμε **κανονικότητα** όσον αφορά στη συμπεριφορά των επισκεπτών του κόμβου. Η κανονικότητα αυτή των χρηστών αποτυπώνεται στους κανόνες συσχέτισης οι οποίοι εξάγονται για τη σειρά προσπέλασης στις διάφορες σελίδες που μας αφορούν.

Η μέθοδος που ακολουθήσαμε παράγει ικανοποιητικό αριθμό κανόνων συσχέτισης, γεγονός που αποδεικνύει ότι η επεξεργασία των δεδομένων ήταν επιτυχής. Επιπλέον, οι κανόνες που παράγονται έχουν χαρακτηριστικά που ευνοούν την εφαρμογή τους για το Prefetching. Όσον αφορά τα υψηλά επίπεδα στήριξης και εμπιστοσύνης (support & confidence) που μας ενδιαφέρουν, δηλαδή με αρκετή βεβαιότητα, οι κανόνες που παράγονται είναι σχετικά μικροί σε μέγεθος και επιπλέον λίγοι και κατανοητοί.

Ακόμα, οι κανόνες συσχέτισης ανιχνεύονται πολύ γρήγορα, καθώς η όλη διαδικασία για την επεξεργασία ενός web log αρχείου με μέσο μέγεθος 250 Mb και την εξαγωγή κανόνων συσχέτισης για τις σελίδες που επισκέπτονται οι χρήστες του κόμβου που μελετάμε, διαρκεί αναλογικά πολύ μικρό χρονικό διάστημα της τάξεως των 40-60 δευτερολέπτων. Κατά συνέπεια μπορεί κανείς να πειραματιστεί και να έχει διαφορετικά επίπεδα στατιστικά ισχυρών κανόνων, με μικρό κόστος από την πλευρά του χρόνου που καταναλώθηκε. Ο χρόνος επεξεργασίας είναι πολύ σημαντικό στοιχείο και για τον επιπλέον λόγο ότι σε πολλές περιπτώσεις θα χρειάζεται η επανάληψη της διαδικασίας επαναπροσδιορισμού των κανόνων σε σύντομα χρονικά διαστήματα. Ειδικά στην περίπτωση του ηλεκτρονικού κόμβου του ειδησεογραφικού οργανισμού που εξετάζουμε (Μ.Π.Ε.), όπου οι σελίδες του κόμβου έχουν μικρή σχετικά κύκλο ζωής και ταυτόχρονα υψηλή επισκεψιμότητα ημερησίως, ο μικρός χρόνος επεξεργασίας επιτρέπει την συχνή εκτέλεση της διαδικασίας ανακάλυψης γνώσης, έτσι ώστε να προσαρμόζεται στα νέα κάθε φορά δεδομένα.

Ως γενικότερο συμπέρασμα της ανάλυσής μας, μπορούμε να ισχυρισθούμε ότι η Ανακάλυψη γνώσης μπορεί να εφαρμοστεί με επιτυχία και στο πεδίο της βελτίωσης της απόδοσης σε κατανεμημένα συστήματα. Οι κανόνες συσχέτισης που μας χρησίμευσαν κυρίως για τη μελέτη της περίπτωσης του Παγκόσμιου Ιστού, δείξαμε ότι μπορούν να τροφοδοτήσουν σε ικανοποιητικό βαθμό το μηχανισμό Prefetching, με απώτερο στόχο τη βελτίωση των χρόνων αναμονής των χρηστών και την ομαλότερη λειτουργία του δικτύου. Η μελέτη του βαθμού της βελτίωσης που επιφέρει η μέθοδος που χρησιμοποιήθηκε, αφορά σε περισσότερο τεχνικά θέματα και αποτελεί αντικείμενο για τη μελλοντική ολοκλήρωση της έρευνάς μας.

7.2 Προτάσεις για μελλοντική εργασία

Στην παρούσα διπλωματική εργασία εξετάσαμε τη δυνατότητα εφαρμογής της διαδικασίας ανακάλυψης γνώσης με υλοποίηση της μεθόδου εξόρυξης κανόνων συσχέτισης, σε πραγματικά δεδομένα επισκέψεων σε έγγραφα του παγκόσμιου ιστού και αναπτύξαμε το αντίστοιχο λογισμικό. Η μελλοντική εργασία που προτείνεται αφορά στην έρευνα σχετικά με το κατά πόσο άλλες τεχνικές εξόρυξης δεδομένων μπορούν να εφαρμοσθούν και να έχουν επιτυχή αποτελέσματα. Θα διαχωρίσουμε τις προτάσεις για μελλοντική έρευνα ανάλογα με τα δύο κατανεμημένα συστήματα που μας απασχόλησαν, αρχικά σε αυτές που αφορούν τον παγκόσμιο ιστό και τέλος σε εκείνες που σχετίζονται με τα κινητά περιβάλλοντα υπολογισμών.

7.2.1 Παγκόσμιος Ιστός – Web Prefetching

Το ευρύ πλαίσιο λειτουργίας του παγκόσμιου ιστού αφήνει πολλά περιθώρια για έρευνα σχετικά με τις δυνατότητες εξόρυξης γνώσης από τα αρχεία καταγραφής των επισκέψεων των χρηστών σε ηλεκτρονικούς κόμβους. Μπορούμε να χρησιμοποιήσουμε τις ευρύτατα διαδεδομένες τεχνικές της ομαδοποίησης (clustering) και της ταξινόμησης (classification) των διαδρομών των επισκεπτών, με σκοπό τον καθορισμό ομάδων ή συνόλων από χρήστες με κοινά ενδιαφέροντα ή σελίδες που ζητούνται στην ίδια διαδρομή. Η τεχνική που επιλέξαμε να διερευνήσουμε στην παρούσα εργασία είναι η εξόρυξη κανόνων συσχέτισης ανάμεσα στις σελίδες που ζητούνται συχνά σε διαφορετικές διαδρομές.

7.2.1.1 Υπολογισμός απόστασης μεταξύ των αιτήσεων

Ένας παράγοντας που δε λάβαμε υπόψη μας και προτείνεται να εξεταστεί μελλοντικά αφορά στον υπολογισμό του αριθμού των εγγράφων που περιλαμβάνονται στα στοιχεία ενός κανόνα, κατά την εξέταση της τρέχουσας διαδρομής ενός χρήστη. Ο αριθμός των εγγράφων που έχει ζητήσει ο χρήστης (**ενδιάμεσα clicks**) μπορεί να αποτελέσει έναν δείκτη του κατά πόσο έγκυρος είναι ο κανόνας και πόσο αποδοτικά αναμένουμε ότι θα αποδώσει κατά τη διαδικασία του Web Prefetching.

7.2.1.2 Recommendation systems

Ένα δεύτερο αρκετά ενδιαφέρον ζήτημα που προκύπτει, είναι η δημιουργία ενός on-line συστήματος εξόρυξης δεδομένων που να ενσωματώνει ένα μηχανισμό υποδείξεων προς τους χρήστες που θα βασίζεται στη μελέτη των παλαιότερων επισκέψεων στον κόμβο (Recommendation system). Ο μηχανισμός που προτείνεται να υλοποιηθεί, σε πρώτη φάση θα επεξεργάζεται τα web logs που μας ενδιαφέρουν και έπειτα, μελετώντας την τρέχουσα διαδρομή του κάθε επισκέπτη θα μπορεί να κάνει προτάσεις για τις περισσότερο πιθανές επιλογές για την πλοήγησή του. Ένα ζήτημα που τίθεται είναι η συχνότητα με την οποία θα εκτελείται η επεξεργασία των δεδομένων καθώς και ο τρόπος με τον οποίο θα καταγράφεται και θα ταυτοποιείται η τρέχουσα διαδρομή του κάθε χρήστη.

7.2.2 Κινητά περιβάλλοντα υπολογισμών

Η μέθοδος η οποία μας απασχόλησε σχετικά με την έρευνα για την εξόρυξη δεδομένων σε κινητά περιβάλλοντα υπολογισμών ήταν η ομαδοποίηση των χρηστών. Με την κατάλληλη επεξεργασία των διαδρομών των χρηστών σε ένα γεωγραφικό χώρο – πλέγμα, μπορούμε να οργανώσουμε τις τροχιές με βάση την ομοιότητά τους. Σε δεύτερο επίπεδο μπορούμε να ορίσουμε τις περισσότερο αντιπροσωπευτικές από αυτές, με σκοπό να κατατάξουμε σε κατηγορίες τις μελλοντικές τροχιές των χρηστών και να είμαστε σε θέση να κάνουμε ασφαλέστερες προβλέψεις όσον αφορά στην κίνησή τους στο πλέγμα.

Μία άλλη μέθοδος που θα ήταν ενδιαφέρον να διερευνηθεί κατά πόσο μπορεί να οδηγήσει σε θετικά συμπεράσματα αναφορικά με την πρόβλεψη της διαδρομής των χρηστών στο πλέγμα, είναι η **αναζήτηση ακολουθιακών προτύπων** (sequential patterns) στις τροχιές των χρηστών και η χρησιμοποίησή

τους για την πρόβλεψη της κίνησης των χρηστών. Τα ακολουθιακά αυτά πρότυπα μπορούν να είναι αντίστοιχοι με τους κανόνες συσχέτισης που εξετάσαμε στην ανάλυσή μας για τα δεδομένα του παγκόσμιου ιστού. Προτείνεται η υλοποίηση του γνωστού αλγορίθμου Apriori που θα λαμβάνει υπόψη τα κελιά που επισκέφθηκε ο χρήστης στο πλέγμα και με ποια σειρά.

8 Βιβλιογραφία - Αναφορές

8.1 Βιβλιογραφία

[1] R. Agrawal, R. Srikant. “Fast Algorithms for Mining Association Rules in Large Databases”. *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, September 1994, Santiago de Chile, Chile.

[2] I.F. Akyildiz, J. McNair, J.S.M. Ho, H. Uzunalioglu and W. Wang, “Mobility Management in Next-Generation Wireless Systems”. *Proceedings of the IEEE*, Vol. 87, pp. 1347-1384, August 1999.

[3] M.S. Chen, J.S. Park and P.S. Yu. “Efficient Data Mining for Path Traversal Patterns”. *IEEE Transactions on Knowledge and Data Engineering*, Vol.10, No.2, pp.209-221, 1998.

[4] M. Garofalakis, R. Rastogi, S. Seshadri and K. Shim. “Data Mining and the Web: Past, Present and Future”. *Workshop on Web Information and Data Management* (WIDM'99), pp. 43--47, 1999.

[5] D. Katsaros and Y. Manolopoulos. “Cache Management for Web-Powered Databases”, κεφάλαιο από το βιβλίο “Web-Powered Databases”, (Taniar, D. and Rahayu, W.J. eds.), IDEA Group Publishing, pp. 201-242, 2002.

[6] D. Katsaros, A. Nanopoulos, M. Karakaya, G. Yavas, O. Ulusoy and Y. Manolopoulos. “Clustering mobile trajectories for resource allocation in mobile environments”

[7] A. Nanopoulos, D. Katsaros and Y. Manolopoulos. “A Data Mining Algorithm for Generalized Web Prefetching”, *IEEE Transactions on Knowledge and Data Engineering*.

[8] V. Padmanabhan and J. Mogul. “Using predictive prefetching to improve world wide web latency”. *ACM SIGCOMM Computer Communication Review*, July 1996.

[9] J. Han and M. Kamber. “Data Mining: Concepts and techniques”, Morgan Kaufman Publishers, 2000

[10] J. Pei, J. Han, B. Mortazavi-Asl and H. Zhu. “Mining Access Patterns Efficiently from Web Logs”. *PAKDD*, 2000, pp.396-407

[11] U. Fayyad, G. Piatesky-Shapiro, P. Smyth, R. Uthurusamy. “Advances in Knowledge Discovery and Data Mining”, MIT Press, 1996.

[12] Z. Chen, A. Wai-Chee Fu and F. Chi-Hung Tong. “Optimal Algorithms for Finding User Access Sessions from Very Large Web Logs”. *Pacific Asia*

Conference on Knolwedge Discovery and Data Mining 2002 (PAKDD 2002) pp. 290-296

[13] B. Lan, S. Bressan, B. Chin Ooi, K. Tan. “Rule-Assisted Prefetching in Web-Server Caching”. *Conference on Information and Knowledge Management (CIKM 2000)*, pp. 504-511

[14] SPSS Clementine, White paper - Technical Report. “Gaining a competitive edge with Web mining”

[15] T. Yan, M. Jacobsen, H. Garcia-Molina and U. Dayal. “From User Access Patterns to Dynamic Hypertext Linking”. *Computer Networks and ISDN Systems 28*, pp.1007-1014 (1996)

[16] Osmar R. Zaïane, University Notes

8.2 Web references

[1] http://www.acm.org/sigkdd
Ομάδα του ACM με ειδικό ενδιαφέρον στην Ανακάλυψη Γνώσης από Βάσεις Δεδομένων και Εξόρυξη δεδομένων (Knowledge Discovery in Data and Data Mining)

[2] http://www.kdnuggets.com/
(KD: Knowledge Discovery) Η βασικότερη πηγή πληροφόρησης στον παγκόσμιο ιστό σχετικά με Data Mining, Web Mining, Knowledge Discovery και θέματα Λήψης αποφάσεων, που περιλαμβάνει πληροφοριακό υλικό από τον χώρο της Ανακάλυψης Γνώσης σχετικά με ειδήσεις, λύσεις, λογισμικό, εργασία, μαθήματα, εκδόσεις, κ.ά.

[3] http://www.dmg.org/
Η ομάδα Εξόρυξης Δεδομένων (Data Mining Group – DMG) είναι μία ανεξάρτητη αρχή πωλητών που αναπτύσσουν τα πρότυπα για την εξόρυξη δεδομένων.

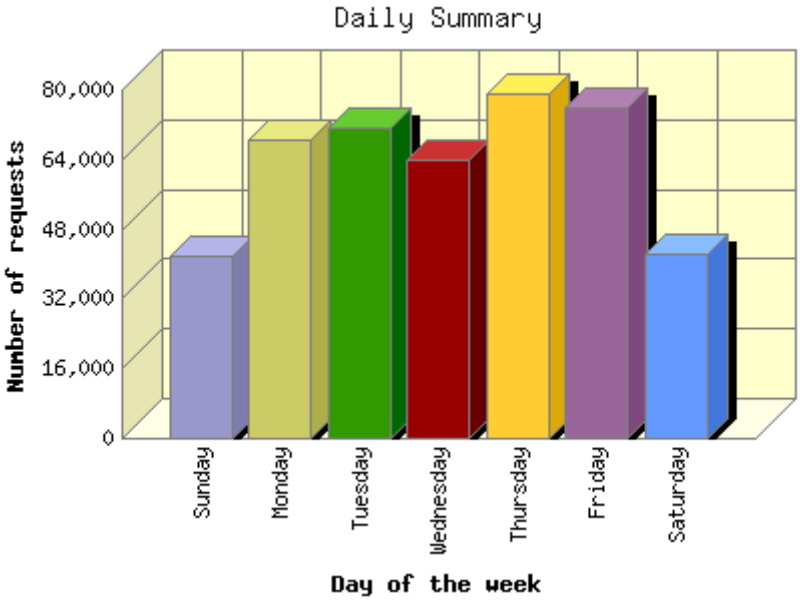
9 Παράρτημα Ι : Ανάλυση δεδομένων εισόδου (συνέχεια)

9.1 Αρχείο Καταγραφής 2: Εβδομάδα: 9η – 15η Αυγούστου 2003

Στον παρακάτω πίνακα παρουσιάζονται τα βασικότερα μεγέθη που αφορούν την ανάλυση του 2ου αρχείου επισκέψεων (log file mpa2, χρονικό διάστημα 9-15/8/2003). Παρουσιάζονται οι αναφορές παρόμοια με το αρχείο mpa1 της πρώτης εβδομάδας.

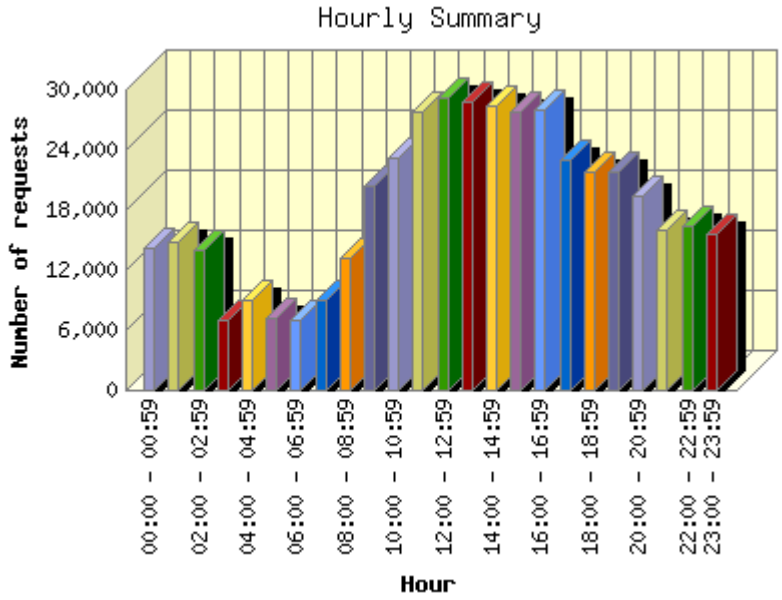
Γενική Αναφορά	
Χρόνος Πρώτης Επίσκεψης	Aug 8, 2002 03:00
Χρόνος Τελευταίας Επίσκεψης	Aug 15, 2002 02:59
Επιτυχημένες αιτήσεις	443086 Requests
Επιτυχημένες αιτήσεις σελίδων	153197 Requests for pages
Αποτυχημένες αιτήσεις	2729 Requests
Αναδρομολογημένες αιτήσεις	1564 Requests
Distinct files requested	10797 Files
Μοναδικοί επισκέπτες	39114 Hosts
Κατεστραμμένες γραμμές αρχείου log file	71 Lines
Μη χρήσιμες καταχωρήσεις αρχείου log file	674654 Lines
Σύνολο δεδομένων που μεταφέρθηκαν	3.193 GB

9.1.1 Αναφορά ημερήσιων επισκέψεων



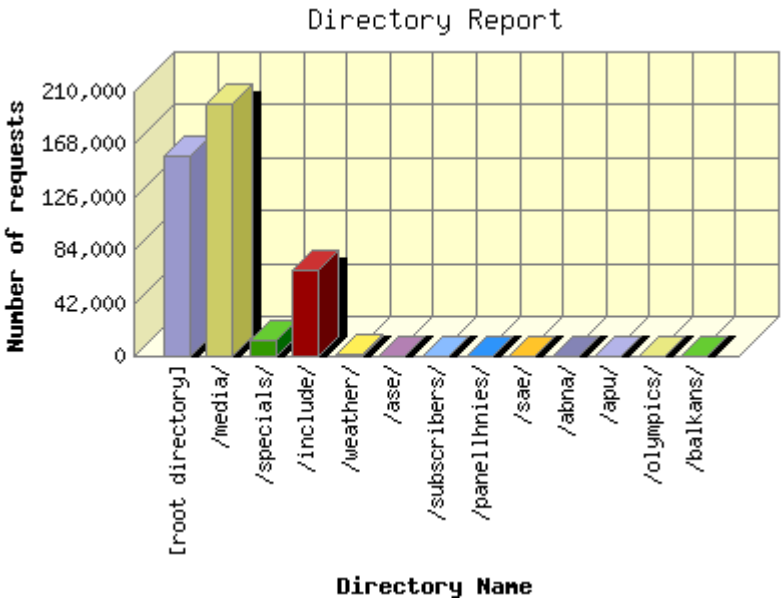
Ημέρα		Αρ. Αιτήσεων	Αρ. Αιτήσεων Σελίδων
1.	Κυριακή	41,888	18,026
2.	Δευτέρα	68,518	23,843
3.	Τρίτη	71,382	23,399
4.	Τετάρτη	63,879	20,099
5.	Πέμπτη	79,254	26,294
6.	Παρασκευή	75,945	24,779
7.	Σάββατο	42,220	16,757

9.1.2 Αναφορά ωριαίων επισκέψεων



9.1.3 Αναφορά λέξεων αναζήτησης

	Search Word	Αρ. Αιτήσεων
1.	http://users.hol.gr/~quasar/greek/astronomy/shootingstars.htm	72
2.	News	58
3.	Macedonian	48
4.	Agency	40
5.	Greek	38
6.	cache:iiyed_9oizmc:www.mpa.gr/	38
7.	Press	37
8.	mpa.gr	20
10.	Mpa	9



Directory Name		Αρ. Αιτήσεων	Percentage of the bytes
1.	[root directory]	159,275	64.78%
2.	/media/	200,620	27.75%
3.	/specials/	12,209	4.13%
4.	/include/	67,669	2.24%
5.	/weather/	1,148	0.35%
6.	/ase/	398	0.16%
7.	/subscribers/	102	0.14%
8.	/panellhnies/	170	0.13%
9.	/sae/	111	0.08%
10.	/abna/	512	0.08%
11.	/apu/	281	0.07%
12.	/olympics/	86	0.04%
13.	/balkans/	447	0.04%

9.1.4 Αναφορά σελίδων

reqs: %bytes: last time: file
-----: ------: ----------------: ----
81936: 43.58%: 15/Aug/02 02:59: /article.html
579: 0.29%: 12/Aug/02 19:29: /article.html?doc_id=286538
478: 0.29%: 15/Aug/02 02:02: /article.html?doc_id=286260

430: 0.25%: 13/Aug/02 03:28: /article.html?doc_id=286233
423: 0.27%: 15/Aug/02 01:58: /article.html?doc_id=286752
388: 0.44%: 15/Aug/02 01:57: /article.html?doc_id=286637
383: 0.19%: 15/Aug/02 01:58: /article.html?doc_id=286581
377: 0.20%: 15/Aug/02 01:58: /article.html?doc_id=286407
331: 0.14%: 15/Aug/02 01:57: /article.html?doc_id=286813
322: 0.16%: 14/Aug/02 00:25: /article.html?doc_id=286175
321: 0.20%: 15/Aug/02 02:02: /article.html?doc_id=286630
316: 0.15%: 15/Aug/02 02:22: /article.html?doc_id=286823

…

.

58926: 49.05%: 15/Aug/02 02:59: /
8045: 7.89%: 15/Aug/02 02:59: /?page=greece
4163: 4.83%: 15/Aug/02 02:59: /?page=balkans
3887: 2.93%: 15/Aug/02 02:58: /?page=english
3429: 2.60%: 15/Aug/02 02:59: /?page=economy
3282: 2.75%: 15/Aug/02 02:59: /?page=home
3161: 2.19%: 15/Aug/02 02:59: /?page=world
3002: 1.69%: 15/Aug/02 02:59: /?page=sports
2246: 1.30%: 15/Aug/02 02:59: /?page=culture
500: 0.36%: 15/Aug/02 02:59: /?page=russian
2447: 1.61%: 15/Aug/02 02:57: /search.html
319: 0.23%: 14/Aug/02 20:28: /search.html?lang=el
683: 1.12%: 14/Aug/02 19:44: /titles.html
630: 0.41%: 15/Aug/02 02:15: /weather/map.html
358: 0.23%: 15/Aug/02 02:28: /ase/
352: 0.10%: 14/Aug/02 14:28: /media/thumb_show.html
321: 0.11%: 15/Aug/02 01:57: /weather/
252: 0.01%: 13/Aug/02 10:10: /specials/elections2000/const.html
221: 0.14%: 15/Aug/02 02:59: /specials/patriarchate/

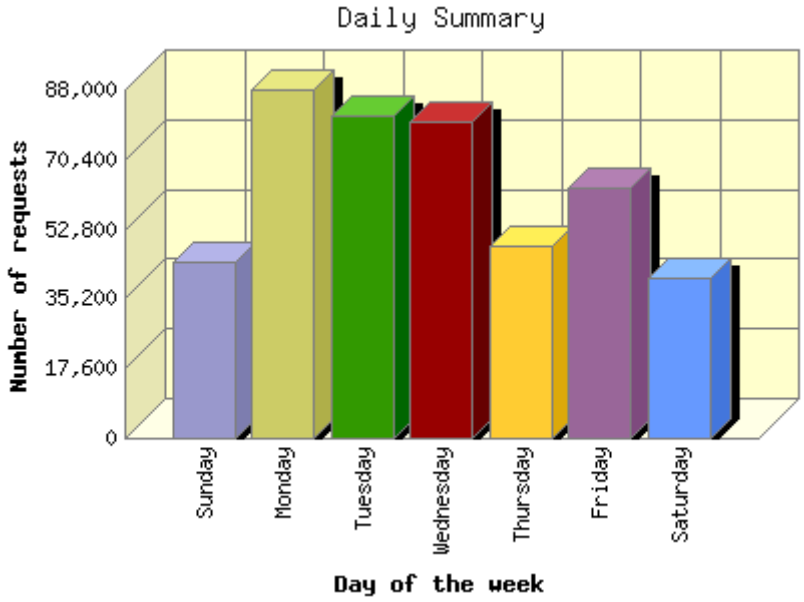
9.2 Αρχείο Καταγραφής 3: Εβδομάδα: 16η – 22η Αυγούστου 2003

Στον παρακάτω πίνακα παρουσιάζονται τα βασικότερα μεγέθη που αφορούν την ανάλυση του 2ου αρχείου επισκέψεων (log file mpa3, χρονικό διάστημα 16-22/8/2003). Παρουσιάζονται οι αναφορές παρόμοια με το αρχείο mpa1 της πρώτης εβδομάδας.

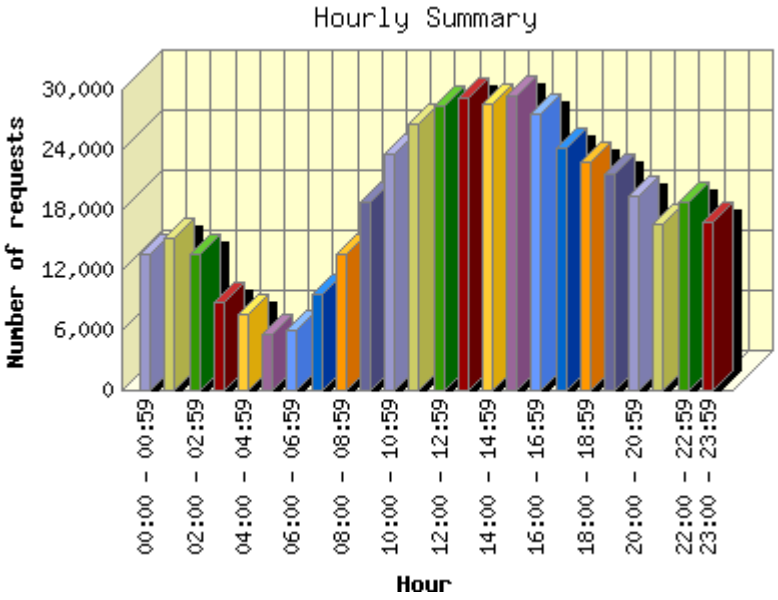
Γενική Αναφορά	
Χρόνος Πρώτης Επίσκεψης	Aug 15, 2002 03:00
Χρόνος Τελευταίας Επίσκεψης	Aug 22, 2002 02:59
Επιτυχημένες αιτήσεις	446253 Requests
Επιτυχημένες αιτήσεις σελίδων	149499 Requests for pages
Αποτυχημένες αιτήσεις	3154 Requests
Αναδρομολογημένες αιτήσεις	1631 Requests
Distinct files requested	6888 Files

Μοναδικοί επισκέπτες	40376 Hosts
Κατεστραμμένες γραμμές αρχείου log file	52 Lines
Μη χρήσιμες καταχωρήσεις αρχείου log file	643321 Lines
Σύνολο δεδομένων που μεταφέρθηκαν	3.329 GB

9.2.1 Αναφορά ημερήσιων επισκέψεων

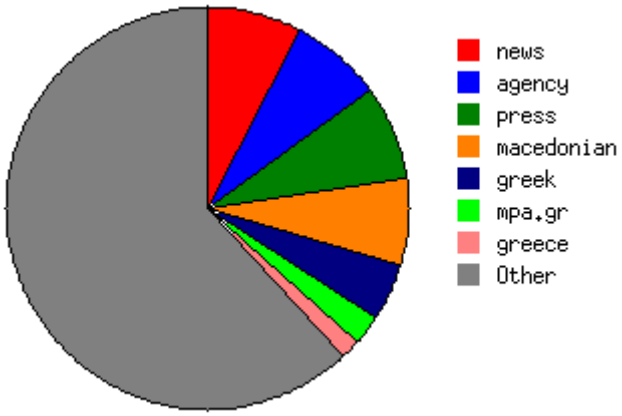


Ημέρα		Αρ. Αιτήσεων	Αρ. Αιτήσεων Σελίδων
1.	Κυριακή	44,374	15,835
2.	Δευτέρα	87,787	27,046
3.	Τρίτη	81,592	25,827
4.	Τετάρτη	80,081	26,072
5.	Πέμπτη	48,526	17,558
6.	Παρασκευή	63,466	22,284
7.	Σάββατο	40,427	14,877

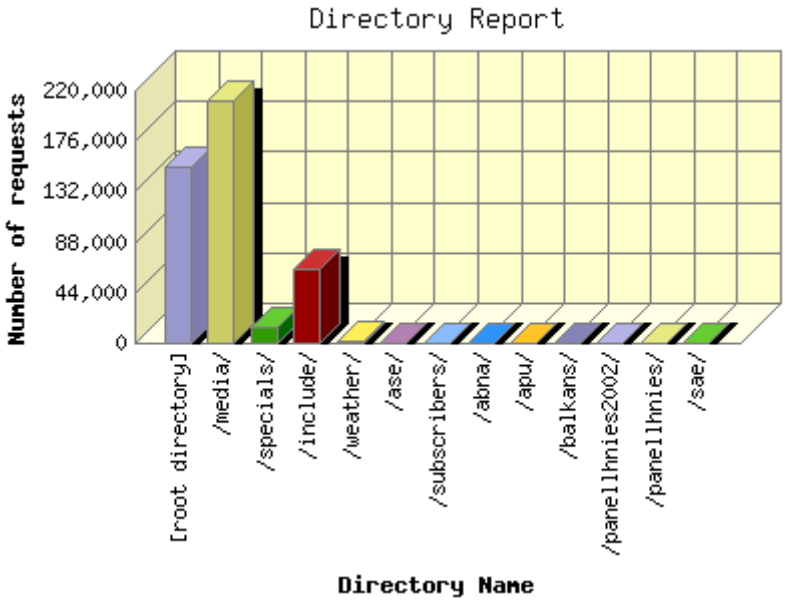


9.2.2 Αναφορά λέξεων αναζήτησης

Search Word		Αρ. Αιτήσεων
1.	News	50
2.	Agency	49
3.	Press	49
4.	macedonian	46
5.	Greek	30
6.	mpa.gr	15
7.	Greece	11
8.	macedonia	6
9.	Mpa	6
10.	www.mpa.gr	6



9.2.3 Αναφορά καταλόγου



Directory Name		Αρ. Αιτήσεων	Percentage of the bytes
1.	**[root directory]**	**153,309**	**58.77%**
2.	**/media/**	**210,856**	**34.17%**
3.	**/specials/**	**13,166**	**4.08%**
4.	**/include/**	**65,356**	**1.98%**
5.	**/weather/**	**1,129**	**0.33%**
6.	**/ase/**	**638**	**0.24%**
7.	**/subscribers/**	**168**	**0.17%**
8.	**/abna/**	**416**	**0.08%**
9.	**/apu/**	**211**	**0.05%**

10.	**/balkans/**	**689**	**0.04%**
11.	**/panellhnies2002/**	**62**	**0.03%**
12.	**/panellhnies/**	**43**	**0.02%**
13.	**/sae/**	**146**	**0.01%**
	[not listed: 8]	**61**	**0.01%**

9.2.4 Αναφορά σελίδων

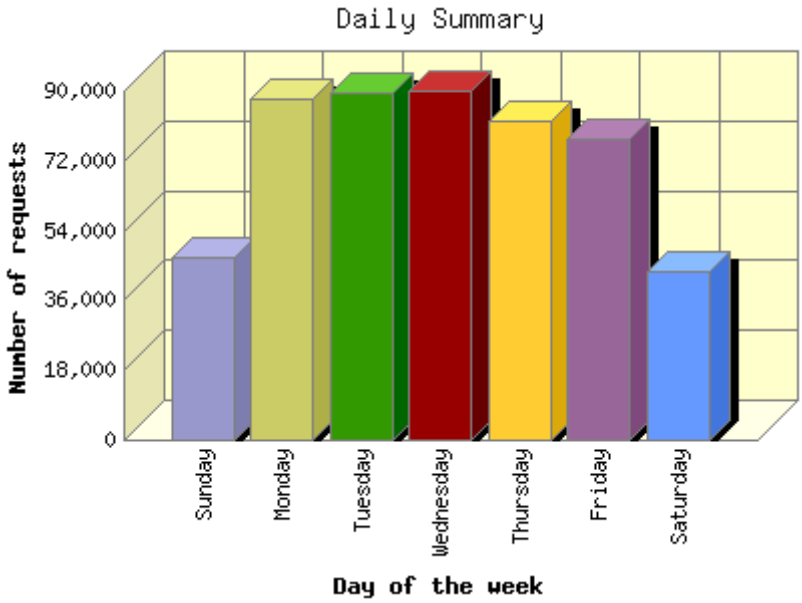
reqs: %bytes: last time: file
-----: ------: ---------------: ----
78477: 42.31%: 22/Aug/02 02:59: /article.html
1012: 0.41%: 22/Aug/02 02:03: /article.html?doc_id=286984
699: 0.39%: 22/Aug/02 01:58: /article.html?doc_id=287184
609: 0.34%: 21/Aug/02 17:03: /article.html?doc_id=287176
441: 0.17%: 22/Aug/02 01:57: /article.html?doc_id=287244
432: 0.23%: 22/Aug/02 02:52: /article.html?doc_id=287513
358: 0.18%: 22/Aug/02 01:58: /article.html?doc_id=287455
353: 0.17%: 22/Aug/02 01:58: /article.html?doc_id=287447
338: 0.17%: 22/Aug/02 01:58: /article.html?doc_id=287130
337: 0.17%: 22/Aug/02 02:38: /article.html?doc_id=287509
329: 0.19%: 21/Aug/02 04:25: /article.html?doc_id=286920
...
57607: 49.64%: 22/Aug/02 02:59: /
7934: 7.65%: 22/Aug/02 02:59: /?page=greece
4269: 5.06%: 22/Aug/02 02:59: /?page=balkans
3704: 2.82%: 22/Aug/02 02:58: /?page=english
3336: 2.64%: 22/Aug/02 02:59: /?page=economy
3188: 2.32%: 22/Aug/02 02:59: /?page=world
3124: 2.77%: 22/Aug/02 02:59: /?page=home
2884: 1.85%: 22/Aug/02 02:59: /?page=sports
2537: 1.46%: 22/Aug/02 02:59: /?page=culture
2628: 1.83%: 22/Aug/02 01:39: /search.html
586: 0.40%: 22/Aug/02 02:06: /weather/map.html
544: 0.37%: 22/Aug/02 02:14: /ase/
479: 0.71%: 22/Aug/02 00:10: /titles.html
387: 0.13%: 21/Aug/02 14:28: /media/thumb_show.html
355: 0.13%: 22/Aug/02 01:57: /weather/
294: 0.06%: 20/Aug/02 02:20: /specials/siatista/
206: 0.13%: 22/Aug/02 00:32: /specials/patriarchate/
168: 0.28%: 21/Aug/02 21:04: /subscribers/

9.3 Αρχείο Καταγραφής 4: Εβδομάδα: 23η – 29η Αυγούστου 2003

Στον παρακάτω πίνακα παρουσιάζονται τα βασικότερα μεγέθη που αφορούν την ανάλυση του 2ου αρχείου επισκέψεων (log file mpa4, χρονικό διάστημα 23-29/8/2003). Παρουσιάζονται οι αναφορές παρόμοια με το αρχείο mpa1 της πρώτης εβδομάδας.

Γενική Αναφορά	
Χρόνος Πρώτης Επίσκεψης	Aug 22, 2002 03:00
Χρόνος Τελευταίας Επίσκεψης	Aug 29, 2002 02:59
Επιτυχημένες αιτήσεις	517564 Requests
Επιτυχημένες αιτήσεις σελίδων	167952 Requests for pages
Αποτυχημένες αιτήσεις	2395 Requests
Αναδρομολογημένες αιτήσεις	1679 Requests
Distinct files requested	8361 Files
Μοναδικοί επισκέπτες	45689 Hosts
Κατεστραμμένες γραμμές αρχείου log file	53 Lines
Μη χρήσιμες καταχωρήσεις αρχείου log file	724022 Lines
Σύνολο δεδομένων που μεταφέρθηκαν	3.868 GB

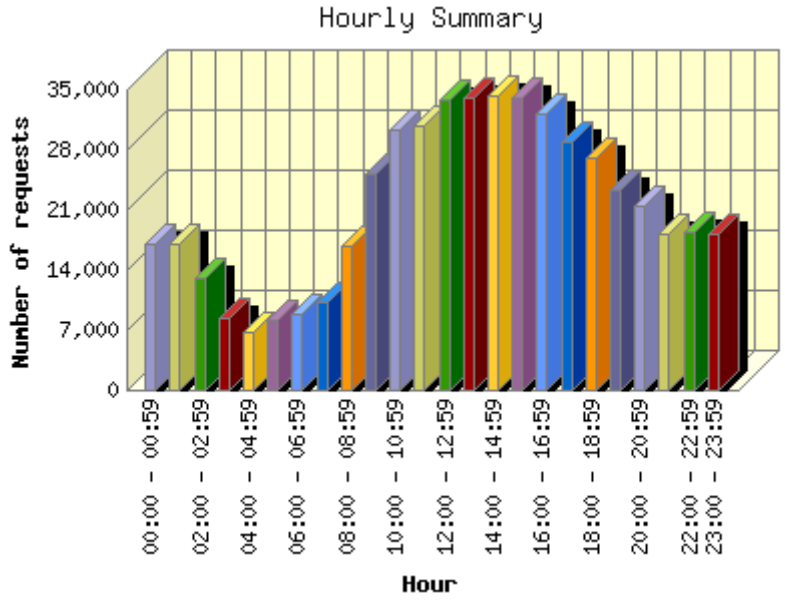
9.3.1 Αναφορά ημερήσιων επισκέψεων



Ημέρα		Αρ. Αιτήσεων	Αρ. Αιτήσεων Σελίδων
1.	Κυριακή	47,324	16,985

2.	Δευτέρα	88,009	26,495
3.	Τρίτη	89,337	27,643
4.	Τετάρτη	89,784	28,835
5.	Πέμπτη	82,107	27,621
6.	Παρασκευή	77,481	25,127
7.	Σάββατο	43,522	15,246

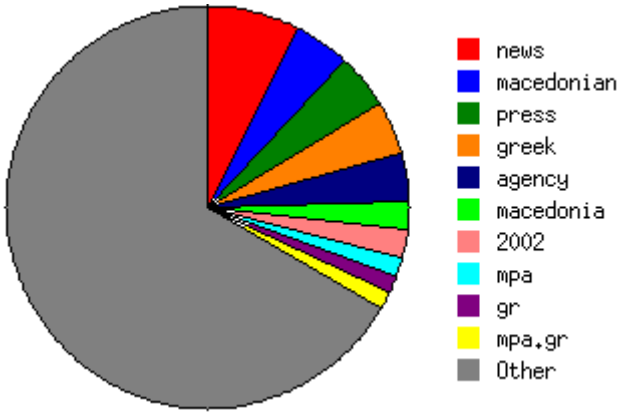
9.3.2 Αναφορά ωριαίων επισκέψεων

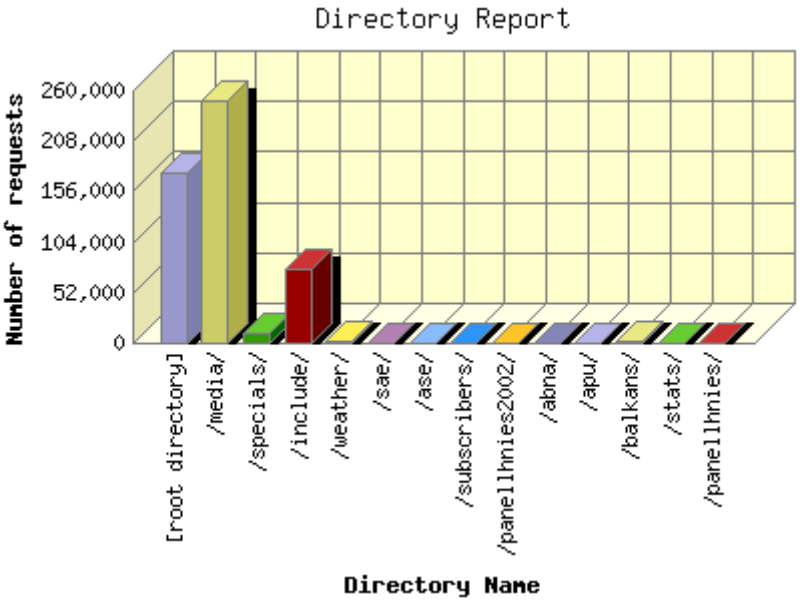


Search Word		Αρ. Αιτήσεων
1.	news	54
2.	macedonian	32
3.	press	32
4.	greek	30
5.	agency	28
6.	macedonia	16
7.	2002	16
8.	mpa	11
9.	gr	10

10.	mpa.gr	10

9.3.3 Αναφορά λέξεων αναζήτησης





9.3.4 Αναφορά σελίδων

reqs: %bytes: last time: file
-----: ------: ----------: ----
90479: 43.46%: 29/Aug/02 02:59: /article.html
426: 0.16%: 29/Aug/02 02:54: /article.html?doc_id=288638
419: 0.22%: 27/Aug/02 15:09: /article.html?doc_id=288191
391: 0.18%: 28/Aug/02 09:13: /article.html?doc_id=288284
365: 0.19%: 29/Aug/02 02:00: /article.html?doc_id=287890
364: 0.19%: 29/Aug/02 01:58: /article.html?doc_id=288084
357: 0.15%: 28/Aug/02 12:18: /article.html?doc_id=288493
341: 1.44%: 29/Aug/02 02:44: /article.html?doc_id=288701
338: 0.18%: 29/Aug/02 01:58: /article.html?doc_id=287761
330: 0.16%: 26/Aug/02 16:43: /article.html?doc_id=288052

325: 0.13%: 29/Aug/02 01:57: /article.html?doc_id=288498
315: 0.15%: 23/Aug/02 14:11: /article.html?doc_id=287712
…
64242: 48.83%: 29/Aug/02 02:59: /
8943: 8.03%: 29/Aug/02 02:58: /?page=greece
4635: 5.13%: 29/Aug/02 02:59: /?page=balkans
3758: 2.77%: 29/Aug/02 02:58: /?page=english
3612: 2.51%: 29/Aug/02 02:59: /?page=economy
3405: 2.20%: 29/Aug/02 02:59: /?page=world
3392: 2.54%: 29/Aug/02 02:58: /?page=home
3270: 1.79%: 29/Aug/02 02:59: /?page=sports
2770: 1.53%: 29/Aug/02 02:59: /?page=culture
3458: 2.03%: 29/Aug/02 02:57: /search.html
489: 0.33%: 28/Aug/02 23:38: /search.html?lang=el
110: 0.07%: 29/Aug/02 02:57: /search.html?lang=en
918: 1.40%: 28/Aug/02 23:34: /titles.html
519: 0.30%: 29/Aug/02 02:59: /ase/
503: 0.29%: 29/Aug/02 01:57: /weather/map.html
379: 0.10%: 28/Aug/02 19:03: /media/thumb_show.html
342: 0.11%: 29/Aug/02 02:59: /weather/
243: 0.13%: 29/Aug/02 01:19: /specials/patriarchate/
195: 0.32%: 29/Aug/02 00:58: /subscribers/

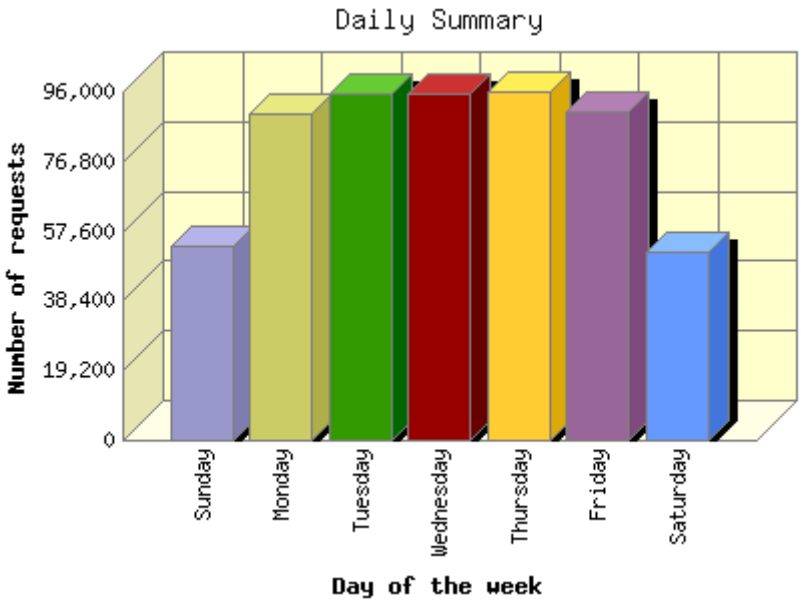
9.4 Αρχείο Καταγραφής 5: Εβδομάδα: 30η – 5η Αυγούστου 2003

Στον παρακάτω πίνακα παρουσιάζονται τα βασικότερα μεγέθη που αφορούν την ανάλυση του 2ου αρχείου επισκέψεων (log file mpa5, χρονικό διάστημα 30/08/2003 - 5/09/2003). Παρουσιάζονται οι αναφορές παρόμοια με το αρχείο mpa1 της πρώτης εβδομάδας.

Γενική Αναφορά	
Χρόνος Πρώτης Επίσκεψης	Aug 29, 2002 03:00
Χρόνος Τελευταίας Επίσκεψης	Sep 5, 2002 02:59
Επιτυχημένες αιτήσεις	571917 Requests
Επιτυχημένες αιτήσεις σελίδων	187499 Requests for pages
Αποτυχημένες αιτήσεις	4042 Requests
Αναδρομολογημένες αιτήσεις	3202 Requests
Distinct files requested	12779 Files
Μοναδικοί επισκέπτες	46939 Hosts
Κατεστραμμένες γραμμές αρχείου log file	90 Lines
Μη χρήσιμες καταχωρήσεις αρχείου log file	819597 Lines

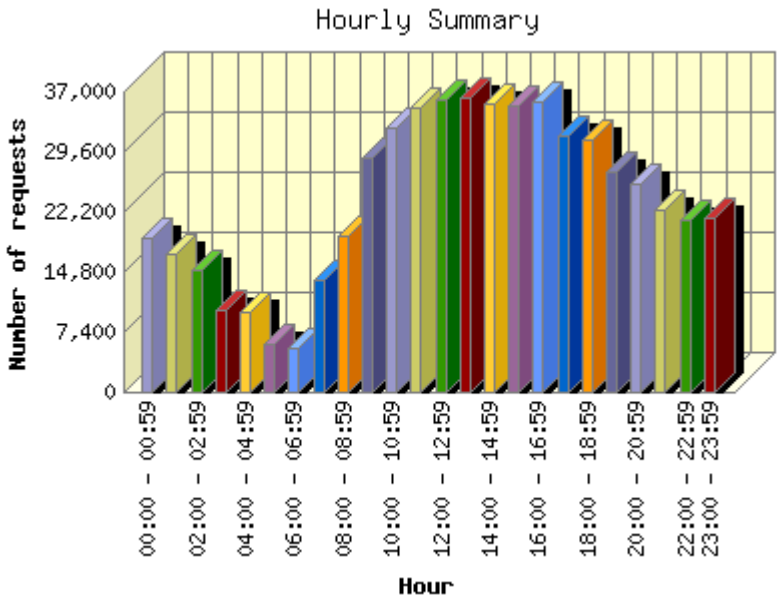
Σύνολο δεδομένων που μεταφέρθηκαν	4.296 GB

9.4.1 Αναφορά ημερήσιων επισκέψεων

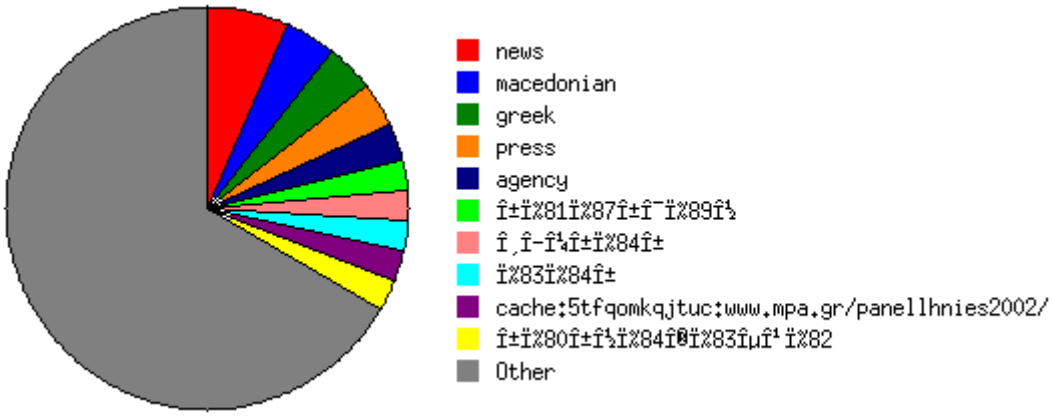


Ημέρα		Αρ. Αιτήσεων	Αρ. Αιτήσεων Σελίδων
1.	Κυριακή	53,333	19,155
2.	Δευτέρα	90,156	28,773
3.	Τρίτη	95,222	30,640
4.	Τετάρτη	95,290	29,456
5.	Πέμπτη	95,742	31,170
6.	Παρασκευή	90,287	29,306
7.	Σάββατο	51,887	18,999

9.4.2 Αναφορά ωριαίων επισκέψεων

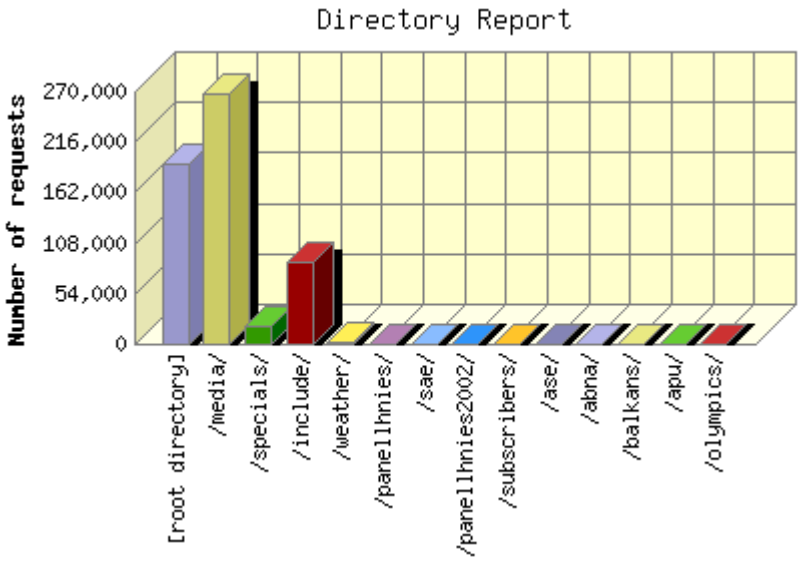


9.4.3 Αναφορά λέξεων αναζήτησης



	Search Word	Αρ. Αιτήσεων
1.	News	63
2.	macedonian	39
3.	Greek	36
4.	Press	34
5.	Agency	28

9.4.4 Αναφορά καταλόγων



Directory Name		Αρ. Αιτήσεων	Percentage of the bytes
1.	**[root directory]**	**193,915**	**60.45%**
2.	**/media/**	**267,731**	**32.76%**
3.	**/specials/**	**18,260**	**3.34%**
4.	**/include/**	**86,846**	**1.80%**
5.	**/weather/**	**1,569**	**0.37%**
6.	**/panellhnies/**	**244**	**0.27%**
7.	**/sae/**	**227**	**0.26%**
8.	**/panellhnies2002/**	**531**	**0.23%**
9.	**/subscribers/**	**176**	**0.20%**
10.	**/ase/**	**492**	**0.15%**
11.	**/abna/**	**569**	**0.07%**
12.	**/balkans/**	**999**	**0.04%**
13.	**/apu/**	**211**	**0.04%**
14.	**/olympics/**	**96**	**0.04%**
	[not listed: 7]	**23**	**0.00%**

9.4.5 Αναφορά σελίδων

reqs: %bytes: last time: file
------: ------: ----------: ----

104125: 44.77%: 5/Sep/02 02:59: /article.html
923: 0.36%: 5/Sep/02 01:21: /article.html?doc_id=288868
794: 0.37%: 4/Sep/02 16:11: /article.html?doc_id=289190
670: 0.28%: 5/Sep/02 00:56: /article.html?doc_id=289504
467: 0.16%: 5/Sep/02 01:57: /article.html?doc_id=288969
428: 0.15%: 5/Sep/02 02:26: /article.html?doc_id=289678
421: 0.16%: 3/Sep/02 15:07: /article.html?doc_id=289125
420: 0.16%: 30/Aug/02 17:02: /article.html?doc_id=288864
350: 0.14%: 3/Sep/02 15:45: /article.html?doc_id=288871
312: 0.12%: 5/Sep/02 02:36: /article.html?doc_id=289882
309: 0.15%: 4/Sep/02 01:54: /article.html?doc_id=289575
308: 0.15%: 5/Sep/02 01:56: /article.html?doc_id=289890
304: 0.16%: 4/Sep/02 04:32: /article.html?doc_id=289217
292: 0.14%: 5/Sep/02 01:36: /article.html?doc_id=289929
...
67950: 47.03%: 5/Sep/02 02:59: /
10033: 8.19%: 5/Sep/02 02:58: /?page=greece
4682: 4.75%: 5/Sep/02 02:58: /?page=balkans
3903: 2.70%: 5/Sep/02 02:58: /?page=english
3709: 2.36%: 5/Sep/02 02:58: /?page=economy
3413: 2.30%: 5/Sep/02 02:58: /?page=home
3323: 1.63%: 5/Sep/02 02:59: /?page=sports
3233: 1.97%: 5/Sep/02 02:58: /?page=world
2591: 1.26%: 5/Sep/02 02:59: /?page=culture
592: 0.44%: 5/Sep/02 02:57: /?page=russian
3845: 2.00%: 5/Sep/02 02:53: /search.html
989: 1.67%: 4/Sep/02 23:47: /titles.html
933: 0.48%: 5/Sep/02 02:55: /weather/map.html
438: 0.23%: 5/Sep/02 02:59: /ase/
438: 0.11%: 4/Sep/02 17:08: /media/thumb_show.html
369: 0.10%: 5/Sep/02 02:59: /weather/
196: 0.06%: 5/Sep/02 02:59: /weather/?lang=el
118: 0.03%: 5/Sep/02 01:55: /weather/?lang=en
316: 0.16%: 5/Sep/02 02:54: /specials/patriarchate/
289: 0.08%: 5/Sep/02 01:57: /specials/deth/

10 Παράρτημα ΙΙ : Περιεχόμενα συνοδευτικού CD

Το περιεχόμενο του συνοδευτικού CD είναι χωρισμένο σε φακέλους με το παρακάτω περιεχόμενο:

- Φάκελος **Source**: Περιέχει τον πηγαίο κώδικα που αναπτύχθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας και διακρίνεται σε τρεις μονάδες λογισμικού:
    1. **Logread**: Ανάγνωση και προ-επεξεργασία των log files.
    2. **Transactions Creator**: Δημιουργία των επισκέψεων των χρηστών και καθορισμός των εγγράφων που περιέχονται σε κάθε επίσκεψη.
    3. **Apriori**: Επεξεργασία των δεδομένων από το προηγούμενο στάδιο επεξεργασίας για το σκοπό της εξόρυξης δεδομένων.

- Φάκελος **Logs**: Περιέχει τα δεδομένα εισόδου που παραχωρήθηκαν από το Μακεδονικό Πρακτορείο Ειδήσεων (log files).

- Φάκελος **Statistics**: Περιέχει τα αρχεία που προέκυψαν από τη στατιστική επεξεργασία των log files που μας μελετήσαμε.

- Φάκελος **Μετρήσεις**: Περιέχει τα αρχεία τύπου MS-Excel που χρησιμοποιήσαμε για να καταγράψουμε τα αποτελέσματα της μεθόδου που αναπτύξαμε.

- Φάκελος **Document**: Περιέχει το παρόν κείμενο της Διπλωματικής εργασίας σε μορφή MS-Word document και Acrobat Reader pdf.

- Φάκελος **Presentations**: Περιέχει τα αρχεία τύπου MS-PowerPoint με τις παρουσιάσεις που έγιναν τον Οκτώβριο του 2002 και το Φεβρουάριο του 2003 σχετικά με την παρούσα διπλωματική εργασία.

- Φάκελος **Other software**: Περιέχει το Ελεύθερο Λογισμικό που χρησιμοποιήσαμε για την εξαγωγή των στατιστικών από τα δεδομένα μας.