ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
Διατμηματικό Μεταπτυχιακό Πρόγραμμα στα Πληροφοριακά Συστήματα (MIS)

Τεχνικές Εξόρυξης Δεδομένων για την βελτίωση της απόδοσης σε Κατανεμημένα Συστήματα

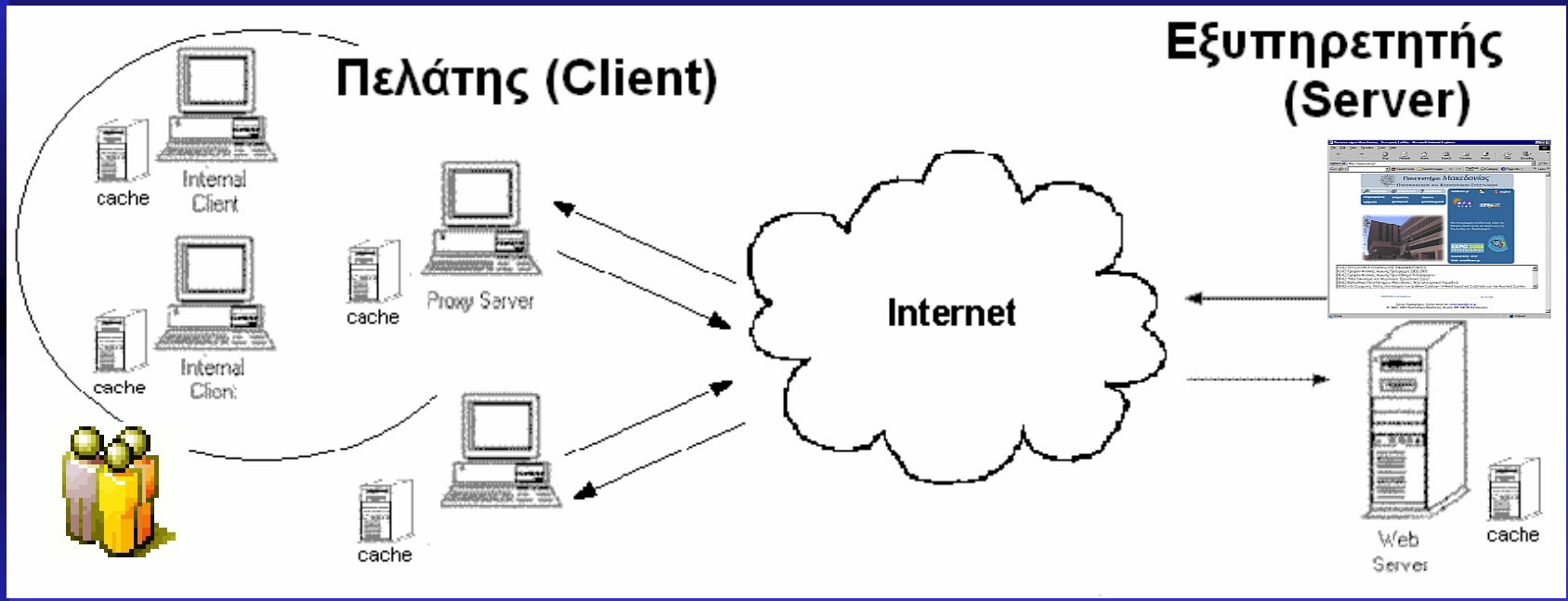
Ζάχος Δημήτριος

Επιβλέποντες: Κ. Μαργαρίτης
Ι. Μανωλόπουλος

Περιεχόμενα παρουσίασης

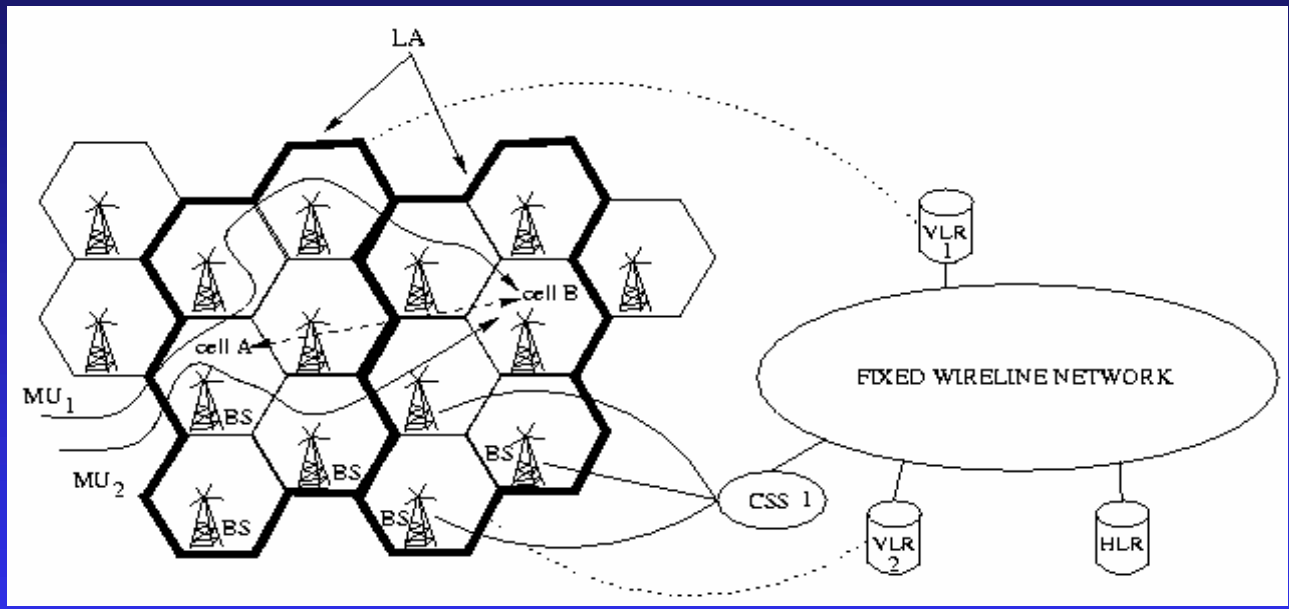
- Κεφάλαιο 1 : Ζητήματα Απόδοσης σε Κατανεμημένα συστήματα
- Κεφάλαιο 2 : Εξόρυξη Δεδομένων
- Κεφάλαιο 3 : Web Prefetching
- Κεφάλαιο 4 : Ανάλυση των Δεδομένων
- Κεφάλαιο 5 : Αξιολόγηση
- Κεφάλαιο 6 : Συμπεράσματα-προτάσεις

Παγκόσμιος Ιστός (World Wide Web)



- WWW = World Wide Wait ?

Κινητά Περιβάλλοντα Υπολογισμών (mobile environments)



- Κατανομή Bandwidth
- Εντοπισμός των χρηστών

Κίνητρο διπλωματικής εργασίας

- Μεγάλος όγκος δεδομένων
- Προβλήματα επίδοσης στα συστήματα

- Πώς μπορώ να αξιοποιήσω τα δεδομένα ?

  - Με τεχνικές εξόρυξης δεδομένων

Ερωτήματα προς διερεύνηση

- Εφαρμογή της διαδικασίας εξόρυξης σε πραγματικά Web δεδομένα
- Μπορεί η παραγόμενη γνώση να αξιοποιηθεί στο prefetching
  - Είναι συνοπτική ?
  - Παράγεται γρήγορα ?

Το πρόβλημα που αντιμετωπίσαμε (παγκόσμιος ιστός)

- Δεδομένα εισόδου: web logs από το Μακεδονικό Πρακτορείο Ειδήσεων

  http://www.mpa.gr

- Output : Εξαγωγή προτύπων για τη συμπεριφορά των επισκεπτών (association rules)

Κινητά περιβάλλοντα υπολογισμού

- Ποια από τις τεχνικές εξόρυξης δεδομένων ταιριάζει στο παράδειγμα ?

Πρόβλεψη κίνησης χρήστη (mobile environments)

- Είσοδος: Οι διαδρομές των χρηστών
- Διαδικασία:
  - Οργάνωση των διαδρομών με βάση την ομοιότητά τους και επιλογή προτύπων - αντιπροσώπων

Πρόβλεψη κίνησης χρήστη (mobile environments)

- Έξοδος: clusters όμοιων διαδρομών
  - Ταίριασμα της διαδρομής του χρήστη με ένα (ή περισσότερα) πρότυπα, ώστε να προβλέψουμε την επόμενη θέση του

Κεφάλαιο 2: Εξόρυξη Δεδομένων

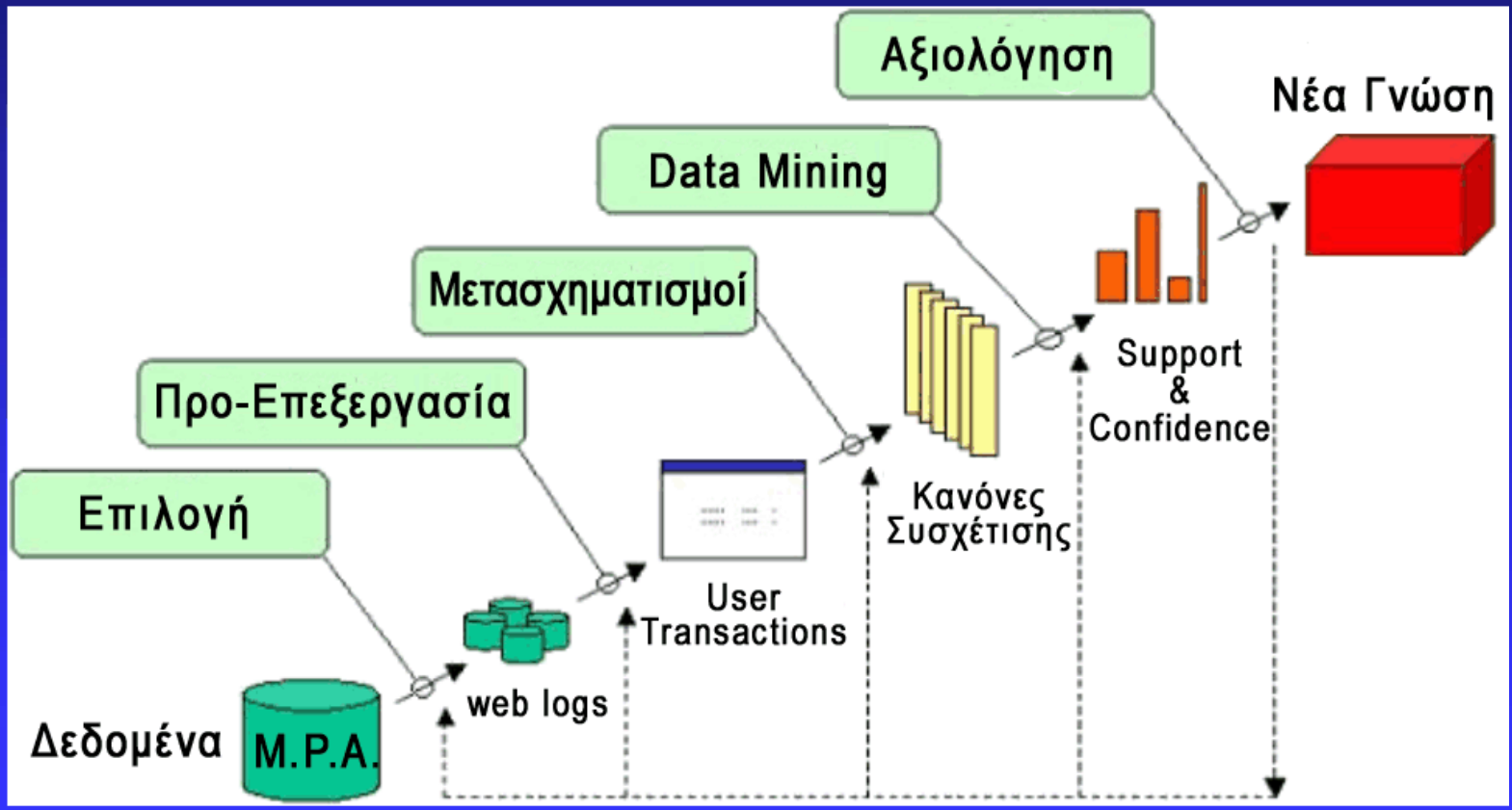
- Ανακάλυψη Γνώσης:

  εύρεση νέων, έγκυρων, χρήσιμων και κατανοητών προτύπων από τα δεδομένα

- Στηρίζεται στα γνωστικά πεδία :
  - Βάσεις δεδομένων
  - Τεχνητή νοημοσύνη
  - Στατιστική

Κεφάλαιο 2: Τεχνικές εξόρυξης

- Στάδια ανάπτυξης της διαδικασίας



Τεχνικές Data Mining επεξεργασίας

- Μελετήσαμε τη χρησιμότητα:
  - Clustering διαδρομών χρήστη (mobile environments)
  - Association rules που συνδέουν σελίδες (web)
  - Sequential Analysis (future work)

Ανακάλυψη Κανόνων Συσχέτισης

- Body → Head [Support, Confidence]

  Επισκέπτεται(X,specials.html) →
  Επισκέπτεται(X,financial.html) [5%,60%]

- Support : στατιστικό μέγεθος (ποσοστό που ισχύουν συγχρόνως A,B)
- Confidence : ποιοτικό (ποσοστό του A όταν ισχύει το B)

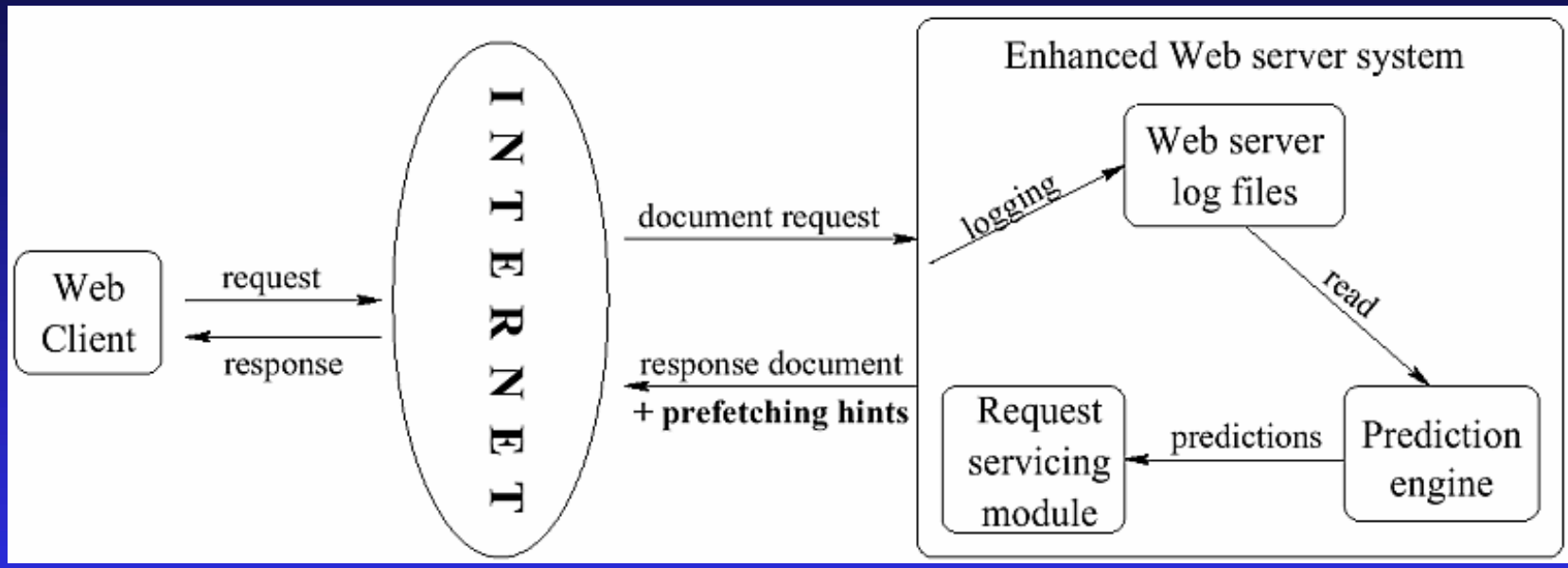
Φάσεις αλγορίθμου Apriori

- Υπολογισμός support για κάθε item
- Δημιουργία frequent itemsets από τα ισχυρά items (support > threshold)
- ...συνέχισε μέχρι να μην υπάρχουν υποψήφια υπερσύνολα

Κεφάλαιο 3: Web Prefetching

- Έννοια του prefetching
  - Συμπληρωματικό στο caching
  - Πρόβλεψη συμπεριφοράς
- Αξιοποιεί :
  - Το χρόνο μεταξύ διαδοχικών αιτήσεων
  - Την “τοπικότητα” των επισκέψεων (αιτήσεις που γίνονται μαζί)

Ο μηχανισμός του prefetching



- Χρησιμότητα (Usefulness)
- Ακρίβεια (Accuracy)
- Φόρτος δικτύου (Network Traffic)

Κεφάλαιο 4: Συγκέντρωση δεδομένων

- Αρχεία επισκέψεων
  http://www.mpa.gr

- Χρονική περίοδος : 01/08 ως 05/09/2003
- Μέγεθος log files : 1.2 Gb (5 εβδ. αρχεία)
- Αριθμός αιτήσεων (HTTP Requests) : ~ 500.000 / εβδομάδα
- Όγκος δεδομένων : 15,5 Gb

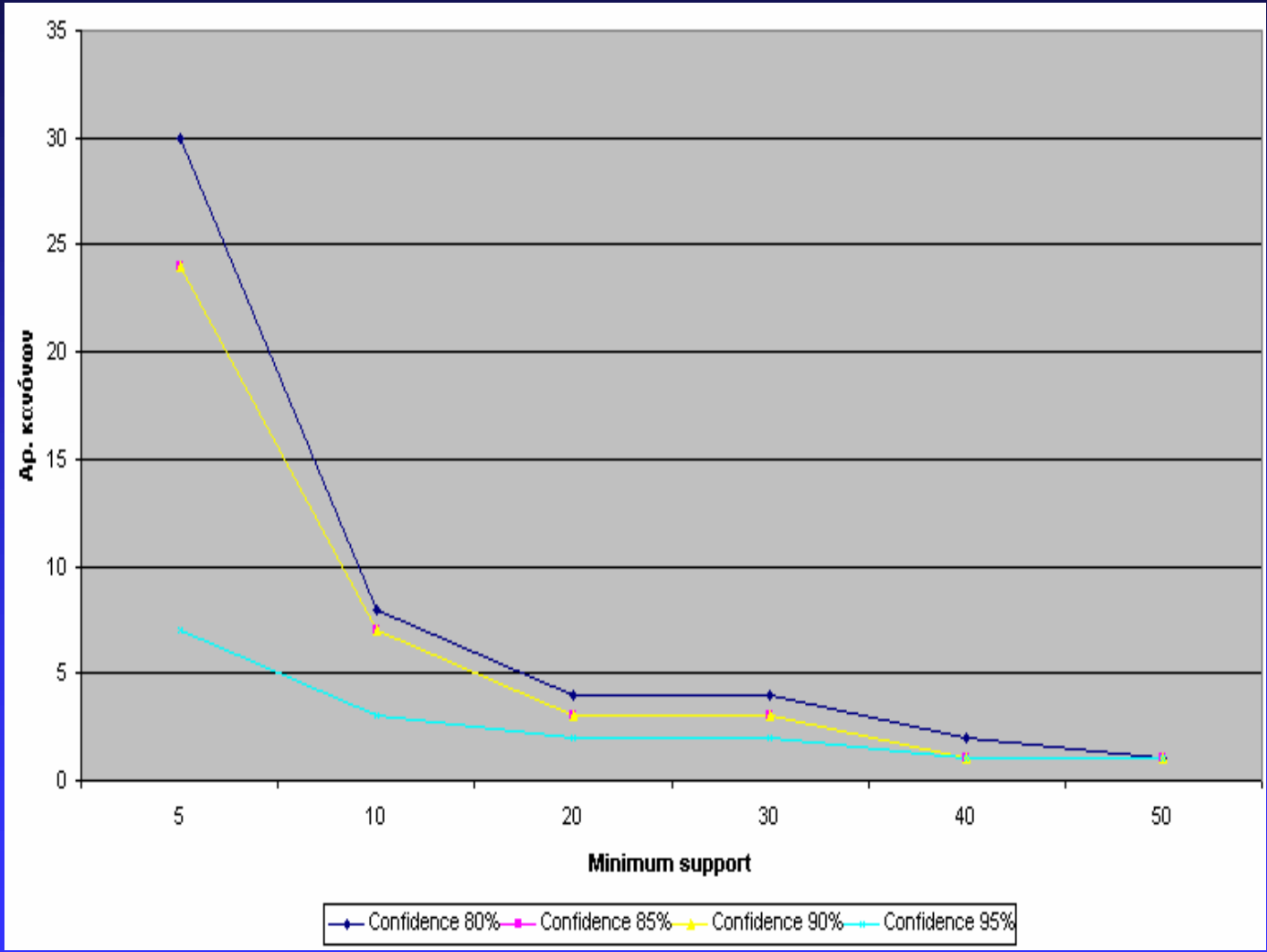
Ανάπτυξη Λογισμικού

- Προ-επεξεργασία log files (www.mpa.gr)
  - Αφαίρεση θορύβου (αιτήσεις για αρχεία εικόνων, cgi, not found pages)
  - Επιλογή δεδομένων (χρήσιμα πεδία του log file: IP,timestamp,URL)
  - Μετασχηματισμοί (mapping IPs', URLs')

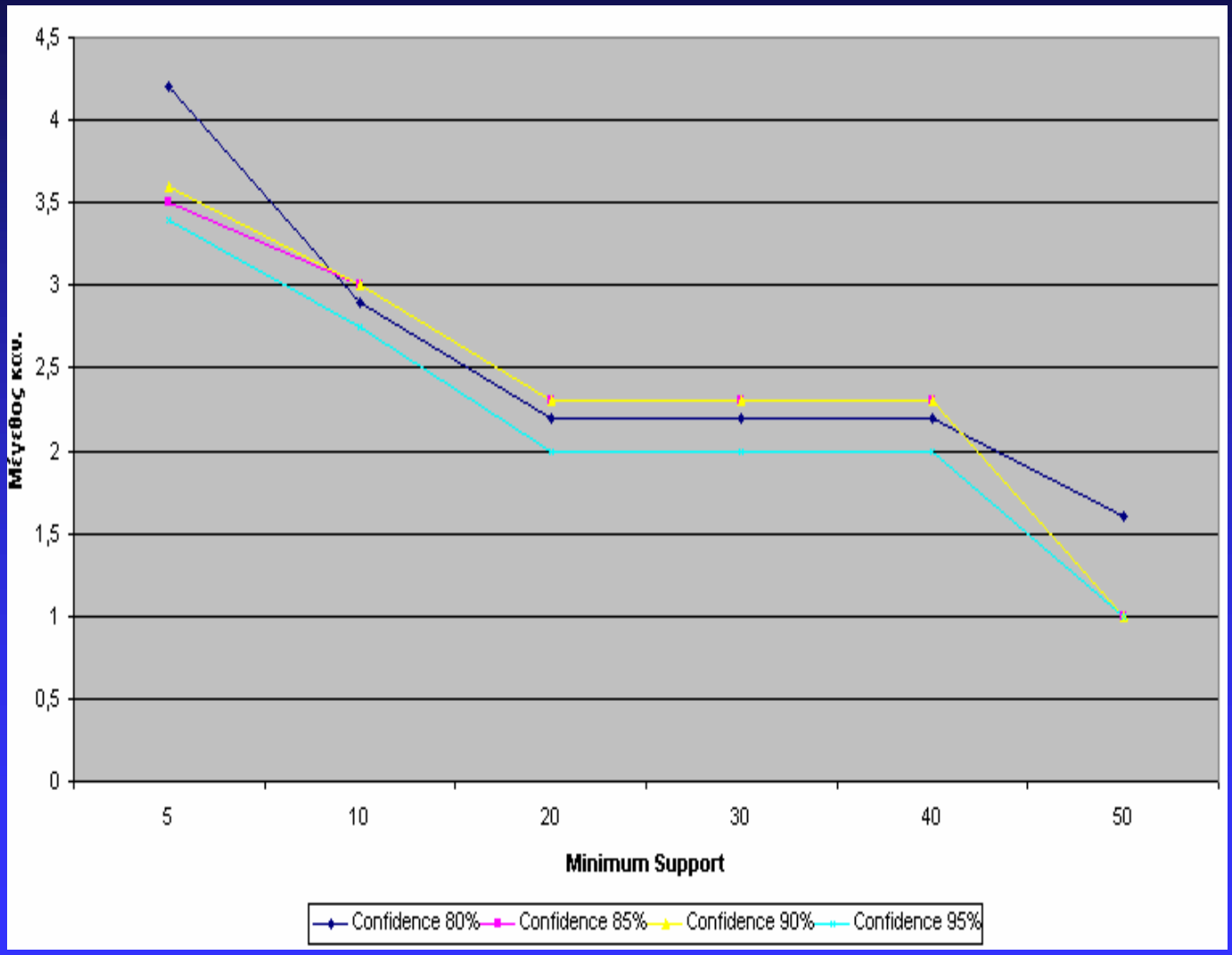
Ανάπτυξη Λογισμικού

- Προσδιορισμός των συνόδων (transactions) των επισκεπτών
  - Maximum 15 σελίδες (λόγω proxy servers)
  - Χρονικό εύρος: 2h
- Data Mining (Υλοποίηση Apriori)
  - Είσοδος : transactions επισκεπτών
  - Έξοδος : κανόνες συσχέτισης σελίδων

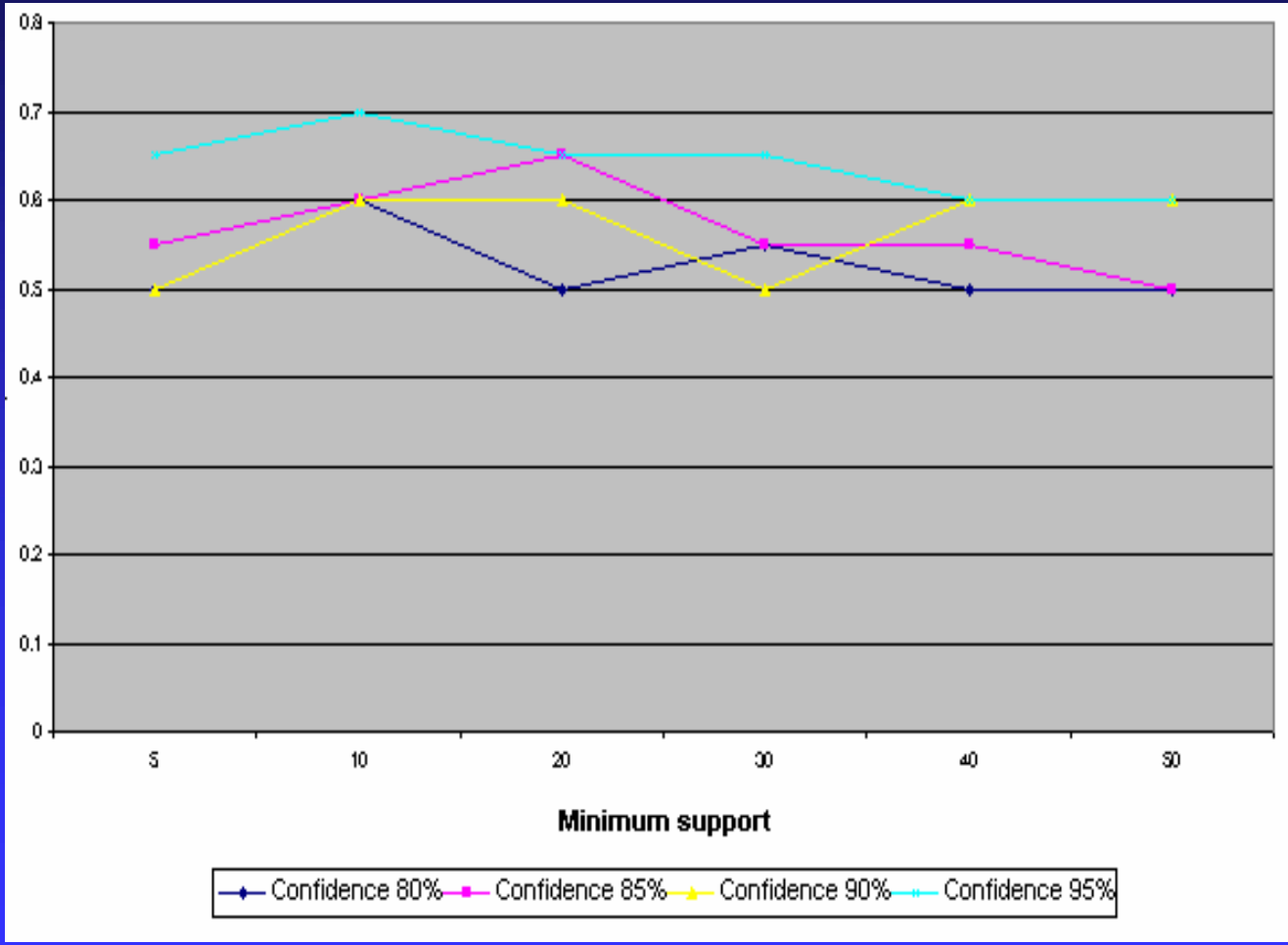
Αριθμός κανόνων –αρχείο mpa2



Μέγεθος κανόνων



Χρόνος επεξεργασίας



Κεφάλαιο 6: Συμπεράσματα

- Υπάρχει κανονικότητα στη συμπεριφορά των επισκεπτών
- Χαρακτηριστικά κανόνων συσχέτισης :
  - Μικρό μέγεθος
  - Λίγοι και κατανοητοί
  - Ανιχνεύονται γρήγορα
- Prefetching → Καλή Εφαρμογή

Γενικό συμπέρασμα

- Η Εξόρυξη Δεδομένων (κανόνες συσχέτισης) εφαρμόζεται με επιτυχία για τη βελτίωση της απόδοσης Κατανεμημένων Συστημάτων

Μελλοντική εργασία

- Web prefetching
  - Εξόρυξη κανόνων, ώστε να λαμβάνεται υπόψη ο αριθμός των clicks μεταξύ των στοιχείων που αποτελούν έναν κανόνα.
- Κινητά περιβάλλοντα
  - Αντί για ομαδοποίηση (clustering), να έχουμε ακολουθιακά πρότυπα (sequential patterns)

Ερωτήσεις ?