ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Πτυχιακή εργασία

A DHT-based approach for frequent pattern mining in web log data

Ευστάθιος Ζάραγκας
2022201600202

Επιβλέπουσα:

Παρασκευή Ραυτοπούλου
ΕΔΙΠ

Τρίπολη, Σεπτέμβριος 2021

Περίληψη

Στο σύγχρονο κόσμο των επιχειρήσεων και των πληροφοριακών συστημάτων, ο πλέον πολύτιμος πόρος για κάθε επιχείρηση, πέρα από το ανθρώπινο δυναμικό, είναι τα δεδομένα και η γνώση που μπορεί να εξαχθεί από αυτά. Η εποχή κατά την οποία κάθε επιχείρηση διέθετε χρόνο και χρήμα με σκοπό την βελτιστοποίηση των εσωτερικών της διαδικασιών έχει παρέλθει. Πρωταρχικός στόχος από τη σύλληψη ακόμη της ίδρυσης μιας εταιρείας είναι το κέρδος. Διάφορα ερωτήματα πυροδοτούνται όπως: πώς λοιπόν θα τροφοδοτείται ένα σύστημα με τις κατάλληλες πληροφορίες οι οποίες θα βεβαιώσουν την ομαλή και επιτυχημένη πορεία μιας επιχείρησης; Πώς θα συλλέγονται, αξιοποιούνται και μετατρέπονται σε χρήσιμες πληροφορίες τα δεδομένα που άπλετα υπάρχουν στον επιχειρηματικό κόσμο και πώς αυτά τα δεδομένα μπορούν να κάνουν μια επιχείρηση ανταγωνιστική; Εύκολα επομένως εξάγεται το συμπέρασμα ότι δύο σημαντικοί στόχοι για κάθε σύγχρονη επιχείρηση είναι η συλλογή δεδομένων και η εκμετάλλευση αυτών. Την απάντηση στα παραπάνω ερωτήματα καλείται λοιπόν να δώσει η επιστήμη της εξόρυξης δεδομένων μέσω των ισχυρότατων εργαλείων και μεθόδων που διαθέτει. Στα πλαίσια της παρούσας εργασίας, στήθηκε και χρησιμοποιήθηκε ο αλγόριθμος FP-Growth σε ένα Chord κατανεμημένο περιβάλλον προκειμένου να εξάγουμε συχνά πρότυπα πάνω από σύνολα δεδομένων. Οι peer-to-peer εφαρμογές έχουν γίνει πολύ δημοφηλείς κυρίως λόγω των εφαρμογών διαμοιρασμού αρχείων, αυτές οι εφαρμογές έχουν πολύ ενδιαφέροντα τεχνικά χαρακτηριστικά, όπως αποκεντρικοποιημένο έλεγχο, αυτοοργάνωση, προσαρμογή στο φυσικό δίκτυο και μεγάλες δυνατότητες κλημάκωσης. Τα peer-to-peer συστήματα μπορούν να χαρακτηριστούν ως κατανεμημένα συστήματα όπου όλοι οι κόμβοι έχουν τις ίδιες δυνατότητες και ευθύνες και όλη η επικοινωνία είναι συμμετρική. Πολλές από αυτές τις εφαρμογές όπως και στη δικη μας βασίζονται στα Distributed Hash Tables (DHTs). Το Chord είναι ένα γενικό peer-to-peer σύστημα για εύρεση τοποθεσίας αντικειμένων και δρομολόγηση, βασισμένο σε ένα αυτόοργανούμενο overlay δίκτυο κόμβων οι οποίοι είναι συνδεδεμένοι στο Internet. Το Chord είναι τελείως αποκεντρικοποιημένο, έχει ανοχή σε λάθη, είναι κλιμακουμενο και αξιόπιστο. Το Chord μπορεί να χρησιμοποιηθεί ως υπόστρωμα για την ανάπτυξη διαφόρων peer-to-peer εφαρμογών όπως διαμοιρασμό αρχείων, αποθήκευση αρχείων, συστημάτων επικοινωνίας και συστημάτων ονοματοδοσίας. Διάφορες εφαρμογές έχουν αναπτυχθεί ήδη πάνω από το Chord.

Abstract

In the modern world of business and IT systems, the most valuable resource for any business, apart from human resources, is data and knowledge that can be extracted from them. The era in which each company had time and money in order to optimize its internal processes has expired. The primary objective of capturing even the founding of a company is profit. How then is a system fed with appropriate information which will assure a smooth and successful a business? How this data can be collected, utilized and converted into useful information and how can this data make a business competitive? Easily therefore, it can be concluded that two of the most important objectives for every modern business are data collection and data exploitation. Towards this direction, data mining, via a wide range of available, with enormous possibilities, and continuously evolving methods, can be proven a valuable tool. For the work presented in this thesis, was applied FP-Growth in a Chord distributed environment to extract frequent pattern over datasets. Peer-to-peer Internet applications have been popularized mostly through file sharing applications, these applications have many interesting technical aspects like decentralized control, self-organization, adaptation and scalability. Peer-to-peer system can be characterized as distributed systems in which all nodes have identical capabilities and responsibilities and all communications are symmetric. Many of these application like ours are based on Distributed Hash Tables (DHTs). Chord is a generic peer-to-peer obejct location and routing scheme, based on self-organizing overlay network of nodes connected to the Internet. Chord is intended as general substrate for the construction of a variety of peer-to-peer Internet application like global file sharing, file storage, group communication and naming system. Several application have been built on top of Chord.

Ευχαριστίες

Μέσα από την εργασία αυτή θα ήθελα να ευχαριστήσω θερμά την καθηγήτρια κα Παρασκευή Ραυτοπούλου, για την εποπτεία και εμπιστοσύνη της για τον σχεδιασμό και υλοποίηση της πτυχιακής μου. Η συμβολή της ήταν αξιοθαύμαστη, καθώς βρισκόταν πάντα σε θέση να με καθοδήγησει και να μου υποδείξει πράγματα όλη αυτή την χρονική περίοδο ακόμα και όταν ο χρόνος της ήταν ελλιπής. Αξίζει να σημειωθεί το γεγονός ότι ήταν γεμάτη ενέργεια και διάθεση να με βοηθήσει και καθοδηγήση ακόμα και στις πιο δύσκολες στιγμές που αντιμετώπισα.

Επίσης, θα ήθελα να ευχαριστήσω την οικογενειά μου για όλη τους την υποστίρηξη και υπομονή καθ' όλη την διάρκεια των σπουδών μου και όχι μονο, που παρά τις όποιες δυσκολίες προέκυπταν ήταν δίπλα μου να τις αντιμετωπίσουμε.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η παρούσα διπλωματική εργασία πραγματεύεται την ανάπτυξη ενός αλγορίθμου ο οποίος θα παρέχει τη δυνατότητα εξαγωγής συχνών προτύπων από ηλεκτρονικές συναλλαγές ή από επισκέψεις χρηστών σε ιστοσελίδες. Το προτεινόμενο σύστημα δίνει έμφαση στην αποτελεσματικότερη (από άποψη χρόνου, κανόνων κ.α.) δημιουργία καλαθιού αγορών στηριζόμενο σε αναζήτηση προϊόντων με τη βοήθεια μεθόδων εξόρυξης δεδομένων. Σε αυτό το κεφάλαιο ορίζουμε τις συγκεκριμένες πτυχές του προβλήματος και συνοψίζουμε εν συντομία την προτεινόμενη λύση μας.

1.1 Περιγραφή του προβλήματος

Η εξόρυξη δεδομένων είναι μια διαδικασία για την απόκτηση δυνητικά χρήσιμων, προηγουμένως άγνωστων και τελικώς κατανοητών γνώσεων από τα δεδομένα. Η εξόρυξη κανόνων συσχέτισης είναι ένα από τα σημαντικά τμήματα της εξόρυξης δεδομένων και χρησιμοποιείται για να βρεθούν ενδιαφέρουσες συσχετίσεις μεταξύ συνόλων στοιχείων σε ένα μεγάλο πλήθος δεδομένων. Η ανακάλυψη συχνών συνόλων ή προτύπων στοιχείων αποτελεί βασική τεχνολογία και σημαντικό βήμα στις εφαρμογές της εξόρυξης κανόνων σύνδεσης.

Ο όγκος των δεδομένων που είναι διαθέσιμα ηλεκτρονικά μεγαλώνει με συνεχώς αυξανόμενο ρυθμό για την τελευταία δεκαετία. Στην επιχειρηματική κοινότητα, οι εταιρείες συλλέγουν κάθε είδους πληροφορίες σχετικά με την επιχειρηματική διαδικασία, όπως οικονομικά στοιχεία, μισθοδοσίες και δεδομένα πελατών, και τα δεδομένα

αυτά είναι συχνά μεταξύ των πιο πολύτιμων περιουσιακών στοιχείων μιας επιχείρησης. Στην επιστημονική κοινότητα, ένα μόνο πείραμα μπορεί να παράγει terabyte δεδομένων. Στη συνέχεια, υπάρχει αυξανόμενη ζήτηση για μεθόδους και εργαλεία που αποθηκεύουν και αναλύουν μεγάλους όγκους δεδομένων. Ωστόσο, ακόμη και η αποθήκευση, πόσο μάλλον η ανάλυση, τόσο τεράστιων ποσοτήτων δεδομένων παρουσιάζει πολλά νέα εμπόδια και προκλήσεις. Μια συχνά χρησιμοποιούμενη μεταφορά «we are drowning in data, and yet starving for knowledge»[1] συνοψίζει τέλεια την κατάσταση. Ο τομέας της εξόρυξης δεδομένων και γνώσης έχει προκύψει από αυτήν την ανάγκη να αναλύουμε τα διαθέσιμά δεδομένα και τελικά να ανακαλύπτουμε πολύτιμες πληροφορίες.

Η εξόρυξη δεδομένων θεωρείται ευρέως ως η διαδικασία «*ανακάλυψης συχνών προτύπων*» από μεγάλες ποσότητες δεδομένων. Ο ορισμός είναι τελείως ασαφής επειδή περιλαμβάνει ένα τεράστιο φάσμα μεθόδων, τεχνικών και αλγορίθμων από διάφορους τομείς όπως βάσεις δεδομένων, μηχανική μάθηση και στατιστική. Για να μπερδευτούν τα πράγματα ακόμη περισσότερο, η εξόρυξη δεδομένων συχνά θεωρείται ότι είναι μόνο ένα βήμα, ή καλύτερα το πιο σημαντικό, στη διαδικασία ανακάλυψης γνώσης. Η διαδικασία ανακάλυψης γνώσης περιλαμβάνει πολλά άλλα στάδια πριν από την εξόρυξη και μετά την εξόρυξη, όπως τον καθαρισμό δεδομένων και την οπτικοποίηση τους.

1.2 Η συνεισφορά μας στο πρόβλημα

Τα τελευταία χρόνια, ένα σημαντικός αριθμός μεθόδων και συστημάτων εξόρυξης δεδομένων και ανακάλυψης συχνών προτύπων έχουν προταθεί με σκοπό να ικανοποιήσουν τις ανάγκες των σύγχρονων εφαρμογών. Οι περισσότερες από αυτές τις δουλειές επικεντρώνονται στους αλγορίθμους εξόρυξης συχνών προτύπων, χωρίς να λαμβάνουν υπόψη τις διαφορετικές οπτικές της αρχιτεκτονικής του συστήματος ή και την αποδοτικότητα από άποψης χρόνου/φόρτου εργασίας. Στην παρούσα δουλεία, έχοντας κατά νου τις σύγχρονες εφαρμογές και τον τεράστιο όγκο δεδομένων που καλούνται αυτές (σε πραγματικό χρόνο) να διαχειριστούν, προτείνουμε μια κατανεμημένη αρχιτεκτονική για την εξόρυξη συχνών προτύπων. Η αρχιτεκτονική μας

βασίζεται στο Chord κατανεμημένο περιβάλλον [6]. Οπότε οι συναλλαγές χωρίζονται σε ομάδες και κάθε ομάδα συναλλαγών αποθηκεύεται σε διαφορετικό υπολογιστικό κόμβο. Στη συνέχεια, κάθε κόμβος του συστήματος, ανεξάρτητα από τους υπόλοιπους, χρησιμοποιεί τον αλγόριθμο FP-Growth[2] για την εξαγωγή συχνών προτύπων πάνω από το σύνολο των δεδομένων για τα οποία είναι υπεύθυνος. Για την επικοινωνία των κόμβων του συστήματος, χρησιμοποιούμε την DHT (Distributed Hash Tables) [7] λογική αναζήτησης και δρομολόγησης που μας παρέχει το δίκτυο Chord.

Πιο συγκεκριμένα, η συνεισφορά αυτής της εργασίας είναι η εξής:

1. Προτείνουμε μια κατανεμημένη αρχιτεκτονική που βασίζεται σε DHT για την εξόρυξη συχνών προτύπων ακολουθώντας τις σύγχρονες τάσεις χρήσης cluster ή grid αρχιτεκτονικών για την διαχείριση μεγάλου όγκου δεδομένων.
2. Κάθε κόμβος του συστήματος είναι υπεύθυνος για ένα υποσύνολο των δεδομένων και άρα μειώνεται ο υπολογιστικός φόρτος και χρόνος εξόρυξης συχνών προτύπων.
3. Τα δεδομένα χωρίζονται και κατανέμονται στους κόμβους ανά προϊόν, οπότε κάθε κόμβος μπορεί ανεξάρτητα και ταυτόχρονα με τους υπόλοιπους να εξάγει συχνά πρότυπα, μειώνοντας στο ελάχιστο την επικοινωνία μεταξύ των κόμβων και ενδεχόμενες καθυστερήσεις.
4. Το σύνολο των κανόνων εξόρυξης που προκύπτουν από το σύστημα περιέχει όλους τους κανόνες που θα μας έδινε και μια κεντρικοποιημένη αρχιτεκτονική, αλλά σε χρόνο μειωμένο μια τάξη μεγέθους.
5. Το σύστημα μας δουλεύει για όλους τους τύπους δεδομένων, ανεξαρτήτως χαρακτηριστικών ή εφαρμογής.

1.3 Οργάνωση κεφαλαίων

Το Κεφάλαιο 1 αποτελεί την εισαγωγή για το αντικείμενο και το περιεχόμενο του συγγράμματος. Στο Κεφάλαιο 2 πραγματοποιείται μια αναφορά στο θεωρητικό

υπόβαθρο των διαφόρων μεθοδολογιών που χρησιμοποιήθηκαν για την υλοποίηση του συστήματος. Στο Κεφάλαιο 3 γίνεται μία περιγραφή των διαφόρων προσεγγίσεων/αλγορίθμων που υλοποιήθηκαν στα πλαίσια της πτυχιακής εργασίας. Στο Κεφάλαιο 4 γίνεται ανάλυση της αρχιτεκτονική και των εργαλείων που εφαρμόστηκαν ενώ το Κεφάλαιο 5 παρουσιάζει τα σύνολα δεδομένων (dataset) που χρησιμοποιήθηκαν αλλά και τα αποτελέσματα που προέκυψαν, καθώς επίσης και μια σύγκριση των προσεγγίσεων.Τέλος, το Κεφάλαιο 6 περιέχει τα τελικά συμπεράσματα που προκύψαν μέσα από την υλοποίηση της πτυχιακής εργασίας.

Κεφάλαιο 2

Εξόρυξη και Τεχνικές Εύρεσης Κανόνων Συσχέτισης

Η εξόρυξη δεδομένων ασχολείται με την ανακάλυψη γνώσης και την εύρεση μοτίβων σε σύνολα δεδομένων μέσω μιας διαδικασίας μοντελοποίησης των δεδομένων. Η μοντελοποίηση, η βασικότερη διαδικασία στην εξόρυξης δεδομένων, είναι η τεχνική και οι αλγόριθμοι που εφαρμόζονται στα δεδομένα για να βρεθούν ομοιότητες, μοτίβα και να ομαδοποιηθούν τα δεδομένα. Η διαδικασία αυτή γίνεται με χρήση αλγορίθμων ταξινόμησης, κανόνων συσχέτισης και αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης των δεδομένων, με στόχο την εύρεση συχνών προτύπων. Η εξόρυξη δεδομένων έχει ένα ευρύ φάσμα εφαρμογών στην επιστήμη και τη μηχανική. Για παράδειγμα, στην πρόγνωση καιρού το μοντέλο ταξινόμησης μπορεί να χρησιμοποιηθεί για την πρόβλεψη του καιρού για την επόμενη μέρα με βάση τα προηγούμενα δεδομένα. Επίσης, η εύρεση συχνών συνόλων χρησιμοποιείται ευρέως σε καταστήματα λιανικής για την πρόβλεψη αγοραστικών συνηθειών των πελατών, καθώς και στη σχεδίαση καταλόγου των διαθέσιμων προϊόντων.

Η ανακάλυψη της γνώσης ως διαδικασία αποτελείται από μια επαναληπτική ακολουθία των ακόλουθων βημάτων:

1. Καθαρισμός δεδομένων / Data cleaning (για την εξάλειψη του θορύβου και των ασυνεπών δεδομένων).

2. Ενσωμάτωση δεδομένων / Data integration (όπου πολλαπλές πηγές δεδομένων

μπορούν να συνδυαστούν).

3. Επιλογή δεδομένων / Data selection (όπου δεδομένα που σχετίζονται με την εργασία ανάλυσης ανακτώνται από τη βάση δεδομένων).
4. Μετασχηματισμός δεδομένων / Data transformation (όπου τα δεδομένα μετατρέπονται ή ενοποιούνται σε μορφή κατάλληλη για εξόρυξη εκτελώντας διαδικασίες περίληψης ή συγκέντρωσης).
5. Εξόρυξη δεδομένων / Data mining (μια βασική διαδικασία όπου εφαρμόζονται έξυπνες μέθοδοι προκειμένου να εξαχθούν συχνά πρότυπα δεδομένων).
6. Αξιολόγηση μοτίβων / Pattern evaluation (για να προσδιοριστούν τα πραγματικά ενδιαφέροντα πρότυπα που αντιπροσωπεύουν τη γνώση με βάση ορισμένα μέτρα ενδιαφέροντος).
7. Παρουσίαση γνώσης / Knowledge presentation (όπου οι τεχνικές απεικόνισης και εκπροσώπησης της γνώσης χρησιμοποιούνται για να παρουσιάσουν την εξόρυξη γνώσης στον χρήστη).

Στην παρούσα εργασία, στο επίκεντρο είναι η ομαδοποίηση δεδομένων και η εύρεση συχνών προτύπων με κατανεμημένο τρόπο, δηλαδή των συνόλων δεδομένων που εμφανίζονται συχνά στην ίδια ομάδα. Παρακάτω παρουσιάζεται η βιβλιογραφία που σχετίζεται με κανόνες και αλγορίθμους εύρεσης προτύπων, όπως επίσης και ο αλγόριθμος Chord που χρησιμοποιήθηκε σε αυτή την εργασία για την κατανομή των δεδομένων και του φόρτου εύρεσης προτύπων σε ανεξάρτητους υπολογιστικούς κόμβους.

2.1 Κανόνες Συσχέτισης

Οι κανόνες συσχέτισης εκφράζουν το αποτέλεσμα της ανάλυσης χιλιάδων καλαθιών αγοράς πελατών. Πιο συγκεκριμένα, με την τεχνική ανάλυσης καλαθιού αγορών (market basket analysis), όπου βασίζεται στην εξέταση της πιθανότητας συσχέτισης μεταξύ ενός προϊόντος με κάποιο άλλο προϊόν ή ένα σύνολο προϊόντων,

το ζητούμενο είναι η ανακάλυψη συσχετίσεων ανάμεσα στα αντικείμενα μιας Βάσης Δεδομένων. Για παράδειγμα, ένας τέτοιος κανόνας είναι ο εξής: «Οι πελάτες που αγοράζουν γάλα, αγοράζουν παράλληλα και ψωμί με ποσοστό 60%». Ο παραπάνω κανόνας γράφεται σύντομα ως «γαλα $\rightarrow$ ψωμί». Η πρόταση αυτή παρουσιάζει ένα αίτιο, αγορά γάλακτος, και το συνδέει με ένα αποτέλεσμα, αγορά ψωμιού. Επίσης, παρέχει μια ένδειξη για το πόσο πιθανό είναι να συμβεί μια τέτοια σχέση αιτίας-αιτιατού μέσω του ποσοστού που δίνεται. Οι κανόνες συσχέτισης επομένως, όπως υποδηλώνει το όνομα τους, είναι κανόνες «if-then» που συσχετίζουν αντικείμενα μεταξύ τους. Το προϊόν ή το σύνολο προϊόντων που αγοράζει ένας πελάτης το ονομάζουμε *itemset*. Το κάθε itemset μπορεί να είναι ένα ή πολλά προϊόντα, που θα εξεταστεί η συσχέτιση μεταξύ τους.

Definition 2.1.1 (Itemset)**.** *Έστω $I = \{i_1, i_2, \ldots, i_m\}$ ένα σύνολο από διακριτά κατηγορήματα που αποκαλούμε items (αντικείμενα). Έστω επίσης, D ένα σύνολο από δοσοληψίες (transaction), όπου κάθε δοσοληψία T είναι ένα σύνολο από αντικείμενα, το οποίο καλειται itemset, και για το οποίο ισχύει $T \subseteq I$. Κάθε δοσοληψία T χαρακτηρίζεται από ένα μοναδικό αναγνωριστικό που καλείται TID.*

Μια δοσοληψία T περιέχει το X, ένα σύνολο από κάποια αντικείμενα του I, αν ισχύει $X \subseteq T$.

Definition 2.1.2 (Κανόνας Συσχέτισης)**.** *Ο κανόνας συσχέτισης είναι μία έκφραση της μορφής $X \rightarrow Y$, όπου X και Y είναι στοιχειοσύνολα $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \oslash$. Το αντικείμενο X το ονομάζουμε προηγηθέν (antecent item) και το αντικείμενο Y το ονομάζουμε συνεπακόλουθο (consequent).*

Μετά τη δημιουργία κανόνων γίνεται η εξέταση του βαθμού αληθείας που περιέχουν. Δηλαδή, μπορεί ένας κανόνας να είναι αληθής χωρίς όμως να είναι τα υποσύνολα απαραίτητα αληθή. Αυτή η σχέση μπορεί να παρασταθεί με τα X AND NOT Y αν για παράδειγμα ισχύει ο κανόνας X και όχι ο Y. Αναλυτικότερα για τις έννοιες support και confidence:

Definition 2.1.3 (Support)**.** *Ως support ενός κανόνα ονομάζουμε το ποσοστό αληθείας που εμπεριέχει. Για παράδειγμα, αν τα X και Y ισχύουν μαζί για τουλάχιστον S% των καλαθιών αγορών, τότε το support του κανόνα είναι το S*

Definition 2.1.4 (Confidence)**.** *Ως confidence ορίζουμε το ποσοστό συσχέτισης μεταξύ δύο προϊόντων ή υποσυνόλων. Δηλαδή, αν έχει ποσοστό εμφάνισης C% του προϊόντος Y σε καλάθια που περιέχουν το προϊόν X, το confidence του κανόνα αυτού είναι C.*

Το support είναι ένα σημαντικό μέγεθος μέτρησης, διότι ένας κανόνας που έχει πολύ μικρή τιμή του support απλά εμφανίζεται από τύχη. Αυτό σημαίνει ότι ένας κανόνας με χαμηλό support δεν έχει ενδιαφέρον γιατί δεν είναι αποδοτικός στην προώθηση προϊόντων, τα οποία οι πελάτες αγοράζουν μαζί. Για αυτό το λόγο το support χρησιμοποιείται για την απομάκρυνση των κανόνων που δεν έχουν κανένα ενδιαφέρον. Σε αντίθεση, το confidence είναι μάλλον ο σημαντικότερος δείκτης, καθώς μας πληροφορεί σε ποιο βαθμό μπορεί μια υπόθεση να είναι αληθής ώστε να έχει αξία για την λήψη κάποιας απόφασης. Σε έναν κανόνα $\mathrm{X} \rightarrow \mathrm{Y}$ όσο πιο ψηλό είναι το confidence τόσο πιο πιθανό είναι να υπάρχει σε μία συναλλαγή το Y όταν περιέχεται το X. Για παράδειγμα, αν εμφανιστεί κάποια συσχέτιση μεταξύ δύο προϊόντων ή δύο ειδών προϊόντων η επιχείρηση μπορεί να αλλάξει τη διαμόρφωση των χώρων πωλήσεων ώστε να ωθήσει τον καταναλωτή στην αγορά και των δύο προϊόντων.

2.2 Επισκόπηση αλγορίθμων εξόρυξης για εύρεση κανόνων συσχέτισης

Ένας αλγόριθμος εξόρυξης δεδομένων είναι μια καλά καθορισμένη διαδικασία που λαμβάνει δεδομένα ως είσοδο και παράγει έξοδο με τη μορφή μοντέλων ή μοτίβων. Για να δημιουργηθεί ένα μοντέλο, ο αλγόριθμος αναλύει πρώτα τα δεδομένα που του παρέχονται αναζητώντας συγκεκριμένους τύπους προτύπων ή τάσεων. Στη συνέχεια, υποβάλει τα αποτελέσματα της ανάλυσης σε πολλές επαναλήψεις για να βρεθούν οι βέλτιστες τιμές παραμέτρων για τη δημιουργία του μοντέλου εξόρυξης. Αυτές οι παράμετροι εφαρμόζονται στη συνέχεια σε ολόκληρο το σύνολο δεδομένων για την εξαγωγή μοτίβων. Το μοντέλο εξόρυξης που δημιουργεί ένας αλγόριθμος από τα δεδομένα σας μπορεί να έχει διάφορες μορφές, όπως:

1. Ένα δέντρο αποφάσεων που προβλέπει ένα αποτέλεσμα και περιγράφει πώς διαφορετικά κριτήρια επηρεάζουν αυτό το αποτέλεσμα.
2. Ένα μαθηματικό μοντέλο που προβλέπει τις πωλήσεις.
3. Ένα σύνολο κανόνων που περιγράφουν τον τρόπο ομαδοποίησης των προϊόντων σε μια συναλλαγή και τις πιθανότητες αγοράς προϊόντων μαζί.

2.2.1 Αλγόριθμος Apriori

Ο αλγόριθμος Apriori [8] έχει προταθεί από τους R.Agrawal και R.Srikant το 1994, και χρησιμοποιείται για την εξόρυξη συχνών συνόλων αντικειμένων (itemset). Από τα βασικά χαρακτηριστικά του αλγορίθμου είναι πως χρησιμοποιεί την ιδιότητα συχνών στοιχειοσυνόλων που περιλαμβάνει το ότι «εάν ένα σύνολο αντικειμένων είναι συχνό τότε όλα τα υποσύνολα του είναι επίσης συχνά». Τα συχνά στοιχειοσύνολα μπορούμε να τα αποκαλέσουμε και κλειστά προς τα κάτω διότι αν κάποιο στοιχειοσύνολο ικανοποιεί τις απαιτήσεις της ελάχιστης υποστήριξης (support) , το ίδιο συμβαίνει και για όλα τα υποσύνολα του. Με βάση την ιδιότητα της αντιθετοαντίστροφής ισχύει και το εξίσου σημαντικό «εάν ένα σύνολο αντικειμένων δεν είναι συχνό τότε και όλα τα υπερσύνολα του είναι επίσης μη συχνά». Ο Apriori υιοθετεί την τεχνική αναζήτηση level-wise, η οποία είναι μια επαναλαμβανόμενη τεχνική που χρησιμοποιεί τα k-itemsets για να κτίσει τα (k+1)-itemsets. Στην αρχή, ο αλγόριθμος βρίσκει τα συχνά εμφανιζόμενα 1-itemset (το σύνολο αντικειμένων με 1 μόνο χαρακτηριστικό). Στη συνέχεια, αναζητά και συναθροίζει τον αριθμό που εμφανίζεται κάθε αντικείμενο - χαρακτηριστικό στη βάση δεδομένων, έπειτα συλλέγει τα αντικείμενα που ικανοποιούντο ελάχιστο support στο σύνολο L1. Κατόπιν, χρησιμοποιώντας το σύνολο L1, χτίζεται το σύνολο L2 το οποίο περιλαμβάνει όλα τα συχνά σύνολα με 2 χαρακτηριστικά (2-itemset), το οποίο και αυτό χρησιμοποιείται για να χτιστεί τα L3, και ούτως κάθε εξής, μέχρι που να μην μπορεί να βρεθεί άλλο σύνολο με k-itemsets, δηλαδή το Lk να είναι κενό. Για τη δημιουργία κάθε επιπέδου με τα συχνά σύνολα αντικειμένων, χρησιμοποιείται η ιδιότητα Apriori Property η οποία μειώνει τον χώρο αναζήτησης και έτσι βελτιώνεται σημαντικά η αποδοτικότητα του αλγορίθμου.

Definition 2.2.1 (Apriori property)**.** *Η ιδιότητα Apriori αναφέρει ότι όλα τα μη κενά υποσύνολα των συχνών συνόλων αντικειμένων πρέπει να είναι επίσης συχνά. Πιο συγκεκριμένα, βασίζεται στο ότι εάν ένα σύνολο αντικειμένων I δεν ικανοποιεί το ελάχιστο όριο support (min_sup), τότε το I δεν είναι συχνό, και ισχύει ότι P(I) < min_sup.*

Παρακάτω παραθέτεται ο ψευδοκώδικας του αλγορίθμου Apriori.

Input:
D: Βάση δεδομένων με δοσοληψίες·
Min-sup: Ο ελάχιστος αριθμός support
Output: Σύνολο με όλα τα συχνά σύνολα αντικειμένων που ανήκουν στο D
Μέθοδος:

1: L_1= find_frequent_1-itemsets(DB);
2: **for** (k=2; $L_{k-1} = \varnothing; k++$) {
3: C_k= Apriori_gen(L_{k-1});
4: **for each** transaction $t \in DB$ { //scan DB for counts
5: C_t = subset(C_k, t); //get the subsets of t that are candidates
6: **for each** candidate $c \in C_t$
7: $c.count++$;
8: }
9: $L_k = \{c \in C_k | c.count \geq min\_sup\}$
10: }
11: return $L = \bigcup_k L_k$;
12: Procedure Apriori gen(L_{k-1}: frequent$(k-1)$-itemsets)

2.2.2 Αλγόριθμος FP-Growth

Ο αλγόριθμος Apriori [8] επιτυγχάνει πολύ καλή απόδοση μειώνοντας μεγάλο αριθμό υποψήφιων προτύπων. Παρόλα αυτά, σε περιπτώσεις μεγάλου αριθμού συχνών προτύπων ή προτύπων τα οποία έχουν μεγάλο μήκος δημιουργείται ένα βασικό πρόβλημα χρόνου, καθώς είναι πολύ κοστοβόρο να διαχειριστεί τον πολύ μεγάλο αριθμό από συχνά πρότυπα. Έτσι αυτό το πρόβλημα πυροδότησε την αρχή για περαιτέρω μελέτες με αντικειμενικό σκοπό να αποφευχθεί η παραγωγή και ο έλεγχος μεγάλου αριθμού υποψήφιων προτύπων. Ως αποτέλεσμα αναπτύχθηκαν τρεις μέθοδοι:

1. Το πρώτο βήμα ήταν η κατασκευή μιας συμπαγούς δομής δεδομένων (compact data structure) που ονομάζεται frequent pattern (FP-tree)[2]. Πρόκειται για ένα δέντρο που περιέχει πληροφορίες για τα συχνά πρότυπα με προθεματικό τρόπο. Ειδικότερα, κόμβους στο δέντρο έχουν μόνο τα συχνά πρότυπα μεγέθους 1, ενώ οι κόμβοι είναι διατεταγμένοι με τέτοιο τρόπο ώστε οι πιο συχνοί να είναι πιο εύκολα προσβάσιμοι. Αυτή η δομή χρησιμοποιείται αντί της βάσης δεδομένων, γιατί αφενός είναι μικρότερης έκτασης αφετέρου περιέχει πολύ πιο ποιοτική πληροφορία.

2. Ως δεύτερο βήμα, δημιουργήθηκε μια μέθοδος σταδιακής εύρεσης των συχνών προτύπων (pattern-fragment growth mining), βασιζόμενη στο FP-tree και ξεκινώντας από τα συχνά πρότυπα μεγέθους 1. Τα πρότυπα αυτά θα αποτελέσουν την κατάληξη επόμενων συχνών προτύπων, ξεχωρίζοντας τα στοιχεία από την βάση δεδομένων με τις συναλλαγές που καταλήγουν στα συχνά πρότυπα μεγέθους 1. Αυτή η διαδικασία έχει ως αποτέλεσμα να προκύψει μια βάση δεδομένων, υποσύνολο της αρχικής, και για την οποία κατασκευάζουμε το αντίστοιχο FP-tree (conditional)[2]. Η διαδικασία αυτή γίνεται επαναληπτικά τοποθετώντας τα αντικείμενα ακριβώς πριν την κατάληξη. Η ύπαρξη των συχνών προτύπων μέσα στις συναλλαγές σημαίνει αυτόματα ότι θα απεικονίζονται οπωσδήποτε και στο FP-tree μέσα από κάποιο διακριτό μονοπάτι, το οποίο σημαίνει ότι η εύρεση των συχνών προτύπων γίνεται με ολοκληρωμένο τρόπο.

3. Ο παραπάνω μηχανισμός κινείται πάνω στην λογική του «διαίρει και βασίλευε», σε αντίθεση με τον Apriori που σε κάθε βήμα διογκώνονταν το μέγεθος και η πολυπλοκότητα του προβλήματος. Εδώ κάθε φορά το conditional FP-tree είναι αρκετά μικρότερο από το προηγούμενο του και το μόνο που γίνεται στο τέλος είναι η επικόλληση της κατάληξης. Έτσι η ανίχνευση συχνών προτύπων ολοένα και μεγαλύτερου μεγέθους οδηγεί σε δομές δεδομένων ολοένα και πιο μικρές.

Οι προσεγγίσεις στη συχνή εξόρυξη αντικειμένων χωρίζονται σε δύο κατευθύνσεις. Αρχικά, στις δουλείες που χρησιμοποιούν την μέθοδο παραγωγής και κλαδέματος συχνών και μη-συχνών αντικειμένων, όπως είναι ο αλγόριθμος Apriori[8], οι οποίες σαρώνουν επανειλημμένα τα δεδομένα εισόδου για να μετρήσουν και να

κλαδέψουν υποψήφια αντικείμενα αυξανόμενης εμφάνισης. Από την άλλη πλευρά υπάρχουν οι προσεγγίσεις ανάπτυξης μοτίβων, οι οποίες σαρώνουν τα δεδομένα εισόδου μόνο μία φορά και δημιουργούν μια συμπαγή αναπαράσταση αυτών. Ο FP-Growth [2], ως μία τέτοια μέθοδος, κανονικοποιεί τις συναλλαγές με βάση την εμφάνιση αντικειμένων και αντιπροσωπεύει τα δεδομένα εισαγωγής ως ένα συμπαγές δέντρο επαυξημένου προθέματος. Τα συχνά σύνολα που προκύπτουν μπορούν στη συνέχεια να προσδιοριστούν αποτελεσματικά διασχίζοντας το FP-tree. Παραλλαγές που έχουν χρησιμοποιηθεί επίσης για συχνή εξόρυξη ακολουθιών είναι το GSP [9] όπου δημιουργεί υποψήφιες k -ακολουθίες ενώνοντας συχνές (k - 1) - ακολουθίες και κλαδεύοντας τις, μέσω σάρωσης των δεδομένων εισόδου. Το FreeSpan [10] μπορεί να θεωρηθεί ως μια πρώιμη προσαρμογή της ανάπτυξης FP στις συχνές ακολουθίες. Το PrefixSpan [11], ο διάδοχός του, χρησιμοποιεί μια προσέγγιση ανάπτυξης προτύπων που βασίζεται σε προβολές βάσης δεδομένων, αλλά χρησιμοποιεί μια κατάτμηση που βασίζεται σε επίθημα[1] του χώρου εξόδου. Το SPADE [12] αναλαμβάνει μια εναλλακτική κατακόρυφη αναπαράσταση των δεδομένων εισόδου, η οποία μπορεί να γίνει κατανοητή ως ένα ανεστραμμένο ευρετήριο που διατηρεί για κάθε στοιχείο τη λίστα των συναλλαγών που περιέχουν το στοιχείο και διασχίζει το (εννοιολογικό) πλέγμα όλων των ακολουθιών.

Σχεδιασμός και κατασκευή FP-Tree

Έστω I = $\{a_1,a_2,a_3,...,a_m\}$ ένα σύνολο από διακριτά αντικείμενα (items). Έστω επίσης D = $\{T_1,T_2,T_3..,T_m\}$ ένα σύνολο από συναλλαγές (transactions) όπου κάθε συναλλαγή T είναι ένα υποσύνολο αντικειμένων του H υποστήριξη (ή αλλιώς συχνότητα εμφάνισης) ενός προτύπου A – όπου A είναι ένα σύνολο από αντικείμενα – είναι ο αριθμός των συναλλαγών που περιέχουν το A. Λέμε ότι ένα πρότυπο A είναι συχνό όταν η υποστήριξή του είναι μεγαλύτερη από κάποιο προκαθορισμένο ελάχιστο κατώφλι. Έτσι λοιπόν με βάση τα παραπάνω το πρόβλημα εύρεσης συχνών προτύπων που αντιμετωπίζουμε έχει ως εξής: *Δεδομένης μιας βάσης συναλλαγών και ενός ελάχιστου κατωφλίου στήριξης, καλούμαστε να βρούμε όλα τα συχνά πρότυ-*

[1] Συμβολοσειρά που αποτελείται από έναν ή περισσότερους συνεχόμενους χαρακτήρες από το τέλος (δεξιό τμήμα) μιάς συμβολοσειράς (string)

πα και όχι μέρος αυτών.

Για καλύτερη κατανόηση του σχεδιασμού, παρακάτω παραθέτουμε ένα παράδειγμα μέσα από το οποίο θα κατασκευάσουμε μια συμπαγής δομή δεδομένων όπως είναι το FP-Tree. Έστω λοιπόν η βάση συναλλαγών που απεικονίζεται στον παρακάτω πίνακα και το ελάχιστο κατώφλι υποστήριξης ορισμένο στην τιμή 3.

TID	**Συναλλαγές**	**Ταξινόμηση συχνών αντικειμένων**
100	f, a, c, d, g, I, m, p	f, c, a, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m, n	f, c, a, m, p

Πίνακας 2.1: Πίνακας Συναλλαγών

Η κατασκευή του FP-Tree βασίζεται στις παρακάτω παρατηρήσεις:

1. Γνωρίζοντας ότι μόνα τα συχνά αντικείμενα έχουν σημαντικό ρόλο στην εξόρυξη συχνών προτύπων, εκτελείται μια σάρωση της βάσης δεδομένων συναλλαγής ώστε να προσδιοριστούν και να απομονωθούν τα συχνά αντικείμενα.

2. Θα πρέπει να γίνει προσπάθεια το σύνολο των συχνών αντικειμένων σε μια συναλλαγή να αποθηκευτεί σε μια συμπαγής δομή ώστε να αποφευχθεί η επαναληπτική ανίχνευση της βάσης που είναι αρκετά κοστοβόρα.

3. Εάν υπάρχουν πολλαπλές συναλλαγές που μοιράζονται ένα σύνολο συχνών αντικειμένων, αυτές μπορούν να συγχωνευτούν και για τα κοινά σύνολα θα αυξάνει το πλήθος των εμφανίσεων τους. Το κλειδί για να είναι η δομή που κατασκευάζουμε συμπαγής είναι η αποθήκευση της πληροφορίας αυτής συνεπτυγμένα που θα υποδηλώνεται με τον αριθμό των συναλλαγών που περιέχουν το σύνολο αυτό. Για να ελέγξουμε αν δύο σύνολα αντικειμένων είναι πανομοιότυπα ταξινομούμε τα αντικείμενα στις συναλλαγές με μια καθορισμένη σειρά (είτε αύξουσα είτε φθίνουσα).

4. Αν δύο συναλλαγές έχουν κοινό πρόθεμα, πάντα με βάση την νέα ταξινόμηση, το πρόθεμα αυτό θα αποθηκευτεί μόνο μία φορά στην δομή, συνοδευόμενη από

τον αριθμό των συναλλαγών που εμφανίζεται. Μάλιστα αν τα αντικείμενα ταξινομούνται με βάση των αριθμό εμφανίσεων τους σε φθίνουσα σειρά, υπάρχουν μεγαλύτερες πιθανότητες τα προθέματα που είναι κοινά να είναι περισσότερα.

Λαμβάνοντας υπόψη τις παραπάνω παρατηρήσεις κατασκευάζουμε το Frequent FP-Tree ακολουθώντας τα παρακάτω βήματα:

1. Πρώτα απ' όλα γίνεται μια σάρωση της Βάσης Δεδομένων ώστε να εντοπιστεί η λίστα με τα συχνά αντικείμενα. Έτσι έχουμε:

   (f:4),(c:4),(a:3),(b:3),(m:3),(p:3)

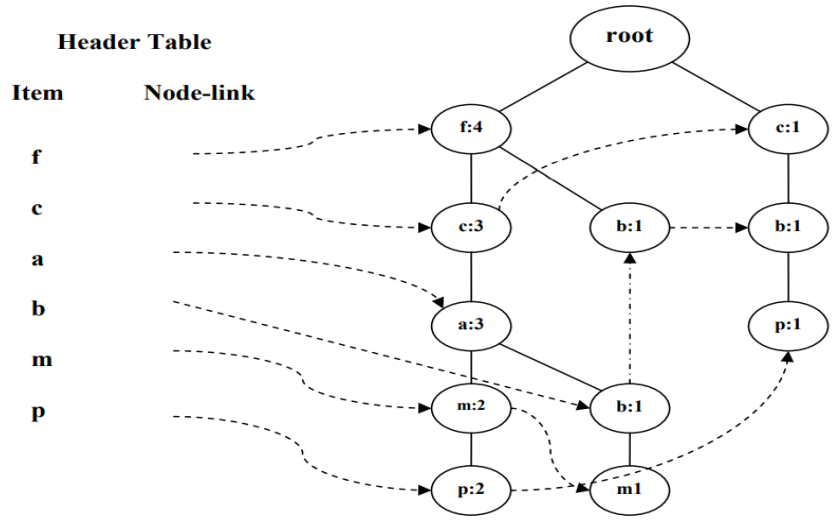
   Όπου κάθε ζεύγος εκπροσωπεί το αντικείμενο και την υποστήριξη (Support). Επίσης να τονίσουμε εδώ ότι τα αντικείμενα ταξινομούνται όπως φαίνεται σε φθίνουσα σειρά. Είναι μια χρήσιμη παρατήρηση που θα χρησιμοποιηθεί στην διαμόρφωση των μονοπατιών του δέντρου, και από δω και πέρα τα αντικείμενα μέσα στις συναλλαγές θα αναφέρονται ταξινομημένα.

2. Στην συνέχεια αρχίζει η κατασκευή του δέντρου. Αρχίζουμε από την ρίζα την οποία και ορίζεται ως "null". Αμέσως μετά κάνουμε και δεύτερο πέρασμα στην βάση συναλλαγών ως εξής:

   - Η πρώτη συναλλαγή, (f:1),(c:1),(a:1), (m:1),(p:1), ξεκινά αμέσως μετά την ρίζα και αποτελούν το πρώτο κλαδί του δέντρου. Κάθε αντικείμενο αποτελεί και έναν κόμβο, στον οποίο αποθηκεύουμε το όνομα και την υποστήριξη.
   - Η δεύτερη συναλλαγή $\{f, c, a, b, m\}$, παρατηρούμε ότι έχει κοινό πρόθεμα με την πρώτη συναλλαγή. Αυτό σημαίνει ότι για το κοινό πρόθεμα θα χρησιμοποιηθούν οι υπάρχοντες κόμβοι με την μόνη διαφορά ότι η υποστήριξη τους θα αυξηθεί κατά 1. Αμέσως μετά θα προστεθούν δύο κόμβοι (ο ένας κάτω από τον άλλο) απόγονοι του (a:2), που θα αντιστοιχούν στα αντικείμενα b, m.
   - Η τρίτη συναλλαγή $\{f, b\}$, έχει κοινό πρόθεμα μόνο το f το κόμβο του οποίου και αυξάνουμε την υποστήριξη κατά ένα (συνολικά δηλαδή 3). Από τον κόμβο f δημιουργούμε απόγονο με τα στοιχεία (b:1).

- Η τέταρτη συναλλαγή $\{c, b, p\}$, δεν παρουσιάζει κανένα κοινό πρόθεμα οπότε ξεκινάμε την κατασκευή του δεύτερου κλαδιού του δέντρου με κόμβους (c:1), (b:1) και (p:1).
- Η τελευταία συναλλαγή είναι πανομοιότυπη με την πρώτη με αποτέλεσμα να μην έχουμε την δημιουργία κανενός νέου κόμβου αλλά την αύξηση των υποστηρίξεων κατά ένα.

Για να διευκολυνθεί η πρόσβαση σε οποιοδήποτε κόμβο του δέντρου παράλληλα κατασκευάζουμε και έναν πίνακα ο οποίος αποθηκεύει τα λεγόμενα αντικείμενα οδηγούς. Πιο συγκεκριμένα για κάθε αντικείμενο (item) ο πίνακας (header table) περιέχει ένα δείκτη στον πρώτο κόμβο με το ίδιο όνομα αντικειμένου όπως απεικονίζεται στην παρακάτω εικόνα 2.1. Επίσης κάθε κόμβος του δέντρου είναι συνδεδεμένος με τον επόμενο κόμβο που έχει το ίδιο όνομα. Με αυτόν τον τρόπο επιτυγχάνεται πλήρης συνδεσιμότητα.



Σχήμα 2.1: FP-Tree [13]

Με βάση λοιπόν την μεθοδολογία που παραθέσαμε μπορούμε να παραθέσουμε τα βήματα αλγορίθμου κατασκευής Fp-Tree ως εξής:

1. Πρώτο πέρασμα στην Βάση συναλλαγών ώστε να προκύψει το σύνολο Φ των

συχνών αντικειμένων, συνοδευόμενα από την υποστήριξη τους. Στην συνέχεια ταξινομούμε σε φθίνουσα σειρά το σύνολο F.

2. Δημιουργούμε την ρίζα του FP-Tree με όνομα T και με την επιγραφή null.
3. Για κάθε συναλλαγή απομονώνουμε μόνο τα συχνά αντικείμενα τα οποία και τα αναδιατάσσουμε ανάλογα με την ταξινόμηση του βήματος 1. Δίνουμε σε κάθε συναλλαγή την μορφή $[p \mid P]$ όπου p είναι το πρώτο αντικείμενο ενώ P είναι η εναπομείνασα λίστα.
4. Καλούμε μια διαδικασία insert($[p \mid P]$, T) η οποία δουλεύει ως ακολούθως:
   - Εάν το T έχει απόγονο N έτσι ώστε N.item-name = p.item-name τότε αύξησε την υποστήριξη του κατά 1.
   - Αν δεν υπάρχει τέτοιος απόγονος τότε δημιουργούμε κόμβο N με αρχική υποστήριξη 1. Ο πατέρας του ορίζεται η ρίζα του δέντρου ενώ φροντίσουμε να δημιουργήσουμε και τις σωστές συνδέσεις με τους συνονόματους κόμβους.
   - Εάν το P δεν είναι άδειο τότε καλούμε επαναληπτικά την διαδικασία insert([P,N]).

Όσον αφορά για την χρονική πολυπλοκότητα του αλγορίθμου έχουμε:

1. Για την κατασκευή του FP-Tree χρειάζονται δύο περάσματα στην βάση δεδομένων. Στο πρώτο πέρασμα, συλλέγονται τα σύνολα συχνών αντικειμένων και στο δεύτερο πέρασμα γίνεται η κατασκευή του FP-Tree.
2. Το κόστος εισαγωγής των συναλλαγών στο FP-Tree είναι ανάλογο του freq(Transaction) δηλαδή O(freq(Transaction)). Όπου freq(Transaction) το σύνολο των συχνών αντικειμένων μέσα στην συναλλαγή (transaction).

Πληρότητα και συμπαγής δομής του FP-Tree

Δοθείσης μιας βάσης συναλλαγών και ενός κατωφλίου υποστήριξης κ ορίζουμε ως F το σύνολο των συχνών αντικειμένων στην βάση. Για κάθε συναλλαγή T ορίζουμε ως freq(T) την προβολή συχνών αντικειμένων και ισούται με το σύνολο των

συχνών αντικειμένων που υπάρχουν στην συναλλαγή. Προφανώς ισχύει freq(T) = $T \cap F$. Με βάση τις αρχές του Απριορι το σύνολο των προβολών των συναλλαγών της βάσης είναι ικανό να μας δώσει το σύνολο των συχνών προτύπων χωρίς καμιά απώλεια και χωρίς διπλοεγγραφές (completeness). Αποδεικνύεται ότι το ολοκληρωμένο σύνολο των προβολών συχνών αντικειμένων των συναλλαγών μπορεί να παραχθεί από το FP-Tree. Από το παραπάνω συνάγεται ότι η δομή FP-Tree είναι η πιο σημαντική στην διαδικασία εξόρυξης προτύπων και για το λόγο αυτό θα ασχοληθούμε με το μέγεθος του. Ισχύει ότι: *Δοθείσης μιας βάσης συναλλαγών και ενός κατωφλίου υποστήριξης κ το μέγεθος του FP-Tree (χωρίς να υπολογίζουμε την ρίζα null) φράσσεται από το* $\Sigma_{T \epsilon DB}|freq(T)|$ *ενώ το ύψος του δέντρου φράσσεται από το* $max_{T \epsilon DB}\{|freq(T)|\}$.

Βασιζόμενοι στον τρόπο κατασκευής του δέντρου για κάθε συναλλαγή T υπάρχει ένα μονοπάτι που ξεκινά από τον κόμβο με επιγραφή του προθέματος της συναλλαγής. Αυτό σημαίνει ότι το σύνολο των κόμβων του μονοπατιού είναι ακριβώς το ίδιο με τα συχνά αντικείμενα που περιέχει η συναλλαγή (δηλαδή το μέγεθος freq(T)). Οπότε στην χειρότερη περίπτωση όπου δεν υπάρχει καθόλου προθεματική επικάλυψη συναλλαγών το μέγεθος του δέντρου (κόμβων) είναι ίσος με $\Sigma_{T \epsilon DB}|freq(T)|$ Ακολουθώντας την ίδια λογική αφού freq(T) είναι ο αριθμός των κόμβων σε μια συναλλαγή τότε το ύψος του δέντρου θα ισούται με το μήκος της συναλλαγής που περιέχει τα περισσότερα συχνά αντικείμενα χωρίς διπλοεγγραφές αντικειμένων. Η απόδειξη αυτή είναι πολύ σημαντική γιατί καταδεικνύει ότι το μέγεθος του FP-Tree δεν ξεπερνά σε καμία περίπτωση το μέγεθος της βάσης συναλλαγών. Στην χειρότερη περίπτωση όπου δεν υπάρχει επικάλυψη συναλλαγών το μέγεθος της δομής ισούται με αυτό της βάσης ενώ συνήθως είναι αρκετά μικρότερο ανάλογα με το ποσοστό της επικάλυψης. Αυτό σε αντιπαράθεση με τον Apriori αλγόριθμο που είναι δυνατόν να δημιουργήσει αριθμό υποψήφιων συχνών προτύπων εκθετικά αυξανόμενο. Ένα επίσης σημαντικό πλεονέκτημα του FP-Tree είναι ο τρόπος ταξινόμησης των συχνών αντικειμένων. Η φθίνουσα ταξινόμηση δεν γίνεται τυχαία. Αυτό γιατί ξεκινώντας από αντικείμενα με την μεγαλύτερη υποστήριξη σημαίνει ότι έχουν μεγαλύτερο αριθμό εμφανίσεων στην βάση και άρα μεγαλύτερες πιθανότητες να μοιράζονται στις διάφορες συναλλαγές. Τέτοια προθέματα όμως στην κατασκευή του δέντρου μας ση-

μαίνει μεγάλη επικάλυψη και άρα ολοένα και πιο συμπαγή δομή. Βέβαια η φθίνουσα ταξινόμηση δεν είναι πάντα η βέλτιστη.

2.3 Κατανεμημένη Εξόρυξη Δεδομένων

Στο πλαίσιο της παρούσας εργασίας, έχουμε χρησιμοποιήσει τον αλγόριθμο FP-Growth σε ένα Chord κατανεμημένο περιβάλλον προκειμένου να εξάγουμε συχνά πρότυπα πάνω από σύνολα δεδομένων. Το κύριο πλεονέκτημα της αρχιτεκτονικής που επιλέγουμε είναι η προσαρμογή της στις απαιτήσεις μιας ακριβούς διαδικασίας εξόρυξης δεδομένων. Για να καταλήξουμε σε αυτήν την αρχιτεκτονική, αξιολογήσαμε άλλες υπάρχουσες επιλογές. Αρχικά, ο μεγάλος όγκος δεδομένων που συνήθως απαιτείται σε μια διαδικασία ανάλυσης και εξόρυξης δεδομένων δρα επιβαρυντικά στην απόδοση των αλγορίθμων, κατάσταση που μπορεί να αποφευχθεί όταν κάθε κόμβος είναι υπεύθυνος για ένα κομμάτι πληροφορίας. Μέχρι πρόσφατα, η έρευνα εξόρυξης δεδομένων επικεντρώθηκε στην κεντρική εξόρυξη δεδομένων όπου όλο το σύνολο δεδομένων αποθηκεύεται σε έναν μόνο κόμβο. Η εξέλιξη της εξόρυξης δεδομένων περιγράφει την ανάπτυξη νέων συνεισφορών τόσο σε νέους αλγορίθμους, θεωρητικά μοντέλα ή τεχνικές εξόρυξης δεδομένων, όσο και σε τεχνολογική και σχεδιαστική έρευνα για νέα συστήματα και αρχιτεκτονικές εξόρυξης δεδομένων. Αν και έχει σημειωθεί μεγάλη πρόοδος όσο αφορά στις τεχνικές και στους αλγορίθμους εξόρυξης δεδομένων, οι περισσότερες δουλείες δεν αφορούσαν κατανεμημένα σύνολα δεδομένων. Παρόλ΄ όλα αυτά, έχουν αναπτυχθεί κάποιοι κατανεμημένοι αλγόριθμοι, βασιζόμενοι στις αρχικές κεντρικοποιημένες εκδοχές. Μερικά παραδείγματα είναι κατανεμημένοι ή παράλληλοι αλγόριθμοι για κανόνες συσχέτισης [14], κανόνες ταξινόμησης [15], μοτίβα ακολουθίας [16] ή αλγόριθμοι ομαδοποίησης [17]. Ωστόσο, η έρευνα πάνω σε νέες κατανεμημένες αρχιτεκτονικές συνεχίζει να είναι σημαντική και ενδιαφέρουσα διότι ασχολείται με την αποτελεσματική χρήση υπολογιστικών πόρων, καθώς και με τεχνικά θέματα που σχετίζονται με την επικοινωνία, τον συγχρονισμό και τον προγραμματισμό.

Υπάρχουν διαφορετικές προσεγγίσεις και αλγόριθμοι που χρησιμοποιούνται για την κατανεμημένη εξόρυξη συχνών αντικειμένων. Για παράδειγμα, το 1995, ο Mueller

πρότεινε δύο παράλληλους αλγόριθμους, όπου ονομάζονται parallel efficient association rules (PEAR) [18] και parallel partition association rules (PPAR) [18]. Οι Park et al. πρότειναν επίσης έναν αλγόριθμο που ονομάζεται parallel data mining (PDM) [19] που βοηθά στην παράλληλη εξόρυξη κανόνων συσχέτισης, και τον αλγόριθμο fast distributed mining (FDM) για κατανεμημένες βάσεις δεδομένων [20]. Αυτές οι δουλείες δίνουν έμφαση στις ιδιότητες της υποστήριξης και της εμπιστοσύνης στο σύνολο αντικειμένων ή στον αριθμό των φορών που θα σαρωθεί (ολόκληρη) η βάση δεδομένων, αλλά αφήνουν σε δεύτερη μοίρα την παραλληλοποίηση ή την κατανομή του φόρτου της εξόρυξης συχνών προτύπων. Πρόσφατα, αλγόριθμοι κατανεμημένης εξόρυξης δεδομένων αναπτύχθηκαν με βάση τις πλατφόρμες Spark ή Hadoop. Το Hadoop είναι μία από τις γνωστές πλατφόρμες που χρησιμοποιούν το MapReduce framework [21], ένα λογισμικό ανοιχτού κώδικα για υπολογισμούς σε μεγάλα δεδομένα πάνω από συστάδες υπολογιστών (cluster) . Για παράδειγμα, μια προσαρμογή του FP-Growth στο MapReduce, είναι ο PFP [22], οποίος χωρίζει μια μεγάλης κλίμακας εξόρυξη σε ανεξάρτητες και παράλληλες διαδικασίες.

Εμείς, στην παρούσα εργασία, προτείνουμε μια διαφορετική προσέγγιση, χρησιμοποιούμε την Chord κατανεμημένη αρχιτεκτονική, κατανέμουμε τα δεδομένα σε υπολογιστικούς κόμβους που δουλεύουν ανεξάρτητα μεταξύ τους, και δείχνουμε ότι μπορούμε να έχουμε αποτελέσματα ισοδύναμα με μια κεντρικοποιημένη προσέγγιση αλλά με σημαντικά μικρότερους χρόνους.

2.3.1 Chord

Ένα από τα βασικότερα προβλήματα που απασχολεί τις peer-to-peer εφαρμογές είναι η εύρεση της τοποθεσίας με αποδοτικό τρόπο του κόμβου δηλαδή, που βρίσκεται αποθηκευμένο το αντικείμενο που αναζητείται. Το Chord [6] είναι ένα κατανεμημένο πρωτόκολλο αποθήκευσης και αναζήτησης το οποίο λύνει αυτό ακριβώς το πρόβλημα. Το Chord παρέχει μια και μόνο υπηρεσία, εάν του δοθεί ένα οποιοδήποτε κλειδί τότε επιστρέφει τον υπεύθυνο κόμβο για αυτό το κλειδί. Η εύρεση της τοποθεσίας των δεδομένων μπορεί να υλοποιηθεί πάνω από το Chord, συσχετίζοντας ένα κλειδί με κάθε κομμάτι δεδομένων και αποθηκεύοντας αυτό το ζευγάρι κλειδιού/δεδομένων στον κόμβο στον οποίο αντιστοιχεί το συγκεκριμένο κλειδί.

Το Chord χρησιμοποιεί συνεπές hashing για να αντιστοιχίσει τα κλειδιά με τους κόμβους. Πιο συγκεκριμένα, τείνει να εξισορροπεί το φορτίο του συστήματος, αφού κάθε κόμβος αναλαμβάνει περίπου τον ίδιο αριθμό κλειδιών και χρειάζεται σχετικά λίγη κινητικότητα κλειδιών όταν κόμβοι συνδέονται ή αποσυνδέονται από το σύστημα. Επιπρόσθετα, στο Chord κάθε κόμβος χρειάζεται να γνωρίζει πληροφορίες δρομολόγησης μόνο για μερικούς άλλους κόμβους. Επειδή, ο πίνακας δρομολόγησης είναι κατανεμημένος, ένας κόμβος του Chord επικοινωνεί με άλλους κόμβους μέσα από τον δικό του πίνακα δρομολόγησης για να ολοκληρώσει μια αναζήτηση. Σε μια σταθερή κατάσταση όπου έχουμε, ένα σύστημα με N κόμβους, κάθε κόμβος διατηρεί πληροφορία για περίπου μόνο O(logN) άλλους κόμβους και ολοκληρώνει όλες τις αναζητήσεις με μόνο O(logN) μηνύματα σε άλλους κόμβους. Το Chord συντηρεί τις πληροφορίες δρομολόγησης όσο κόμβοι εισέρχονται ή εξέρχονται από το σύστημα.

Το Chord διευκολύνει τον σχεδιασμό των peer-to-peer συστημάτων και των εφαρμογών που τρέχουν πάνω από αυτό, λύνοντας τα παρακάτω δύσκολα προβλήματα.

- Εξισορρόπηση φόρτου: Το Chord λειτουργεί σαν μια κατανεμημένη hash συνάρτηση, μοιράζοντας όλα τα κλειδιά ισοδύναμα στους κόμβους.
- Αποκεντρικοποίηση: Το Chord είναι πλήρως κατανεμημένο. Κανένας κόμβος δεν είναι σημαντικότερος από έναν άλλον κόμβο.
- Κλιμάκωση: Το κόστος της αναζήτησης στο Chord μεγαλώνει λογαριθμικά σύμφωνα με τον αριθμό των κόμβων. Με αυτόν τον τρόπο ακόμα και πολύ μεγάλα συστήματα είναι εφικτά. Δεν χρειάζεται κάποια αλλαγή σε κάποια παράμετρο για να επιτευχθεί αυτή η κλιμάκωση.
- Διαθεσιμότητα: Το Chord αυτόματα τροποποιεί τους εσωτερικούς πίνακες δρομολόγησης που χρησιμοποίει, ώστε να είναι πλήρως ενημερωμένοι με όλες τις τελευταίες αλλαγές, δηλαδή με τους καινούριους κόμβους που συνδέθηκαν ή με αυτούς που απέτυχαν. Με αυτόν τον τρόπο εγγυάται, εξαιρώντας βέβαια τρομερές αποτυχίες στο υποκείμενο φυσικό δίκτυο, ότι ο κόμβος ο οποίος είναι υπεύθυνος για κάποιο κλειδί θα μπορεί πάντα να βρεθεί

- Ευέλικτη ονοματολογία: Το Chord δεν θέτει περιορισμούς στην δομή των κλειδιών που αναζητά, ο χώρος κλειδιών του Chord είναι επίπεδος. Αυτή η ιδιότητα δίνει στις εφαρμογές τεράστια ευελιξία στον τρόπο που θα αντιστοιχίσουν τα ονόματα τους σε κλειδιά στο Chord.

Κεφάλαιο 3

Αρχιτεκτονική Συστήματος

Η βάση δεδομένων είναι μια συλλογή δεδομένων που σχετίζονται με κάποιον τρόπο μεταξύ τους. Πολλοί οργανισμοί (εταιρίες, πανεπιστήμια κλπ.) χρησιμοποιούν βάσεις δεδομένων για αποθήκευση, διαχείριση και ανάκτηση δεδομένων με γρήγορο και αποδοτικό τρόπο. Ένα κατανεμημένο σύστημα διαχείρισης βάσης δεδομένων χωρίζει τη βάση δεδομένων σε πολλά αρχεία, κάθε ένα από τα οποία αποθηκεύεται σε διαφορετική τοποθεσία στο δίκτυο. Για τις ανάγκες της παρούσας εργασίας, χρησιμοποιήθηκε μια κατανεμημένη αρχιτεκτονική για την διαχείριση των δεδομένων, οπότε τα δεδομένα χωρίστηκαν και κάθε κομμάτι αποθηκεύτηκε σε έναν διαφορετικό υπολογιστικό κόμβο του δικτύου. Σε αυτό το κεφάλαιο, παρουσιάζουμε τις επιλογές που κάναμε και την προσέγγιση που ακολουθήσαμε σε υψηλό επίπεδο, ώστε τελικά να επιτύχουμε με αποδοτικό τρόπο την εξόρυξη συχνών προτύπων σε μεγάλο όγκο δεδομένων.

3.1 Κατανεμημένη Βάση Δεδομένων

Μια κατανεμημένη βάση δεδομένων (distributed database) μπορεί να οριστεί σαν μια ομάδα από λογικά συνδεόμενες βάσεις δεδομένων που είναι διεσπαρμένες σε ένα δίκτυο υπολογιστών. Η ανάγκη του τεμαχισμού μιας βάσης δεδομένων και της κατανομής της γεωγραφικά στο δίκτυο προέκυψαν ιστορικά από την ανάγκη των μεγάλων οργανισμών να εκμεταλλεύονται το σύνολο των δεδομένων τους που βρίσκουμε αποθηκευμένα σε διάφορα τοπικά μηχανογραφικά συστήματα. Οι λόγοι που καθιέρωσαν

τα κατανεμημένα συστήματα σαν μια αποδεκτή και χρησιμοποιούμενη στην πράξη τεχνολογία, μπορούν να περιγραφούν συνοπτικά ως εξής:

1. Οργανωτικοί και οικονομικοί λόγοι, καθώς αποδείχθηκε, πολλές φορές στην πράξη, ότι η συντήρηση κατανεμημένων συστημάτων μπορεί να είναι πιο αποδοτική από τη συντήρηση ενός κεντρικού συστήματος.

2. Λόγοι αξιοποίησης της πληροφορίας, καθώς οι κατανεμημένες βάσεις δεδομένων προκύπτουν σαν μια φυσική λύση σε γεωγραφικά κατανεμημένα δεδομένα που προσπελαύνονται από κεντρικές εφαρμογές.

3. Λόγοι σταδιακής αύξησης του οργανισμού, η οποία μπορεί να γίνει πιο εύκολα σε ένα κατανεμημένο, παρά σε ένα κεντρικό περιβάλλον, προσθέτοντας νέα χαμηλού κόστους συστήματα.

4. Λόγοι μείωσης του τηλεπικοινωνιακού φορτίου.

5. Λόγοι απόδοσης στην αποτίμηση ερωτήσεων.

6. Λόγοι αξιοπιστίας και διαθεσιμότητας της πληροφορίας, μέσω πολλαπλών κατανεμημένων αντιγράφων της.

Βέβαια, όλα αυτά τα θετικά χαρακτηριστικά έρχονται μαζί με την ανάγκη διαχείρισης και συντήρησης μιας κατανεμημένης και πολύπλοκης εφαρμογής καθώς και με το κόστος του επιπλέον λογισμικού και υλικού υποστήριξης. Τα κατανεμημένα συστήματα βάσεων δεδομένων κλήθηκαν να καλύψουν τις απαιτήσεις των διαφόρων οργανισμών με την ίδια αξιοπιστία και ταχύτητα που τα κλασικά συστήματα βάσεων δεδομένων ανταποκρίνονταν στις αντίστοιχες απαιτήσεις.

Οι βασικές απαιτήσεις και τα χαρακτηριστικά ενός κατανεμημένου συστήματος βάσεων δεδομένων είναι τα εξής:

1. Έλεγχος του συστήματος. Ο έλεγχος του συστήματος επιτελείται και από τοπικούς διαχειριστές (administrators) και καθολικά, από κάποιον επιβλέποντα διαχειριστή.

2. Ανεξαρτησία από τα δεδομένα. Βασικό χαρακτηριστικό των κατανεμημένων βάσεων δεδομένων είναι ο βαθμός στον οποίο η φυσική οργάνωση των δεδομένων είναι διαφανής στον προγραμματιστή.

3. Ύπαρξη αντιγράφων. Σε ένα κατανεμημένο σύστημα, η μείωση του πλεονασμού των δεδομένων (που ήταν από τα βασικά ζητούμενα στις μη-κατανεμημένες βάσεις δεδομένων) δεν είναι πλέον τόσο σημαντικό. Αντιθέτως, η ύπαρξη τοπικών αντιγράφων της πληροφορίας είναι ως ένα βαθμό και επιδιωκόμενη, για την αύξηση της απόδοσης του συστήματος. Θα πρέπει όμως το σύστημα να υποστηρίζει μηχανισμούς ενημέρωσης των αντιγράφων με τα πιο πρόσφατα δεδομένα.

4. Η βελτιστοποίηση των ερωτήσεων σε περιβάλλοντα κατανεμημένων βάσεων δεδομένων δεν γίνεται μόνο τοπικά, αλλά προηγείται και ένα στάδιο καθολικής βελτιστοποίησης, οπότε τα ερωτήματα να μπορούν να εκτελούνται αποδοτικά (από άποψη χρόνου εκτέλεσης, αλλά και εύρεσης όλων των διαθέσιμων εγγράφων) ανεξάρτητα από την γεωγραφική κατανομή των δεδομένων.

5. Η αξιοπιστία, ο έλεγχος συντονισμού, η ανάνηψη των συστημάτων, καθώς και η δυνατότητα ορισμού δικαιωμάτων στους χρήστες, είναι από τα βασικά χαρακτηριστικά και των κατανεμημένων συστημάτων βάσεων δεδομένων.

Στον Πίνακα 3.1 παρουσιάζεται συνοπτικά τα πλεονεκτήματα και μειονεκτήματά χρήσης κατανεμημένων βάσεων δεδομένων.

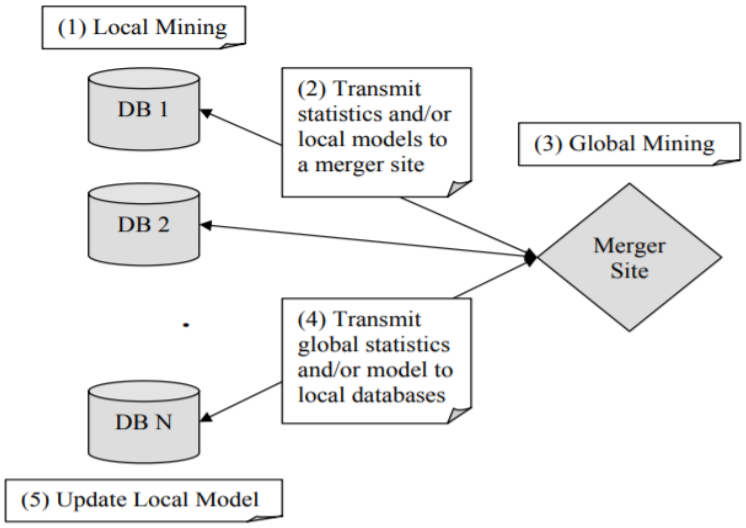
Πλεονεκτήματα	**Μειονεκτήματα**
Μεγαλύτερη αξιοπιστία και διαθεσιμότητα	Απαίτηση επιπλέον λογισμικού
Ευκολότερη επέκταση	Αυξημένη πιθανότητα για σφάλματα λογισμικού (bugs)
Αυτονομία(Τοπική)	Έλλειψη προτύπων(εργαλεία-μεθοδολογίες)
Μια μεμονωμένη βλάβη δεν επηρεάζει τις επιδόσεις του συστήματος	Επιπλέον επεξεργασία-συντήρηση
Προστασία των πολύτιμων δεδομένων	Αυξημένη πολυπλοκότητα
Μεγαλύτερη απόδοση	Συγχρονισμός ελέγχου
Μικρότερο κόστος	Ελλιπής ασφάλεια

Πίνακας 3.1: Χαρακτηριστικά κατανεμημένων βάσεων δεδομένων

3.2 Αρχιτεκτονική Κατανεμημένης Εξόρυξης

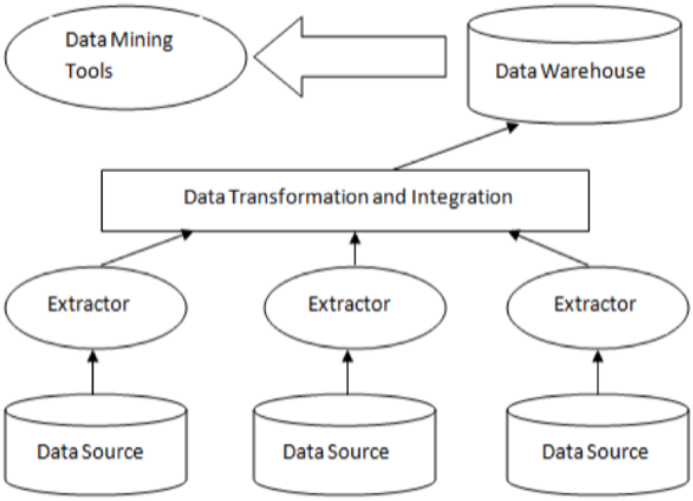
Οι κύριοι παράγοντες που οδήγησαν στην εξέλιξη της κατανεμημένης εξόρυξης δεδομένων είναι η μείωση του κόστους μετάδοσης, του κόστους υπολογισμού και του κόστους μνήμης. Κι ένας από τους βασικούς στόχους είναι η εξαγωγή χρήσιμων πληροφοριών από δεδομένα που βρίσκονται σε ετερογενείς τοποθεσίες. Στα κατανεμημένα συστήματα βάσεων δεδομένων, η βάση δεδομένων αποθηκεύεται σε διάφορους υπολογιστές που επικοινωνούν ο ένας με τον άλλον μέσω διαφόρων μέσων επικοινωνίας, όπως δίκτυα υψηλής ταχύτητας, τηλεφωνικές γραμμές ή τοπικά δίκτυα. Δεν μοιράζονται κύρια μνήμη ή δίσκους. Οι υπολογιστές σε ένα κατανεμημένο σύστημα μπορεί να διαφέρουν σε μέγεθος και λειτουργία.

Μια τυπική αρχιτεκτονική μιας προσέγγισης κατανεμημένης εξόρυξης δεδομένων απεικονίζεται στο Σχήμα 3.1. Η πρώτη φάση συνήθως περιλαμβάνει την ανάλυση σε τοπικό επίπεδο κάθε επιμέρους βάσης δεδομένων. Στη συνέχεια, η ανακαλυφθείσα γνώση μεταδίδεται σε έναν ιστότοπο συγχώνευσης, όπου πραγματοποιείται η συνένωση των τοπικών μοντέλων σε ένα γενικό μοντέλο εξόρυξης που να καλύπτει το σύνολο των δεδομένων του συστήματος. Τα αποτελέσματα μεταδίδονται πίσω στις κατανεμημένες βάσεις δεδομένων, έτσι ώστε όλοι οι ιστότοποι να ενημερώνονται. Σε ορισμένες προσεγγίσεις, αντί για έναν ιστότοπο συγχώνευσης, τα τοπικά μοντέλα μεταδίδονται σε όλους τους άλλους ιστότοπους, έτσι ώστε κάθε ιστότοπος να μπορεί παράλληλα να εκτελεί υπολογισμούς.

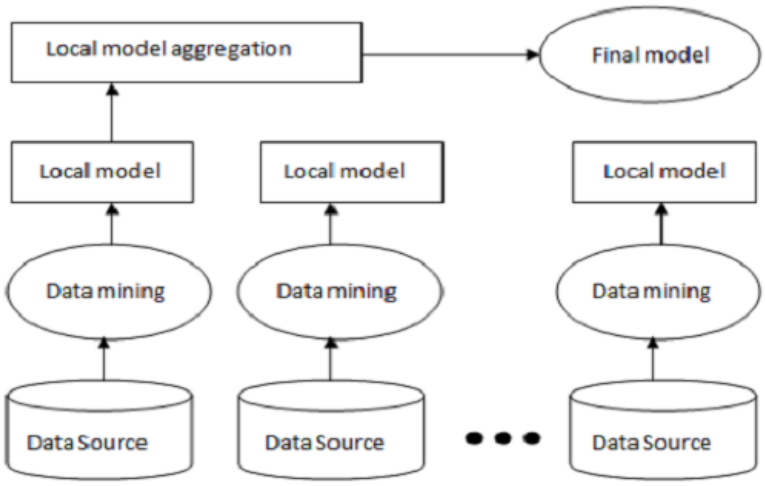


Σχήμα 3.1: Τυπική αρχιτεκτονική κατανεμημένης εξόρυξης δεδομένων [23]

Στο διάγραμμα που απεικονίζεται παρακάτω, στο Σχήμα 3.2 φαίνεται μια κλασσική/παραδοσιακή κεντρικοποιημένη (centralised) προσέγγιση για την αποθήκευση των δεδομένων και την εξόρυξη γνώσης από αυτά. Πιο συγκεκριμένα, τα δεδομένα, ανεξάρτητα από που προέρχονται, καταλήγουν σε μια βάση δεδομένων, όπου εφαρμόζονται κατάλληλα εργαλεία και τεχνικές για την διαδικασία εξόρυξης. Αυτή η προσέγγιση έχει, όλα τα μειονεκτήματα μιας οποιαδήποτε κεντρικοποιημένης αρχιτεκτονικής όπως μεγάλους χρόνους απόκρισης, μη αποδοτική χρήση των κατανεμημένων πόρων, μεταφορά δεδομένων πάνω από το δίκτυο, κοκ. Η λύση σε αυτά τα προβλήματα είναι μια κατανεμημένη εφαρμογή που ζητά την επεξεργασία κατανεμημένων δεδομένων εκμεταλλευόμενη όλους τους διαθέσιμους πόρους. Όπως φαίνεται στο Σχήμα 3.3, ο στόχος της κατανεμημένης εξόρυξης δεδομένων είναι να εκτελέσει τις εργασίες εξόρυξης δεδομένων με βάση τη διαθεσιμότητα πόρων και αποφεύγοντας την μετακίνηση των δεδομένων από τις φυσικές πηγές τους σε μια κεντρική τοποθεσία. Οποιοσδήποτε ιστότοπος επιλέγεται για πρόσβαση στα δεδομένα και στη συνέχεια εκτελεί τις λειτουργίες τοπικά. Οι ιστότοποι επιλέγονται σύμφωνα με την ικανότητα αποθήκευσης, υπολογισμού και επικοινωνίας.



Σχήμα 3.2: Κεντρικοποιημένη αρχιτεκτονική εξόρυξης δεδομένων [28]



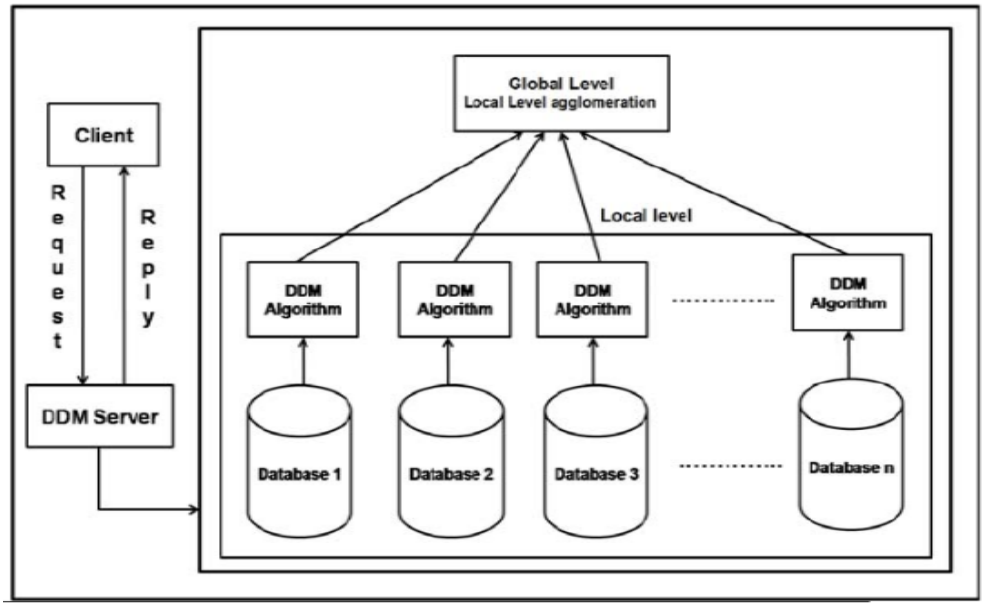
Σχήμα 3.3: Κατανεμημένη αρχιτεκτονική εξόρυξης δεδομένων [28]

Η κατανεμημένη εξόρυξη περιλαμβάνει δύο παραλλαγές αρχιτεκτονικών μοντέλων: αρχιτεκτονική βασισμένη σε διακομιστή-πελάτη (client-server) και αρχιτεκτονική βασισμένη σε πράκτορες (agents) με βάση τη συλλογή / επεξεργασία δεδομένων.

3.3 Διαθέσιμες προσεγγίσεις

3.3.1 Αρχιτεκτονική διακομιστή-πελάτη

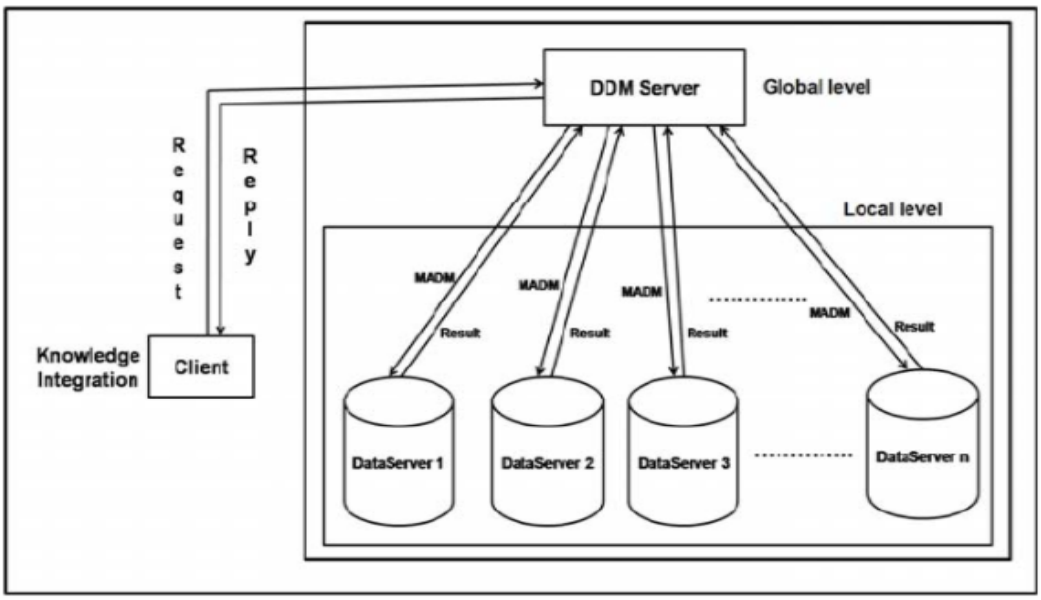
Η αρχιτεκτονική που βασίζεται σε διακομιστή-πελάτη φαίνεται στο Σχήμα 3.4 όπου ο πελάτης στέλνει ένα αίτημα σε διακομιστή, ο οποίος με τη σειρά του εξουσιοδοτεί τη συλλογή των συσσωρευμένων δεδομένων σε τοπικό επίπεδο. Τα συσσωρευμένα δεδομένα από ετερογενείς ιστότοπους (τοπικό επίπεδο) πρέπει να υποβληθούν σε επεξεργασία σε κεντρικό επίπεδο. Η αρχιτεκτονική πελάτη-διακομιστή έχει πολυπλοκότητα μεταφοράς των δεδομένων (μετάδοση όλων των δεδομένων για εκτέλεση εξόρυξης σε κεντρικό επίπεδο) και επομένως αυτό αυξάνει το εύρος ζώνης δικτύου και τον λανθάνοντα χρόνο δικτύου [26].



Σχήμα 3.4: Client-Server based DDM [24]

3.3.2 Αρχιτεκτονική βασισμένη σε πράκτορες

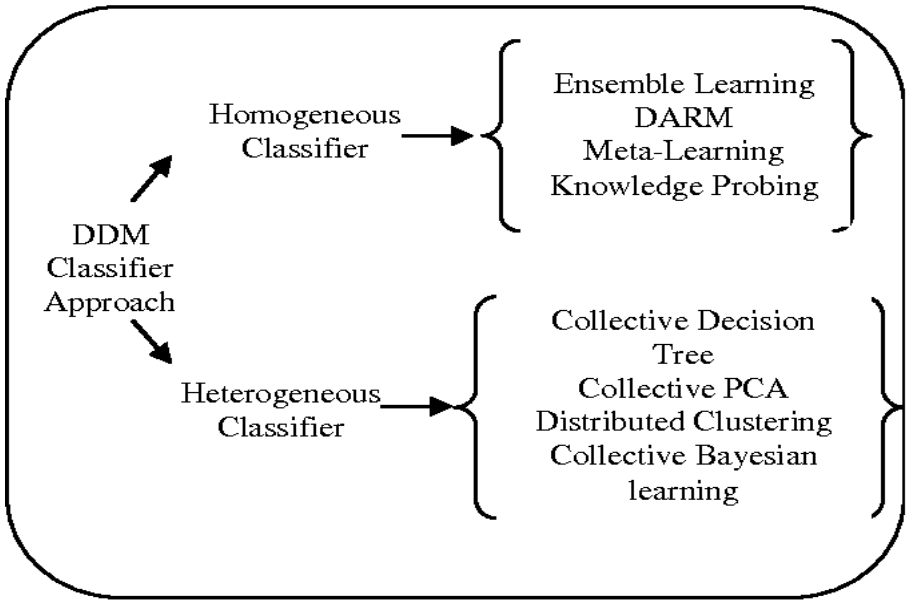
Η αρχιτεκτονική που βασίζεται σε πράκτορες φαίνεται στο Σχήμα 3.5. Ο πελάτης δημιουργεί πολλούς πράκτορες εξόρυξης δεδομένων με βάση το Mobile-Agent (MADM) όπου οι πράκτορες είναι οι οντότητες που μπορούν να μεταφέρονται σε τοποθεσίες-υπολογιστές σε ένα περιβάλλον δικτύου για την εκπλήρωση των στόχων τους. Πιο συγκεκριμένα, οι κινητοί πράκτορες είναι προγράμματα που μπορούν να μεταναστεύσουν από κόμβο σε κόμβο μέσα σε ένα δίκτυο. Η κατάσταση του τρέχοντας προγράμματος σώζεται, στέλνεται στο νέο κόμβο και αποκαθίσταται, έτσι το πρόγραμμα μπορεί να συνεχίσει από το σημείο που σταμάτησε. Το κυριότερο πλεονέκτημά ενός κινητού πράκτορα είναι η εξοικονόμησης εύρους ζώνης, δηλαδή έχουν την δυνατότητα να μετακινούνται μεταξύ διαφορετικών συστημάτων για να ενισχύσουν την αποδοτικότητα του υπολογισμού και να μειώσουν το κυκλοφοριακό των δικτύων, με αυτόν τον τρόπο επιτυγχάνεται ελαχιστοποίηση της κατανάλωσης των πόρων του δικτύου και η ακεραιότητα των πληροφοριών που λαμβάνει ο πράκτορας δεν επηρεάζεται από τυχόν προβλήματα επικοινωνίας με τον εξυπηρετητή [26].



Σχήμα 3.5: Agent-based DDM [24]

3.4 Προσεγγίσεις στους ταξινομητές δεδομένων

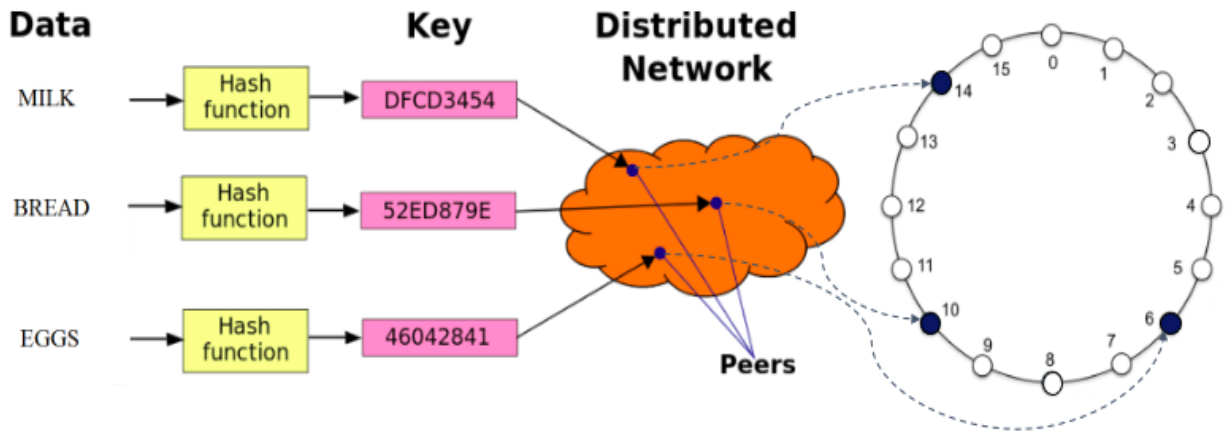
Με βάση τα εξεταζόμενα σύνολα δεδομένων, η προσέγγιση ταξινομητή μπορεί να χωριστεί σε δύο προσεγγίσεις: ομοιογενείς ταξινομητές (εξόρυξη κατανεμημένων τοποθεσιών δεδομένων που περιλαμβάνουν παρόμοια χαρακτηριστικά) και ετερογενής ταξινομητές (εξόρυξη κατανεμημένων τοποθεσιών δεδομένων που περιλαμβάνουν διακριτά χαρακτηριστικά). Η ταξινόμηση δεδομένων ορίζεται γενικά ως η διαδικασία οργάνωσης δεδομένων από σχετικές κατηγορίες, έτσι ώστε να μπορούν να χρησιμοποιούνται και να προστατεύονται πιο αποτελεσματικά. Σε βασικό επίπεδο, η διαδικασία ταξινόμησης διευκολύνει τον εντοπισμό και την ανάκτηση δεδομένων. Το Σχήμα 3.6 απεικονίζει τις δύο προσεγγίσεις ταξινομητή κατανεμημένης εξόρυξης δεδομένων (ομοιογενής και ετερογενής ταξινομητής) μαζί με τις 4 τεχνικές κάθε προσέγγισης. Οι ομοιογενείς τεχνικές ταξινόμησης επικεντρώνονται στην εξόρυξη δεδομένων με παρόμοιες ιδιότητες. Αντίθετα, οι ετερογενείς ταξηνομητές εφαρμόζονται στην εξόρυξη δεδομένων με διαφορετικές ιδιότητες.



Σχήμα 3.6: Τεχνικές ομοιογενούς ταξινομητή και τεχνική ετερογενούς ταξινομητή [29]

3.5 Η προτεινόμενη προσέγγιση

Οι κύριοι παράγοντες όπου οδήγησαν στην επιλογή της προτεινόμενης προσέγγισης ήταν ώστε να αντιμετωπιστούν κάποια βασικά προβλήματα αρκετών σύγχρονων αρχιτεκτονικών. Πιο συγκεκριμένα, σε διάφορες αρχιτεκτονικές για να πραγματοποιηθεί εξόρυξη δεδομένων σε δεδομένα όπου βρίσκονται διάσπαρτα σε διαφορετικούς εξυπηρέτες έπρεπε να προωθηθούν σε μία κεντρική βάση δεδομένων, αυτό είχε ως αποτέλεσμα να δημιουργούνται σημεία συμφόρησης. Ορμώμενοι ενός τέτοιου προβλήματος υλοποιήσαμε έναν κατανεμημένο τρόπο επεξεργασίας όπου τα δεδομένα διαμοιράζονται ανάμεσα στους εξυπηρέτες και ο καθένας επεξεργάζεται ένα τμήμα των δεδομένων. Μια τέτοια διαμέριση δεδομένων επιτρέπει την βέλτιστη κατανομή των δεδομένων στους κόμβους και την ομοιόμορφη κατανομή του φόρτου εργασίας. Ειδικότερα, υλοποιήθηκε ένα δομημένο Peer-to-Peer σύστημα όπου ονομάζεται Chord, τα δομημένα συστήματα χαρακτηρίζονται από μια συγκεκριμένη δομή με την έννοια ότι υπάρχει κάποιος κανόνας για τις συνδέσεις μεταξύ των κόμβων καθώς και τα δεδομένα δεν τοποθετούνται τυχαία σε αυτούς άλλα σε προκαθορισμένες τοποθεσίες, γεγονός που βοηθά σημαντικά στην απόδοση του δικτύου. Σε κάθε έναν κόμβο εφαρμόζεται ο αλγόριθμος FP-Growth και εξάγονται συχνά σύνολα. Εκτός ότι βελτιώνει το βασικό πρόβλημα συμφόρησης αλλά έχει και σαν στόχο να μειώσει τον χρόνο εύρεσης συχνών συνόλων όταν αναφερόμαστε για ένα τεράστιο όγκο δεδομένων.



Σχήμα 3.7: Αρχιτεκτονική της προτεινόμενης προσέγγισης

Πιο αναλυτικά, στην Εικόνα 3.7 που απεικονίζεται παραπάνω είναι ένα παράδειγμα του πως κατασκευάστηκε η αρχιτεκτονική. Αρχικά, έχουμε κάποια προϊόντα τα οποία περνάνε από μια συνάρτηση κατακερματισμού. Αυτό έχει ως αποτέλεσμα να εξαχθεί μια συμβολοσειρά, όπου θα μας βοηθήσει να κατανείμουμε σωστά τα προϊόντα στους κόμβους του δικτύου. Ο κύκλος απεικονίζει τον τρόπο με τον οποίο είναι σχεδιασμένο το δίκτυο. Σε κάθε κόμβο του δικτύου αποθηκεύονται τελικά όλες η συναλλαγές που περιέχουν το συγκεκριμένο προϊόν.

Κεφάλαιο 4

Ανάλυση καλαθιού αγορών

Η τεχνική ανάλυσης καλαθιού αγορών (Market Basket Analysis) αν κάποιος πελάτης αγοράσει κάποιο συγκεκριμένο προϊόν (ή σύνολο προϊόντων), τότε είναι πολύ πιθανό (ή αντίστοιχα ελάχιστα πιθανό) να αγοράσει και ένα άλλο προϊόν (ή σύνολο προϊόντων). Το σύνολο των προϊόντων που αγοράζει ένας πελάτης κατά την διάρκεια μιας συγκεκριμένης αγοράς του ονομάζεται itemset (Ορισμός 2.1.1).

Η τεχνική ανάλυσης καλαθιού αγορών έχει σαν κύριο στόχο την ανάλυση των δεδομένων που προκύπτουν από τις αγορές των πελατών, με σκοπό την ανακάλυψη συσχετίσεων μεταξύ των διαφόρων προϊόντων [30].

Definition 4.1 (Ανάλυση καλαθιού αγορών)**.** *Ο όρος Market Basket Analysis (MBA), ή ανάλυση καλαθιού αγορών, αφορά στην ανάλυση διαφόρων υποσυνόλων αντικειμένων (προϊόντων), τα οποία επιλέχθηκαν μέσα από κάποιον μεγαλύτερο πληθυσμό αντικειμένων. Ένα παράδειγμα κανόνα είναι το $A \rightarrow B$, όπου υποδηλώνει ότι: «εάν το αντικείμενο A υπάρχει στο καλάθι αγορών (market basket), τότε υπάρχει και το αντικείμενο B». Το A ονομάζεται προηγηθέν αντικείμενο (antecedent item), ενώ το B συνεπακόλουθο (consequent).*

Για την καλύτερη κατανόηση της ανάλυσης καλαθιού αγορών θα παραθέσουμε το παράδειγμα ενός καλαθιού αγορών, το οποίο περιέχει διάφορα προϊόντα ύστερα από μια αγορά σε ένα κατάστημα. Το καλάθι αυτό περιέχει προϊόντα όπως αλεύρι, ψωμί, γάλα και αυγά. Κάθε καλάθι μας δίνει την πληροφορία για το τι ένας πελάτης αγόρασε εκείνη την συγκεκριμένη στιγμή. Όμως μια πλήρη λίστα τέτοιων καλαθιών

αγορών από όλους τους πελάτες μας δίνει περισσότερη πληροφορία.Ο κάθε πελάτης αγοράζει διαφορετικό σύνολο από προϊόντα σε διαφορετικές ποσότητες και σε διαφορετικούς χρόνους. Η ανάλυση καλαθιού αγορών στοχεύει να μελετηθεί/αναλυθεί η πληροφορία σχετικά με το τι αγοράζουν οι πελάτες ώστε για παράδειγμα να γνωρίζουμε κατά προφορίασέγγιση προϊόντα που έχουν την τάση να αγοράζονται μαζί ή ποία προϊόντα πρέπει να προωθηθούν περισσότερο. Αυτές οι πληροφορίες αρκούν από μόνες τους να διαμορφώσουν νέες διατάξεις σε καταστήματα, να καθορίσουν ποιο προϊόν θα μπει σε ειδική προσφορά, να υποδείξουν πότε θα γίνει χρήση κουπονιών και πολλά άλλα. Η τεχνική εξόρυξης δεδομένων, που είναι στενά συνδεδεμένη με την ανάλυση καλαθιού αγορών, είναι η αυτόματη δημιουργία κανόνων συσχέτισης. Οι κανόνες συσχέτισης απεικονίζουν πρότυπα στα δεδομένα χωρίς συγκεκριμένο στόχο. Το κατά πόσο τα πρότυπα έχουν νόημα εξαρτάται από την ίδια την ερμηνεία του ανθρώπου. Οι κανόνες συσχέτισης προέρχονται από τα δεδομένα των σημείων πώλησης και περιγράφουν ποια προϊόντα πωλούνται μαζί. Παρόλο που οι ρίζες τους είναι στην ανάλυση των συναλλαγών στα σημεία πώλησης, μπορούν να εφαρμοστούν και εκτός του λιανικού εμπορίου και συγκεκριμένα στην εύρεση σχέσεων μεταξύ άλλων τύπων καλαθιού.

Μερικά παραδείγματα είναι τα ακόλουθα:

1. Εμπορεύματα που αγοράζονται μέσω πιστωτικής κάρτας όπως ενοικίαση αυτοκινήτων και δωματίων ξενοδοχείου παρέχουν πληροφορίες για το επόμενο προϊόν που πιθανόν θα αγοράσουν οι πελάτες.

2. Τραπεζικές υπηρεσίες που χρησιμοποιούνται από πελάτες λιανικής (λογαριασμοί αγορών, επενδυτικές υπηρεσίες, δάνεια αυτοκινήτων) μπορούν να προσδιορίζουν τους πελάτες που πιθανόν να θέλουν και άλλες υπηρεσίες.

3. Ασυνήθιστοι συνδυασμοί ασφαλιστικών απαιτήσεων μπορεί να είναι σημάδι απάτης και να πυροδοτήσει περαιτέρω έρευνα.

4. Ιατρικά ιστορικά ασθενών μπορεί να δώσουν ενδείξεις για πιθανές επιπλοκές που βασίζεται σε ορισμένους συνδυασμούς θεραπευτικών αγωγών.

5. Σελίδες στο Διαδίκτυο που επισκέπτονται οι χρήστες μπορεί να δώσουν πληρο-

φορίες σε προωθητικές εταιρίες για τις προτιμήσεις των χρηστών.

Η ανάλυση καλαθιού αγορών δεν παραπέμπει σε μία μοναδική τεχνική. Αναφέρεται σε μια σειρά από προσεγγίσεις που μπορούν να ακολουθηθούν ώστε να γίνουν κατανοητά τα δεδομένα των συναλλαγών στα διάφορα σημεία πώλησης. Η πιο κοινή τεχνική είναι οι κανόνες συσχέτισης

4.1 Οφέλη της ανάλυσης καλαθιού αγοράς

Η Market Basket Analysis (MBA) μπορεί να αποδειχθεί πολύ χρήσιμη βοήθεια για τις κύριες προκλήσεις που αντιμετωπίζουν σήμερα οι έμποροι λιανικής πώλησης, απαντώντας σε μια σειρά εμπορικών ζητημάτων. Οι κορυφαίοι λιανοπωλητές χρησιμοποιούν MBA για να κάνουν τις επιχειρήσεις τους πιο κερδοφόρες με τον εντοπισμό συσχετίσεων των προϊόντων στην πάροδο του χρόνου. Ένα από τα κύρια πλεονεκτήματα της ανάλυσης καλαθιού αγοράς είναι ότι είναι ιδανική για μη-κατευθυνόμενη[1] εξόρυξη δεδομένων. Η γνώση των προϊόντων που πωλούνται μαζί μπορεί να είναι πολύ χρήσιμη για κάθε επιχείρηση. Το πιο προφανές αποτέλεσμα είναι η αύξηση των πωλήσεων που μπορεί να επιτύχει ένα κατάστημα με την οργάνωση των προϊόντων του έτσι ώστε τα πράγματα που πωλούνται μαζί να βρίσκονται μαζί. Αυτό διευκολύνει την ώθηση αγοράς. Επιπλέον, αυτό έχει ως αποτέλεσμα της βελτίωσης της ικανοποίησης των πελατών. Όμως η χρήση της ανάλυσης καλαθιού αγοράς μπορεί να φανεί πάρα πολύ χρήσιμη και στην ερευνητική κοινότητα. Παρακάτω παραθέτουμε σχετικές έρευνες που διεξήχθησαν χρησιμοποιώντας την τεχνική της ανάλυσης καλαθιού αγοράς για να βελτιωθούν ζητήματα πέρα του λιανεμπορίου.

1. Οι χειρουργοί συνταγογραφούν αντιβιοτικά μετά από χειρουργική επέμβαση ανοιχτής καρδιάς, μειώνοντας έτσι τη συχνότητα μόλυνσης μετά τη χειρουργική επέμβαση. Χρησιμοποιώντας την ανάλυση καλαθιού αγοράς κατάφεραν να κατανοήσουν ποια σύνολα αντιβιοτικών συσχετίστηκαν με υψηλότερα και χαμηλότερα ποσοστά μόλυνσης μετά από χειρουργική επέμβαση καρδιάς, οδηγώντας σε

[1] Στη μη-κατευθυνόμενη ή ελεύθερη εξόρυξη δεδομένων (undirected data-mining), δεν επιλέγεται κάποια μεταβλητή ως στόχος. Ο σκοπός της είναι να δημιουργηθεί κάποια σχέση ανάμεσα σε όλες τις μεταβλητές.

συστάσεις σχετικά με τον τρόπο βελτίωσης των πρακτικών συνταγογράφησης μετά τη χειρουργική επέμβαση [31].

2. Η ασφάλεια σε περιοχές που υπάρχουν σταθμοί της πυρηνική ενέργειας είναι ζωτικής σημασίας λόγω και της πυκνότητας του πληθυσμού και της σεισμικής δραστηριότητας. Έχοντας συγκεντρωθεί μεγάλος όγκος δεδομένων από πυρηνικούς σταθμούς, οι Hibino και Niwa, χρησιμοποιώντας την ανάλυση καλαθιού αγοράς, κατάφεραν να χαρτογραφήσουν πληροφορίες σχετικά με πυρηνικά ατυχήματα από τα αρχεία πυρηνικών σταθμών της Ιαπωνίας. Τα δεδομένα που προέκυψαν από την έρευνα δείχνουν τον τρόπο βελτίωσης με τον οποίο οι πληροφορίες ασφάλειας κοινοποιούνται στο κοινό [32].

3. Τα φυτικά προϊόντα στην Κίνα καταναλώνονται ευρέως παρά την έλλειψη φαρμακοεπιδημιολογικών πληροφοριών σχετικά με αυτά. Σε αντίθεση με άλλα φάρμακα, τα φυτικά προϊόντα δεν παρακολουθούνται με αποτέλεσμα να αποσύρονται από την αγορά όταν ανακαλύπτονται προβλήματα. Οι Hsieh et al. χρησιμοποίησαν την ανάλυση καλαθιού αγοράς σε ένα σύνολο δεδομένων που παρέχεται από το εθνικό σύστημα ασφάλισης υγείας της Ταϊβάν, το οποίο περιελάμβανε χιλιάδες άτομα που είχαν λάβει συνταγή κινεζικών βοτάνων. Τα αποτελέσματα της μελέτης τους ήταν η εξαγωγή μοτίβων συγγραφής που απαιτούν περαιτέρω διερεύνηση για να διασφαλιστεί η ασφαλής κατανάλωσης βοτάνων [33].

4. Άτομα με αλλεργίες σε διάφορα τρόφιμα έχει ως αποτέλεσμα σε ορισμένες περιπτώσεις να έχουν απειλητικές για τη ζωή τους επιπτώσεις. Οι Kanagawa et al. χρησιμοποιώντας την ανάλυση καλαθιού αγοράς κατάφεραν και εντόπισαν κοινά μεμονωμένα αλλεργιογόνα τρόφιμα καθώς και συνδυασμούς αλλεργιογόνων που έχουν πανομοιότυπες πρωτεΐνες ώστε να αποφευχθούν σοβαρές για την υγεία επιπτώσεις. Επίσης, απέκλεισαν ορισμένα φερόμενα αλλεργιογόνα που δεν βρέθηκαν να σχετίζονται μεταξύ τους [34].

5. Η πρόβλεψη των κυκλώνων δεν είναι καθόλου τετριμμένη λόγω της τάσης τους για ταχεία εντατικοποίηση [2] (RI), η οποία οδηγεί σε υψηλά ποσοστά σφάλματος.

[2] Η ταχεία εντατικοποίηση είναι μια μετεωρολογική κατάσταση όπου ένας τροπικός κυκλώνας εντείνεται δραματικά σε σύντομο χρονικό διάστημα.

Οι Yang et al. χρησιμοποίησαν την ανάλυση καλαθιού αγορών σε ένα μεγάλο σύνολο δεδομένων από διάφορες εμφανίσεις κυκλώνων ώστε να εξαχθούν κατάλληλες υποθέσεις σχετικά με το ποιες είναι οι καλύτεροι παράμετροι για να εφαρμοστούν και να βγουν αποτελεσματικότερα δεδομένα [35].

4.1.1 Πλεονεκτήματα που προσφέρει η ανάλυση καλαθιού αγοράς

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν μερικές πρακτικές εφαρμογές της ανάλυσης καλαθιού αγορών όπου με το *καλάθι αγορών* νοείται μια ευρεία γκάμα δεδομένων που ανάλογα με την εφαρμογή μπορεί να εμφανίζονται (συχνά) μαζί. Παρακάτω παρουσιάζονται τα πλεονεκτήματα μιας τέτοιας ανάλυσης.

1. Περισσότερο κερδοφόρα διαφήμιση. Οι έμποροι λιανικής πώλησης χρησιμοποιούν την τεχνική MBA για να κάνουν πιο στοχευμένη διαφήμιση και προωθητικές ενέργειες ώστε οι αγοραστές να ανταποκρίνονται σε διαφορετικές προσφορές. Για παράδειγμα, η τεχνική MBA μπορεί να βοηθήσει τους λιανοπωλητές να αποφευχθούν άσκοπες εκπτώσεις, όταν και όπου οι εκπτώσεις δεν αυξάνουν συνολικά το περιθώριο μικρού κέρδους. Οι έμποροι λιανικής πώλησης, επίσης, θέλουν να διαχωρίσουν τις τάσεις των πωλήσεων από την επίδραση της διαφήμισης για να καταλάβουν τη μετατόπιση των εσόδων.

2. Η ακριβέστερη στόχευση των προσφορών βελτιώνει την απόδοση της επένδυσης ROI[3]. Το MBA χρησιμοποιείται για τη βελτιστοποίηση των καμπανιών και των προωθήσεων για περιθώρια κέρδους και αύξηση πωλήσεων με ακριβέστερη στόχευση. Για παράδειγμα, η αυξημένη ακρίβεια στη στόχευση στις προσφορές οδηγεί σε υψηλότερα ποσοστά κέρδους και επιτρέπει την πρόβλεψη των προτιμήσεων ώστε να προωθείται το κατάλληλο μείγμα προϊόντος στο σωστό πελάτη, τη σωστή στιγμή.

3. Ανάλυση των προτιμήσεων των καταναλωτών σε βάθος χρόνου. Η χρήση MBA για μεγάλο διάστημα επιτρέπει στους λιανοπωλητές να παρατηρούν την αγοραστική συμπεριφορά των πελατών στην πάροδο του χρόνου, αξιοποιώντας αυτή

[3]Απόδοση επένδυσης είναι ένας δείκτης που χρησιμοποιείται για την αξιολόγηση της απόδοσης μιας επένδυσης ή για να συγκρίνει την αποδοτικότητα διαφορετικών επενδύσεων.

τη γνώση για την καλύτερη κατανόηση των πελατών τους. Οι έμποροι λιανικής πώλησης χρησιμοποιούν τις τεχνικές MBA για να λάβουν τα δεδομένα του κύκλου ζωής του πελάτη, έτσι ώστε να μπορούν να αναλύσουν τη συμπεριφορά της αγοραστικής ζωής του πελάτη, όπως η συχνότητα αγορών ή η περίοδος αυξημένων αγορών.

4. Μπορεί να αυξηθεί το μέγεθος και η αξία του καλαθιού αγοράς. Με τα δεδομένα της πιστωτικής κάρτας, οι έμποροι λιανικής μπορούν να δουν πόσες φορές ο πελάτης ήταν στο κατάστημα και τα περιεχόμενα του καλαθιού του, και στη συνέχεια, να αξιοποιηθεί αυτή τη γνώση με στόχο την αύξηση του μεγέθους του καλαθιού. Με το MBA, μπορούν να εντοπίζουν και να στοχεύουν προωθητικές ενέργειες σε πελάτες οι οποίοι, για παράδειγμα, αγοράζουν όλες τις ανάγκες τους εκτός από συγκεκριμένα προϊόντα.

5. Μπορεί να καθοριστεί το σημείο των τέλειων τιμών για ένα κατάστημα. Σήμερα, με τη χρήση παραδοσιακών εργαλείων συλλογής πληροφοριών, η βελτιστοποίηση των τιμών μπορεί να πάρει δύο ή τρεις εβδομάδες. Οι έμποροι λιανικής πώλησης θέλουν να είναι σε θέση να χρησιμοποιούν on-demand MBA ώστε να κάνει αυτές τις αποφάσεις σε σχεδόν πραγματικό χρόνο.

6. Προσαρμογή των προϊόντων στα χαρακτηριστικά των καταναλωτών. Η κάθε επιχείρηση θα πρέπει να γνωρίζει τα χαρακτηριστικά ενός πληθυσμού στον οποία καλείται να πωλήσει ένα προϊόν. Οι προτιμήσεις ποικίλλουν ανάλογα με το κλίμα, τη μόδα ή τα δημογραφικά στοιχεία, όπως το εισόδημα, η ηλικία ή η αστική τάξη έναντι των αγροτών.

Όλα τα πλεονεκτήματα που παρουσιάστηκαν παραπάνω βρίσκουν εφαρμογή τόσο σε φυσικά όσο και σε ψηφιακά καταστήματα αγορών.

4.2 Πως ενδυναμώνεται η επιχείρηση μέσω της χρήσης ανάλυσης καλαθιού αγοράς

Οι έμποροι πρέπει να δουν τις τάσεις πιο μακροπρόθεσμα για να αποφασίσουν πόσο να αγοράσουν και πως κάποιο προϊόν ταιριάζει στο επιχειρηματικό μοντέλο.

Εδώ παραθέτονται μερικοί τρόποι πως κορυφαίοι λιανοπωλητές χρησιμοποιούν την τεχνική MBA. Οι λιανοπωλητές χρειάζονται καλύτερα εργαλεία για να βοηθήσουν τους σχεδιαστές και τους εμπόρους. Ο προγραμματισμός είναι ένα από τα ταχύτερα αναπτυσσόμενα μέρη του λιανικού εμπορίου. Ένας καλός σχεδιασμός ρωτά: «Πόσο θα πρέπει να αγοράσουν οι καταναλωτές, πώς θα πρέπει να εμφανίζεται ένα προϊόν και ποια είναι η διάρκεια του κύκλου ζωής του προϊόντος» ή «Μπορούμε να πουλήσουμε αρκετά γρήγορα για να καλυφθεί το ύψος των δαπανών·». Αυτά τα υψηλής αξίας ερωτήματα και άλλες αποφάσεις υψηλού κινδύνου μπορούν να απαντηθούν ικανοποιητικά με τη χρήση MBA. Η ανάλυση MBA δίνει τη δυνατότητα στους εμπόρους να αγοράζουν πιο έξυπνα και να ενισχύουν τη διαπραγματευτική θέση τους με τους προμηθευτές, παρέχοντας τους εμπόρους καλύτερες πληροφορίες σχετικά με την αγοραστική συμπεριφορά των πελατών. Ενώ ορισμένοι λιανοπωλητές προτιμούν να περιορίζουν τη χρήση του MBA στους σχεδιαστές, οι λιανοπωλητές πειραματίζονται όλο και περισσότερο με την παροχή στους εμπόρους δομημένων αλλά εύχρηστων εργαλείων MBA. Οι έμποροι επικεντρώνονται στην ανάγκη για την αγορά των αποθεμάτων, το σχεδιασμό και την διάθεση, αλλά μπορεί επίσης να ασχολούνται με τη διαφήμιση και προωθητικές ενέργειες. Η τεχνική MBA μπορεί να βοηθήσει στη βελτίωση των αποφάσεων διάθεσης και αποθήκευσης, αλλά και στην καλύτερη κατανόηση της εποχικής ζήτησης.

4.3 Προβλήματα της μεθόδου ανάλυσης καλαθιού αγοράς

Όλες οι τεχνικές έχουν τα δικά τους πλεονεκτήματα και μειονεκτήματα. Αυτή η ενότητα παρέχει μερικά από τα μειονεκτήματα των αλγορίθμων, αλλά και τεχνικές για την υπέρβαση αυτών των δυσκολιών.

Μεταξύ των μεθόδων που συζητούνται για την εξόρυξη δεδομένων, ο αλγόριθμος Apriori [8] θεωρείται ο καλύτερος για την εξόρυξη κανόνων συσχετίσεων. Όμως υπάρχουν διάφορες δυσκολίες που αντιμετωπίζει ο αλγόριθμος Apriori, οι οποίες είναι:

1. Σαρώνει τη βάση δεδομένων πολλές φορές ξεκινώντας από τα μεμονωμένα στοιχεία που εμφανίζονται συχνά στη βάση δεδομένων και σε κάθε επανάληψη δη-

μιουργεί όλο και μεγαλύτερα σύνολα στοιχείων που εμφανίζονται συχνά μαζί. Η εξερεύνηση υποσυνόλων από πάνω προς τα κάτω και η πολλαπλή σάρωση της βάσης δεδομένων μειώνει σημαντικά την συνολική απόδοση της διαδικασίας.

2. Με την δημιουργία μεγάλου πλήθους υποσυνόλων, υποθέτοντας επίσης ότι όλα τα δεδομένα είναι στην κύρια μνήμη, τόσο η χρονική όσο και η χωρική πολυπλοκότητα του αλγορίθμου είναι πολύ ψηλή (αυξάνεται εκθετικά ως προς το πλήθος των δεδομένων).

Αυτά τα μειονεκτήματα για να μπορούν να ξεπεραστούν έπρεπε να γίνει μια αποτελεσματική τροποποίηση του αλγορίθμου Apriori που είχε ως αποτέλεσμα την δημιουργία του αλγορίθμου FP-Growth [2].

4.4 Χρήση ανάλυση καλαθιού αγοράς στην προτεινόμενη προσέγγιση

Η ανάλυση καλαθιού αγοράς όπως αναφέραμε και προηγούμενός στα παραπάνω υποκεφάλαια μπορεί να βρει εφαρμογή σε πολλούς τομείς με μεγάλη επιτυχία. Έτσι και εμείς χρησιμοποιήσαμε την ανάλυση καλαθιού αγοράς για να μπορέσουμε να ανακαλύψουμε συνδέσεις σε προϊόντα όπου είναι διαθέσιμα σε καταστήματα λιανικής. Ειδικότερα, έχοντας στην κατοχή μας σύνολα δεδομένων όπου ήταν διαθέσιμα online από αγορές που έχουν πραγματοποιηθεί, εφαρμόσαμε τον αλγόριθμο FP-Growth με αποτέλεσμα να εξαχθούν συχνά σύνολα προϊόντων τα οποία έχουν αγοραστεί μαζί. Στη συνέχεια, με την βοήθεια των συχνών συνόλων δεδομένων όπου έχουμε στην διάθεσή μας και έχουν προκύψει μετά από την εφαρμογή του αλγορίθμου, μπορούν ληφθούν αποφάσεις για μελλοντικές στρατηγικές κινήσεις μιας εταιρίας αλλά και για την βελτίωση της κατανόησης αντιλήψεων των καταναλωτών για τα προϊόντα, όπου και συνεπάγεται η εφαρμογή πολλών νέων αλλαγών ενός επιχειρησιακού πλάνου μιας επιχείρησης.

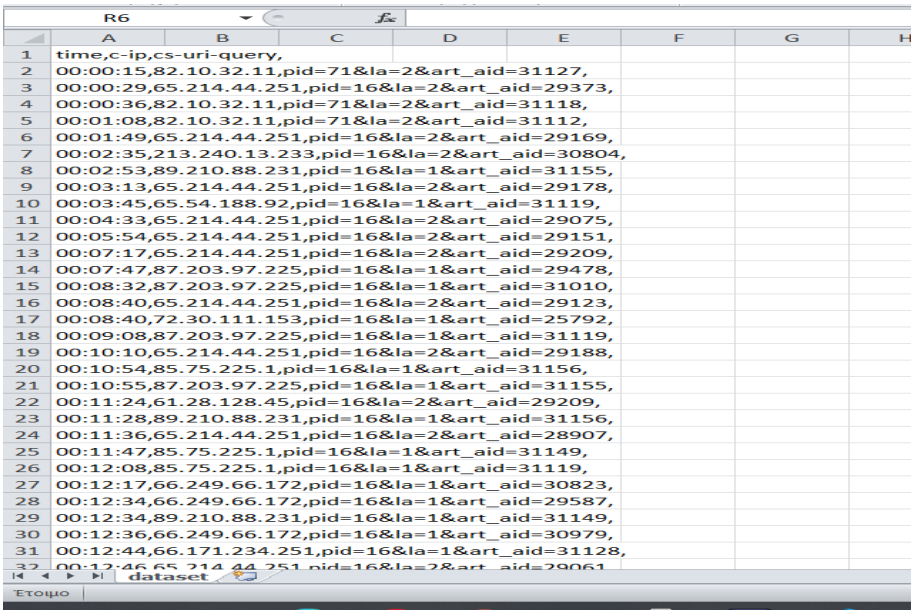
Κεφάλαιο 5

Πειράματα, αποτελέσματα και συμπεράσματα

Στο παρόν κεφάλαιο παρουσιάζουμε την πειραματική αξιολόγηση των προτεινόμενων αλγορίθμων. Πιο συγκεκριμένα, παρουσιάζουμε τα δεδομένα τα οποία χρησιμοποιήθηκαν και στη συνέχεια τις διαφορετικές προσεγγίσεις με τις οποίες κατανεμήθηκαν στους κόμβους του συστήματος που αναπτύχθηκε θέλοντας να εξάγουμε συχνά πρότυπα. Για όλες τις προσεγγίσεις παίρνουμε μετρικές, τις οποίες στη συνέχεια συγκρίνουμε, σχολιάζουμε και αξιολογούμε.

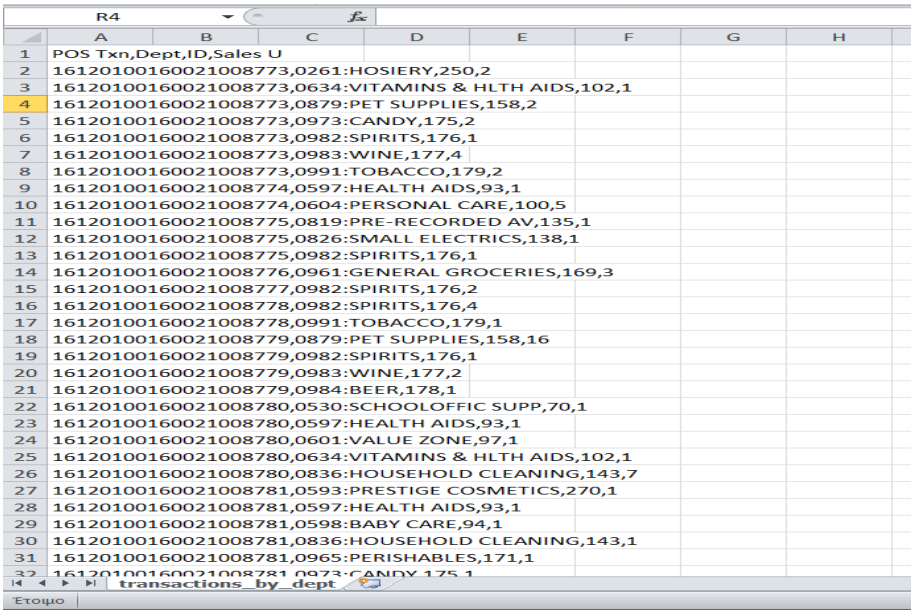
5.1 Σύνολα δεδομένων

Τα δεδομένα που θα επεξεργαστούμε και θα αναλύσουμε έχουν συλλεχθεί από on-line αγορές (π.χ. Σουπερ Μαρκετ) στα οποία εφαρμόσαμε τον αλγόριθμο FP-Growth σε ένα Chord κατανεμημένο περιβάλλον προκειμένου να εξάγουμε συχνά πρότυπα. Παρακάτω παρουσιάζουμε δύο εικόνες από τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία.



	A
1	time,c-ip,cs-uri-query,
2	00:00:15,82.10.32.11,pid=71&la=2&art_aid=31127,
3	00:00:29,65.214.44.251,pid=16&la=2&art_aid=29373,
4	00:00:36,82.10.32.11,pid=71&la=2&art_aid=31118,
5	00:01:08,82.10.32.11,pid=71&la=2&art_aid=31112,
6	00:01:49,65.214.44.251,pid=16&la=2&art_aid=29169,
7	00:02:35,213.240.13.233,pid=16&la=2&art_aid=30804,
8	00:02:53,89.210.88.231,pid=16&la=1&art_aid=31155,
9	00:03:13,65.214.44.251,pid=16&la=2&art_aid=29178,
10	00:03:45,65.54.188.92,pid=16&la=1&art_aid=31119,
11	00:04:33,65.214.44.251,pid=16&la=2&art_aid=29075,
12	00:05:54,65.214.44.251,pid=16&la=2&art_aid=29151,
13	00:07:17,65.214.44.251,pid=16&la=2&art_aid=29209,
14	00:07:47,87.203.97.225,pid=16&la=1&art_aid=29478,
15	00:08:32,87.203.97.225,pid=16&la=1&art_aid=31010,
16	00:08:40,65.214.44.251,pid=16&la=2&art_aid=29123,
17	00:08:40,72.30.111.153,pid=16&la=1&art_aid=25792,
18	00:09:08,87.203.97.225,pid=16&la=1&art_aid=31119,
19	00:10:10,65.214.44.251,pid=16&la=2&art_aid=29188,
20	00:10:54,85.75.225.1,pid=16&la=1&art_aid=31156,
21	00:10:55,87.203.97.225,pid=16&la=1&art_aid=31155,
22	00:11:24,61.28.128.45,pid=16&la=2&art_aid=29209,
23	00:11:28,89.210.88.231,pid=16&la=1&art_aid=31156,
24	00:11:36,65.214.44.251,pid=16&la=2&art_aid=28907,
25	00:11:47,85.75.225.1,pid=16&la=1&art_aid=31149,
26	00:12:08,85.75.225.1,pid=16&la=1&art_aid=31119,
27	00:12:17,66.249.66.172,pid=16&la=1&art_aid=30823,
28	00:12:34,66.249.66.172,pid=16&la=1&art_aid=29587,
29	00:12:34,89.210.88.231,pid=16&la=1&art_aid=31149,
30	00:12:36,66.249.66.172,pid=16&la=1&art_aid=30979,
31	00:12:44,66.171.234.251,pid=16&la=1&art_aid=31128,

Σχήμα 5.1: Σύνολο δεδομένων 1



	A
1	POS Txn,Dept,ID,Sales U
2	16120100160021008773,0261:HOSIERY,250,2
3	16120100160021008773,0634:VITAMINS & HLTH AIDS,102,1
4	16120100160021008773,0879:PET SUPPLIES,158,2
5	16120100160021008773,0973:CANDY,175,2
6	16120100160021008773,0982:SPIRITS,176,1
7	16120100160021008773,0983:WINE,177,4
8	16120100160021008773,0991:TOBACCO,179,2
9	16120100160021008774,0597:HEALTH AIDS,93,1
10	16120100160021008774,0604:PERSONAL CARE,100,5
11	16120100160021008775,0819:PRE-RECORDED AV,135,1
12	16120100160021008775,0826:SMALL ELECTRICS,138,1
13	16120100160021008775,0982:SPIRITS,176,1
14	16120100160021008776,0961:GENERAL GROCERIES,169,3
15	16120100160021008777,0982:SPIRITS,176,2
16	16120100160021008778,0982:SPIRITS,176,4
17	16120100160021008778,0991:TOBACCO,179,1
18	16120100160021008779,0879:PET SUPPLIES,158,16
19	16120100160021008779,0982:SPIRITS,176,1
20	16120100160021008779,0983:WINE,177,2
21	16120100160021008779,0984:BEER,178,1
22	16120100160021008780,0530:SCHOOLOFFIC SUPP,70,1
23	16120100160021008780,0597:HEALTH AIDS,93,1
24	16120100160021008780,0601:VALUE ZONE,97,1
25	16120100160021008780,0634:VITAMINS & HLTH AIDS,102,1
26	16120100160021008780,0836:HOUSEHOLD CLEANING,143,7
27	16120100160021008781,0593:PRESTIGE COSMETICS,270,1
28	16120100160021008781,0597:HEALTH AIDS,93,1
29	16120100160021008781,0598:BABY CARE,94,1
30	16120100160021008781,0836:HOUSEHOLD CLEANING,143,1
31	16120100160021008781,0965:PERISHABLES,171,1

Σχήμα 5.2: Σύνολο δεδομένων 2

Αρχικά στην πρώτη Εικόνα 5.1 τα πεδία τα οποία αναφέρονται είναι τα:

1. Το time όπου προσδιορίζεται η ώρα αγοράς που επιλέχθηκε το προϊόν.
2. Το c-ip είναι το αναγνωριστικό κάθε πελάτη.
3. Το cs-uri-querly είναι το αναγνωριστικό του κάθε προϊόντος.

Αντίστοιχά, στην Εικόνα 5.2 τα πεδία που αναφέρονται είναι τα:

1. Το POS Txn όπου μας δείχνει τον αριθμό συναλλαγής (βάση του οποίου διαφοροποιούνται οι πελάτες).
2. Το Dept μας περιγράφει το αντικείμενο.
3. Το ID είναι ο κωδικός του κάθε αντικειμένου.
4. Το Sales U αναφέρεται στις μονάδες που αγοράστηκαν από το αντικείμενο.

Αυτές οι πληροφορίες/χαρακτηριστικά που μας δίνονται από τα δεδομένα μας, βοηθάνε να σκεφτούμε και να βρούμε πώς θα κατανεμηθούν καταλλήλως σε κόμβους τα δεδομένα τα οποία έχουμε ώστε να εξάγουν όσο το δυνατόν πιο έγκυρα αποτελέσματα όταν μπουν σε διαδικασία εύρεσης συχνών συνόλων. Πιο συγκεκριμένα, επιλέγοντας τον τρόπο όπου κάθε κόμβος αναλαμβάνει τα δεδομένα με βάση το προϊόν θα έχουμε ως αποτέλεσμα ο κόμβος να είναι υπεύθυνος για ένα συγκεκριμένο προϊόν και να περιέχει όλα τα προϊόντα που επιλέχθηκαν και αγοράστηκαν μαζί. Έτσι, το κάθε καλάθι αγορών (κόμβος) θα περιέχει μόνο τις χρήσιμες πληροφορίες που μας ενδιαφέρουν, δηλαδή το ποια προϊόντα προτιμούνται και αγοράζονται με ένα συγκεκριμένο προϊόν (το προϊόν που είναι υπεύθυνος ο κόμβος). Αντίθετα, σε διαφορετική περίπτωση που επιλέξουμε να χωρίσουμε τα δεδομένα σε κόμβους με κάποιο άλλο πεδίο, για παράδειγμα, το αναγνωριστικό του κάθε πελάτη, το καλάθι αγοράς θα γεμίσει με τα προϊόντα που έχει αγοράσει και προτιμά ένας συγκεκριμένος πελάτης (μεμονωμένα), αυτό θα έχει ως αποτέλεσμα να λάβουμε μη έγκυρα αποτελέσματα για να ληφθεί μια σημαντική απόφαση για τις προτιμήσεις του γενικού συνόλου των καταναλωτών. Γι αυτόν τον λόγο είναι πολύ σημαντική η απόφαση που

θα παρθεί πως θα χωριστούν τα δεδομένα διότι ασχέτως πόσο καλός και αποτελεσματικός είναι ο αλγόριθμος που θα επιλέξουμε εάν εισάγουμε εσφαλμένα δεδομένα ως είσοδο θα λάβουμε και παραπλανητικά αποτελέσματα.

5.2 Παράμετροι

Επιπλέον, προκειμένου να εξασφαλίσουμε ένα ποιοτικό σύνολο δεδομένων απαλλαγμένο από προβλήματα, όπως για παράδειγμα ελλειπή δεδομένα (μη συμπληρωμένα ή διαγραμμένα πεδία), λανθασμένα δεδομένα (λανθασμένες τιμές ή ακραίες τιμές), ασυνέπειες δεδομένων κ.α., θα πρέπει να γίνουν μια σειρά από ενέργειες και να ληφθούν υπόψη κάποιοι παράμετροι. Ένα απαραίτητο κομμάτι της ανάλυσης του συνόλου δεδομένων αποτελεί και ο καθαρισμός των δεδομένων, κατά τον οποίο εντοπίζονται και διορθώνονται ή αφαιρούνται οι ανακριβείς και κατεστραμμένες τιμές από το σύνολο των δεδομένων. Στην συγκεκριμένη περίπτωση έπρεπε να γίνει ένας αυτοματοποιημένος έλεγχος μέσα στο πρόγραμμα που υλοποιήθηκε ώστε οι μεταβλητές όπου δίνονται χωρίς συγκεκριμένη μορφολογία να σπάνε και να εισάγονται σωστά ώστε να μην υπάρχουν λανθασμένες καταχωρίσεις.

Όλα τα παραπάνω ως γνωστόν δύναται να επηρεάσουν αρνητικά την πορεία της ανάλυσης και να οδηγήσουν σε ψευδείς κανόνες κατά την διαδικασία της κατηγοριοποίησης. Από το αρχικό σύνολο δεδομένων αφαιρέθηκαν τα πεδία που δεν μας βοηθούσαν στην ολοκλήρωση του συνόλου όπως για παράδειγμα στο πρώτο σύνολο δεδομένων (Εικόνα 5.1) το time διότι δεν μας ενδιαφέρει ο χρόνος αγοράς του προϊόντος ενώ από το δεύτερο σύνολο δεδομένων (Εικόνα 5.2) αφαιρέθηκαν τα ID και Sales U, δεν ασχολούμαστε με το ID αλλά με το όνομα του προϊόντος καθ αυτού. Στην συνέχεια, προκειμένου να εξαχθούν χρήσιμοι κανόνες θα πρέπει να εξασφαλίσουμε κάποια ελάχιστη υποστήριξη στο υπάρχον σύνολο δεδομένων. Δηλαδή, στην περίπτωση που εξαχθεί κάποιος κανόνας να έχουμε ικανοποιητικό αριθμό περιπτώσεων στις οποίες να εφαρμόζεται αυτός ο κανόνας. Άλλος ένας τρόπος που εξετάστηκε ήταν να σπάσουμε ολόκληρο το σύνολο των δεδομένων σε ποσοστά (π.χ. 20%, 40%, ...). Πιο συγκεκριμένα, έχοντας τον συνολικό αριθμό των εγγραφών του αρχείου υπολογίζουμε πόσες εγγραφές θα περιέχονται στο 20%, 40% κ.ο.κ., έτσι

κάθε φορά λαμβάνουμε υπόψιν μας μόνο τις εγγραφές που έχουν προκύψει από τον υπολογισμό των ποσοστών, αυτό μας βοηθάει να παρατηρήσουμε την διαφορά που θα είχε στην εξαγωγή των κανόνων συσχέτισης στην περίπτωση που τα δεδομένα μας ήταν ελλιπής και κατά πόσο το δυνατόν τα αποτελέσματα που θα μας δοθούν θα είναι αντιπροσωπευτικά για να χρησιμοποιηθούν και να ληφθούν αποφάσεις. Παρακάτω, μπορούμε να δούμε τα αποτελέσματα που εξήχθησαν από διάφορα ποσοστά του συνολικού αρχείου.

```
192
==========================
{'275'}
{'275', '192'}
{'37', '275', '192'}
{'275', '37'}
{'192'}
{'37', '192'}
{'37'}


===========================

[{'275'}, {'192'}, 1.0]
[{'192'}, {'275'}, 0.775]
[{'37'}, {'275', '192'}, 0.8723404255319149]
[{'275'}, {'192', '37'}, 0.9919354838709677]
[{'192'}, {'275', '37'}, 0.76875]
[{'275', '37'}, {'192'}, 1.0]
[{'192', '37'}, {'275'}, 0.8723404255319149]
[{'275', '192'}, {'37'}, 0.9919354838709677]
[{'275'}, {'37'}, 0.9919354838709677]
[{'37'}, {'275'}, 0.8723404255319149]
[{'37'}, {'192'}, 1.0]
[{'192'}, {'37'}, 0.88125]
===========================
```

Σχήμα 5.3: Αποτελέσματα του 20% απ΄ το σύνολο δεδομένων

```
192
==========================
{'37'}
{'37', '153'}
{'153'}


==========================

[{'37'}, {'153'}, 0.7352941176470589]
[{'153'}, {'37'}, 0.6410256410256411]
==========================
```

Σχήμα 5.4: Αποτελέσματα του 40% απ΄ το σύνολο δεδομένων

```
192
==========================
{'192'}
{'192', '275'}
{'37', '192', '275'}
{'37', '192'}
{'275'}
{'37', '275'}
{'37'}


==========================

[{'275'}, {'192'}, 1.0]
[{'275'}, {'37', '192'}, 0.984]
[{'37', '275'}, {'192'}, 1.0]
[{'192', '275'}, {'37'}, 0.984]
[{'37'}, {'192'}, 1.0]
[{'275'}, {'37'}, 0.984]
==========================
```

Σχήμα 5.5: Αποτελέσματα του 60% απ΄ το σύνολο δεδομένων

```
192
==========================
{'275'}
{'275', '37'}
{'37'}


==========================

[{'275'}, {'37'}, 0.9851301115241635]
[{'37'}, {'275'}, 0.9013605442176871]
==========================
```

Σχήμα 5.6: Αποτελέσματα του 80% απ΄ το σύνολο δεδομένων

```
192
==========================
{'163'}
{'163', '153'}
{'163', '156', '153'}
{'163', '156'}
{'156'}
{'153', '156'}
{'153'}


==========================

[{'163'}, {'153'}, 0.6210526315789474]
[{'153'}, {'163'}, 0.6781609195402298]
[{'163', '156'}, {'153'}, 0.7692307692307693]
[{'163', '153'}, {'156'}, 0.847457627118644]
[{'153', '156'}, {'163'}, 0.7352941176470589]
[{'163'}, {'156'}, 0.6842105263157895]
[{'156'}, {'163'}, 0.7647058823529411]
[{'153'}, {'156'}, 0.7816091954022989]
[{'156'}, {'153'}, 0.8]
==========================
```

Σχήμα 5.7: Αποτελέσματα από ολόκληρο το σύνολο δεδομένων

Αναλύοντας τις παραπάνω εικόνες που βλέπουμε έχουμε ότι, στο πάνω μέρος αναγράφεται το προϊόν όπου έχει αναλάβει κάποιος κόμβος του συστήματος (στην προκείμενη περίπτωση το 192). Στη συνέχεια, παραθέτονται τα πιο συχνά σύνολα δεδομένων που εμφανίζονται κατά το πέρασμα του αλγορίθμου στον κόμβο (στα σημεία που απεικονίζεται μόνο ένα προϊόν βασίζεται στο γεγονός εμφάνισης των δεδομένων από τον αλγόριθμο διότι είναι ένα προϊόν με μεγάλο ποσοστό εμφάνισης). Στο τελευταίο μέρος που διακρίνουμε ποσοστά δίπλα από τα σύνολα δεδομένων μας υποδεικνύει πόσο πιθανόν είναι να προκύψει αυτό το καλάθι αγορών. Παρατηρώντας και τις τέσσερις εικόνες (Εικόνα 5.3, 5.4, 5.5, 5.6 και 5.7) διακρίνουμε εμφανείς διαφορές στα συχνά σύνολα και αυτό οφείλεται στο γεγονός ότι έχουμε κόψει εγγραφές από ολόκληρο το σύνολο με αποτέλεσμα να μην είναι απολύτως έγκυρα τα αποτελέσματα, έτσι το βέλτιστο θα ήταν να έχουμε όσο το δυνατόν λιγότερες ελλείψεις. Βέβαια, μια χρήση που θα είχε το σπάσιμο τον δεδομένων θα ήταν σε περίπτωση που θα θέλαμε να εξάγουμε σύνολο κανόνων βασισμένο σε συγκεκριμένη χρονική περίοδο.

5.3 Αποτελέσματα

Στην μέχρι τώρα πορεία της εργασίας επικεντρωθήκαμε στην εύρεση «συχνών» συνόλων αντικειμένων (frequent itemsets) μέσα από συναλλαγές (transactions), με υλοποίηση του αλγορίθμου FP-Growth σε ένα Chord κατανεμημένο περιβάλλον προκειμένου να εξάγουμε συχνά πρότυπα πάνω από σύνολα δεδομένων. Το επόμενο βήμα που προκύπτει από την υλοποίηση του συστήματος είναι η εξόρυξη κανόνων συσχέτισης, δηλαδή η εύρεση μορφών $X \rightarrow \Upsilon$ για τα οποία θα ισχύει $\text{conf}(X \rightarrow Y) > \text{minconf}$. Ως minconf ορίζουμε ένα κατώφλι ελάχιστης εμπιστοσύνης πάνω από το οποίο γίνεται αποδεκτός ένας κανόνας. Πιο τυπικά αν $I = \{i_1,i_2,i_3,..,i_m\}$ ένα σύνολο από διακριτά αντικείμενα (items) και D ένα σύνολο από δοσοληψίες (transactions) όπου κάθε δοσοληψία T είναι ένα υποσύνολο αντικειμένων του I τότε ένας κανόνας συσχέτισης ορίζεται ως εξής: $X \rightarrow Y$ όπου τα X, Y είναι υποσύνολο του I και $X \cap Y = \varnothing$. Ο συνολικός αριθμός των πιθανών κανόνων που εξάγονται από ένα σύνολο

δεδομένων είναι

$$R = 3^d - 2^d + 1 + 1$$

όπου d ο αριθμός των αντικειμένων. Όπως γίνεται εύκολα αντιληπτό ότι πρόκειται για ένας όγκος κανόνων, ο οποίος δεν είναι καθόλου εύκολο να διαχειριστούμε [38] [39]. Από αυτά εμείς θα πρέπει να ξεχωρίσουμε τα πιο συχνά εμφανιζόμενα υποσύνολα, τα οποία και θα αποτελέσουν τα συστατικά μέρη των κανόνων μας. Για το σκοπό αυτό χρησιμοποιούμε τις παρακάτω ποσότητες:

- Υποστήριξη (Support) ενός κανόνα. Ορίζεται ως η υποστήριξη του συνόλου των items του κανόνα δηλαδή $\sup(X \rightarrow Y) = \sup(X \cup Y)$. Ουσιαστικά μας δείχνει πόσες φορές εμφανίζεται το σύνολο αντικειμένων του κανόνα.
- Εμπιστοσύνη (Confidence) ενός κανόνα. Ορίζεται ως $\text{conf}(X \rightarrow Y) = \sup(X \cup Y) \ / \ \sup(X)$ και καθορίζει τον αριθμό των εμφανίσεων του Υ στα transaction που περιέχουν το X. Ουσιαστικά μας δείχνει την ισχύ της συνεπαγωγής του κανόνα που εξετάζουμε.

Η διαδικασία εύρεσης κάθε δυνατού συνδυασμού X, Y μέσα από το dataset συνιστά μια πολύπλοκη και χρονοβόρα διαδικασία. Με βάση τον ορισμό της εμπιστοσύνης ισχύει ότι $\text{conf}(X \rightarrow Y) = \sup(X \cup Y) \ / \ \sup(X)$. Στην προκειμένη περίπτωση ανάλογα με το πλήθος των προϊόντων που αναλαμβάνει κάθε κόμβος (όλα τα προϊόντα που αγοράστηκαν μαζί με το προϊόν που είναι υπεύθυνος ο κάθε κόμβος) υπήρχε και διαφορά στις τιμές υποστήριξης. Πιο συγκεκριμένα, για μπορέσουν να προκύψουν αρκετά συχνά αντικείμενα από ένα σύνολο δεδομένων πρέπει να υπάρχει και ένα σωστό όριο κατώτατου κατωφλίου. Στην περίπτωση που έχουμε πολύ μικρή τιμή κατωφλίου σημαίνει πολλά συχνά αντικείμενα άρα και πολλά συχνά πρότυπα, στην περίπτωση που ένας κόμβος δεν περιέχει πολλά προϊόντα, (δεν είναι πολύ δημοφιλές το προϊόν) θα πρέπει να εισάγουμε στις τιμές υποστήριξης ένα χαμηλό ποσοστό ώστε ο αλγόριθμος FP-Growth να καταφέρει να εξάγει ένα επαρκές πλήθος συχνών προτύπων για να γίνει μια πιο έγκυρη έρευνα και να έχουμε όσο το δυνατόν πιο έγκυρα αποτελέσματα. Στη συνέχεια, παραθέτουμε τα αποτελέσματα που προέκυψαν από τα σύνολα δεδομένων που αναφέραμε παραπάνω (Εικόνα 5.1 και 5.2)

```
337
==========================
{'337'}
{'337', '5'}
{'337', '5', '330'}
{'337', '330'}
{'5'}
{'330', '5'}
{'330'}


==========================

[{'5'}, {'337'}, 1.0]
[{'5'}, {'337', '330'}, 0.8389423076923077]
[{'337', '5'}, {'330'}, 0.8389423076923077]
[{'330', '5'}, {'337'}, 1.0]
[{'330'}, {'337'}, 1.0]
[{'5'}, {'330'}, 0.8389423076923077]
==========================
```

Σχήμα 5.8: Αποτελέσματα πρώτου συνόλου δεδομένων

```
OUTDOOR LIVING
==========================
{'SLEEPWEARLOUNGEWEAR'}
{'OUTDOOR LIVING', 'SLEEPWEARLOUNGEWEAR'}
{'HEALTH AIDS'}
{'HEALTH AIDS', 'OUTDOOR LIVING'}
{'OUTDOOR LIVING'}


==========================

[{'OUTDOOR LIVING'}, {'SLEEPWEARLOUNGEWEAR'}, 0.2727272727272727]
[{'SLEEPWEARLOUNGEWEAR'}, {'OUTDOOR LIVING'}, 1.0]
[{'HEALTH AIDS'}, {'OUTDOOR LIVING'}, 1.0]
[{'OUTDOOR LIVING'}, {'HEALTH AIDS'}, 0.2727272727272727]
==========================
```

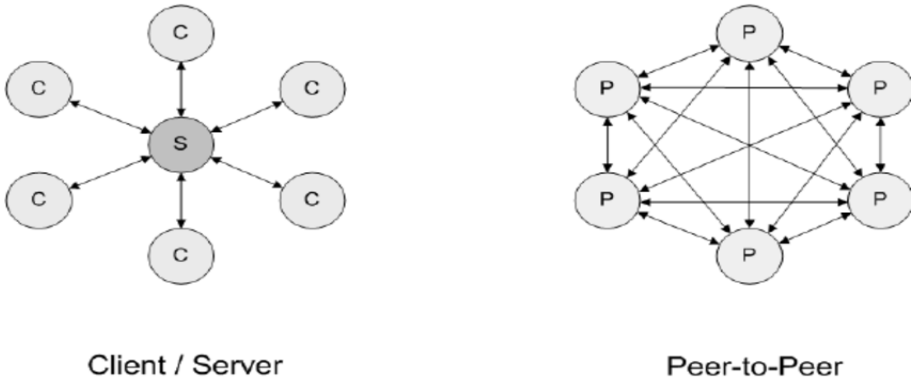
Σχήμα 5.9: Αποτελέσματα δεύτερου συνόλου δεδομένων

Στις Εικόνες 5.8 και 5.9 αρχικά αναγράφεται το προϊόν (που είναι υπεύθυνος ο εκάστοτε κόμβος), στη συνέχεια εμφανίζονται τα πιο συχνά προϊόντα που συναντά ο αλγόριθμος κατά τις σαρώσεις στην βάση δεδομένων και τέλος βλέπουμε την πιθανότητα διαμόρφωσης από τα βασικά προϊόντα που θα έχει το καλάθι αγορών. Εδώ παρατηρούμε ότι στις τελευταίες γραμμές όπου αναγράφονται τα ποσοστά εμφάνισης του καλαθιού αγοράς μπορεί φαινομενικά να είναι ίδιες λόγο του ότι περιέχουν τα ίδια προϊόντα αλλά είναι τελείως διαφορετικές, δηλαδή σε κάθε διαφορετική γραμμή εμφανίζεται η πιθανότητα έχοντας ο καταναλωτής ξεκινήσει και εν τέλει επιλέξει για αγορά διαφορετικής σειράς προϊόντα ασχέτως που η τελική μορφή του καλαθιού αγοράς φαίνεται ίδια.

5.4 Σύγκρισή κατανεμημένης με κεντρικοποιημένης προσέγγισης

Η κεντρική βάση δεδομένων βασίζεται στην τεχνική όπου όλα τα δεδομένων αποθηκεύονται και διατηρούνται σε μία θέση. Αυτή είναι η παραδοσιακή προσέγγιση για την αποθήκευση και διαχείριση δεδομένων. Αντίθετα, η κατανεμημένη βάση δεδομένων έχει ως κύριο χαρακτηριστικό να αποθηκεύει όλα τα δεδομένα σε συσκευές αποθήκευσης που δεν βρίσκονται στην ίδια φυσική θέση, αλλά η βάση δεδομένων ελέγχεται χρησιμοποιώντας ένα κεντρικό σύστημα διαχείρισης βάσεων δεδομένων. Τα κεντρικά συστήματα διαχείρισης βάσεων δεδομένων υλοποιούνται όταν τα στοιχεία τους βρίσκονται, αποθηκεύονται και διατηρούνται σε μια ενιαία θέση. Αυτή η θέση είναι πολύ συχνά ένας κεντρικός υπολογιστής. Η κατανεμημένη προσέγγιση υλοποιήθηκε με χρήση των συστημάτων P2P όπου έχουν υιοθετηθεί από ποικίλες εφαρμογές, διαφόρων κατηγοριών όπου και περιλαμβάνουν κατανεμημένες βάσεις δεδομένων στους οποίους οι κόμβοι οργανώνονται σε ένα δομημένο γράφο. Στα αντικείμενα που εισάγονται στο σύστημα αντιστοιχίζεται ένα κλειδί και η τοποθέτησή τους στους κόμβους γίνεται με προκαθορισμένο τρόπο έτσι ώστε να διευκολύνεται η αναζήτησή τους και να επιτυγχάνεται κλιμάκωση. Η χρήση ενός Peer-to-Peer (P2P) δικτύου υπολογιστών ή ενός δικτύου ομότιμων τερματικών υπάρχει για να χρησιμοποιηθεί πολλαπλή συνδεσιμότητα μεταξύ των συμμετεχόντων στο δίκτυο, σε αντίθεση με

το παραδοσιακό μοντέλο όπου ένας μικρός αριθμός διακομιστών (servers) παρέχει τους υπολογιστικούς πόρους για την παροχή υπηρεσιών. Ένα τέτοιο δίκτυο μπορούμε να το χρησιμοποιήσουμε τόσο για μεταφορά αρχείων (file sharing) αλλά και για μεταφορά πραγματικού χρόνου δεδομένων (real-time data-telephony traffic). Σε ένα P2P δίκτυο δεν υπάρχει η έννοια του client ή του server, παρά μόνο υπάρχουν ισοδύναμοι Peers, οι λεγόμενοι nodes, οι οποίοι συμπεριφέρονται ταυτόχρονα σαν clients και servers για τους άλλους nodes του δικτύου. Βλέπουμε τη διαφορά των δύο εννοιών γραφικά στη παρακάτω εικόνα.



Σχήμα 5.10: Μοντέλο client/server και peer-to-peer

Τα P2P συστήματα είναι κατανεμημένες αρχιτεκτονικές που αποτελούνται από διασυνδεδεμένους κόμβους οι οποίοι είναι σε θέση να αυτό οργανωθούν σε τοπολογίες δικτύων με σκοπό το διαμοιρασμό πόρων (περιεχομένου, αποθηκευτικού χώρου, κύκλων CPU) με απ' ευθείας ανταλλαγή δεδομένων χωρίς τη μεσολάβηση κάποιου ενδιάμεσου. Επίσης χαρακτηρίζονται από την ικανότητά τους να προσαρμόζονται σε τυχών αποτυχίες και πληθυσμιακές αλλαγές κόμβων που μπορεί να προκύψουν, διατηρώντας παράλληλα ικανοποιητικά επίπεδα συνδεσιμότητας και απόδοσης. Τα P2P δίκτυα δεν βασίζονται σε ένα κεντρικό υπολογιστή ώστε να ελέγχει το σύστημα δικτύωσης, με αποτέλεσμα να παρέχονται ασταμάτητα δεδομένα μεταξύ των χρηστών χωρίς να επηρεάζεται η λειτουργία του δικτύου από τυχών υπάρξεις βλάβης σε κόμβο με πολλούς χρήστες, καθώς και ότι δεν υπάρχει κανένα data traffic λόγω της αρχιτεκτονικής του. Αυτό το καθιστά αρκετά ισχυρό σε Dos[1] επιθέσεις. Ένα σημαντικό

[1]Επιθέσεις άρνησης εξυπηρέτησης ονομάζονται γενικά οι επιθέσεις εναντίον ενός υπολογιστή, ή μιας υπηρεσίας που παρέχεται, οι

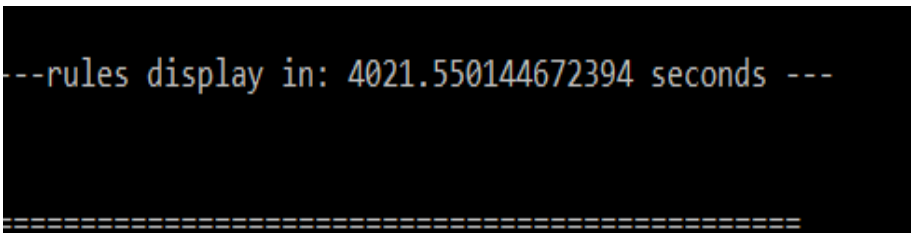
στοιχείο των ομότιμων δικτύων είναι πως μπορούν να αποτελούνται από προσωπικούς υπολογιστές με αποτέλεσμα η ισχύς τους να είναι σε πολύ υψηλό επίπεδο σε σχέση με άλλα δίκτυα που χρησιμοποιούν servers. Επιπλέον χαρακτηριστικό αυτών των δικτύων είναι πως η ανταλλαγή δεδομένων μεταξύ των χρηστών μπορεί να γίνει με πολύ μεγάλες ταχύτητες όσο μεγαλώνει ο αριθμός των κόμβων.

Παρακάτω παραθέτονται οι τρόποι υλοποίησης των προσεγγίσεων. Αρχικά, ο τρόπος που αναπτύχθηκε και σχεδιάστηκε η κεντρική βάση δεδομένων ήταν αρκετά τετριμμένος. Πιο συγκεκριμένα, χρειάστηκε να ομαδοποιηθούν καθώς αποθηκεύονται όλα τα αναγνωριστικά των πελατών και τα ονόματα των προϊόντων σε μια κεντρική βάση δεδομένων. Στη συνέχεια, για να εξάγουμε συχνά σύνολα δεδομένων έπρεπε να γίνονται συνεχείς σαρώσεις με βάση το προϊόν που θέλαμε να μάθουμε πληροφορίες και παράλληλα κάνοντας χρήση του αναγνωριστικού του πελάτη μπορούσαμε και γνωρίζαμε ποια προϊόντα αγοράστηκαν μαζί, αυτό εφαρμόζεται για όλους τους πελάτες μέχρι την τελευταία εγγραφή του αρχείου, έτσι προέκυπταν όλα τα σχετιζόμενα προϊόντα όπου και τα αποθηκεύαμε σε μια μορφή δομής δεδομένων. Αντιθέτως, στην κατανεμημένη προσέγγιση ό τρόπος σχεδίασης και ανάπτυξης ήταν αρκετά περίπλοκος. Αυτό διότι, αρχικά έπρεπε να δημιουργηθούν οι κόμβοι του συστήματος για να μπορέσουν να αποθηκευτούν τα δεδομένα. Σε αυτό το σημείο θα αναλύσουμε παραπάνω την δημιουργία των κόμβων. Η τοποθέτηση ενός κόμβου ή ενός αντικειμένου στον δακτύλιο του Chord, δηλαδή η εύρεση του αναγνωριστικού τους, γίνεται κατακερματίζοντας την IP διεύθυνση ή το κλειδί τους αντίστοιχα (όπου στην προκειμένη περίπτωση το κλειδί που χρησιμοποιήθηκε ήταν το προϊόν), με μία consistent συνάρτηση κατακερματισμού. Όπου μία consistent συνάρτηση κατακερματισμού εγγυάται ότι με πολύ μεγάλη πιθανότητα όλοι οι κόμβοι θα αναλάβουν τον ίδιο αριθμό κλειδιών. Η ανάθεση των κλειδιών στους κόμβους εφαρμόστηκε ως εξής: αφού βρεθεί το αναγνωριστικό του κλειδιού, αυτό ανατίθεται στον κόμβο που βρίσκεται στη θέση του δακτυλίου με αυτό το αναγνωριστικό. Αν σε περίπτωση δεν υπάρχει κόμβος σε εκείνη τη θέση, τότε το κλειδί ανατίθεται στον κόμβο που έχει την ακριβώς επόμενη θέση στο δακτύλιο. Αν υποθέσουμε ότι στο δίκτυο συμμε-

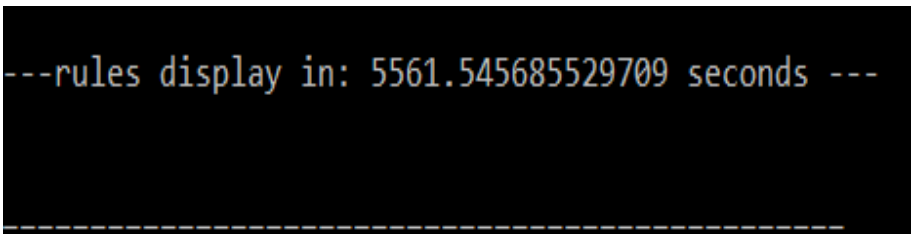
οποίες έχουν ως σκοπό να καταστήσουν τον υπολογιστή ή την υπηρεσία ανίκανη να δεχτεί άλλες συνδέσεις και έτσι να μην μπορεί να εξυπηρετήσει άλλους πιθανούς πελάτες

τέχουν N κόμβοι τότε κάθε κόμβος κρατάει πληροφορίες δρομολόγησης (δείκτες) για άλλους O(logN) κόμβους και επιλύει τις αναζητήσεις σε O(logN) hops. Για κάθε κόμβο, αυτοί οι δείκτες δείχνουν σε επόμενους από αυτόν κόμβους οι οποίοι βρίσκονται σε αποστάσεις που ισούνται με τις δυνάμεις του δύο. Επίσης, έπρεπε να γίνετε έλεγχος σε μία εισαγωγή ή μία διαγραφή ενός κόμβου διότι οι πληροφορίες δρομολόγησης πρέπει να κρατούνται συνεπείς. Αυτό επιτυγχάνεται με ανταλλαγή O(log2N) μηνυμάτων. Επίσης για λόγους αξιοπιστίας και ανοχής σε πολλαπλές αποτυχίες, κάθε κόμβος πρέπει να κρατάει και μια λίστα με μερικούς άμεσους διαδόχους (που βρίσκονται συνεχόμενα στον δακτύλιο). Βασιζόμενοι στον παραπάνω τρόπο δημιουργίας κόμβων, γίνεται μια σάρωση όλου του συνόλου δεδομένων και κάθε όνομα προϊόντος περνάει μέσα από μια συνάρτηση κατακερματισμού ώστε να αποθηκευτεί στον εκάστοτε κόμβο. Ξέροντας το σύστημα πόσους κόμβους θα περιέχονται κάνει και τους αντίστοιχους υπολογισμούς ώστε η έξοδος της συνάρτησης κατακερματισμού να μην περνάει τον αριθμό τον κόμβων. Παράλληλα με την σάρωση του αρχείου καθώς αποθηκεύονται τα προϊόντα στους κόμβους κρατούνται και τα αναγνωριστικά των πελατών (στην περίπτωση που υπάρχουν πολλά ίδια προϊόντα με διαφορετικό αναγνωριστικό πελατών αποθηκεύονται και αυτά στον κόμβο) όπου θα μας βοηθήσουν στο διαμοιρασμό και την αποθήκευση των συνόλων δεδομένων σε κόμβους. Εφόσον, έχουν δημιουργηθεί οι κόμβοι και έχει αναλάβει κάθε κόμβος ένα προϊόν γίνεται άλλη μια σάρωση στο σύνολο δεδομένων όπου συλλέγονται και αποθηκεύονται τα σχετιζόμενα προϊόντα. Πιο αναλυτικά ο διαμοιρασμός γίνετε, καθώς εξετάζεται μία μία εγγραφή του αρχείου ελέγχετε σε κάθε κόμβο εάν το προϊόν έχει το ίδιο αναγνωριστικό πελάτη με ένα από όλα τα αποθηκευμένα, όταν βρεθεί ένα αναγνωριστικό πελάτη που ταιριάζει στέλνεται το προϊόν στον κόμβο. Υλοποιήσαμε και τις δύο προσεγγίσεις, χρησιμοποιήσαμε και στις δύο περιπτώσεις τα σύνολα δεδομένων και την παραμετροποίηση που παρουσιάσαμε στα προηγούμενα κεφάλαια και παρακάτω παρουσιάζουμε τα αποτελέσματα εξόρυξης συχνών προτύπων και τη συγκριτική αξιολόγηση των προσεγγίσεων αυτών. Στις παρακάτω εικόνες παρατηρούμε τους χρόνους που χρειάστηκαν οι δύο προσεγγίσεις για εξάγουν κανόνες με το αλγόριθμο FP-Growth οι Εικόνες 5.11 και 5.12 απεικονίζουν τους χρόνους του πρώτου συνόλου δεδομένων (Εικόνα 5.1) ενώ οι Εικόνες 5.13 και 5.14 τους χρόνους

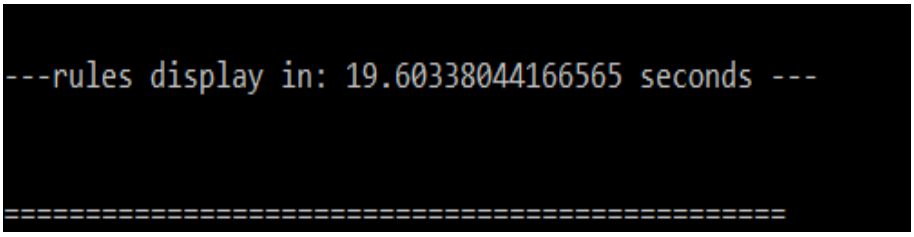
του δεύτερου συνόλου δεδομένων (Εικόνα 5.1).



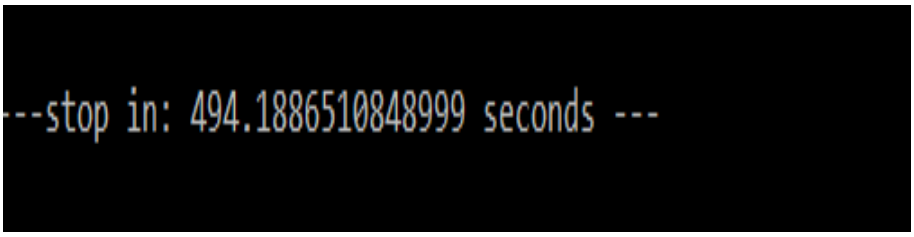
Σχήμα 5.11: Χρόνος απόκρισης κατανεμημένης προσέγγισης (πρώτου συνόλου δεδομένων)



Σχήμα 5.12: Χρόνος απόκρισης κεντρικοποιημένης προσέγγισης (πρώτου συνόλου δεδομένων)



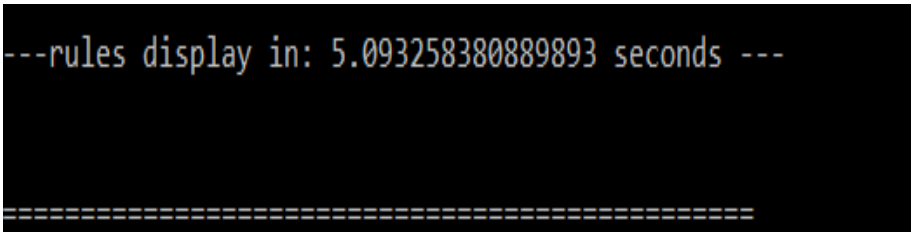
Σχήμα 5.13: Χρόνος απόκρισης κατανεμημένης προσέγγισης (δεύτερου συνόλου δεδομένων)



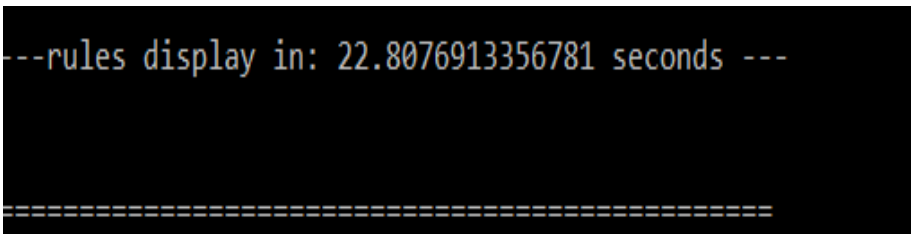
Σχήμα 5.14: Χρόνος απόκρισης κεντρικοποιημένης προσέγγισης (δεύτερου συνόλου δεδομένων)

Στις παραπάνω εικόνες (Σχήματα 5.11 - 5.14) απεικονίζεται ο χρόνος σε δευτερόλεπτα που απαιτείται από την εκάστοτε προσέγγιση (και για τα δύο διαφορετικά σύνολα δεδομένων) να εξάγει συχνά πρότυπα στο σύνολο των δεδομένων. Στην περίπτωση της κατανεμημένη λύσης, και θεωρώντας ότι το δίκτυο αποτελείται από 256 κόμβους, ο χρόνος που απαιτείται για την εξαγωγή συχνών προτύπων είναι σε κάθε περίπτωση μικρότερος από τον χρόνο που απαιτείται όταν όλη την δουλεία την κάνει μόνο ένας εξυπηρέτης. Στην περίπτωση δε του ενός από τα σύνολα δεδομένων (Σχήματα 5.13 - 5.14), όπου τα δεδομένα φαίνεται να ευνοούν τον FP-Growth αλγόριθμο, η κατανεμημένη προσέγγιση τα καταφέρνει με μια τάξη μεγέθους ταχύτερα.

Επίσης, για μπορέσουμε να έχουμε ως το δυνατόν πιο αντικειμενική άποψη σχετικά με τον χρόνο που χρειάζεται για να εξαχθούν τα σύνολα δεδομένων από την κατανεμημένη προσέγγιση εξετάσαμε και την περίπτωση που υπάρχουν πάρα πολύ κόμβοι και αντίστοιχά πάρα πολύ λίγοι, επιπλέον υπολογίσαμε και τον μέσο όρο των χρόνων. Οπότε, εξετάσαμε την περίπτωση που στο δίκτυο υπήρχαν 32 κόμβοι και σε μια άλλη περίπτωση όπου υπήρχαν 2048 κόμβοι. Παρακάτω, βλέπουμε του χρόνους αυτών των δύο περιπτώσεων.



Σχήμα 5.15: Χρόνος εξαγωγής συνόλων με μεγάλο αριθμό κόμβων (2048 κόμβους)



Σχήμα 5.16: Χρόνος εξαγωγής συνόλων με μικρό αριθμό κόμβων (32 κόμβους)

Στα Σχήματα 5.15 και 5.16 μπορούμε να παρατηρήσουμε μια πολύ μεγάλη απόκλιση στους χρόνους εξαγωγής προτύπων. Αυτό οφείλεται στο γεγονός ότι υπάρχει και μεγάλη διαφορά στον αριθμό των κόμβων του συστήματος. Πιο συγκεκριμένα, στην περίπτωση που το σύστημα έχει πάρα πολλούς κόμβους (2048), τα δεδομένα φαίνεται πως κατανέμονται στο σύνολο των κόμβων του δικτύου, οπότε κάθε κόμβος αναλαμβάνει μόνο τα δεδομένα που αφορούν σε ένα συγκεκριμένο προϊόν ή σε πολύ μικρό αριθμό προϊόντων. Σε αυτή την περίπτωση, (ισο-)κατανέμεται μεταξύ των κόμβων ο φόρτος επεξεργασίας των δεδομένων και εξαγωγής συχνών προτύπων, οπότε κάθε κόμβος χρειάζεται μόλις 5 δευτερόλεπτα για να εξάγει τα σύνολα προ-

τύπων. Από την άλλη, όταν το σύστημα διαθέτει έναν μικρό αριθμό από κόμβους σε σχέση με το πλήθος των υπό επεξεργασία προϊόντων (32 κόμβους στην περίπτωση που φαίνεται στο Σχήμα 5.16), τότε κάθε κόμβος γίνεται υπεύθυνος για μεγαλύτερο πλήθος προϊόντων και συνεπώς χρειάζεται περισσότερο χρόνο να επεξεργαστεί τα δεδομένα του και να εξάγει συχνά σύνολα προτύπων (4 φορές περισσότερο χρόνο σε σύγκριση με το μεγαλύτερο δίκτυο του Σχήματος 5.15). Στην ακραία περίπτωση, το δίκτυο έχει έναν κόμβο/εξυπηρέτη οπότε και αναλαμβάνει το σύνολο των δεδομένων κάνοντας πολλαπλάσιο χρόνο να εξάγει συχνά σύνολα (Σχήμα 5.14).

Στη συνέχεια, προκειμένου να έχουμε μια καλύτερη εκτίμηση του χρόνου που απαιτείται στην κατανεμημένη προσέγγιση, κάναμε το ίδιο πείραμα σε 10 διαφορετικά δίκτυα Chord. Δηλαδή χρησιμοποιήσαμε το 2ο σύνολο δεδομένων (Σχήμα 5.2) και ένα δίκτυο 256 κόμβων κάθε φορά για την εξαγωγή συχνών προτύπων. Σε κάθε δίκτυο η ανάθεση δεδομένων στους κόμβους ήταν τυχαία και εξαρτώμενη αποκλειστικά από την συνάρτηση κατακερματισμού. Το αποτέλεσμα του μέσου χρόνου απόκρισης (21 seconds) ήταν αυτό που περιμέναμε. Συμπερασματικά, η απόδοση της κατανεμημένης προσέγγισης σε κάθε περίπτωση είναι εξαιρετική, συνδεδεμένη σαφώς, όπως σε κάθε πρόβλημα μηχανικής μάθησης, με τα δεδομένα, αλλά σίγουρα ταχύτερη (αν και με τα ίδια αποτελέσματα εξαγωγής συχνών προτύπων) από την περίπτωση του ενός εξυπηρέτη.

Όπως μπορούμε να παρατηρήσουμε από τις παραπάνω εικόνες που παραθέτονται οι χρόνοι εξαγωγής κανόνων η κατανεμημένη προσέγγιση υπερτερεί σε έναν πάρα πολύ μεγάλο βαθμό όσο αφορά τον χρόνο εξαγωγής συχνών αντικειμένων. Τα datasets που επεξεργαστήκαμε αποτελούνταν από 4540 έως 345611 εγγραφές. Σε αυτό το σημείο, είναι σημαντικό να αναφερθεί ότι άσχετα με τις υπολογιστικές δυνατότητες η κεντρικοποιημένη προσέγγιση είχε πολύ μεγάλη καθυστέρηση σε σύγκριση με την κατανεμημένη προσέγγιση. Επιπλέον, παρακάτω παραθέτουμε και τους κανόνες που εξάγει κάθε προσέγγιση.

```
JUVENILLE FURNITURE
==========================
{'MENS FURNISHINGS'}
{'JUVENILLE FURNITURE', 'MENS FURNISHINGS'}
{'INFANT APPAREL'}
{'INFANT APPAREL', 'MENS FURNISHINGS'}
{'INFANT APPAREL', 'JUVENILLE FURNITURE'}
{'INFANT APPAREL', 'JUVENILLE FURNITURE', 'MENS FURNISHINGS'}
{'JUVENILLE FURNITURE'}


==========================

[{'JUVENILLE FURNITURE'}, {'MENS FURNISHINGS'}, 0.3333333333333333]
[{'MENS FURNISHINGS'}, {'JUVENILLE FURNITURE'}, 1.0]
[{'INFANT APPAREL'}, {'MENS FURNISHINGS'}, 1.0]
[{'MENS FURNISHINGS'}, {'INFANT APPAREL'}, 1.0]
[{'INFANT APPAREL'}, {'JUVENILLE FURNITURE'}, 1.0]
[{'JUVENILLE FURNITURE'}, {'INFANT APPAREL'}, 0.3333333333333333]
[{'INFANT APPAREL'}, {'JUVENILLE FURNITURE', 'MENS FURNISHINGS'}, 1.0]
[{'JUVENILLE FURNITURE'}, {'INFANT APPAREL', 'MENS FURNISHINGS'}, 0.3333333333333333]
[{'MENS FURNISHINGS'}, {'INFANT APPAREL', 'JUVENILLE FURNITURE'}, 1.0]
[{'INFANT APPAREL', 'JUVENILLE FURNITURE'}, {'MENS FURNISHINGS'}, 1.0]
[{'INFANT APPAREL', 'MENS FURNISHINGS'}, {'JUVENILLE FURNITURE'}, 1.0]
[{'JUVENILLE FURNITURE', 'MENS FURNISHINGS'}, {'INFANT APPAREL'}, 1.0]
==========================
```

Σχήμα 5.17: Αποτελέσματα κατανεμημένης προσέγγισης

```
JUVENILLE FURNITURE
==========================
{'MENS FURNISHINGS'}
{'JUVENILLE FURNITURE', 'MENS FURNISHINGS'}
{'INFANT APPAREL'}
{'INFANT APPAREL', 'MENS FURNISHINGS'}
{'INFANT APPAREL', 'JUVENILLE FURNITURE'}
{'INFANT APPAREL', 'JUVENILLE FURNITURE', 'MENS FURNISHINGS'}
{'JUVENILLE FURNITURE'}


==========================

[{'JUVENILLE FURNITURE'}, {'MENS FURNISHINGS'}, 0.3333333333333333]
[{'MENS FURNISHINGS'}, {'JUVENILLE FURNITURE'}, 1.0]
[{'INFANT APPAREL'}, {'MENS FURNISHINGS'}, 1.0]
[{'MENS FURNISHINGS'}, {'INFANT APPAREL'}, 1.0]
[{'INFANT APPAREL'}, {'JUVENILLE FURNITURE'}, 1.0]
[{'JUVENILLE FURNITURE'}, {'INFANT APPAREL'}, 0.3333333333333333]
[{'INFANT APPAREL'}, {'JUVENILLE FURNITURE', 'MENS FURNISHINGS'}, 1.0]
[{'JUVENILLE FURNITURE'}, {'INFANT APPAREL', 'MENS FURNISHINGS'}, 0.3333333333333333]
[{'MENS FURNISHINGS'}, {'INFANT APPAREL', 'JUVENILLE FURNITURE'}, 1.0]
[{'INFANT APPAREL', 'JUVENILLE FURNITURE'}, {'MENS FURNISHINGS'}, 1.0]
[{'INFANT APPAREL', 'MENS FURNISHINGS'}, {'JUVENILLE FURNITURE'}, 1.0]
[{'JUVENILLE FURNITURE', 'MENS FURNISHINGS'}, {'INFANT APPAREL'}, 1.0]
==========================
```

Σχήμα 5.18: Αποτελέσματα κεντρικοποιημένης προσέγγισης

Μπορούμε να κατανοήσουμε σε αυτό το σημείο ότι με την κατανεμημένη προσέγγιση εξάγουμε ακριβώς τα ίδια αποτελέσματα με την κεντρικοποιημένη αλλά σε έναν πολύ καλύτερο και προτιμότερο χρόνο απόκρισης. Αυτό έχει ως αποτέλεσμα να κάνει την κατανεμημένη προσέγγιση ιδανική η οποία βελτιώνει αισθητά τον χρόνο και ίσως σε μεγαλύτερο βαθμό από ότι φαίνεται αφού τα δεδομένα δεν είναι αρκετά μεγάλα. Παρόλα αυτά η υπεροχή της μπορεί να φανεί και στα μικρότερα dataset.

Κεφάλαιο 6

Συμπεράσματα

Η παρούσα διπλωματική εργασία αποτελεί ένα παράδειγμα αξιοποίησης των αλγορίθμων εξόρυξης κανόνων συσχέτισης από βάσεις δεδομένων που περιλαμβάνουν πληροφορίες δοσοληψιών που διαμορφώνουν τα καλάθια αγοράς των πελατών και που είναι πολύ πιθανόν να βοηθήσουν επιχειρήσεις και γενικότερα ανθρώπους που εμπλέκονται στη διαχείριση μιας πελατειακής βάσης δεδομένων ή στη διοίκηση μιας επιχείρησης οπότε και καλούνται να πάρουν σημαντικές αποφάσεις σχετικά με τον τρόπο λειτουργίας της. Μέσα από τη μελέτη μας διαπιστώσαμε πως η εξόρυξη αυτών των δεδομένων περιλαμβάνει αρκετούς παράγοντες που είναι σημαντικό να ληφθούν σοβαρά υπόψη. Αυτοί οι παράγοντες αφορούν αφενός την αξιοπιστία των κανόνων συσχέτισης οι οποίοι εξάγονται από τη βάση δεδομένων και αφετέρου την απόρριψη εκείνων των κανόνων οι οποίοι δεν οδηγούν σε κάποιο ουσιαστικό συμπέρασμα.

Με γνώμονα τα παραπάνω, έγινε σχεδίαση και ανάπτυξη ενός αλγόριθμου ο οποίος θα παρέχει τη δυνατότητα εξαγωγής συχνών προτύπων από ηλεκτρονικές συναλλαγές ή από επισκέψεις χρηστών σε ιστοσελίδες. Αποδείχθηκε, πως ο αλγόριθμος FP-Growth σε ένα Chord κατανεμημένο περιβάλλον προκειμένου να εξαχθούν συχνά πρότυπα πάνω από σύνολα δεδομένων δύναται να είναι μια πάρα πολύ καλή προσέγγιση, αξιοποιώντας κατάλληλα τους πόρους του συστήματος και να δίνει γρήγορα και έγκυρα αποτελέσματα. Δεχόμενοι τις κατάλληλες παραμετροποίησεις, ώστε να μην περιορίζονται στην παραγωγή κανόνων, αλλά με την παρέμβαση του χρήστη από την επιχείρηση να εξάγουν κανόνες που είναι αξιόπιστοι, θέτοντας κάθε φορά όρια για τα μεγέθη της υποστήριξης και της εμπιστοσύνης. Επίσης, μέσα από την μελέτη των

αλγορίθμων Apriori και FP-Grwoth φτάσαμε σε ένα συμπέρασμα πως εξάγουν σε έναν μεγάλο βαθμό τον ίδιο αριθμό κανόνων. Ωστόσο, ο χρόνος που απαιτεί ο Apriori για την εξαγωγή κανόνων είναι αισθητά μεγαλύτερος του FP-Growth για ορισμένες τιμές της υποστήριξης. Αυτό οφείλεται στην συμπαγή δομή του FP-Growth όπου εξαλείφει την επαναλαμβανόμενη ανίχνευση των συναλλαγών. Επιπρόσθετα, ένα πολύ σημαντικό εργαλείο που μας βοήθησε στην υλοποίηση ήταν η επιλογή του πρωτοκόλλου Chord. Αυτό διότι, το Chord προσφέρει την πολύ δυνατή ιδιότητα ότι δίνοντας του κάποιο κλειδί, επιστρέφει με πολύ αποδοτικό τρόπο τον κόμβο ο οποίος είναι υπεύθυνος για την αποθήκευση αυτού του κλειδιού. Σε μια σταθερή κατάσταση σε ένα δίκτυο N κόμβων, κάθε κόμβος διατηρεί πληροφορία δρομολόγησης μόνο για $O(\log N)$ άλλους και ολοκληρώνει όλες τις αναζητήσεις με $O(\log N)$ μηνύματα προς τους άλλους κόμβους. Επιπλέον, τα δυνατά χαρακτηριστικά του Chord είναι, η απλότητα του, η αποδεδειγμένη ορθότητα του και η αποδεδειγμένη απόδοση του ακόμα και όταν υπάρχουν ταυτόχρονες είσοδοι και έξοδοι κόμβων από το σύστημα. Συνεχίζει να λειτουργεί σωστά, αν και με μειωμένη απόδοση, ακόμα και όταν οι πληροφορίες δρομολόγησης που διατηρούν οι κόμβοι είναι μερικώς σωστές.

Μέσα από τους κανόνες συσχέτισης εξάγεται χρήσιμη γνώση που με κατάλληλη αξιοποίηση της μπορεί να βελτιώσει την ανταγωνιστική θέση της επιχείρησης. Σε αυτό το σημείο αξίζει να αναφέρουμε πως οι κανόνες συσχέτισης μπορούν να βρουν εφαρμογή και σε άλλους τομείς, πέραν αυτού της επιχειρηματικότητας, όπως στα ιατρικά δεδομένα. Επαγωγικά λοιπόν, σκεπτόμενοι, καταλήγουμε στο συμπέρασμα πως κάθε μέθοδος εξόρυξης γνώσης από μια βάση δεδομένων μπορεί να βοηθήσει, αλλά σαν διαδικασία επιβάλλεται να γίνεται με στρατηγική και προσεκτική εκ προοιμίου μελέτη. Αυτό διότι κάθε βάση δεδομένων περιέχει εξ΄ ορισμού μεγάλο όγκο πληροφορίας και, πολύ εύκολα, μπορεί να εξαχθεί γνώση που μπορεί να οδηγήσει σε μη ασφαλή συμπεράσματα.

Βιβλιογραφία

[1] https://ericbrown.com/drowning-in-data-starved-for-information.htm

[2] Han J, Pei J, Yin Y (2000). Han J, Pei J, Yin Y, Mao R. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min Knowl Discov 2004, 8(1):53–87. https://www.cs.sfu.ca/ jpei/publications/sigmod00.pdf

[3] https://www.mygreatlearning.com/blog/understanding-fp-growth-algorithm/

[4] https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/

[5] K. Suguna, K. Nandhini, PhD, "Frequent Pattern Mining of Web Log Files Working Principles", International Journal of Computer Applications (0975 – 8887) Volume 157 – No 3, January 2017.

[6] Stoica. I, Morris. R, Liben-Nowell. D, David R.Karger, , M. Frans Kaashoek, Frank Dabek, Hari Balakrishnan. Chord: a scalable peer-to-peer lookup protocol forInternet applications" IEEE J NET, 2003, 11, 17-32 https://pdos.csail.mit.edu/papers/ton:chord/paper-ton.pdf

[7] Ganesan, P.; Gummadi, K. and Garcia-Molina, H.
Canon in G Major: Designing DHTs with Hierarchical Structure Technical Report.", Stanford InfoLab, Stanford, 2003

[8] Agrawal R, Imielinski T, Swami A. Mining associa-tion rules between sets of items in large database. In:Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington,DC, USA, 1993, 207–216. http://www.rakesh.agrawal-family.com/papers/sigmod93assoc.pdf

[9] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In EDBT, pages 3–17, 1996.

[10] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu. FreeSpan: Frequent pattern-projected sequential pattern mining. In KDD, pages 355–359, 2000.

[11] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining sequential patterns by pattern-growth: The PrefixSpan approach. IEEE Transactions on Knowledge and Data Engineering, 16:1424–1440, 2004.

[12] M. J. Zaki. SPADE: An efficient algorithm for mining frequent sequences. Mach. Learn., 42(1-2):31–60, 2001.

[13] `https://dzone.com/articles/machinex-understanding-fp-tree-construction`

[14] T. Shintani and M. Kitsuregawa. Parallel algorithms for mining association rule mining onlarge scale PC cluster. In Zaki and Ho. in conjunction with ACM SIGKDD InternationalConference on Knowledge Discovery and Data Mining (KDD99).

[15] M. J. Zaki, C. T. Ho, and R. Agrawal. Parallel classification for data mining on shared-memory multiprocessors. In Proceedings International Conference on Data Engineering,March 1999.

[16] Mohammed J. Zaki. Parallel sequence mining on SMP machines a data clustering algorithm ondistributed memory machines a data clustering algorithm on distributed memory machines. InZaki and Ho. in conjunction with ACM SIGKDD International Conference on KnowledgeDiscovery and Data Mining (KDD99).

[17] H. Kargupta, W. Huang, S. Krishnamoorthy, and E. Johnson. Distributed clustering usingcollective principal component analysis. Knoledge and Information Systems Journal SpecialIssue on Distributed and Parallel Knowledge Discovery, 2000.

[18] A Mueller, Fast sequential and parallel algorithms forassociation rule mining: a comparison. TechnicalReport, University of Maryland at CollegePark, 1995.

[19] Park JS, Chen MS, Yu PS. Efficient parallel data min-ing for association rules. In: Proceedings of theACM Int Conf Inf Knowl Manage, Baltimore, MD,USA, 1995:31–36.

[20] Cheung DW, Han J, Ng VT, Fu AW, Fu Y. A fast distributed algorithm for mining association rules. In:Proceedings of the Int Conf Parallel Distrib Inf Syst,Miami Beach, Florida, USA, 1996:31–42

[21] Dean J, Ghemawat S. MapReduce: a flexible data processing tool. Commun ACM 2010, 53(1):72–77.

[22] Li H, Wang Y, Zhang D, Zhang M, Chang EY. PFP:parallel FP-growth for query recommendation. In:Proceedings of the ACM Conf Recommender Syst,Lousanne, Switzerland, 2008:107–114.

[23] Grigorios Tsoumakas, Ioannis Vlahavas: Distributed Data Mining. `http://lpis.csd.auth.gr/publications/tsoumakas-dwm2.pdf`

[24] Urmela, S. and Nandhini, M. (2017) Approaches and Techniques of Distributed Data Mining: A Comprehensive Study. International Journal of Engineering and Technology, 9, 63-76.

[25] Miliaraki, I., Berberich, K., Gemulla, R. & Zoupanos, S. (2013). Mind the gap: large-scale frequent sequence mining. In K. A. Ross, D. Srivastava & D. Papadias (eds.), SIGMOD Conference (p./pp. 797-808), : ACM. ISBN: 978-1-4503-2037-5

[26] Vinaya Sawant and Ketan Shah, "A review of Distributed Data Mining using agents", International Journal of Advanced Technology & Engineering Research (IJATER), vol. 3, no. 5, pp. 27-33,2013.

[27] Park, B-H., & Kargupta, H. (2003). Distributed data mining: Algorithms, systems and applications. In N. Ye (Ed.), The handbook of data mining (pp. 341–358). Lawrence Erlbaum Associates.

[28] W. Gan, J. C.-W. C. H.-C. Lin and J. Zhan, "Data mining in Distributed Environment: A Survey," Wiley Interdiscriplinary Reviews: Data Mining and Knowledge Discovery , vol. 7, no. 6, pp. 1-19, 2017

[29] Hillol Kargupta, "An Introduction to Distributed Data Mining", `http://eric.univ-lyon2.fr/~pkdd2000/Download/T1.pdf`

[30] M. Kaur and S. Kang, "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining," Procedia Comput. Sci., vol. 85, no. Cms, pp. 78–85, 2016.

[31] Cerrito, P. B. 2007. Choice of antibiotic in open heart surgery. Intelligent Decision Technologies, 1: 63-69

[32] Hibino, A., Niwa, Y. 2008. Graphical representation of nuclear incidents/accidents by associating network in nuclear technical communication. Journal of Nuclear Science and Technology, 45: 369-377

[33] Hsieh, S.-C., Lai, J.-N., Lee, C.-F., Hu, F.-C., Tseng, W.-L., Wang, J.-D. 2008. The prescribing of Chinese herbal products in Taiwan: A cross-sectional analysis of the national health insurance reimbursement database. Pharmacoepidemiology and Drug Safety, 17: 609-619

[34] Kanagawa, Y., Matsumoto, S., Koike, S., & Imamura, T. 2009. Association analysis of food allergens. Pediatric Allergy and Immunology, 20: 347-352

[35] Yang, R., Tang, J., Kafatos, M. 2007. Improved associated conditions in rapid intensifications of tropical cyclones. Geophysical Research Letters, 34: 1-5

[36] X. Su, "Intertemporal Pricing with Strategic Customer Behavior," Manage. Sci., vol. 53, no. 5, pp. 726–741, 2007.

[37] E. Sherman, A. Mathur, and R. B. Smith, "Store Environment and Consumer Purchase Behavior: Mediating Role of Consumer Emotions," Psychol. Mark., vol. 14, no. 4, pp. 361–378, 1997.

[38] http://www.dei.unipd.it/~capri/DM/MATERIALE/AssocAnalysis1617.pdf

[39] https://www-users.cs.umn.edu/~kumar001/dmbook/ch5_association_analysis.pdf