ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ & ΠΛΗΡΟΦΟΡΙΚΗΣ

Διδακτορική Διατριβή

“Εφαρμογή Τεχνικών Data Mining σε
Συστήματα Ηλεκτρονικού Εμπορίου”

Κουρής Ν. Γιάννης

ΠΑΤΡΑ 2006

«Αφιερωμένο στους γονείς μου Νικογιάννη και Ευγενία, στα
αδέρφια μου Διονύση και Μπέσσυ αλλά και σε όσους ήταν μαζί
μου αυτά τα χρόνια...»

«Ένα ευχαριστώ και στον Μακρή Χρήστο.»

Τριμελής Επιτροπή

Αθανάσιος Τσακαλίδης (Επιβλέπων), Καθηγητής
Ιωάννης Γαροφαλάκης, Αναπληρωτής Καθηγητής
Ιωάννης Χατζηλυγερούδης, Επίκουρος Καθηγητής

ΠΡΟΛΟΓΟΣ

Η συγκεκριμένη διατριβή κινείται στο χώρο της εξόρυξης γνώσης από τις βάσεις δεδομένων (knowledge discovery in databases) και ειδικότερα στην περιοχή της ανακάλυψης κανόνων συσχετίσεων. Ένας ευρέως χρησιμοποιούμενος ορισμός της ανακάλυψης γνώσης είναι ο εξής: *η σύνθετη διαδικασία της εξαγωγής μη τετριμμένης, προηγούμενα άγνωστης, υποκρυπτόμενης και πιθανώς χρήσιμης πληροφορίας από αποθηκευμένα δεδομένα*. Η εξαγόμενη γνώση μπορεί να εφαρμοστεί σε μια πληθώρα τομέων όπως διαχείριση πληροφορίας, λήψη επιχειρηματικών αποφάσεων, έλεγχος διαδικασιών και σε πολλές άλλες εφαρμογές. Το θέμα του Data Mining αποτελεί ένα από τα πιο ενεργά θέματα τόσο της ακαδημαϊκής όσο και της επιχειρηματικής έρευνας και έχει συγκεντρώσει μεγάλο ενδιαφέρον από ερευνητές από πολλά διαφορετικά πεδία όπως τεχνητή νοημοσύνη, βάσεις δεδομένων, στατιστική, μάρκετινγκ, οπτικοποίηση δεδομένων κ.α.. Ανάμεσα στις πιο γνωστές διεργασίες που μπορούν να θεωρηθούν ότι ανήκουν άμεσα στον τομέα του Data Mining είναι η κατηγοριοποίηση (classification), η συσταδοποίηση (clustering) και οι κανόνες συσχετίσεων (association rules).

Η παρούσα διατριβή ασχολήθηκε με την εφαρμογή τεχνικών data mining σε συστήματα ηλεκτρονικού εμπορίου. Για να είμαστε πιο ακριβείς επικεντρωθήκαμε στην εύρεση κανόνων συσχετίσεων από δεδομένα, και κύρια δεδομένα που είχαν να κάνουν με βάσεις συναλλαγών. Η βασική ιδέα ενός κανόνα συσχετίσεως είναι να αναπτύξει μια συστηματική μέθοδο με την οποία ένας χρήστης μπορεί να προβλέψει την εμφάνιση κάποιων αντικειμένων, δοσμένης της ύπαρξης κάποιων άλλων σε μια συναλλαγή, και συνήθως αποτελούν συνεπαγωγές της μορφής $X \Rightarrow Y$. Παράδειγμα ενός τέτοιου κανόνα είναι: “οι πελάτες που αγοράζουν κινητά τηλέφωνα και handsfree αγοράζουν και θήκη για το κινητό τους”. Τα τελευταία χρόνια είχε γίνει κοινός τόπος όλων των μελετών και των ερευνητών οι αδυναμίες και τα μειονεκτήματα του μοντέλου εύρεσης κανόνων συσχετίσεων. Στόχος μας ήταν να επιλύσουμε υπάρχοντα προβλήματα αλλά και να εκθέσουμε και να αντιμετωπίσουμε κάποια νέα. Σαν σύγγραμμα η παρούσα διατριβή μπορεί να χωριστεί σε τρία κομμάτια. Το πρώτο είναι τα τρία πρώτα κεφάλαια, τα οποία και αποτελούν εισαγωγικά κεφάλαια απαραίτητα για την υποστήριξη και κατανόηση της δουλειάς μας. Ακολούθως τα κεφάλαια 4 έως 8 αποτελούν το δεύτερο και κύριο κομμάτι της παρούσας διατριβής, και περιγράφουν διάφορες τεχνικές και προτάσεις μας, αποτελέσματα της ερευνάς μας. Το τρίτο και τελευταίο κομμάτι της διατριβής, αναφορικά το Κεφάλαιο 9, αποτελεί την σύνοψη ολόκληρης της εργασίας μας όπου παραθέτουμε εν συντομία την τελική προσφορά μας στο χώρο, δίνουμε πιθανές εφαρμογές των προτάσεων μας, και τέλος προτείνουμε μελλοντικές κατευθύνσεις της έρευνας σε ανοιχτά πεδία – προβλήματα. Η συνολική συνεισφορά μας στο χώρο μπορεί να συνοψιστεί στα παρακάτω σημεία:

Καταφέραμε κατ’ αρχήν να επιλύσουμε το πρόβλημα της ανάθεσης των καταλλήλων τιμών υποστήριξης στα αντικείμενα μιας βάσης [85], δίνοντας μεγαλύτερη έμφαση σε βάσεις αποτελούμενες από δεδομένα συναλλαγών λιανεμπορίου (retail data). Το μέχρι στιγμής κυρίαρχο μοντέλο ουσιαστικά όλων των εργασιών στους κανόνες συσχετίσεων βασιζόταν στη λογική του trial and error, όπου δηλαδή αν τελικά παρήγαμε πάρα πολλούς ή πάρα πολύ λίγους κανόνες σε σχέση με το αναμενόμενο αποτέλεσμα (πάντα με βάση εντελώς υποκειμενικά και απλοϊκά κριτήρια) τότε ξεκινούσαμε την όλη διαδικασία από την αρχή με νέες τιμές υποστήριξης και βεβαιότητας. Η πρακτική αυτή είχε το προφανές μειονέκτημα ότι κατασπαταλούσε πόρους και επιβάρυνε το σύστημα μας άσκοπα, ενώ στην περίπτωση που χρησιμοποιούσαμε πολλαπλές τιμές υποστήριξης και είχαμε μεγάλο πλήθος αντικειμένων η όλη διαδικασία κατέληγε πρακτικά ανεπίλυτη. Με την προσέγγιση μας καταφέραμε να λάβουμε υπόψη και να συνδυάσουμε μια σειρά παραγόντων – παραμέτρων που καθόριζαν την σημασία ενός αντικειμένου σε μια βάση, σε σχέση τόσο με την φύση των αντικειμένων αυτών καθ’ εαυτών αλλά και σε σχέση με τα υπόλοιπα αντικείμενα της βάσης. Η όλη διαδικασία ακολουθεί μια αυτοκαθοριζόμενη λογική, όπου ο χρήστης (data miner) δεν πρέπει να ορίσει κάτι προκαταβολικά αλλά αφήνει τα αντικείμενα να λάβουν τις δικές τους τιμές υποστήριξης

σύμφωνα με μια σειρά παραγόντων όπως οι εμφανίσεις τους στη βάση καθώς και σύμφωνα με τις εμφανίσεις όλων των άλλων αντικειμένων. Η φύση της προσέγγισης που προτείνουμε μπορεί να εφαρμοστεί με μικρές αλλαγές σε οποιαδήποτε εφαρμογή και σε οποιαδήποτε συλλογή δεδομένων.

Ένα άλλο μεγάλο πρόβλημα των μεθόδων και των προσεγγίσεων που χρησιμοποιούνταν έως τώρα ήταν ότι ανακάλυπταν σχέσεις μεταξύ αντικειμένων οι οποίες προϋπήρχαν για μεγάλο διάστημα αδυνατώντας να ανακαλύψουν τις νέες. Όλες οι προσεγγίσεις έως τώρα ακολουθούσαν παρά συμβάδιζαν με τις πωλήσεις ή γενικότερα τις εμφανίσεις των αντικειμένων. Όμως εμείς στην πράξη χρειαζόμαστε μια προσέγγιση η οποία θα βρίσκει ταυτόχρονα με τις ήδη υπάρχουσες, τις διαφαινόμενες - ανερχόμενες τάσεις εν τη γενέσει τους ακόμα προτού αυτές καθιερωθούν. Κατ' αυτόν τον τρόπο σε κάθε περίπτωση είτε προβλέπουμε αρκετά νωρίς την μελλοντική συμπεριφορά ενός αντικειμένου (αν αυτό τελικά εδραιωθεί κατά τις επόμενες περιόδους σαν ένα συχνό αντικείμενο), είτε καταφέρνουμε να το λάβουμε υπόψη και να το χειριστούμε κατάλληλα την περίοδο ακριβώς όπου αυτό έγινε ενδιαφέρον (ακόμα και αν η μελλοντική του συμπεριφορά μας δείξει τελικά ότι ήταν κάτι στιγμιαίο). Έτσι στο [82] εισαγάγαμε την έννοια των λεγομένων “καυτών” αντικειμένων και προτείνουμε μια αποτελεσματική μέθοδο για την εξεύρεση και διαχείριση τους. Σαν καυτά θεωρούμε τα αντικείμενα αυτά που παρουσιάζουν μια απροσδόκητα εκρηκτική (bursty) συμπεριφορά σε σχέση με τα υπόλοιπα αντικείμενα της βάσης μας και έτσι θα πρέπει να τύχουν μιας ιδιαίτερης μεταχείρισης και προσοχής. Ένα αντικείμενο το οποίο παρουσιάζει τόσες πολλές εμφανίσεις σε τόσο μικρή περίοδο πρέπει να είναι προφανώς πολύ ενδιαφέρον, και μάλιστα τέτοιες “εκρηκτικές” και απρόβλεπτες συμπεριφορές είναι πολλές φορές πολύ πιο ενδιαφέρουσες από τις συνηθισμένες. Με αυτόν τον τρόπο καταστήσαμε την διαδικασία εξεύρεσης κανόνων εξαιρετικά πιο άμεση και αποτελεσματική ως προς τον τελικό χρήστη.

Η χρήση πολλαπλών τιμών υποστήριξης για την εύρεση κανόνων συσχετίσεων αποτελεί ένα πραγματικό εργαλείο για την επίλυση των σημαντικών μειονεκτημάτων που παρουσίαζε το μοντέλο της χρήσης μιας μοναδικής – ενιαίας τιμής υποστήριξης για όλα τα αντικείμενα της βάσης μας [89]. Για παράδειγμα με τη χρήση μιας ενιαίας τιμής υποστήριξης είχαμε εμφάνιση του λεγόμενου “διλήμματος του σπάνιου αντικειμένου” (rare itemset dilemma) [93], όπου αν θέλαμε να συμπεριλάβουμε στους τελικούς κανόνες και κάποια αντικείμενα τα οποία είχαν όμως λίγες εμφανίσεις θα έπρεπε να μειώσουμε κατά πολύ το κατώφλι υποστήριξης. Αυτό όμως σήμαινε αυτόματα ότι θα είχαμε μια δραματική αύξηση του πλήθους των συχνών αντικειμένων, και σαν αποτέλεσμα την δημιουργία πάρα πολλών άχρηστων κανόνων συσχετίσεων. Συνδυάζοντας έτσι στο [79] δύο πολύ γνωστούς αλγόριθμους και κάνοντας ταυτόχρονα και τις κατάλληλες τροποποιήσεις καταφέραμε να δημιουργήσουμε ένα νέο βελτιωμένο αλγόριθμο για την εξόρυξη κανόνων συσχετίσεων με πολλαπλές τιμές

υποστήριξης. Ο προτεινόμενος αλγόριθμος με την ονομασία KTM, ο οποίος βασίζεται στους αλγορίθμους DIC [27] και MSAPriori, αποδείχτηκε τελικά αποτελεσματικότερος και από τους δύο (λαμβάνοντας υπόψη και ότι ο κάθε αλγόριθμος επίλυε ένα ξεχωριστό πρόβλημα).

Πολύ σημαντική υπήρξε και η θεώρηση του προβλήματος εξεύρεσης κανόνων συσχετίσεων αλλά και των προτάσεων αντικειμένων στους χρήστες γενικότερα (σε μια λογική παρόμοια με αυτή των collaborative filtering συστημάτων), μέσω ενός εντελώς διαφορετικού πρίσματος [81]. Πιο συγκεκριμένα προτείναμε ένα σύστημα το οποίο δεν ψάχνει απλά όλη βάση, αλλά αντίθετα λειτουργεί σαν μια μηχανή αναζήτησης ειδικά δημιουργημένη για να κάνει προτάσεις στους χρήστες χρησιμοποιώντας τεχνικές δανεισμένες από την ανάκτηση πληροφορίας. Έτσι αντί να υποβαλλόμαστε στην εξαιρετικά χρονοβόρα διαδικασία του ελέγχου όλων των δυνατών συνδυασμών υποψηφίων αντικειμένων η οποία σημειωτέον είχε πάρα πολλά προβλήματα απλά αφήνουμε την βάση μας να αποκαλύψει τα μυστικά της καθώς οι χρήστες θα την χρησιμοποιούν. Το προτεινόμενο σύστημα χρησιμοποιεί είτε ερωτήματα διάταξης είτε δυαδικά ερωτήματα, και προσφέρει στους χρήστες μια πληθώρα υπηρεσιών.

Τέλος διαπιστώσαμε την ύπαρξη των λεγομένων τοπικά συχνών προϊόντων (locally frequent items), δώσαμε έναν τυπικό ορισμό τους και προτείναμε μια σειρά μεθόδων για την διαχείριση τους [23]. Η κύρια ιδέα είναι να χωρίσουμε την βάση μας σε διαμερίσεις σύμφωνα με τις ανάγκες του προβλήματος και να ψάχνουμε και για αντικείμενα που είναι συχνά σε σχέση με κάποια διαμέριση, πέρα από αυτά που είναι συχνά σε ολόκληρη τη βάση. Σαν τοπικά συχνά προϊόντα καλούνται αυτά τα οποία παρουσιάζουν μια εξόχως τοπικιστική συμπεριφορά σε τμήματα ή διαμερίσεις της βάσης μας, και αποτελούν τη μεγαλύτερη και πιο συχνά εμφανιζόμενη κατηγορία σπάνιων αλλά ενδιαφερόντων αντικειμένων ειδικότερα σε εφαρμογές λιανεμπορίου. Για παράδειγμα κάποιο αντικείμενο μπορεί στο σύνολο της βάσης μας να μην συγκεντρώνει επαρκή αριθμό εμφανίσεων αδυνατώντας να ξεπεράσει το κατώφλι υποστήριξης, αλλά θεωρούμενο ως προς κάποιο τμήμα της βάσης να εμφανίζει συγκριτικά υπερβολικά μεγάλο αριθμό εμφανίσεων. Τα τμήματα αυτά μπορούν να ορίζονται είτε ως προς κάποιον χρονικό παράγοντα (π.χ. εποχή ή μήνας που συλλέχθηκαν τα εκάστοτε δεδομένα) είτε ως προς κάποιον χωρικό παράγοντα (π.χ. δεδομένα συλλεγμένα από διαφορετικές τοποθεσίες, τμήματα, εφαρμογές κτλ.) είτε ως προς οποιαδήποτε άλλη διαμέριση επιθυμεί ο χρήστης.

Κεφάλαιο 1 ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγή

Το Data Mining αποτελεί τα τελευταία χρόνια ένα εντελώς αυτόνομο πεδίο έρευνας, και παρ' όλες τις αλληλεπιδράσεις και τις αλληλεξαρτήσεις που μοιραία παρουσιάζει με πάρα πολλούς άλλους τομείς της επιστήμης των υπολογιστών (βλέπε την Εικόνα 1) τόσο το εύρος των εφαρμογών – απαιτήσεων που δημιουργεί καθώς και το τεράστιο ενδιαφέρον που εμφανίζει, τόσο επιστημονικό όσο και οικονομικό, το κατέστησαν μια ξεχωριστή θεματική ενότητα. Απευθείας υιοθέτηση μεθόδων και τεχνικών που έχουν εφαρμοστεί σε σχετικά πεδία όπως η μηχανική μάθηση, η στατιστική, και τα συστήματα βάσεων δεδομένων δεν μπορούν να λύσουν τα πολλά προκλητικά ερευνητικά θέματα που προκύπτουν. Είναι απαραίτητο να πραγματοποιηθούν στοχευμένες έρευνες προκειμένου να ανακαλύψουμε νέες μεθόδους ή να αναπτύξουμε ολοκληρωμένες τεχνικές για αποτελεσματικό και αποδοτικό Data Mining. Σαν θέμα αποτελεί ένα από τα πιο ενεργά πεδία στις ημέρες μας και θα συνεχίσει να αποτελεί για πάρα πολλά χρόνια ακόμα, καθώς ο όγκος των δεδομένων θα συνεχίσει να αυξάνει με ακόμα μεγαλύτερους ρυθμούς (κάθε χρόνο παράγονται και αποθηκεύονται τεράστιοι όγκοι δεδομένων από οργανισμούς και επιχειρήσεις, της τάξεως των 1-2 exa-bytes) και η ανάγκη των χρηστών για ανάλυση και χρήση αυτών των δεδομένων θα συνεχίσει να είναι μεγάλη. Αυτό μπορεί να γίνει εύκολα ορατό αν κάποιος κοιτάξει το πλήθος και την συχνότητα των συνεδρίων και των δημοσιεύσεων που πραγματοποιούνται στο θέμα, το πλήθος των διδακτορικών διατριβών παγκοσμίως, τα εκδιδόμενα περιοδικά, καθώς και τις εμπλεκόμενες εταιρείες του χώρου.

Η παρούσα διατριβή ασχολήθηκε με την εφαρμογή τεχνικών data mining σε συστήματα ηλεκτρονικού εμπορίου. Ειδικότερα επικεντρωθήκαμε στην εύρεση κανόνων συσχετίσεων από δεδομένα, και κύρια δεδομένα που είχαν να κάνουν με βάσεις συναλλαγών. Η διατριβή σαν κείμενο μπορεί να χωριστεί σε τρία κομμάτια. Το πρώτο είναι τα τρία πρώτα κεφάλαια, τα οποία και αποτελούν εισαγωγικά κεφάλαια απαραίτητα για την υποστήριξη και κατανόηση της δουλειάς μας. Πιο συγκεκριμένα το παρόν κεφάλαιο αποτελεί μια γενικότερη εισαγωγή πάνω στο θέμα του Data Mining, δίνοντας στον αναγνώστη κάποιες γενικές έννοιες και μια σύντομη περιγραφή του χώρου. Στο κεφάλαιο 2 κάνουμε μια επισκόπηση των κυριότερων τεχνικών κανόνων συσχετίσεων, αφού αυτό αποτέλεσε και το βασικό αντικείμενο μελέτης μας. Τέλος στο κεφάλαιο 3 κάνουμε μια σύντομη παράθεση τεχνικών και ορισμών από την ανάκτηση πληροφορίας (Information Retrieval) μιας και χρησιμοποιήθηκαν εκτενώς σε ορισμένα σημαντικά σημεία της ερευνάς μας. Ακολούθως τα κεφάλαια 4 έως 8 αποτελούν το δεύτερο και κύριο κομμάτι της παρούσας διατριβής, και περιγράφουν διάφορες τεχνικές και προτάσεις μας, αποτελέσματα της ερευνάς μας.

Εικόνα 1: Οι κυριότεροι τομείς αλληλεπίδρασης του data mining

Έτσι στο κεφάλαιο 4 εισάγουμε την έννοια των “καυτών προϊόντων” (hot items), αιτιολογούμε την ύπαρξη και την χρησιμότητα τους και δίνουμε έναν αλγόριθμο για την αποτελεσματική εξεύρεση τους. Στο κεφάλαιο 5 εξετάζουμε τα αντικείμενα δεδομένων βάσεων λιανεμπορίου καθώς και τους παραγόμενους κανόνες συσχετίσεων μέσα από το κατάλληλο μικροοικονομικό πλαίσιο και προτείνουμε ένα μηχανισμό απόδοσης βαρών ο οποίος δίνει τις κατάλληλες τιμές υποστήριξης σε όλα τα αντικείμενα, αυτόματα βρίσκει τα πιο σημαντικά και τελικά αποδεικνύεται αποτελεσματικός. Το κεφάλαιο 6 έχει να κάνει με την πρόταση ενός νέου αλγορίθμου εύρεσης κανόνων συσχετίσεων με χρήση πολλαπλών τιμών υποστήριξης χρησιμοποιώντας σαν βάση του τον συνδυασμό δύο υπαρχόντων αλγορίθμων. Στο κεφάλαιο 7 προτείνεται μια επαναστατική τεχνική που έχει να κάνει με την χρήση μεθόδων δανεισμένων από την ανάκτηση πληροφορίας στην διαδικασία της ανακάλυψης γνώσης. Τέλος κλείνουμε με το κεφάλαιο 8 όπου και παρουσιάζουμε την ιδέα των τοπικά συχνών αντικειμένων (locally frequent items) σε μια βάση, και αντιστοίχως προτείνουμε δύο διαφορετικούς αλγορίθμους για την εξεύρεση τους. Το τρίτο και τελευταίο κομμάτι της διατριβής, αναφορικά το Κεφάλαιο 9, αποτελεί την σύνοψη ολόκληρης της εργασίας μας όπου παραθέτουμε εν συντομία την τελική προσφορά μας στο χώρο, δίνουμε πιθανές εφαρμογές των προτάσεων μας, και τέλος προτείνουμε μελλοντικές κατευθύνσεις της έρευνας σε ανοιχτά πεδία – προβλήματα.

1.2 Τι είναι το Data Mining

Πολλοί θεωρούν το Data Mining σαν συνώνυμο ενός άλλου πολύ γνωστού και διαδεδομένου όρου, αναφορικά αυτόν της Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD). Εναλλακτικά, άλλοι βλέπουν το Data Mining σαν ένα βασικό βήμα στην διαδικασία της ανακάλυψης γνώσης από βάσεις δεδομένων. Η δεύτερη αυτή θεώρηση είναι και η πιο ακριβής κατά τη γνώμη μας, και είναι και αυτή που θα χρησιμοποιήσαμε στην παρούσα διατριβή. Για να διαφοροποιούμε την ανακάλυψη γνώσης από το Data Mining θα χρησιμοποιούμε από εδώ και στο εξής τον όρο εξόρυξη δεδομένων όταν αναφερόμαστε στο Data Mining. Η ανακάλυψη γνώσης σαν διαδικασία αποτελείται ουσιαστικά από μια επαναληπτική ακολουθία των παρακάτω βημάτων:

1. Καθαρισμός δεδομένων (data cleaning), όπου ουσιαστικά απομακρύνουμε τον θόρυβο και ακατάλληλα δεδομένα.
2. Ενοποίηση δεδομένων (data integration), όπου πιθανώς να έχουμε πολλαπλές πηγές δεδομένων οι οποίες θα πρέπει να συνδυαστούν.
3. Επιλογή δεδομένων (data selection), όπου δεδομένα σχετικά με την διαδικασία της ανάλυσης μας θα πρέπει να επιλεγθούν και να ανακτηθούν από τη βάση μας.
4. Μετατροπή δεδομένων (data transformation), όπου τα δεδομένα μας θα πρέπει να μετατραπούν σε μια ενιαία μορφή κατάλληλη προς επεξεργασία.
5. Εξόρυξη δεδομένων (Data Mining), μια ουσιαστική διαδικασία όπου εφαρμόζονται ευφρείς μέθοδοι προκειμένου να εξαγάγουμε μοτίβα – πρότυπα από τα δεδομένα μας.
6. Αξιολόγηση μοτίβων (Pattern evaluation), η διαδικασία κατά την οποία αναγνωρίζουμε και ξεχωρίζουμε τα πραγματικά ενδιαφέροντα μοτίβα με χρήση μετρικών ενδιαφέροντος (interestingness measures)
7. Αναπαράσταση γνώσης (knowledge presentation), όπου εφαρμόζουμε τεχνικές οπτικοποίησης και αναπαράστασης γνώσης προκειμένου να παρουσιάσουμε καλύτερα την εξαγόμενη γνώση στους χρήστες.

Τα βήματα αυτά καθώς και ολόκληρη η διαδικασία ανακάλυψης γνώσης φαίνονται στην Εικόνα 2. Ο εκάστοτε χρήστης μπορεί να επανέλθει σε οποιοδήποτε βήμα αν κάτι δεν πήγε καλά ή αν απλά δεν είναι ευχαριστημένος από κάποιο αποτέλεσμα όπως επίσης μπορεί να ξεκινήσει την όλη διαδικασία όχι από την αρχή της αλλά από οποιοδήποτε ενδιάμεσο βήμα.

Εικόνα 2: Η διαδικασία ανακάλυψης γνώσης

1.3 Σε ποια είδη βάσεων – δεδομένων μπορούμε να κάνουμε εξόρυξη δεδομένων

Η εφαρμογή τεχνικών εξόρυξης δεδομένων μπορεί θεωρητικά να εφαρμοστεί σε οποιαδήποτε είδος δεδομένων. Στην πράξη όμως ορισμένα είδη παρουσιάζουν και το μεγαλύτερο ενδιαφέρον από τους χρήστες. Ας δούμε ποια είναι αυτά:

1.3.1 Σχεσιακές βάσεις δεδομένων

Ίσως το μεγαλύτερο μέρος των εμπορικών εφαρμογών, και ως εκ τούτου των πιο ενδιαφερόντων δεδομένων από οικονομικής απόψεως, είναι αποθηκευμένα σε σχεσιακές βάσεις δεδομένων. Μια σχεσιακή βάση δεδομένων είναι ουσιαστικά μια συλλογή από πίνακες, κάθε ένας από τους οποίους έχει ένα μοναδικό όνομα. Κάθε πίνακας αποτελείται από ένα σύνολο πεδίων (συνήθως στήλες) και σε αυτόν βρίσκονται αποθηκευμένα ένας μεγάλος αριθμός δεδομένων (εγγραφών). Κάθε εγγραφή σε έναν σχεσιακό πίνακα αναπαριστά ένα αντικείμενο και χαρακτηρίζεται από ένα μοναδικό “κλειδί”. Τα σχεσιακά δεδομένα μπορούν να επεξεργαστούν ή να αναλυθούν μέχρι κάποιο βαθμό με χρήση ερωτημάτων γραμμένων σε γλώσσα SQL ή με χρήση γραφικών περιβαλλόντων. Παράδειγμα τέτοιων ερωτημάτων θα μπορούσαν να ήταν “Δώσε μου τις πωλήσεις των τελευταίων 2 μηνών ανά κατάστημα” ή “ποια μετοχή είχε τη μεγαλύτερη μεταβολή το τελευταίο έτος”. Με την χρήση τεχνικών εξόρυξης δεδομένων τώρα κάποιος μπορεί να εισχωρήσει βαθύτερα στα δεδομένα και να ψάξει για μοτίβα ή τάσεις σε αυτά. Για παράδειγμα, ένα τέτοιο σύστημα μπορεί να αναλύσει τα δεδομένα των πελατών και να προβλέψει μελλοντικές συμπεριφορές βασισμένο σε προηγούμενα δεδομένα.

1.3.2 Βάσεις δεδομένων συναλλαγών

Γενικά μια βάση δεδομένων συναλλαγών αποτελείται από ένα αρχείο όπου κάθε εγγραφή αναπαριστά μια συναλλαγή. Μια συναλλαγή συνήθως περιλαμβάνει έναν μοναδικό αριθμό, και μια λίστα των αντικειμένων που αποτελούν την συναλλαγή (όπως τα προϊόντα που αγοράζονται σε ένα κατάστημα). Μια τέτοια βάση μπορεί να έχει επιπρόσθετους πίνακες συσχετισμένους με αυτή, οι οποίοι περιέχουν πρόσθετες πληροφορίες σχετικά με την πώληση, όπως την ημερομηνία και την ώρα πραγματοποίησης της συναλλαγής, τον κωδικό αριθμό του πελάτη, τον κωδικό αριθμό του πωλητή, το κατάστημα στο οποίο πραγματοποιήθηκε η συναλλαγή, κ.ο.κ. Με χρήση ενός συστήματος εξόρυξης δεδομένων μπορούμε να πραγματοποιούμε ανάλυση δεδομένων και να βρίσκουμε στοιχεία όπως π.χ. ποια προϊόντα αγοράζονται συνήθως μαζί και να προγραμματίζουμε έτσι καλύτερα την προώθησή τους.

Αντίθετα ένα απλό σύστημα ανάκτησης πληροφορίας δεν είναι σε θέση να προχωρήσει σε τέτοιο βάθος στην ανάλυση και εξόρυξη των δεδομένων από τέτοιες βάσεις.

1.3.3 Χωρικές και Χρονικές Βάσεις

Οι χωρικές βάσεις δεδομένων περιέχουν δεδομένα που καθορίζονται – περιλαμβάνουν μια χωρική διάσταση. Τέτοιες βάσεις είναι οι γεωγραφικές βάσεις (χάρτες), βάσεις σχετικά με την σχεδίαση VLSI κυκλωμάτων, ιατρικές εικόνες καθώς και εικόνες δορυφόρων. Για παράδειγμα σε μια βάση που έχει καταχωρημένη την κατανομή πλούτου σε σχέση με μια γεωγραφική περιοχή μπορούμε να ανακαλύψουμε τάσεις συγκεντρώσεων ή αραιώσεων πληθυσμών. Οι συγκεκριμένες βάσεις έχουν μια πληθώρα εφαρμογών όπως οικολογία, logistics, χωροταξία κ.α. Αναφορικά ορισμένες εργασίες σχετικές και με την εξόρυξη δεδομένων που χρησιμοποιούν χωρικά δεδομένα μπορούν να βρεθούν στα [95] [96].

Οι χρονικές βάσεις έχουν, όπως εύκολα μπορεί να γίνει κατανοητό, δεδομένα τα οποία περιέχουν και τη χρονική διάσταση. Η διάσταση αυτή μπορεί να είναι απλά η ημερομηνία ή ώρα πραγματοποίησης ενός γεγονότος ή η καταχώρηση πολλαπλών τιμών χρονικής καταγραφής κάποιων παραμέτρων. Σε αυτού του είδους τις βάσεις οι τεχνικές εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν προκειμένου να βρουν μεταβολές σε σχέση με το χρόνο, ή τάσεις μεταβολής διαφόρων αντικειμένων. Τέτοιες πληροφορίες μπορεί να είναι ιδιαιτέρως χρήσιμες στην λήψη αποφάσεων ή στην χάραξη στρατηγικής σε επιχειρήσεις. Για παράδειγμα οι μεταβολές των τιμών μετοχών σε σχέση με το χρόνο μπορεί να μας αποκαλύψουν πότε είναι η κατάλληλη περίοδος για αγορά ή πώληση μιας μετοχής. Στον τομέα της εξόρυξης δεδομένων έχουν γίνει κάποιες εργασίες πάνω σε χρονικά δεδομένα [38][120].

Τέλος στις χωροχρονικές βάσεις έχουμε συνδυασμό της χωρικής και της χρονικής διάστασης στα δεδομένα μας. Παράδειγμα σε μια βάση συναλλαγών αυτό μπορεί να σήμαινε ότι καταγράφαμε και λαμβάναμε υπόψη την ακριβή χρονική στιγμή όπου ελάμβανε χώρα μια αγορά καθώς και την ακριβή τοποθεσία του αντικειμένου στο κατάστημα μας. Μια τέτοια θεώρηση καθώς και μια προσέγγιση προτάθηκε στο [80] όπου τα δεδομένα στη βάση επεξεργάζονταν με τη χρήση τεχνικών γράφων [35]. Άλλα ερωτήματα – πληροφορίες που μπορούν να εξαχθούν από αυτού του είδους τις βάσεις μπορεί να είναι: εύρεση εταιριών με παρόμοιες τάσεις ανάπτυξης, προϊόντα με παρόμοιες κατανομές πωλήσεων, μετοχές με παρόμοιες κινήσεις ως προς την αξία τους, εικόνες μετεωρολογίας με παρόμοιες κινήσεις, γεωλογικά χαρακτηριστικά, οικολογικές καταστροφές, ή μοτίβα στον τομέα της αστροφυσικής.

1.3.4 Βάσεις κειμένων και πολυμεσικές Βάσεις

Οι βάσεις κειμένων είναι βάσεις οι οποίες περιέχουν λέξεις ή ολόκληρα κείμενα, ή εναλλακτικά που περιέχουν λεκτικές περιγραφές αντικειμένων. Αυτές οι περιγραφές μπορούν να κυμαίνονται από απλές λέξεις κλειδιά, μέχρι ολόκληρες προτάσεις, όπως για παράδειγμα περιγραφές προϊόντων, απαντήσεις σε ερωτήματα – παράπονα χρηστών σε ένα call-center κ.α. Η πληροφορία που μπορεί να ανακαλύψει κάποιος από τέτοιες βάσεις είναι ανεξάντλητη (παράδειγμα [62][63][64][65]). Παράδειγμα από μια βάση κειμένων μπορεί να δημιουργήσει έναν θησαυρό λέξεων, ή μια λίστα συνωνύμων. Από μια βάση παραπόνων - απαντήσεων χρηστών σε ένα call – center μπορεί πάλι να δημιουργήσει μια λίστα σχετικών αυτοματοποιημένων απαντήσεων σε αντίστοιχα ερωτήματα.

Οι βάσεις πολυμέσων αποθηκεύουν ήχο, στατική και κινούμενη εικόνα, καθώς και κείμενο και έχουν ποικίλες εφαρμογές. Οι συγκεκριμένες βάσεις είναι συνήθως πάρα πολύ μεγάλες σε μέγεθος λόγω της φύσεως των δεδομένων που αποθηκεύουν. Η χρήση των τεχνικών εξόρυξης δεδομένων μπορεί να μας απαλλάξει από διάφορα προβλήματα και δυσκολίες που συναντάμε στις συγκεκριμένες εφαρμογές, όπως την εύρεση και την εξαγωγή πολλαπλών χαρακτηριστικών από τα πολυμεσικά δεδομένα, εύρεση με βάση κάποια μετρική ομοιότητας κ.α.

1.3.5 Ο Παγκόσμιος Ιστός

Ο παγκόσμιος ιστός και οι σχετιζόμενες καταναεμημένες υπηρεσίες πληροφοριών παρέχουν μια υποδομή βάσεων όπου τα δεδομένα είναι αλληλοσυνδεδεμένα μεταξύ τους προκειμένου να διευκολύνουν την αλληλεπιδραστική πρόσβαση. Οι χρήστες του παγκόσμιου ιστού περιπλανιόνται από τη μια ιστοσελίδα στην άλλη ψάχνοντας για τις πληροφορίες που τους ενδιαφέρουν. Ο παγκόσμιος ιστός έτσι λόγω του όγκου της πληροφορίας που υπάρχει διαθέσιμη αλλά και λόγω του πλήθους των χρηστών που το επισκέπτονται δίνει πολλές δυνατότητες και ευκαιρίες εφαρμογών εξόρυξης δεδομένων. Για παράδειγμα μπορούμε να κατανοήσουμε καλύτερα τις συνήθειες περιπλάνησης των χρηστών (path traversal patterns) στις διάφορες σελίδες και έτσι να τους κατευθύνουμε καλύτερα ή να κατηγοριοποιήσουμε σελίδες σύμφωνα με διάφορες παραμέτρους.

1.4 Τι είδη τεχνικών μπορούν να εφαρμοστούν

Όπως η εξόρυξη γνώσης μπορεί να εφαρμοστεί σε διάφορους τύπους δεδομένων, αντίστοιχα μπορούμε να εφαρμόσουμε διαφορετικές τεχνικές και να ανακαλύψουμε αρκετούς τύπους μοτίβων (patterns) από τα δεδομένα μας. Παρακάτω δίνουμε εν συντομία τις σημαντικότερες από αυτές:

1.4.1 Κανόνες συσχετίσεων.

Η σημαντικότερη και δημοφιλέστερη απ' όλες τις διεργασίες εξόρυξης γνώσης, τόσο από οικονομικής όσο και από επιστημονικής απόψεως, είναι αδιαμφισβήτητα οι κανόνες συσχετίσεων. Οι κανόνες συσχετίσεων συνίστανται στο να ανακαλύψουμε από μια βάση δεδομένων ένα σύνολο ισχυρών κανόνων της μορφής $X \Rightarrow Y$, όπου τα X και Y αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα υπό ανάλυση δεδομένα. Παράδειγμα ενός τέτοιου κανόνα είναι:

Φορητός_Υπολογιστής, τσάντα $\rightarrow$ ADSL [sup=5%, conf=80%]

Ο κανόνας αυτός μας υποδυκνείει ότι οι πελάτες οι οποίοι αγοράζουν φορητό υπολογιστή και τσάντα μεταφοράς του αγοράζουν και σύνδεση ADSL με ένα ποσοστό υποστήριξης για 5% και βεβαιότητας 80% για τον κανόνα. Με άλλα λόγια το 80% των πελατών που αγοράζουν φορητό υπολογιστή και τσάντα αγοράζουν επίσης σύνδεση ADSL, και το 5% επι του συνόλου όλων των πελατών τα αγοράζουν όλα αυτά μαζί. Το πρώτο κομμάτι ενός κανόνα αποκαλείται συνήθως υπόθεση του κανόνα ενώ το δεύτερο συμπέρασμα. Ο κανόνας $X \Rightarrow Y$ ισχύει στο σύνολο των δεδομένων με βεβαιότητα (confidence) c , αν $c\%$ των εγγραφών που περιέχουν το X περιέχουν επίσης και το Y , ενώ λέμε ότι έχει υποστήριξη (support) s , αν $s\%$ των εγγραφών περιέχουν την $X \cup Y$. Το πρόβλημα της εύρεσης κανόνων συσχετίσεων μπορεί να διαχωριστεί στα ακόλουθα υπό-προβλήματα:

1. Εύρεση όλων των αντικειμένων που έχουν υποστήριξη πάνω από το κατώφλι υποστήριξης. Αυτά καλούνται συχνά αντικείμενα. Όλα τα υπόλοιπα καλούνται σπάνια.
2. Παραγωγή όλων των κανόνων συσχετίσεων που ικανοποιούν το κατώφλι βεβαιότητας χρησιμοποιώντας σαν είσοδο τα συχνά αντικείμενα.

Η αρχική και κύρια εφαρμογή των κανόνων συσχετίσεων ήταν τα δεδομένα λιανεμπορίου, και ειδικότερα αυτά που αφορούσαν στις πωλήσεις σουπερμάρκετ γι' αυτό και συχνά αναφέρονται και σαν πρόβλημα των καλαθιών αγορών (market basket analysis). Αυτή η πληροφορία μπορεί να είναι ιδιαίτερα χρήσιμη στην λήψη διαφόρων αποφάσεων όπως προωθήσεις προϊόντων, χωροθέτηση προϊόντων σε καταστήματα, στόχευση πελατών, διαχείριση αποθεμάτων κ.α. Βέβαια κάθε σύνολο δεδομένων που ταιριάζει σ' αυτό το υπόδειγμα μπορεί να αναλυθεί με χρήση των κανόνων συσχετίσεων. Στο επόμενο κεφάλαιο (Κεφάλαιο 2) κάνουμε μια λεπτομερή περιγραφή των κανόνων συσχετίσεων.

1.4.2 Κατηγοριοποίηση

Μια άλλη πολύ σημαντική εφαρμογή της εξόρυξης γνώσης είναι η δυνατότητα να πραγματοποιούμε κατηγοριοποίηση (classification) σε τεράστιο όγκο δεδομένων. Αυτό αναφέρεται εναλλακτικά και σαν εξόρυξη κανόνων κατηγοριοποίησης (mining classification rules). Η κατηγοριοποίηση προσπαθεί να κατηγοριοποιήσει ένα σύνολο δεδομένων με βάση τις τιμές τους σε συγκεκριμένες παραμέτρους σε κατηγορίες που έχουμε ορίσει εκ των προτέρων. Για παράδειγμα, μια αντιπροσωπεία αυτοκινήτων μπορεί να επιθυμεί να κατηγοριοποιήσει τους πελάτες της σύμφωνα με τις προτιμήσεις τους για αυτοκίνητα έτσι ώστε οι πωλητές της να γνωρίζουν ποιόν και με ποιο τρόπο θα προσεγγίσουν και τι θα του προτείνουν έτσι ώστε να μεγιστοποιούν τις πιθανότητες επιτυχίας. Η βασική εργασία της κατηγοριοποίησης είναι η δημιουργία ενός μοντέλου το οποίο θα χρησιμοποιείται για να κατηγοριοποιεί δεδομένα τα οποία δεν έχουμε κατηγοριοποιήσει. Η κατηγοριοποίηση είναι μια διαδικασία δύο βημάτων, αναφορικά:

1. Εκμάθηση (Learning). Με χρήση ενός μέρους των δεδομένων μας, τα οποία ονομάζονται δεδομένα εκπαίδευσης (training data), χτίζουμε ένα μοντέλο περιγράφοντας ένα προκαθορισμένο σύνολο από κατηγορίες δεδομένων.
2. Κατηγοριοποίηση (Classification). Έχοντας το μοντέλο που προέκυψε από το προηγούμενο βήμα προσπαθούμε με χρήση δοκιμαστικών παραδειγμάτων (training samples) να επιβεβαιώσουμε την ακρίβεια του. Αν έχει τελικά μια αποδεκτή ακρίβεια τότε θα χρησιμοποιηθεί για την κατηγοριοποίηση νέων δεδομένων αλλά και δεδομένων τα οποία δεν ανήκουν σε κάποια κατηγορία.

Οι πιο γνωστές μέθοδοι κατηγοριοποίησης είναι η Bayesian, τα δέντρα απόφασης (decision trees) [94], τα νευρωνικά δίκτυα (neural networks), η κατηγοριοποίηση κοντινότερων γειτόνων (nearest neighbor) [53][46][59], και τα Support Vector Machines [47][71].

1.4.3 Συσταδοποίηση

Ένας τυπικός ορισμός της συσταδοποίησης (clustering) είναι ο εξής: Συσταδοποίηση καλείται η εργασία καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων (clusters). Αντίθετα με την κατηγοριοποίηση, η συσταδοποίηση δεν βασίζεται σε προκαθορισμένες κατηγορίες αλλά τα δεδομένα ομαδοποιούνται με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους. Ακολουθώς εξαρτάται από τον χρήστη αλλά και από το προσδοκώμενο αποτέλεσμα η σημασία που θα δώσουμε στις προκύπτουσες συστάδες. Παράδειγμα μιας τέτοιας εφαρμογής θα μπορούσε να ήταν σε μια βάση με δεδομένα πωλήσεων λιανεμπορίου να ομαδοποιήσουμε τους πελάτες που εμφανίζουν παρόμοιες

αγοραστικές συνήθειες. Η αρχή με την οποία πραγματοποιείται η συσταδοποίηση είναι η εξής: μεγιστοποίησε την ενδό-συσταδική (intra-class) ομοιότητα και ελαχιστοποίησε την δια-συσταδική (inter-class) ομοιότητα [69]. Η όλη διαδικασία μπορεί να χωριστεί στα ακόλουθα βήματα:

1. Επιλογή χαρακτηριστικών γνωρισμάτων, όπου επιλέγουμε τα γνωρίσματα (attributes) αυτά που θα μας συμπεριλάβουν τελικά την πληροφορία που χρειαζόμαστε.
2. Αλγόριθμος συσταδοποίησης. Επιλογή του καταλληλότερου αλγορίθμου συσταδοποίησης για την εργασία μας, με βάση το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης που χρησιμοποιεί.
3. Επικύρωση αποτελεσμάτων. Έλεγχος της ακρίβειας των τελικών αποτελεσμάτων με χρήση διαφόρων μετρικών και κριτηρίων.
4. Ερμηνεία και παρουσίαση αποτελεσμάτων. Παρουσίαση και περαιτέρω ανάλυση της εξαχθείσας γνώσης με ειδικούς και από άλλους τομείς προκειμένου να χρησιμοποιηθεί κατά τον βέλτιστο τρόπο.

Να αναφέρουμε τέλος ότι η συσταδοποίηση έχει μελετηθεί και σε άλλους τομείς* όπως η στατιστική [36][70], η μηχανική μάθηση [50][51], οι χωρικές βάσεις δεδομένων [20], και στην εξόρυξη γνώσης [97][128].

1.5 Απαιτήσεις και προκλήσεις της Εξόρυξης Δεδομένων

Προκειμένου να πραγματοποιήσουμε αποτελεσματικά εξόρυξη δεδομένων, χρειάζεται να εξετάσουμε πρώτα τι χαρακτηριστικά πρέπει να έχει ένα τέτοιο σύστημα στην πράξη και τι προκλήσεις – ιδιαιτερότητες μπορεί να αντιμετωπίσει κάποιος κατά την ανάπτυξη τεχνικών εξόρυξης δεδομένων [37].

1. Διαχείριση διαφορετικών ειδών δεδομένων

Επειδή υπάρχουν πολλά είδη δεδομένων και βάσεων που χρησιμοποιούνται σε διάφορες εφαρμογές, κάποιος θα περίμενε ένα σύστημα ανακάλυψης γνώσης να είναι σε θέση να πραγματοποιεί αποτελεσματικά εξόρυξη δεδομένων σε διαφορετικά είδη δεδομένων. Αφού οι περισσότερες διαθέσιμες βάσεις είναι σχεσιακές, είναι κρίσιμο ένα σύστημα εξόρυξης δεδομένων να μπορεί εργάζεται τόσο αποτελεσματικά όσο και αποδοτικά σε σχεσιακά δεδομένα. Επιπρόσθετα, πολλές βάσεις περιέχουν σύνθετους τύπους δεδομένων, όπως

* Γι' αυτό και είναι γνωστή με διάφορα ονόματα όπως μη εποπτευόμενη μάθηση (unsupervised learning), αριθμητική ταξονομία (numerical taxonomy), τμηματοποίηση (partition).

δομημένα δεδομένα και πολύπλοκα αντικείμενα, υπερκείμενα και πολυμεσικά δεδομένα, χωρικά, χρονικά και χωροχρονικά δεδομένα, δεδομένα συναλλαγών, οικονομικά δεδομένα κοκ. Έτσι ένα δυνατό σύστημα θα έπρεπε να είναι σε θέση να πραγματοποιεί αποτελεσματικά εξόρυξη δεδομένων σε τέτοια πολύπλοκα δεδομένα. Παρόλ' αυτά, η ανομοιομορφία των τύπων δεδομένων και των διαφορετικών στόχων της εξόρυξης δεδομένων κάνουν μη ρεαλιστική την επιδίωξη και την προσδοκία για ένα ενιαίο σύστημα το οποίο θα μπορεί να διαχειρίζεται όλων των ειδών τα δεδομένα. Ειδικά συστήματα εξόρυξης δεδομένων θα πρέπει να δημιουργηθούν για την ανακάλυψη γνώσης ανάλογα με το είδος των δεδομένων, όπως συστήματα αφιερωμένα στην ανακάλυψη γνώσης από σχεσιακά δεδομένα, δεδομένα συναλλαγών, χωρικές βάσεις, πολυμεσικές βάσεις κοκ.

2. Αποτελεσματικότητα και κλιμάκωση (scalability) των αλγορίθμων εξόρυξης δεδομένων.

Προκειμένου να εξαγάγουμε αποτελεσματικά πληροφορίες και γνώση από τεράστιες ποσότητες δεδομένων, οι χρησιμοποιούμενοι αλγόριθμοι πρέπει να είναι αποτελεσματικοί και κλιμακούμενοι ιδιαίτερα σε μεγάλες βάσεις. Αυτό σημαίνει, ότι ο χρόνος εκτέλεσης ενός αλγορίθμου εξόρυξης δεδομένων πρέπει να είναι προβλέψιμος και αποδεκτός σε μεγάλες βάσεις δεδομένων. Αλγόριθμοι με εκθετική ή ακόμα και μεσαίας τάξης πολυωνυμική πολυπλοκότητα δεν έχουν πρακτική χρήση.

3. Χρησιμότητα, βεβαιότητα και εκφραστικότητα αποτελεσμάτων εξόρυξης δεδομένων.

Η ανακαλυπτόμενη γνώση θα πρέπει να απεικονίζει αποτελεσματικά τα περιεχόμενα της βάσης και να είναι χρήσιμη για συγκεκριμένες εφαρμογές. Οι ατέλειες θα πρέπει να εκφράζονται με διαφόρους τρόπους αλλά και μετρικές αβεβαιότητας, όπως για παράδειγμα με χρήση κατά προσέγγιση κανόνων ή ποσοτικών κανόνων. Ο θόρυβος και ειδικά δεδομένα θα πρέπει να διαχειρίζονται ορθά από τα συστήματα εξόρυξης δεδομένων. Αυτό επίσης μας παρακινεί για μια συστηματική μελέτη της ποιότητας της ανακαλυπτόμενης γνώσης, συμπεριλαμβανομένου του ενδιαφέροντος και της αξιοπιστίας, δημιουργώντας μοντέλα και εργαλεία προσομοίωσης, αναλυτικά και στατιστικά.

4. Διατύπωση διαφόρων ειδών αποτελεσμάτων εξόρυξης δεδομένων.

Διάφορα είδη γνώσης μπορούν να αποκαλυφτούν από μεγάλες ποσότητες δεδομένων. Επίσης, κάποιος μπορεί να θέλει να εξετάσει την γνώση που έχει βρει από διαφορετικές απόψεις και να την παρουσιάσει με διαφορετικούς τρόπους. Αυτό απαιτεί να εκφράσουμε τόσο τις ανάγκες των απαιτήσεων της ίδιας της εξόρυξης δεδομένων όσο και την ανακαλυπτόμενη γνώση με γλώσσες υψηλού επιπέδου και γραφικά περιβάλλοντα έτσι ώστε το έργο του εκάστοτε χρήστη να μπορεί να προδιαγραφεί και από μη ειδικούς και η γνώση

που ανακαλύπτουμε να είναι κατανοητή και άμεσα χρησιμοποιήσιμη από τους χρήστες. Αυτό επίσης προϋποθέτει το σύστημα ανακάλυψης γνώσης να υιοθετεί εκφραστικές τεχνικές αναπαράστασης γνώσης.

5. Ανακάλυψη γνώσης από διαφορετικές πηγές δεδομένων

Τα ευρέως διαδεδομένα τοπικά και ευρεία δίκτυα υπολογιστών, συμπεριλαμβανομένου του Ίντερνετ, συνδέουν πολλές πηγές δεδομένων από τεράστιες κατανεμημένες και ανομοιογενείς βάσεις. Η ανακάλυψη γνώσης από διαφορετικές πηγές μορφοποιημένων και μη δεδομένων με διαφορετικές υποστάσεις και έννοιες θέτει νέες προκλήσεις στην εξόρυξη δεδομένων. Από την άλλη, η εξόρυξη δεδομένων μπορεί να βοηθήσει στην αποκάλυψη των υψηλού επιπέδου ομοιοτήτων σε ανομοιογενείς βάσεις οι οποίες πολύ δύσκολα μπορούν να αποκαλυφθούν από απλά συστήματα δημιουργίας ερωτημάτων (query systems). Επιπρόσθετα, το τεράστιο μέγεθος της βάσης, ο μεγάλος καταμερισμός των δεδομένων, και η υπολογιστική πολυπλοκότητα ορισμένων μεθόδων εξόρυξης δεδομένων δρουν ως κίνητρα για την ανάπτυξη παράλληλων και κατανεμημένων αλγορίθμων εξόρυξης δεδομένων.

6. Διαφύλαξη της ιδιωτικότητας και της ασφάλειας των δεδομένων

Όταν τα δεδομένα μπορούν να εξετάζονται από διαφορετικές σκοπιές και σε διαφορετικά επίπεδα, διακυβεύεται ο σκοπός της προστασίας της ασφάλειας των δεδομένων και η διαφύλαξη τους από παραβίαση της ιδιωτικότητας. Είναι σημαντικό να εξετάσουμε πότε η ανακάλυψη γνώσης μπορεί να οδηγήσει σε τέτοιες περιπτώσεις, και τι μέτρα ασφαλείας μπορούν και θα πρέπει να προβλεφθούν και να υλοποιηθούν για την αποτροπή τυχόν αποκάλυψης ευαίσθητων πληροφοριών.

Σημειωτέον ορισμένες από αυτές τις απαιτήσεις μπορεί να δημιουργούν αντικρουόμενους στόχους. Για παράδειγμα, ο στόχος της προστασίας της ασφάλειας των δεδομένων μπορεί να έρχεται σε σύγκρουση με την απαίτηση της διαδραστικής εξόρυξης δεδομένων ή της απόκτησης πολλαπλής γνώσης από διαφορετικές γωνίες (για περισσότερες πληροφορίες στο θέμα ο αναγνώστης μπορεί να ανατρέξει στα [122][12][39][43][48]).

1.6 Κατηγοριοποίηση τεχνικών Εξόρυξης Δεδομένων

Έχει πραγματοποιηθεί μεγάλη προσπάθεια στην έρευνα και την εξέλιξη της εξόρυξης δεδομένων, και πολλές τεχνικές και συστήματα έχουν αναπτυχθεί. Για την κατηγοριοποίηση των διαφόρων τεχνικών και συστημάτων εξόρυξης δεδομένων αλλά και των χρηστών τους μπορούν να χρησιμοποιηθούν διάφορα σχήματα βασισμένα σε ποια είδη βάσεων εργαζόμαστε,

τα είδη της γνώσης που μπορούν να αποκαλυφθούν, τις χρησιμοποιούμενες τεχνικές, και τέλος τις εφαρμογές στις οποίες υλοποιούνται όπως φαίνεται παρακάτω.

▮ Σε ποια είδη βάσεων εργαζόμαστε.

Ένα σύστημα εξόρυξης δεδομένων μπορεί να κατηγοριοποιηθεί σύμφωνα με τα είδη των βάσεων όπου εφαρμόζεται. Για παράδειγμα, ένα σύστημα είναι σχεσιακό αν ανακαλύπτει γνώση από σχεσιακές βάσεις δεδομένων. Γενικότερα τα πιο γνωστά είδη βάσεων που χρησιμοποιούνται και μπορούν να κατηγοριοποιηθούν μια τέτοια τεχνική ή σύστημα είναι: σχεσιακές βάσεις, βάσεις συναλλαγών, αντικειμενοστραφείς βάσεις, χωρικές βάσεις, χρονικές βάσεις, χωροχρονικές βάσεις, πολυμεσικές βάσεις, ανομοιογενείς βάσεις, βάσεις δεδομένων διαδικτύου κ.α..

▮ Τι είδη γνώσης μπορούν να ανακαλυφθούν.

Διάφοροι τύποι γνώσης μπορούν να ανακαλυφθούν από τους χρήστες, συμπεριλαμβανομένων των κανόνων συσχέτισεων, των χαρακτηριστικών κανόνων, των κανόνων ταξινόμησης (classification rules) και την ταξινόμηση, των κανόνων διαφοροποίησης (discriminant rules), την κατηγοριοποίηση (clustering), και την ανάλυση απόκλισης (deviation analysis). Επιπρόσθετα, τα συστήματα αυτά μπορούν επίσης να κατηγοριοποιηθούν σύμφωνα με την μορφή και το είδος της ανακαλυπτόμενης γνώσης σε συστήματα γενικευμένης γνώσης, στοιχειώδους-βασικής γνώσης, και πολυεπίπεδης γνώσης. Ένα ευέλικτο σύστημα εξόρυξης δεδομένων μπορεί σε γενικές γραμμές να ανακαλύψει γνώση σε πολλαπλά επίπεδα.

▮ Ποια είδη τεχνικών χρησιμοποιούνται.

Μια άλλη κατηγοριοποίηση είναι σύμφωνα με τις χρησιμοποιούμενες τεχνικές. Για παράδειγμα, μπορούν να χαρακτηριστούν σαν αυτόνομα συστήματα, συστήματα οδηγούμενα από τα δεδομένα, συστήματα οδηγούμενα από τα διάφορα ερωτήματα, και διαδραστικά συστήματα. Αντίστοιχα μπορούν να χαρακτηριστούν και σαν συστήματα βασισμένα σε στατιστικά μοντέλα, σε μαθηματικά μοντέλα, σε συστήματα εύρεσης μοτίβων, τάσεων κ.α.

▮ Ποιες εφαρμογές χρησιμοποιούνται.

Μια τελευταία πιθανή κατηγοριοποίηση συστημάτων εξόρυξης δεδομένων είναι ανάλογα με τις εφαρμογές που υλοποιούν. Για παράδειγμα, μπορούν να υπάρχουν συστήματα που έχουν δημιουργηθεί αποκλειστικά για δεδομένα οικονομικής φύσεως, τηλεπικοινωνίες, ανάλυση DNA, χρηματαγορές, ηλεκτρονικό ταχυδρομείο κ.α. Φυσικά στην πράξη διαφορετικές εφαρμογές συχνά απαιτούν την δημιουργία και ενσωμάτωση ειδικά

δημιουργημένων μεθόδων. Έτσι ένα γενικής χρήσης, πολλαπλών εφαρμογών σύστημα μπορεί να μην ικανοποιεί συγκεκριμένες απαιτήσεις.

Κεφάλαιο 2 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΕΩΝ

ΠΕΡΙΛΗΨΗ

Στο παρόν κεφάλαιο παρουσιάζουμε έναν τυπικό ορισμό του προβλήματος της ανακάλυψης κανόνων συσχετίσεων από τα δεδομένα αφού αυτό αποτέλεσε και το κύριο πεδίο της ερευνάς μας. Παράλληλα κάνουμε μια σύντομη αναδρομή ορισμένων εκ των κυριότερων προσπαθειών που έχουν γίνει στο συγκεκριμένο πεδίο, από την εισαγωγή του μέχρι σήμερα. Κλείνουμε δίνοντας μια σύντομη περιγραφή των κυριότερων αλγορίθμων που έχουν προταθεί μέχρι σήμερα στον τομέα αυτό καθώς και αυτών που έχουν άμεση σχέση με την δουλειά μας με τον ένα ή τον άλλο τρόπο, όπως επίσης και με μια περιγραφή της γεννήτριας συνθετικών δεδομένων.

Λέξεις Κλειδιά: *Εξόρυξη Δεδομένων; Κανόνες Συσχετίσεων; Αλγόριθμοι*

2.1 Κανόνες Συσχετίσεων

Μετά την αρχική πρόταση των Agrawal, Imielinski και Swani στο [8], το πρόβλημα της ανακάλυψης κανόνων συσχετίσεων μελετήθηκε εκτενώς από πολλούς ερευνητές και μια σειρά παραλλαγών προτάθηκαν. Σε μια μετέπειτα δουλειά οι Agrawal και Srikant [10] συζήτησαν πως ο βασικός αλγόριθμος για την εύρεση των συχνών αντικειμένων θα μπορούσε να βελτιωθεί εισάγοντας μια τεχνική απομάκρυνσης αντικειμένων (pruning itemsets) που μειώνει το μέγεθος του συνόλου των υποψήφιων αντικειμένων C_k . Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί το τέχνασμα ότι όλα τα υποσύνολα ενός συχνού αντικειμένου πρέπει να είναι επίσης συχνά. Έτσι, αν κάποιο $(k-1)$ -υποσύνολο ενός αντικειμένου $I \subseteq C_k$ δεν ανήκει στο L_{k-1} † σύνολο, τότε το συγκεκριμένο αντικείμενο μπορεί να απομακρυνθεί από την περαιτέρω εξέταση. Οι μετέπειτα δουλειές στην εύρεση των συχνών αντικειμένων επικεντρώθηκαν κύρια στα παρακάτω θέματα:

1. Βελτίωση του I/O κόστους μειώνοντας τον αριθμό των περασμάτων από τη βάση.
2. Βελτίωση της υπολογιστικής απόδοσης της διαδικασίας.
3. Πρόταση αποτελεσματικών αλγορίθμων που εργάζονταν παράλληλα.
4. Πρόταση τεχνικών δειγματοληψίας για την βελτίωση του I/O και του υπολογιστικού κόστους της εύρεσης των συχνών αντικειμένων.
5. Επεκτάσεις της μεθόδου εύρεσης συχνών αντικειμένων σε άλλα προβλήματα όπως ποσοτικοποιημένοι κανόνες συσχετίσεων, γενικευμένες συσχετίσεις, και κυκλικοί-επαναλαμβανόμενοι κανόνες κ.α.. ([15][4][112][115][11]).
6. Εύρεση μεθόδων για την online δημιουργία κανόνων συσχετίσεων χρησιμοποιώντας την φιλοσοφία των OLAP – On Line Analytical Processing τεχνικών (προεπεξεργασία της βάσης μια φορά για την μετέπειτα ευκολότερη πολλαπλή εφαρμογή ερωτημάτων σε αυτή).
7. Χρήση εναλλακτικών μετρικών ενδιαφέροντος, πέρα από την υποστήριξη και την βεβαιότητα ([102][27][6][18]).

Η εύρεση κανόνων συσχετίσεων μπορεί να εφαρμοστεί σε διάφορους άλλους τύπους δεδομένων, όπως αρχεία κειμένων ([62][63][64]), δεδομένα απογραφών [27], τηλεπικοινωνιακά δεδομένα [75], αρχεία σχετικά με τις εγγραφές φοιτητών [92], ακόμα και

† Όπου L_{k-1} το σύνολο των συχνών αντικειμένων του προηγούμενου περάσματος. Παράδειγμα αν ψάχνουμε για τα πιθανά συχνά τρισύνολα τότε το L_{k-1} είναι το σύνολο των επιβεβαιωμένων συχνών δισυνόλων.

ημι-δομημένα αρχεία κειμένων [113] κ.α. Ουσιαστικά οποιαδήποτε βάση η οποία αποτελείται από “καλάθια” (baskets) τα οποία περιέχουν πολλά “αντικείμενα” μπορεί να ταιριάζει σε αυτό το μοντέλο.

2.2 Εύρεση Κανόνων Συσχετίσεων

Η ενότητα αυτή βασίζεται κυρίως στην πρώτη εισαγωγή του προβλήματος που έγινε στο [8] και στο [10]. Τυπικά το πρόβλημα μας μπορεί να οριστεί ως εξής: Έστω $I = \{i_1, i_2, \dots, i_m\}$ ένα σύνολο διαφορετικών στοιχείων που καλούμε αντικείμενα (items). Έστω T ένα σύνολο συναλλαγών (η βάση δεδομένων), όπου κάθε συναλλαγή t είναι ένα σύνολο αντικειμένων τέτοιο ώστε $t \subseteq I$. Ένας κανόνας συσχέτισης είναι μια συνεπαγωγή της μορφής $X \Rightarrow Y$, όπου $X \subset I$, $Y \subset I$, και $X \cap Y = \emptyset$. Η βασική ιδέα ενός κανόνα συσχέτισης είναι να αναπτύξει μια συστηματική μέθοδο με την οποία ένας χρήστης μπορεί να προβλέψει την εμφάνιση κάποιων αντικειμένων, δοσμένης της ύπαρξης κάποιων άλλων σε μια συναλλαγή. Αυτή η πληροφορία είναι ιδιαίτερα χρήσιμη στην λήψη διαφόρων αποφάσεων όπως προωθήσεις προϊόντων, χωροθέτηση προϊόντων σε καταστήματα, στόχευση πελατών κ.α. Ο κανόνας $X \Rightarrow Y$ ισχύει στο σύνολο των συναλλαγών T με βεβαιότητα (confidence) c , αν $c\%$ των συναλλαγών στη T που περιέχουν το X περιέχουν επίσης και το Y . Ο κανόνας έχει υποστήριξη (support) s στη T , αν $s\%$ των συναλλαγών στη T περιέχουν την $X \cup Y$. Έτσι όταν λέμε ότι ένας κανόνας έχει 90% βεβαιότητα τότε αυτό σημαίνει ότι το 90% των συναλλαγών που περιέχουν το X περιέχουν επίσης το Y . Γενικά, ένα σύνολο στοιχείων (όπως το αρχικό ή το απορρέον κομμάτι ενός κανόνα) καλείται αντικείμενο. Ο αριθμός των στοιχείων ενός αντικειμένου καλείται μήκος του αντικειμένου. Τα αντικείμενα μήκους k καλούνται k -αντικείμενα. Για ένα αντικείμενο $X \cdot Y$, αν το Y είναι ένα m -αντικείμενο τότε το Y καλείται μια m -επέκταση του X .

Δοσμένου ενός συνόλου συναλλαγών T (η βάση μας), το πρόβλημα της εξόρυξης ή απλούστερα της εύρεσης κανόνων συσχέτισης είναι να ανακαλύψουμε όλους τους κανόνες που έχουν υποστήριξη μεγαλύτερη από μια ελάχιστη τιμή υποστήριξης (ορισμένη από τον χρήστη) και που καλείται κατώφλι υποστήριξης (minsup) και βεβαιότητα μεγαλύτερη από μια ελάχιστη τιμή βεβαιότητας που καλείται κατώφλι βεβαιότητας (minconf). Το πρόβλημα αυτό μπορεί να διαχωριστεί στα ακόλουθα υπό-προβλήματα:

1. Εύρεση όλων των αντικειμένων που έχουν υποστήριξη πάνω από το κατώφλι υποστήριξης. Αυτά καλούνται συχνά αντικείμενα (πρακτικά τα αντικείμενα αυτά τα οποία εμφανίζονται έναν ικανοποιητικό αριθμό φορές στη βάση μας προκειμένου να θεωρηθούν σημαντικά). Όλα τα υπόλοιπα καλούνται σπάνια.

2. Παραγωγή όλων των κανόνων συσχετίσεων που ικανοποιούν το κατώφλι βεβαιότητας χρησιμοποιώντας σαν είσοδο τα συχνά αντικείμενα.

Το δεύτερο πρόβλημα, δηλ. αυτό της δημιουργίας των κανόνων συσχετίσεων από τα συχνά αντικείμενα είναι μια σχετικά απλή διαδικασία. Βέβαια υπό κάποιες συνθήκες μπορεί και αυτή να δημιουργήσει δυσκολίες, όπως για παράδειγμα όταν ο αριθμός των παραγομένων κανόνων είναι πολύ μεγάλος. Για να επιλύσουμε το πρόβλημα αυτό χρησιμοποιούμε την εξής διαδικασία. Για κάθε συχνό αντικείμενο r , βρες όλα τα κενά υποσύνολα του. Για κάθε υποσύνολο p του r , ανακάλυψε και παρουσίασε τον κανόνα $p \Rightarrow (r - p)$ αν η βεβαιότητα του κανόνα η οποία ισούται με $\frac{\text{sup}(r)}{\text{sup}(p)}$ είναι μεγαλύτερη ή ίση με το κατώφλι βεβαιότητας. Το πρώτο υπό-πρόβλημα όμως, η εύρεση όλων των συχνών αντικειμένων καθώς και του πλήθους των εμφανίσεων τους είναι ένα πολύπλοκο πρόβλημα το οποίο επιδεινώνεται ιδιαίτερα αν το πλήθος των στοιχείων $|I|$, και η βάση T είναι μεγάλα. Για παράδειγμα, αν $|I| = m$, ο αριθμός των δυνατών διαφορετικών αντικειμένων είναι 2^m . Το πρόβλημα είναι να εντοπίσουμε ποια από αυτά τα αντικείμενα έχουν την ελάχιστη επιτρεπτή υποστήριξη για την συγκεκριμένη βάση. Για πολύ μικρές τιμές του m , είναι δυνατόν να δημιουργήσουμε 2^m μετρητές, έναν για κάθε διαφορετικό αντικείμενο και να μετρήσουμε την υποστήριξη κάθε αντικειμένου διαβάζοντας την βάση μια φορά. Όμως σε πραγματικές εφαρμογές το m μπορεί να είναι αρκετά μεγαλύτερο του 1.000. Έτσι μια τέτοια προσέγγιση προφανώς δεν είναι πραγματοποιήσιμη. Επίσης πρέπει να σημειώσουμε ότι τελικά μόνο ένα πολύ μικρό ποσοστό αυτού του εκθετικά μεγάλου αριθμού αντικειμένων θα έχει την ελάχιστη υποστήριξη. Έτσι δεν είναι ανάγκη να βρούμε την υποστήριξη κάθε αντικειμένου. Ακόμα όμως και αν κάτι τέτοιο ήταν πρακτικά δυνατό, η εύρεση της υποστήριξης όλων των αντικειμένων ανεξαιρέτως αποτελεί τεράστια σπατάλη. Προκειμένου να μειώσουμε τον πιθανό χώρο ψαξίματος, όλοι οι αλγόριθμοι εκμεταλλεύονται την ακόλουθη ιδιότητα, η οποία ονομάζεται downward ή subset closure property: κάθε υποσύνολο ενός συχνού αντικειμένου πρέπει να είναι επίσης συχνό. Για παράδειγμα αν μια συναλλαγή περιέχει το αντικείμενο ABCD, τότε επίσης περιέχει τα A, AB, BC, ABC, κ.ο.κ. Αντίστοιχα όλες οι προεκτάσεις ενός σπάνιου αντικειμένου είναι επίσης σπάνιες. Έτσι, αν σε κάποιο βήμα βρούμε ότι π.χ. το αντικείμενο ADE είναι σπάνιο, τότε κανένα από τα αντικείμενα τα οποία είναι προεκτάσεις του όπως για παράδειγμα το ADEF ή το ADEFG κτλ. δεν χρειάζεται να ελεγχθούν αν έχουν την ελάχιστη υποστήριξη.

Θα ξεκινήσουμε με μια περιγραφή του αλγορίθμου Apriori, μιας και αποτέλεσε τον πρώτο αποδοτικό αλγόριθμο εύρεσης κανόνων συσχετίσεων αλλά σημείο αναφοράς για την ανάπτυξη και την σύγκριση ουσιαστικά όλων των μετέπειτα αλγορίθμων. Ακολούθως δίνουμε και

άλλους γνωστούς αλγορίθμους, και κλείνουμε με την γεννήτρια συνθετικών δεδομένων, ένα σημαντικό βοήθημα στην έρευνα για τους κανόνες συσχετίσεων.

2.3 Ο Αλγόριθμος Apriori

Ο αλγόριθμος Apriori, ίσως ο γνωστότερος ανάμεσα στους αλγορίθμους εύρεσης κανόνων συσχετίσεων, μελετήθηκε και επεκτάθηκε στους αλγορίθμους AprioriTid και AprioriHybrid [10]. Ο αλγόριθμος Apriori δημιουργεί υποψήφια αντικείμενα και ανακαλύπτει από αυτά τα συχνά. Το πιο σημαντικό χαρακτηριστικό του Apriori είναι ότι, όταν δημιουργεί τα υποψήφια αντικείμενα, χρησιμοποιεί τα αντικείμενα αυτά που έχουν ήδη βρεθεί συχνά και όχι όλα τα πιθανά αντικείμενα [10]. Προτού ξεκινήσουμε να περιγράφουμε τον αλγόριθμο ας δούμε κάποιους ορισμούς και παραδοχές που χρησιμοποιούνται. Το σύνολο των συχνών αντικειμένων περιλαμβάνει τα αντικείμενα αυτά των οποίων η υποστήριξη είναι μεγαλύτερη ή ίση από το κατώφλι υποστήριξης (minsup). Ένα σύνολο k -συχνών αντικειμένων, όπου k είναι το πλήθος των στοιχείων στα αντικείμενα, αναπαρίσταται με L_k . Ένα σύνολο υποψηφίων αντικειμένων περιλαμβάνει τα αντικείμενα που έχουν την πιθανότητα να είναι μέρος των συχνών αντικειμένων, και χρησιμοποιείται για να ανακαλύψει τα συχνά αντικείμενα. Ένα σύνολο k -υποψηφίων αντικειμένων αναπαρίσταται ως C_k , όπου k είναι το πλήθος των στοιχείων σε κάθε αντικείμενο. Ισχύει ότι $L_k \subseteq C_k$, δηλαδή ότι το σύνολο των συχνών αντικειμένων μήκους k είναι υποσύνολο των αντίστοιχων υποψηφίων αντικειμένων. Σε κάθε συναλλαγή όπως σε κάθε αντικείμενο θεωρούμε ότι τα στοιχεία είναι διατεταγμένα με κάποιο τρόπο (συνήθως λεξικογραφικά). Έτσι παράδειγμα αν έχουμε ένα k -αντικείμενο X το οποίο αποτελείται από τα στοιχεία $X[1].X[2].X[3]...X[k]$, τότε θα ισχύει $X[1]<X[2]<X[3]<...<X[k]$. Σε κάθε αντικείμενο αποθηκεύουμε εκτός από τη λίστα με τα στοιχεία που το απαρτίζουν, επίσης ένα μετρητή ο οποίος καταγράφει το πλήθος των εμφανίσεων του. Ο μετρητής αυτός αρχικοποιείται στο 0 όταν πρωτοδημιουργηθεί το συγκεκριμένο αντικείμενο και κάθε φορά που το συναντάμε σε κάποια συναλλαγή τον αυξάνουμε κατά 1. Η τελική του τιμή μας δίνει και την υποστήριξη του συγκεκριμένου αντικειμένου. Ένα αντικείμενο καλείται συχνό αν η τελική τιμή υποστήριξης του είναι μεγαλύτερη ή ίση από το κατώφλι υποστήριξης ενώ στην αντίθετη περίπτωση καλείται σπάνιο.

Ο αλγόριθμος Apriori ξεκινά κάνοντας ένα πέρασμα από τη βάση όπου και βρίσκεται το σύνολο των συχνών στοιχείων (1-items), απομακρύνοντας αυτά τα οποία είναι σπάνια. Τα στοιχεία αυτά αποτελούν το σύνολο των συχνών αντικειμένων L_1 , με βάση το οποίο δημιουργούμε το σύνολο των υποψηφίων αντικειμένων μήκους 2 (2-itemsets) C_2 . Η βάση μας ξαναδιαβάζεται εκ νέου προκειμένου να μετρήσουμε τις εμφανίσεις όλων των υποψηφίων αντικειμένων στο C_2 και να καταλήξουμε στο L_2 . Σε κάθε νέο πέρασμα τώρα

πραγματοποιούνται δύο βήματα. Έστω έτσι ότι βρισκόμαστε στο πέρασμα k . Στο πρώτο βήμα παράγουμε το σύνολο των υποψηφίων k -αντικειμένων C_k , χρησιμοποιώντας σαν είσοδο τα συχνά $(k-1)$ -αντικείμενα L_{k-1} που βρέθηκαν στο προηγούμενο πέρασμα. Στο δεύτερο βήμα τώρα κάνουμε πάλι ένα πέρασμα από τη βάση προκειμένου να υπολογίσουμε τον αριθμό εμφανίσεων του κάθε αντικειμένου. Έτσι για κάθε υποψήφιο αντικείμενο το οποίο βρίσκεται σε μια συναλλαγή ο μετρητής του αυξάνεται κατά 1. Ο αλγόριθμος Apriori καθώς και οι κοντινοί του συγγενείς μπορούν να συνοψιστούν σαν μια μέθοδος που παράγει υποψήφια αντικείμενα σε κάθε φάση, διαβάζει την βάση προκειμένου να υπολογίσει την υποστήριξη των αντικειμένων αυτών και εξάγει τελικά τα συχνά. Τα βήματα επεξεργασίας του αλγορίθμου Apriori επαναλαμβάνονται ενώ σταδιακά αυξάνουμε τον αριθμό των στοιχείων που αποτελούν τα υποψήφια αντικείμενα. Ο Apriori επαναλαμβάνει την όλη διαδικασία μέχρι να μην υπάρχουν πλέον άλλα υποψήφια αντικείμενα. Στην Εικόνα 3 παρουσιάζεται ο ψευδοκώδικας του αλγορίθμου Apriori.

2.3.1 Δημιουργία υποψηφίων αντικειμένων

Όπως είδαμε και προηγούμενα σε κάθε πέρασμα του αλγορίθμου Apriori από την βάση μας πραγματοποιούνται δύο βήματα. Ας δούμε πως λειτουργεί το βήμα δημιουργίας υποψηφίων αντικειμένων:

Το συγκεκριμένο βήμα περιλαμβάνει δύο κομμάτια, αναφορικά το κομμάτι της συνένωσης (join-phase) και το κομμάτι της διαγραφής (prune-phase). Έστω έτσι ότι έχουμε βρει όλα τα συχνά $(k-1)$ -αντικείμενα, έχουμε δηλαδή ουσιαστικά το σύνολο L_{k-1} . Με βάση τα συχνά αντικείμενα στο κομμάτι της συνένωσης γίνεται η ένωση οποιωνδήποτε $(k-1)$ -αντικειμένων έχουν ακριβώς $k-2$ κοινά στοιχεία. Το νέο αντικείμενο που θα προκύψει θα αποτελείται από τα κοινά $k-2$ στοιχεία των δύο αντικειμένων μαζί με τα δύο μη κοινά στοιχεία τους, φτάνοντας έτσι τελικά στα k στοιχεία. Παράδειγμα έστω ότι έχουμε τα αντικείμενα $\langle 1,2,3 \rangle$ και $\langle 1,2,4 \rangle$. Αυτά έχουν τα στοιχεία 1 και 2 κοινά. Το νέο αντικείμενο που θα προκύψει θα είναι έτσι το $\langle 1,2,3,4 \rangle$. Αφού δημιουργήσουμε κατ' αυτόν τον τρόπο όλα τα υποψήφια αντικείμενα περνάμε στο κομμάτι της διαγραφής, όπου ουσιαστικά απομακρύνουμε πριν ξεκινήσουμε την καταμέτρηση τους τα αντικείμενα αυτά τα οποία λανθασμένα συμπεριλάβαμε σαν υποψήφια. Στο προηγούμενο παράδειγμα για να είναι τελικά υποψήφιο το αντικείμενο $\langle 1,2,3,4 \rangle$ και να

- 1) *Database* = set of transactions;
- 2) *Items* = set of items;
- 3) transaction = $\langle TID; \{x/x \in Items\} \rangle$;
- 4) **Comment:** F_1 is a set of frequent 1-itemsets
- 5) $F_1 = \emptyset$;

```

6) Comment: Scan the transactions to count the occurrences of each item
7) foreach transaction  $t \in Database$  do begin
8)   foreach item  $x$  in  $t$  do
9)      $x.count++$ ;
10) end
11) Comment: Form the set of frequent 1-itemsets
12) foreach item  $i \in Items$  do
13)   if  $i.count/|Database| \geq minsup$ 
14)     then  $F_1 = F_1 \cup i$ ;
15) Comment: Find  $F_k$ , the set of frequent  $k$ -itemsets, where  $k \geq 2$ 
16) for ( $k = 2$ ;  $F_{k-1} \neq \emptyset$ ;  $k++$ ) do begin
17)   Comment:  $C_k$  is the set of candidate  $k$ -itemsets
18)    $C_k = \emptyset$ ;
19)   Comment:  $F_{k-1} * F_{k-1}$  is the natural join of  $F_{k-1}$  and  $F_{k-1}$  on the first  $k-2$  items
20)   foreach  $x \in \{F_{k-1} * F_{k-1}\}$  do
21)     if  $\neg \exists y \mid y = (k-1)\text{-subset of } x \wedge y \notin F_{k-1}$ 
22)       then  $C_k = C_k \cup x$ ;
23)   Comment: Scan the transactions to count the occurrences of candidate  $k$ -itemsets
24)   foreach transaction  $t \in Database$  do begin
25)     foreach  $k$ -itemset  $x$  in  $t$  do
26)       if  $x \in C_k$ 
27)         then  $x.count++$ ;
28)   end
29)    $F_k = \emptyset$ ;
30)   foreach  $x \in C_k$  do
31)     if  $x.count/|Database| \geq minsup$ 
32)       then  $F_k = F_k \cup x$ ;
33) end
34)  $Answer = \bigcup_k F_k$ ;

```

Εικόνα 3: Ο ψευδοκώδικας του αλγορίθμου Apriori

συμπεριληφθεί στα υποψήφια αντικείμενα της επόμενης φάσης θα πρέπει όλα του τα υποσύνολα (αναφορικά τα $\langle 1,2,3 \rangle$, $\langle 1,2,4 \rangle$, $\langle 1,3,4 \rangle$ και $\langle 2,3,4 \rangle$) να είναι επίσης συχνά. Με άλλα λόγια απορρίπτονται τα αντικείμενα αυτά για τα οποία υπάρχει ακόμα και ένα $(k-1)$ υποσύνολο τους που να μην ανήκει στο σύνολο L_{k-1} .

2.4 Ο αλγόριθμος Dynamic Itemset Counting - DIC

Ο αλγόριθμος DIC ο οποίος προτάθηκε από τους Brin et. al στο [27] ουσιαστικά προσπαθούσε να μειώσει τον αριθμό των περασμάτων που πραγματοποιούνταν πάνω στα δεδομένα σε

σύγκριση με τους άλλους αλγορίθμους. Όλοι οι αλγόριθμοι που δουλεύουν βήμα-βήμα πραγματοποιούν τόσα περάσματα από τα δεδομένα όσος είναι και ο αριθμός των στοιχείων στο μεγαλύτερης τάξης υποψήφιο αντικείμενο. Παράδειγμα αν το μεγαλύτερης τάξης υποψήφιο αντικείμενο αποτελείται από 15 στοιχεία τότε απαιτούνται περίπου 15 διαδοχικά περάσματα από τη βάση. Η βασική ιδέα του αλγορίθμου DIC είναι ότι δεν πρέπει να περιμένουμε μέχρι το τέλος του πρώτου περάσματος προτού ξεκινήσουμε να μετράμε αντικείμενα μεγαλύτερου μεγέθους. Για παράδειγμα μπορούμε να ξεκινήσουμε να μετράμε δισύνολα αντικειμένων (2-items) ακόμα προτού τελειώσουμε να μετράμε τα μονοσύνολα (1-items). Η λειτουργία του DIC συγκρίνεται με αυτή ενός τραίνου. Ο αλγόριθμος είναι το τρένο ενώ η διαδρομή τα δεδομένα. Κάθε M συναλλαγές το τρένο πραγματοποιεί μια στάση. Σε κάθε στάση οι επιβαίνοντες (δηλαδή τα αντικείμενα στη βάση μας) επιτρέπεται να επιβιβαστούν (δηλ. να ξεκινήσουμε να τα μετράμε) ή να αποβιβαστούν (δηλ. να σταματήσουμε το μέτρημα τους). Όταν το τρένο φτάσει στο τέλος της διαδρομής και υπάρχουν ακόμα αντικείμενα τα οποία θα πρέπει να καταμετρηθούν τότε απλά ξανά-ξεκινάει την διαδρομή του από την αρχή μέχρι να τα μετρήσει σε ολόκληρη τη βάση και να μην υπάρχουν άλλα υποψήφια αντικείμενα.

Εικόνα 4: Η μείωση των περασμάτων από τη βάση που επιτυγχάνεται με χρήση του αλγορίθμου DIC σε σχέση με τον αλγόριθμο Apriori

Κάθε αντικείμενο χαρακτηρίζεται με έναν διαφορετικό τρόπο ο οποίος αντικατοπτρίζει την τρέχουσα κατάσταση του. Πιο συγκεκριμένα δείχνει αν ένα αντικείμενο έχει καταμετρηθεί σε όλη τη βάση ή όχι καθώς επίσης και αν έχει υποστήριξη κάτω ή πάνω από το κατώφλι υποστήριξης. Ο αλγόριθμος DIC επίσης χρησιμοποιεί μια ενιαία τιμή κατωφλίου υποστήριξης

για όλα τα αντικείμενα στη βάση και χρησιμοποιεί την ίδια μέθοδο για να απορρίπτει τα μη συχνά αντικείμενα όπως ο Apriori (δηλ. για να είναι ένα αντικείμενο συχνό πρέπει όλα τα υποσύνολα του να είναι επίσης συχνά). Αυτή η ιδιότητα είναι καθοριστικής σημασίας προκειμένου ο DIC να καταφέρει να ξεκινήσει να μετράει νωρίτερα από άλλους αλγορίθμους τα μεγαλύτερης τάξης αντικείμενα. Έτσι αν όλα τα υποσύνολα ενός αντικειμένου έχουν τιμή υποστήριξης πάνω από το κατώφλι υποστήριξης τότε είναι σίγουρο ότι θα είναι συχνά και μπορούμε να αρχίσουμε να μετράμε το εν λόγω αντικείμενο. Εν κατακλείδι ο DIC πραγματοποιεί συνολικά λιγότερα περάσματα από τα δεδομένα και παράλληλα παράγει λιγότερα υποψήφια αντικείμενα από άλλους αλγορίθμους όπως π.χ. ο αλγόριθμος Sampling [118].

2.5 Ο αλγόριθμος DHP (Direct Hashing and Pruning)

Ο αλγόριθμος DHP - Direct Hashing and Pruning [98], χρησιμοποιεί μια τεχνική κατακερματισμού (hashing technique) προκειμένου να απομακρύνει μη αναγκαία στοιχεία για τη δημιουργία υποψηφίων αντικειμένων. Κάθε $(k+1)$ -αντικείμενο στις συναλλαγές καταχωρείται σε μια συγκεκριμένη θέση στον πίνακα κατακερματισμού ενώ οι εμφανίσεις των υποψηφίων k -αντικειμένων καταμετρώνται διαβάζοντας τις συναλλαγές. Αν η υποστήριξη μιας συγκεκριμένης θέσης του πίνακα κατακερματισμού είναι μικρότερη από το κατώφλι υποστήριξης, τότε όλα τα $(k+1)$ -αντικείμενα που έχουν καταχωρηθεί σ' αυτή τη θέση δεν θα περιληφθούν στο σύνολο των υποψηφίων $(k+1)$ -αντικειμένων στο επόμενο πέρασμα. Η απομάκρυνση υποψηφίων αντικειμένων με βάση τις τιμές υποστήριξης των θέσεων του πίνακα αυτού είναι ασφαλής αφού επιτρέπει την ύπαρξη λανθασμένων πιθανολογούμενων συχνών αντικειμένων – false positives (δηλ. τα υποψήφια αντικείμενα του επόμενου βήματος τα οποία τελικά δεν είναι συχνά) αλλά δεν επιτρέπει την απομάκρυνση λανθασμένων αρνητικών αντικειμένων (false negatives).

Μέθοδοι περιορισμού της εξεταζόμενης βάσης αλλά και μείωσης κάποιων συναλλαγών προτάθηκαν επίσης μαζί με τον αλγόριθμο DHP, έτσι ώστε το μέγεθος της βάσης που θα πρέπει να διαβαστεί προκειμένου να μετρήσουμε τις εμφανίσεις όλων των υποψηφίων αντικειμένων να μειώνεται σε κάθε πέρασμα. Η μείωση των συναλλαγών ή/και η διαγραφή τους πραγματοποιείται με βάση την ιδιότητα subset closure των συχνών αντικειμένων; με άλλα λόγια, κάθε υποσύνολο ενός συχνού αντικειμένου πρέπει να είναι επίσης συχνό. Αυτή η ιδιότητα θεωρεί ότι μια συναλλαγή δύναται να έχει ένα υποψήφιο $(k + 1)$ -αντικείμενο μόνο αν περιέχει $k+1$ υποψήφια k -αντικείμενα που είχαν βρεθεί στο προηγούμενο πέρασμα. Έτσι, καθώς μια συναλλαγή διαβάζεται προκειμένου να βρούμε τις εμφανίσεις των υποψηφίων k -αντικειμένων, μπορούμε να αποφασίσουμε αν αυτή η συναλλαγή μπορεί να διαγραφεί από την

βάση μας κατά το επόμενο πέρασμα. Από την άλλη, αν μια συναλλαγή περιέχει ένα συχνό $(k+1)$ -αντικείμενο, τότε οποιοδήποτε αντικείμενο περιέχεται σε αυτό το $(k+1)$ -αντικείμενο θα πρέπει να εμφανίζεται σε τουλάχιστον k από τα υποψήφια k -αντικείμενα που περιέχονται σε αυτή τη συναλλαγή. Έτσι, καταμετρώντας πόσες φορές κάθε αντικείμενο σε μια συναλλαγή περιλαμβάνεται στα υποψήφια k -αντικείμενα σε αυτή τη συναλλαγή, μπορούμε να αποφασίσουμε αν το αντικείμενο αυτό μπορεί να διαγραφεί από τη συναλλαγή κατά το επόμενο πέρασμα. Μια συναλλαγή μειώνεται ξαναγράφοντας την αλλά χωρίς τα αντικείμενα που δεν θα συνεισφέρουν στην δημιουργία συχνών αντικειμένων στο επόμενο πέρασμα.

Το πόσο μπορεί ο DHP να μειώσει τον αριθμό των υποψήφιων αντικειμένων εξαρτάται από τον αριθμό των false positives. Τα false positives δημιουργούνται όταν η θέση στην οποία θα πρέπει να αποθηκευτούν κάποια αντικείμενα είναι ίδια για ένα σύνολο υποψηφίων αντικειμένων (τα hash synonyms), και ενώ οι συχνότητες εμφανίσεων τους ξεχωριστά είναι μικρότερες από το κατώφλι υποστήριξης η τελική τιμή της θέσης του πίνακα κατακερματισμού στην οποία και έχουν όλα μαζί αποθηκευτεί είναι μεγαλύτερη από αυτό το κατώφλι. Ο αριθμός των υποψηφίων αντικειμένων που έχουν την ίδια τιμή κατακερματισμού εξαρτάται άμεσα από το μέγεθος του πίνακα κατακερματισμού. Δυστυχώς αυτός ο πίνακας δεν μπορεί να αυξηθεί αλόγιστα καθώς οι απαιτήσεις του σε μνήμη έρχονται σε αντίθεση με το δέντρο που χρησιμοποιείται για να κρατάει τις εμφανίσεις των υποψηφίων αντικειμένων. Βέβαια θα μπορούσαν να εφαρμοστούν και άλλες τεχνικές όπως η perfect hashing ή η καλύτερη επιλογή συνάρτησης κατακερματισμού αλλά με τη σειρά τους αυτές έρχονται αντιμέτωπες με την απαίτηση για καλύτερους χρόνους τρεξίματος και απόκρισης του τελικού αλγορίθμου.

2.6 Ο αλγόριθμος Partition

Το πλεονέκτημα της διατήρησης της βάσης στη κύρια μνήμη αποφεύγοντας έτσι το δαπανηρό κόστος που συνεπάγεται η χρησιμοποίηση της δευτερεύουσας μνήμης παρακίνησε τους Savasere, Omiecinski και Navathe στο [109] να προτείνουν τον αλγόριθμο Partition. Τόσο ο Apriori όσο και ο DHP, πραγματοποιούν αρκετά περάσματα από τη βάση προκειμένου να βρουν τα συχνά αντικείμενα διαφόρων μεγεθών. Το άνω όριο στον αριθμό των περασμάτων από μια βάση μπορεί να είναι το μέγιστο μέγεθος του επιθυμητού συχνού αντικειμένου, ή όταν δεν υπάρχει άλλο υποψήφιο αντικείμενο σε ένα πέρασμα. Ας θυμηθούμε ότι δεν μπορεί να υπάρξει συχνό $(k+1)$ -αντικείμενο χωρίς όλα του τα k -αντικείμενα να είναι συχνά. Αντίθετα, ο αλγόριθμος Partition πραγματοποιεί μόλις δύο περάσματα από τη βάση. Ας δούμε πώς.

Κατ' αρχήν, η βάση χωρίζεται σε τόσα τμήματα όσα απαιτούνται έτσι ώστε όλες οι συναλλαγές σε κάθε τμήμα να χωράνε στη κύρια μνήμη. Στη πρώτη φάση, κάθε τμήμα εξετάζεται ξεχωριστά προκειμένου να βρούμε τα συχνά αντικείμενα στο συγκεκριμένο τμήμα.

Κατόπιν όλα τα συχνά αντικείμενα από όλα τα τμήματα ενώνονται προκειμένου να δημιουργήσουν ένα σύνολο καθολικών υποψηφίων αντικειμένων. Στη δεύτερη φάση, όλα τα τμήματα επεξεργάζονται σε σχέση με το καθολικό σύνολο υποψηφίων αντικειμένων που βρέθηκαν στη πρώτη φάση, και οι ακριβείς εμφανίσεις τους καταγράφονται. Ας υποθέσουμε ότι υπάρχουν m τμήματα. Αν το κατώφλι υποστήριξης για τα αντικείμενα στη βάση μας είναι s , τότε το κατώφλι υποστήριξης για τα αντικείμενα προκειμένου να είναι συχνά σε ένα τμήμα είναι επίσης s . Είναι ξεκάθαρο ότι προκειμένου ένα αντικείμενο να έχει υποστήριξη s , θα πρέπει να έχει υποστήριξη τουλάχιστον s σε τουλάχιστον ένα από τα m τμήματα. Ας σημειώσουμε ότι η ιδιότητα της subset closure των συχνών αντικειμένων είναι αρκετά χρήσιμη. Αν ένα αντικείμενο δεν έχει την ελάχιστη απαιτούμενη υποστήριξη σε ένα τμήμα, τότε όλες του οι επεκτάσεις δεν θα έχουν επίσης την ελάχιστη απαιτούμενη υποστήριξη, και έτσι κανένα από αυτά τα αντικείμενα δεν χρειάζεται να καταμετρηθεί σ' αυτό το τμήμα.

Η δεύτερη φάση χρειάζεται προκειμένου να απομακρύνει τα αντικείμενα αυτά τα οποία λανθασμένα θεωρήθηκαν σαν πιθανά συχνά αντικείμενα (false positives). Όταν ολοκληρωθεί η πρώτη φάση, όλα τα τμήματα θα έχουν επεξεργαστεί και θα υπάρχει ένα μοναδικό σύνολο υποψηφίων αντικειμένων. Ορισμένα από αυτά τα υποψήφια αντικείμενα μπορεί να είναι false positives, αλλά δεν υπάρχουν false negatives. Στη δεύτερη φάση, καταμετρώνται όλες μαζί οι εμφανίσεις για κάθε υποψήφιο αντικείμενο. Τα αντικείμενα αυτά των οποίων οι εμφανίσεις δεν είναι λιγότερες από το κατώφλι υποστήριξης διατηρούνται, ενώ τα υπόλοιπα είναι false positives και απομακρύνονται.

Η πρώτη φάση του αλγόριθμου Partition μπορεί εύκολα να γίνει παράλληλα αφού κάθε τμήμα μπορεί να επεξεργαστεί από διαφορετική μονάδα. Η δεύτερη φάση μπορεί επίσης να πραγματοποιηθεί παράλληλα ανταλλάσσοντας πληροφορίες σχετικά με τους μετρητές ανάμεσα στις διάφορες μονάδες. Για κάθε αντικείμενο σε ένα τμήμα, διατηρούμε μια λίστα των αναγνωριστικών συναλλαγών (TIDs – Transaction Identifiers) των συναλλαγών που περιέχουν το αντικείμενο. Οι λίστες με τα TIDs χρησιμοποιούνται προκειμένου να αποφανθούμε πόσες συναλλαγές περιέχουν ένα συγκεκριμένο αντικείμενο. Για παράδειγμα έστω A και B δύο στοιχεία, το καθένα με μια λίστα TIDs. Τότε η υποστήριξη για το αντικείμενο $\{A, B\}$ είναι απλά ο αριθμός των TIDs που είναι κοινά και στις δύο λίστες.

2.7 Ο Αλγόριθμος MSApriori

Διάφορες τεχνικές και αλγόριθμοι έχουν προταθεί προκειμένου να αντιμετωπίσουν την ανεπάρκεια της χρήσης ενιαίας τιμής υποστήριξης για όλα τα δεδομένα στη βάση μας, όταν τα εξεταζόμενα δεδομένα δεν είναι της ίδιας φύσης και οι αντίστοιχες συχνότητες εμφάνισης τους παρουσιάζουν μεγάλες διακυμάνσεις. Μια είναι να χωρίσουμε τα δεδομένα σε ορισμένα μπλοκ

ανάλογα με τις εμφανίσεις των δεδομένων σε αυτά και κατόπιν να ψάξουμε για κανόνες σε κάθε μπλοκ ξεχωριστά με μια διαφορετική τιμή υποστήριξης [88]. Μια άλλη προσέγγιση είναι να ομαδοποιήσουμε ορισμένα συγγενή σπάνια αντικείμενα μαζί σε ένα μεγαλύτερης τάξης αντικείμενο, έτσι ώστε το αντικείμενο αυτό να είναι τελικά συχνό [88], ή όπως χρησιμοποιήθηκε με έναν παρόμοιο τρόπο στο [73]. Παρόλ' αυτά η πρώτη προσέγγιση αποτυγχάνει να βρει και να παράγει με έναν απλό τρόπο κανόνες για προϊόντα που βρίσκονται σε διαφορετικά μπλοκ, ενώ η δεύτερη αποτυγχάνει να παράγει τους κανόνες που αφορούν μόνο τους τα συχνά αντικείμενα που σχηματίζουν το υψηλότερης τάξης αντικείμενο. Οι Wang, He και Han στο [123] πρότειναν την χρήση περιορισμών υποστήριξης στην διαδικασία δημιουργίας αντικειμένων του Apriori προκειμένου να αντιμετωπίσουν την ανομοιογενή φύση της μετρικής της υποστήριξης σε πραγματικές εφαρμογές. Η πρώτη δουλειά που γνωρίζουμε και η οποία έκανε χρήση πολλαπλών τιμών υποστήριξης είναι αυτή των Liu Hsu και Ma στο [89].

Στην πράξη, τα περισσότερα αντικείμενα σε πολλές εφαρμογές συνήθως εμφανίζονται με αρκετά μεγάλες διακυμάνσεις ως προς την συχνότητα εμφάνισέως τους. Επίσης δεν είναι όλα τα συχνά αντικείμενα σημαντικά και αντίστροφα δεν είναι όλα τα σπάνια αντικείμενα ασήμαντα. Μάλιστα τα σπάνια αντικείμενα μπορούν να είναι το ίδιο σημαντικά ή αρκετές φορές ακόμη πιο σημαντικά από τα αντίστοιχα συχνά αντικείμενα. Οι αλγόριθμοι τύπου Apriori που χρησιμοποιούν ένα ενιαίο κατώφλι υποστήριξης για όλα τα αντικείμενα στην βάση δεν μπορούν να αντιμετωπίσουν όλες αυτές τις διαφορές αποτελεσματικά, διότι βάζοντας μια μεγάλη τιμή στο κατώφλι αυτό αποκλείει όλα τα σπάνια αντικείμενα ενώ δίνοντας του μια χαμηλή τιμή, εμφανίζονται εκτός από τα σημαντικά σπάνια αντικείμενα και όλα τα άλλα αντικείμενα που ικανοποιούν το κατώφλι αυτό. Έτσι έχουμε την εμφάνιση του λεγόμενου προβλήματος του “σπάνιου αντικειμένου” (rare itemset) όπου όλα τα αντικείμενα που ικανοποιούν το κατώφλι υποστήριξης συνδέονται το ένα με το άλλο [93]. Οι Liu et al. στο [89] πρότειναν τον αλγόριθμο MSApriori προκειμένου να βρίσκουν όλα τα σπάνια αλλά σημαντικά αντικείμενα μαζί με τα συχνά χωρίς να αντιμετωπίζουν το πρόβλημα που μόλις περιγράψαμε.

Έτσι προκειμένου να λάβουν υπόψη την σημαντικότητα των αντικειμένων, ο αλγόριθμος MSApriori χρησιμοποιεί την ελάχιστη τιμή υποστήριξης (minimum item support - MIS), κάθε αντικειμένου. Ένα πολύ σημαντικό αντικείμενο θα απολαμβάνει μια χαμηλή MIS τιμή ενώ ένα λιγότερο σημαντικό μια μεγαλύτερη. Με αυτόν τον τρόπο προσπαθούμε να εξασφαλίσουμε ότι τα σημαντικά αντικείμενα θα βρίσκονται άσχετα από το ποια είναι η συχνότητα εμφάνισέων τους σε σύγκριση με αυτές των άλλων αντικειμένων στη βάση μας. Η MIS τιμή κάθε αντικειμένου εκφράζεται σαν MIS (i), και υπολογίζεται χρησιμοποιώντας την παρακάτω εξίσωση:

$$MIS(i) = \begin{cases} M(i) & M(i) > LS \\ LS & Otherwise \end{cases} \quad \text{Εξίσωση 1}$$

$$M(i) = \beta \cdot f(i)$$

Όπου $f(i)$ είναι η ακριβής συχνότητα εμφανίσεων ενός αντικειμένου στη βάση, LS είναι μια οριζόμενη από το χρήστη κατώτερη ελάχιστη επιτρεπόμενη τιμή υποστήριξης και β είναι μια παράμετρος η οποία ελέγχει πως θα σχετίζονται οι MIS τιμές με τις αντίστοιχες συχνότητες εμφανίσεων των αντικειμένων. Ο αλγόριθμος $MSApriori$ δουλεύει επίσης βήμα-βήμα, δηλαδή βρίσκοντας τα συχνά αντικείμενα σε κάθε βήμα και κατόπιν χρησιμοποιώντας τα για να παράγει τα υποψήφια αντικείμενα του επόμενου βήματος, βρίσκοντας εκ νέου τα συχνά αντικείμενα κ.ο.κ. Τώρα όλα τα μεγαλύτερης τάξης αντικείμενα (κ -αντικείμενα, όπου $\kappa > 1$) παίρνουν σαν MIS τιμή την ελάχιστη MIS τιμή ανάμεσα στις τιμές που έχουν όλα τα αντικείμενα που απαρτίζουν το μεγαλύτερης τάξης αντικείμενο. Αναφορικά, η ελάχιστη τιμή υποστήριξης του αντικειμένου $I_{1,2,\dots,m}$ ισούται με $\min[MIS(I_1), MIS(I_2), \dots, MIS(I_m)]$. Έτσι αν όλα τα αντικείμενα μέσα σε ένα μεγαλύτερης τάξης αντικείμενο παρουσιάζονται συχνά τότε θα αποδοθεί μια υψηλή MIS τιμή, ενώ αν υπάρχει έστω και ένα αντικείμενο το οποίο παρουσιάζεται σπάνια τότε θα αποδοθεί μια χαμηλή MIS τιμή έτσι ώστε όλοι οι κανόνες που αντιστοιχούν στο συγκεκριμένο αντικείμενο να βρεθούν. Φυσικά η απόδοση σε κάθε αντικείμενο μια διαφορετικής MIS τιμής σημαίνει ότι η ιδιότητα subset closure που περικόπτει τα αντικείμενα στους αλγορίθμους με μια ενιαία τιμή υποστήριξης δεν ισχύει πλέον. Το πρόβλημα αυτό λύνεται εισάγοντας και χρησιμοποιώντας την sorted closure ιδιότητα, όπου τα αντικείμενα ταξινομούνται με βάση τις MIS τιμές τους σε αύξουσα σειρά και επίσης όλα τα υπόλοιπα αντικείμενα μέσα σε κάθε αντικείμενο ταξινομούνται σύμφωνα με τις MIS τιμές τους.

Ο $MSApriori$ αντιμετωπίζει επιτυχώς το πρόβλημα της ύπαρξης σπάνιων και συχνών αντικειμένων σε μια βάση, αλλά αφήνει αναπάντητο ένα μεγάλο ερώτημα. Πιο συγκεκριμένα δεν απαντά το ερώτημα πως μπορούν αυτά τα αντικείμενα να βρεθούν και ποια θα πρέπει να είναι η καλύτερη MIS τιμή που θα τους αποδοθεί. Σύμφωνα με τους δημιουργούς του ένας χρήστης έχει δύο επιλογές για την απόδοση τιμών MIS στα δεδομένα του. Η πρώτη επιλογή είναι χρησιμοποιώντας την Εξίσωση 1. Αλλά αν η παράμετρος β στην συγκεκριμένη εξίσωση τεθεί ίση με το 0 τότε ο αλγόριθμος $MSApriori$ μετατρέπεται σε έναν οποιονδήποτε αλγόριθμο που χρησιμοποιεί ένα ενιαίο κατώφλι υποστήριξης. Αν από την άλλη το β τεθεί κοντά στο 1, τότε κάθε αντικείμενο με αριθμό εμφανίσεων πάνω από ένα οριζόμενο από τον χρήστη ελάχιστο κατώφλι υποστήριξης - LS (lowest minimum support) θεωρείται σημαντικό και λαμβάνεται υπόψη με όλα τα άλλα αντικείμενα πρακτικά να απορρίπτονται (αφού δεν μπορούν να είναι συχνά). Ακολούθως και στις δύο περιπτώσεις ο συγκεκριμένος αλγόριθμος λειτουργεί

παρόμοια με έναν οποιονδήποτε αλγόριθμο ο οποίος χρησιμοποιεί ένα ενιαίο κατώφλι υποστήριξης με την μόνη διαφορά στην δεύτερη περίπτωση όπου κάθε αντικείμενο με αριθμό εμφανίσεων πάνω από το ελάχιστο κατώφλι υποστήριξης έχει την δική του τιμή MIS. Στην πράξη όλα τα αντικείμενα θεωρούνται της ίδιας σημαντικότητας, και το μόνο πράγμα που τα ξεχωρίζει είναι ο αριθμός των εμφανίσεων τους. Αν κάποιο αντικείμενο θα πρέπει να τύχει ιδιαίτερης προσοχής αυτό δεν λαμβάνεται υπόψη με κάποιο τρόπο. Ουσιαστικά οι κανόνες που δημιουργούνται με αυτή την επιλογή καθορίζονται με βάση την παράμετρο β παρά με βάση την συχνότητα εμφανίσεων του κάθε αντικειμένου. Η άλλη επιλογή για να αποδώσουμε MIS τιμές σε όλα τα μονοσύνολα (1-items) χωρίς να χρησιμοποιήσουμε τους παραπάνω τύπους θα ήταν απλά μαντεύοντας τις. Φυσικά εάν κάποιος δεν είναι ικανοποιημένος με το τελικό αποτέλεσμα θα πρέπει να τρέξει ξανά και ξανά τον αλγόριθμο, μαντεύοντας κάθε φορά νέες MIS τιμές μέχρι να ικανοποιηθεί από το αποτέλεσμα, σπαταλώντας έτσι άσκοπα χρόνο και πόρους. Σε κάθε περίπτωση πάντως δεν προτείνεται κάποια μέθοδος για να ανακαλύπτουμε τα σπάνια αλλά σημαντικά αντικείμενα.

2.8 Η Γεννήτρια Συνθετικών Δεδομένων

Η γεννήτρια συνθετικών δεδομένων[‡], η οποία θεωρείται ότι προσομοιάζει την αγοραστική συμπεριφορά των πελατών σε μια επιχείρηση λιανεμπορίου έχει γίνει το ντε φάκτο σημείο αναφοράς για σχεδόν όλες τις προσεγγίσεις στον τομέα των κανόνων συσχετίσεων. Σε πολλά σημεία της ερευνάς μας χρησιμοποιήσαμε είτε την γεννήτρια αυτή καθ' εαυτή είτε έτοιμες συνθετικές βάσεις δεδομένων παραγμένες από την γεννήτρια προκειμένου να ελέγξουμε τις προτάσεις μας, και έτσι θεωρούμε σκόπιμο να δώσουμε μια σύντομη περιγραφή του τρόπου λειτουργίας της.

Το χρησιμοποιούμενο μοντέλο λειτουργίας της γεννήτριας συνθετικών δεδομένων υιοθετεί τη λογική ότι στην πραγματικότητα οι χρήστες τείνουν να αγοράσουν μαζί κάποια σύνολα αντικειμένων. Κάθε τέτοιο σύνολο είναι ένα πιθανό συχνό αντικείμενο. Ένα παράδειγμα ενός τέτοιου συνόλου μπορεί να είναι σεντόνια, μαξιλάρια και κουβέρτες. Παρόλ' αυτά ορισμένοι χρήστες μπορεί να αγοράσουν μόνο ένα μέρος των αντικειμένων ενός τέτοιου συνόλου. Για παράδειγμα, ορισμένοι μπορεί να αγοράσουν μόνο σεντόνια και μαξιλάρια, και κάποιοι άλλοι μόνο σεντόνια. Μια συναλλαγή μπορεί να περιέχει περισσότερα από ένα συχνά αντικείμενα. Για παράδειγμα, ένας χρήστης μπορεί να αγοράσει ένα πουκάμισο και ένα σακάκι μαζί με σεντόνια και μαξιλάρια, όπου το πουκάμισο με το σακάκι σχηματίζουν μαζί ένα άλλο συχνό

[‡] Η γεννήτρια, έτοιμες συνθετικές βάσεις δεδομένων παραγμένες από αυτή καθώς και πληροφορίες – δημοσιεύσεις σχετικά με την λειτουργία και τη χρήση της μπορούν να βρεθούν στις ηλεκτρονικές διευθύνσεις: <http://www.almaden.ibm.com/cs/quest/syndata.html> ή στο http://www.cs.umbc.edu/~cgiannel/assoc_gen.html

αντικείμενο. Τα συνήθη μεγέθη των συχνών αντικειμένων εμφανίζονται επίσης συγκεντρωμένα κατά βάση γύρω από ένα μέσο όρο, με πολύ λίγα συχνά αντικείμενα να έχουν μεγάλο πλήθος συχνών στοιχείων. Προκειμένου να δημιουργήσουμε ένα τέτοιο σύνολο συνθετικών συναλλαγών, η γεννήτρια συνθετικών δεδομένων χρησιμοποιεί τις παραμέτρους που φαίνονται στον Πίνακα 1.

Πίνακας 1: Οι παράμετροι που χρησιμοποιούνται για την δημιουργία των συνθετικών δεδομένων

$ D $	Αριθμός συναλλαγών
$ T $	Μέσο μήκος συναλλαγής
$ I $	Μέσος αριθμός μέγιστων πιθανών συχνών αντικειμένων
$ L $	Αριθμός μέγιστων πιθανών συχνών αντικειμένων
N	Αριθμός στοιχείων

Η πρώτη ενέργεια είναι ο καθορισμός του μεγέθους της επόμενης συναλλαγής. Το μέγεθος επιλέγεται από μια κατανομή Poisson με μέση τιμή για το μ ίση με $|T|$. Ας σημειωθεί ότι αν κάθε στοιχείο επιλέγεται με την ίδια πιθανότητα ρ , και υπάρχουν N στοιχεία, το αναμενόμενο πλήθος στοιχείων σε μια συναλλαγή δίνεται από μια δυνωμική κατανομή χρησιμοποιώντας τις παραμέτρους N και ρ , και προσεγγίζει μια κατανομή Poisson με μέση τιμή $N\rho$.

Κατόπιν αποδίδουμε σε κάθε συναλλαγή στοιχεία (items). Κάθε συναλλαγή λαμβάνει μια σειρά πιθανών συχνών στοιχείων. Αν το τρέχον συχνό στοιχείο δεν ταιριάζει σε μια συναλλαγή, στις μισές περιπτώσεις τοποθετείται στη συναλλαγή έτσι και αλλιώς και στις υπόλοιπες τοποθετείται στην επόμενη συναλλαγή. Τα συχνά στοιχεία επιλέγονται από ένα σύνολο T_α τέτοιων στοιχείων. Το πλήθος των στοιχείων στο T τίθεται ίσο με $|L|$. Υπάρχει μια αντίστροφη σχέση ανάμεσα στο $|L|$ και την μέση υποστήριξη (support) των πιθανών συχνών αντικειμένων. Ένα αντικείμενο στο T_α δημιουργείται διαλέγοντας κατ' αρχήν το μέγεθος του από μια κατανομή Poisson με μέση τιμή για το μ ίση με $|I|$. Τα στοιχεία στο πρώτο αντικείμενο επιλέγονται τυχαία. Προκειμένου να μοντελοποιήσουμε το φαινόμενο όπου τα συχνά αντικείμενα συχνά έχουν κοινά στοιχεία, ένα ποσοστό των στοιχείων στα επόμενα αντικείμενα επιλέγεται από το αντικείμενο που δημιουργήθηκε προηγούμενα. Για την επιλογή αυτού του ποσοστού χρησιμοποιείται μια τυχαία μεταβλητή η οποία ακολουθεί μια εκθετική κατανομή με μέση τιμή ίση με το επίπεδο συσχέτισης (correlation level). Τα υπόλοιπα στοιχεία επιλέγονται τυχαία.

Σε κάθε αντικείμενο στο T έχει αποδοθεί ένα βάρος, που αντιστοιχεί στην πιθανότητα το συγκεκριμένο αντικείμενο να επιλεγεί από τους καταναλωτές. Το βάρος αυτό επιλέγεται από μια εκθετική κατανομή με μέση τιμή 1, και κατόπιν κανονικοποιείται έτσι ώστε το άθροισμα

όλων των βαρών για όλα τα αντικείμενα στο T να είναι 1. Το επόμενο αντικείμενο που θα τοποθετηθεί στη συναλλαγή επιλέγεται από το T ρίχνοντας ένα νόμισμα με βάρη που έχει $|L|$ -πλευρές, και όπου το βάρος μιας πλευράς είναι η πιθανότητα να επιλέξουμε το συγκεκριμένο αντικείμενο.

Για να μοντελοποιήσουμε τώρα το φαινόμενο ότι όλα τα στοιχεία που αποτελούν ένα συχνό αντικείμενο δεν αγοράζονται πάντοτε μαζί, αποδίδουμε σε κάθε αντικείμενο στο T ένα ποσοστό αλλοίωσης (contortion) c . Όταν προσθέτουμε ένα στοιχείο σε μια συναλλαγή, συνεχίζουμε να αποκλείουμε ένα στοιχείο από το αντικείμενο για όσο διάστημα ένας αριθμός ο οποίος ακολουθεί μια ομοιόμορφη κατανομή στο διάστημα 0 έως 1 είναι μικρότερος του c . Έτσι για ένα αντικείμενο μεγέθους 1, τις $1-c$ φορές θα προσθέσουμε 1 στοιχεία στην συναλλαγή, $1-1$ στοιχεία τις $c(1-c)$ φορές, $1-2$ τις $c^2(1-c)$ φορές κ.ο.κ. Ο βαθμός αλλοίωσης για ένα αντικείμενο είναι σταθερός και προκύπτει από μια κανονική κατανομή με μέση τιμή 0.5 και απόκλιση (variance) 0.1.

Κεφάλαιο 3 ΤΕΧΝΙΚΕΣ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Περίληψη

Η λύση που παρουσιάζουμε στο Κεφάλαιο 7 χρησιμοποιεί σαν βάση της τεχνικές από τον τομέα της ανάκτησης πληροφορίας. Έτσι για να διευκολύνουμε αλλά και να υποστηρίξουμε καλύτερα την παρουσίαση αυτής της λύσης πρέπει πρώτα να περιγράψουμε τα κυριότερα στοιχεία που δανειζόμαστε από τον συγκεκριμένο τομέα. Σε αυτό το κεφάλαιο παρουσιάζουμε εν συντομία τα ευρετήρια και τις τεχνικές δημιουργίας τους που χρησιμοποιούμε στο σύστημα μας, τον τρόπο με τον οποίο επιλύονται τα δυαδικά αλλά και τα ερωτήματα διάταξης, και τέλος την λειτουργία του διανυσματικού μοντέλου. Για περισσότερες πληροφορίες και λεπτομέρειες πάνω σε αυτά τα θέματα ο αναγνώστης μπορεί να συμβουλευτεί το [126].

Λέξεις Κλειδιά: *Ανάκτηση πληροφορίας; Τεχνικές Δεικτοδότησης; Ευρετήρια; Ανεστραμμένα αρχεία; Συμπίεση; Δυαδικά ερωτήματα; Ερωτήματα διάταξης; Μοντέλο διανυσματικού χώρου*

3.1 Ευρετήρια

Προκειμένου να αντιμετωπίσουμε το ζητούμενο του πώς θα πρέπει να οργανώνονται τα δεδομένα μέσα στη βάση έτσι ώστε τα διάφορα ερωτήματα να μπορούν να απαντώνται τόσο αποδοτικά όσο και αποτελεσματικά, ειδικά σε μια προσέγγιση πραγματικού χρόνου (online), υιοθετήσαμε τεχνικές που χρησιμοποιούνται στην κλασσική θεωρία ανάκτησης πληροφορίας. Πιο συγκεκριμένα χρησιμοποιήσαμε την έννοια του ευρετηρίου (index) και των τεχνικών δεικτοδότησης (indexing). Όταν αναφερόμαστε σε ευρετήρια και στις αντίστοιχες τεχνικές τους αναφερόμαστε σε έναν μηχανισμό για την εύρεση ενός δοσμένου αντικειμένου μέσα στα δεδομένα (αντίστοιχα σε ένα κλασσικό σύστημα ανάκτησης πληροφορίας θα ψάχναμε για έναν όρο ή μια λέξη μέσα στο κείμενο). Αυτό μπορεί να επιτευχθεί με διαφόρους τρόπους, χρησιμοποιώντας μεθόδους όπως signature files, bitmaps και ανεστραμμένα αρχεία (inverted files). Εμείς χρησιμοποιήσαμε τα ανεστραμμένα αρχεία, καθώς για τις περισσότερες περιπτώσεις αποτελούν την καλύτερη λύση.

3.1.1 Ευρετήριο Ανεστραμμένων Αρχείων

Προκειμένου να δημιουργήσουμε ένα ευρετήριο ανεστραμμένων αρχείων πρέπει πρώτα να δημιουργήσουμε το αντίστοιχο λεξικό για τα δεδομένα μας. Το λεξικό δεν είναι τίποτε άλλο από μια λίστα όλων των στοιχείων που εμφανίζονται στη βάση μας. Στην περίπτωση μας το λεξικό είναι ήδη έτοιμο αφού ξέρουμε εκ των προτέρων όλα τα προϊόντα που πωλούνται και δεν υπάρχει περίπτωση ένα προϊόν που δεν περιέχεται σε αυτό να εμφανίζεται στη βάση και αντίστροφα. Αυτό που ουσιαστικά πραγματοποιεί το λεξικό είναι μια αντιστοιχία ανάμεσα στα στοιχεία και στις αντίστοιχες ανεστραμμένες λίστες τους και στην απλούστερη μορφή του είναι και το ίδιο μια λίστα, πιθανά διατεταγμένη, που περιέχει αλφαριθμητικά ή διευθύνσεις κύριας ή δευτερεύουσας μνήμης. Ένα ευρετήριο ανεστραμμένων αρχείων περιέχει επίσης για κάθε στοιχείο στο λεξικό, μια ανεστραμμένη λίστα που αποθηκεύει μια σειρά δεικτών προς όλες τις εμφανίσεις του στοιχείου στη βάση, όπου κάθε δείκτης είναι στην ουσία ο αριθμός της συναλλαγής στην οποία εμφανίζεται αυτό το στοιχείο. Η δομή ενός τέτοιου ευρετηρίου μπορεί να φανεί καλύτερα στο ακόλουθο παράδειγμα. Έχουμε μια βάση που περιέχει 7 συναλλαγές και 8 στοιχεία, τα οποία παρουσιάζονται στον Πίνακα 2. Σε κάθε συναλλαγή αποδίδουμε έναν μοναδικό αριθμό που καλείται αναγνωριστικό συναλλαγής – TID (Transaction Identifier). Χρησιμοποιώντας τώρα την βάση αυτή σαν είσοδο δημιουργούμε το αντίστοιχο ευρετήριο ανεστραμμένου αρχείου, το οποίο θα μοιάζει με αυτό που φαίνεται στον Πίνακα 3. Δίπλα από κάθε στοιχείο υπάρχει και η αντίστοιχη ανεστραμμένη λίστα του.

Πίνακας 2: Μια βάση η οποία περιέχει 7 συναλλαγές, όπου κάθε γραμμή είναι και μια συναλλαγή

TID	ΣΤΟΙΧΕΙΑ
1	1,3,6
2	2,4,6,7
3	1,4,5,8
4	2,4
5	3,4
6	1,4,5
7	2,3,6,8

Η λίστα για κάθε στοιχείο είναι της μορφής $\{f_i, tid_1, tid_2, tid_3, \dots, tid_f\}$, όπου f_i είναι ο αριθμός των συναλλαγών στις οποίες εμφανίζεται ένα στοιχείο, ακολουθούμενο από τα αναγνωριστικά των συναλλαγών (τα TIDs) τα οποία δείχνονται εδώ σαν tid_k . Για παράδειγμα το στοιχείο 2 εμφανίζεται σε 3 συναλλαγές, αναφορικά στις συναλλαγές 2, 4 και 7. Η πρώτη στήλη του Πίνακας 3 μπορεί να θεωρηθεί σαν το λεξικό και η δεύτερη οι ανεστραμμένες λίστες. Βέβαια συνήθως ο αριθμός εμφανίσεων κάθε στοιχείου δεν αποθηκεύεται στην ανεστραμμένη λίστα όπως φαίνεται εδώ, αλλά στο λεξικό μαζί με το ίδιο το στοιχείο.

Πίνακας 3: Το ανεστραμμένο αρχείο για την βάση του Πίνακας 2

Στοιχείο	Συναλλαγές
1	{3; 1, 3, 6}
2	{3; 2, 4, 7}
3	{3; 1, 5, 7}
4	{5; 2, 3, 4, 5, 6}
5	{2; 3, 6}
6	{3; 1, 2, 7}
7	{1; 2}
8	{2; 3, 7}

3.1.2 Συμπίεση του Ευρετηρίου

Όταν το ευρετήριο έχει δημιουργηθεί είναι πολύ πιθανό ότι χρειάζεται να συμπιεστεί έτσι ώστε να μπορεί να διατηρείται και να επεξεργάζεται από την κύρια μνήμη. Η διατήρηση του στην κύρια μνήμη καθιστά την προσπέλαση του ταχύτερη, και την επεξεργασία οποιουδήποτε ερωτήματος πρακτικά άμεση. Φυσικά αυτό δεν είναι πάντα αναγκαίο μιας και το ευρετήριο μας μπορεί να είναι αρκετά μικρό ή η διαθέσιμη μνήμη αρκετά μεγάλη και έτσι να μην υπάρχει ανάγκη να το συμπιέσουμε. Από την άλλη, η συμπίεση του ευρετηρίου σημαίνει κάποιο πρόσθετο χρόνο να πραγματοποιηθεί, και κάποια πρόσθετη προσπάθεια και χρόνο προκειμένου να επεξεργαστούμε κάθε φορά κάποιο ερώτημα. Παρόλα αυτά τα κόστη είναι αμελητέα συγκρινόμενα με τα οφέλη που επιτυγχάνουμε διατηρώντας ολόκληρο το ευρετήριο στην κύρια μνήμη.

Μια ανεστραμμένη λίστα μπορεί να αποθηκευτεί σαν μια αύξουσα ακολουθία ακεραίων. Τώρα ας υποθέσουμε ότι έχουμε μια βάση με ορισμένες συναλλαγές, και ένα συγκεκριμένο στοιχείο εμφανίζεται σε επτά από αυτές – έστω στις συναλλαγές 5, 8, 9, 10, 53, 54 και 55. Τότε το συγκεκριμένο στοιχείο περιγράφεται από μια λίστα σαν την παρακάτω:

$$\{7; 5, 8, 9, 10, 53, 54, 55\}$$

Όλα τα TIDs στη ανεστραμμένη λίστα ενός αντικειμένου είναι αποθηκευμένα σε αύξουσα σειρά έτσι ώστε να ισχύει $tid_k < tid_{k+1}$. Λόγω αυτής της διάταξης, και επειδή η επεξεργασία πραγματοποιείται σειριακά από την αρχή κάθε λίστας, κάθε λίστα μπορεί να αποθηκευτεί σαν μια αρχική θέση ακολουθούμενη από μια λίστα αντιστοιχών tid -διάκενων, όπου ένα διάκενο είναι η διαφορά $tid_{k+1} - tid_k$. Η παραπάνω λίστα μπορεί να αποθηκευτεί σαν:

$$\{7; 5, 3, 1, 1, 43, 1, 1\}$$

Μπορούμε πάντα να ανακτήσουμε τα αρχικά TIDs απλά υπολογίζοντας σε κάθε βήμα τα αθροίσματα των διακένων.

3.2 Επίλυση Δυαδικών Ερωτημάτων

Όταν ένα ερώτημα που περιλαμβάνει οποιονδήποτε από τους τρεις κύριους δυαδικούς τελεστές τίθεται στο σύστημα μας, η διαδικασία που ακολουθείται προκειμένου να απαντηθούν έχει ως εξής: Πρώτα ψάχνουμε το λεξικό για κάθε όρο. Κατόπιν, σύμφωνα με το που δείχνει το λεξικό για κάθε όρο, οι αντίστοιχες ανεστραμμένες λίστες προσπελαύνονται, ανακτώνται και αποκωδικοποιούνται (decoded). Τώρα είμαστε έτοιμοι να ενώσουμε λίστες, παίρνοντας την τομή τους (AND – περίπτωση), την ένωση τους (OR – περίπτωση) ή το συμπλήρωμα τους (NOT – περίπτωση). Αν ο χρήστης επιθυμεί να δει επίσης και τις συναλλαγές που περιέχουν τα συγκεκριμένα αντικείμενα, τότε οι ίδιες οι συναλλαγές ανακτώνται και παρουσιάζονται σε μια

μορφή που αυτός θεωρεί βέλτιστη. Παρόλα αυτά στην περίπτωση μας συνήθως θέλουμε να ξέρουμε μόνο σε πόσες συναλλαγές παρουσιάζεται ένα αντικείμενο.

Ας δούμε πως ένα ερώτημα που περιλαμβάνει το τελεστή AND και μερικά στοιχεία επιλύεται, καθώς πέραν του γεγονότος ότι είναι η πιο συνηθισμένη περίπτωση είναι ίσως η πιο πολύπλοκη, ενδιαφέρουσα και πιο απαιτητική. Ας υποθέσουμε ότι επιθυμούμε να βρούμε σε πόσες συναλλαγές εμφανίζονται τα στοιχεία A, C, F, ή με άλλα λόγια θέτουμε το εξής ερώτημα.

A AND C AND F

Αυτό που κάνουμε κατ' αρχήν είναι ο εντοπισμός του κάθε στοιχείου στο λεξικό, το οποίο όπως είπαμε κ προηγούμενα βρίσκεται στην κύρια μνήμη για γρηγορότερη πρόσβαση και έλεγχο. Κατόπιν, εφόσον όλα τα στοιχεία έχουν βρεθεί τα διατάσσουμε σε αύξουσα σειρά σύμφωνα με την συχνότητα εμφανίσεών τους. Η διάταξη αυτή είναι απλή αφού στο λεξικό έχουμε ήδη αποθηκευμένα το πλήθος των εμφανίσεων κάθε στοιχείου (f_i). Κατόπιν η ανεστραμμένη λίστα του λιγότερα συχνού στοιχείου αποσυμπίεζεται και διαβάζεται στην κύρια μνήμη. Αυτή θα είναι και η λίστα η οποία θα αποτελέσει το σύνολο των υποψήφιων συναλλαγών (πιο συγκεκριμένα τα αντίστοιχα TIDs), που θα ελεγχθούν σε σύγκριση με τις λίστες των άλλων στοιχείων (πάντα σε αύξουσα σειρά συχνότητας). Το μέγεθος της λίστας που θα αποτελεί την απάντηση μπορεί να είναι το πολύ το ίδιο μέγεθος με αυτό της πρώτης υποψήφιας λίστας, και συνήθως θα είναι πολύ μικρότερη ή μερικές φορές ακόμα και άδεια. Έτσι βλέπουμε ότι η εργασία που κυριαρχεί κατά την επίλυση ενός ερωτήματος που περιλαμβάνει τελεστές AND είναι η αναζήτηση παρά η ένωση λιστών.

Η περίπτωση του OR υλοποιείται με έναν πολύ απλούστερο τρόπο. Αυτό που χρειάζεται να κάνουμε ξανά είναι να εντοπίσουμε κάθε στοιχείο στο λεξικό, και να επεξεργαστούμε τις αντίστοιχες ανεστραμμένες λίστες ενώνοντας παρά ψάχνοντας. Αφότου όλες οι λίστες θα έχουν συνενωθεί αφαιρούμε όλα τα διπλότυπα και παρουσιάζουμε το αποτέλεσμα. Τέλος η περίπτωση του NOT υλοποιείται σχεδόν με τον ίδιο τρόπο όπως και η περίπτωση του AND. Η μόνη διαφορά είναι ότι όταν ελέγχουμε τις λίστες όλων των άλλων στοιχείων σε σχέση με τη λίστα των υποψηφίων, απομακρύνουμε αυτές τις συναλλαγές που εμφανίζονται τόσο στην λίστα με τα υποψήφια όσο και σε οποιαδήποτε άλλη λίστα.

3.3 Το Μοντέλο του Διανυσματικού Χώρου

Το ευρετήριο που δημιουργείται όπως περιγράφηκε παραπάνω μπορεί να δίνει απαντήσεις μόνο σε δυαδικά ερωτήματα. Όταν χρησιμοποιούμε ένα ευρετήριο για να κάνουμε ερωτήματα διάταξης χρησιμοποιούμε ένα μοντέλο που καλείται μοντέλο διανυσματικού χώρου. Θα

περιγράφουμε την λειτουργία αυτού του μοντέλου μέσω της πιο συνηθισμένης εφαρμογής του, τις συλλογές κειμένων. Σε ένα διανυσματικό μοντέλο, τα κείμενα παρουσιάζονται σαν διανύσματα σε έναν πολυδιάστατο Ευκλείδειο χώρο. Κάθε άξονας σε αυτό το χώρο αντιστοιχεί σε έναν όρο. Η λειτουργία του διανυσματικού μοντέλου μπορεί να χωριστεί σε τρεις λογικές φάσεις. Η πρώτη φάση συνίσταται στη δημιουργία ενός ευρετηρίου, βάση μιας συλλογής αρχείων (συνήθως κειμένων). Η δεύτερη λογική φάση είναι η απόδοση ενός βάρους σε όλα τα στοιχεία στο ευρετήριο και η τρίτη και τελευταία ουσιαστικά είναι η χρησιμοποίηση κάποιας μετρικής εγγύτητας προκειμένου να κατατάξουμε τα κείμενα που αποτελούν την απάντηση σε ένα ερώτημα. Φυσικά αυτές οι φάσεις, και ειδικότερα η πρώτη και η δεύτερη δεν είναι πραγματικά διαφορετικές αλλά τις θεωρούμε έτσι για απλότητα. Ας δούμε πως δουλεύει αυτό το μοντέλο.

Στην φάση της δημιουργίας του ευρετηρίου η συλλογή διαβάζεται από την βάση και το ευρετήριο δημιουργείται. Αυτή η συλλογή μπορεί να είναι πρακτικά οτιδήποτε, από ένα μόνο αρχείο κειμένου έως αρκετές χιλιάδες αρχεία με οποιοδήποτε περιεχόμενο και μορφή. Η συγκεκριμένη φάση ουσιαστικά είναι η ίδια με αυτή που παρουσιάστηκε στην ενότητα 3.1.1.

Στην δεύτερη φάση τώρα σε όλα τα στοιχεία αποδίδεται ένα βάρος [131]. Ένα συνηθισμένο σχήμα απόδοσης βαρών είναι αυτό που καλείται TFIDF. Σύμφωνα με αυτό, κάθε όρος λαμβάνει ένα βάρος ανάλογα με το πόσες φορές εμφανίζεται σε ένα κείμενο, διαβαθμισμένο με διάφορους τρόπους προκειμένου να κανονικοποιήσουμε το μήκος του κειμένου (όπως το άθροισμα του αριθμού των όρων στο κείμενο). Αυτό το βάρος καλείται παράγοντας TF (Term Frequency). Το TFIDF σχήμα προσπαθεί επίσης να μειώσει την σημασία των όρων που εμφανίζονται σε πολλά κείμενα, ή είναι μικρής σημασίας για τους χρήστες. Αυτό επιτυγχάνεται με χρήση του παράγοντα IDF (Inverse Document Frequency). Μια ευρέως χρησιμοποιούμενη μέθοδος για τον καθορισμό του παράγοντα IDF είναι η αρχή του Zipf [130]. Σύμφωνα με αυτή την αρχή η συχνότητα ενός όρου τείνει να είναι αντιστρόφως ανάλογη της σημαντικότητας του και δίνεται από τον εξής τύπο: $w_i = 1/f_i$. Οι παράγοντες TF και IDF συνδυάζονται σε ένα πλήρες διανυσματικό μοντέλο ως εξής: Κάθε όρος αντιστοιχεί σε μια διάσταση στο διανυσματικό χώρο. Τα κείμενα όπως και τα ερωτήματα αναπαριστώνται σαν διανύσματα σ' αυτόν τον χώρο. Η συντεταγμένη ενός κειμένου στην κατεύθυνση ενός άξονα t δίνεται από τη σχέση $d_i = TF(d, t) \cdot IDF(t)$. Κάθε ερώτημα q θεωρείται σαν έναν κείμενο και μετατρέπεται επίσης σε έναν άξονα στον ίδιο διανυσματικό χώρο. Αυτό το σχήμα έχει χρησιμοποιηθεί εκτεταμένα με πολύ καλά αποτελέσματα ιδιαίτερα για συλλογές κειμένων.

Η τελευταία και πιο σημαντική φάση είναι να μετρήσουμε την εγγύτητα ανάμεσα στο ερώτημα q και όλα τα σχετικά κείμενα στη συλλογή, η διάταξη των απαντήσεων και η παρουσίαση στο χρήστη των πιο σχετικών. Η εγγύτητα ανάμεσα σε ένα ερώτημα και στα

κείμενα της συλλογής μας μετράται χρησιμοποιώντας διάφορες μετρικές εγγύτητας, όπως το εσωτερικό γινόμενο, ο συντελεστής Jaccard, ο συντελεστής Dice ή η πιο δημοφιλής μετρική του συνημίτονου (cosine measure) [131]. Τα αποτελέσματα διατάσσονται και τελικά στο χρήστη παρουσιάζονται τα πιο σχετικά κείμενα σε φθίνουσα σειρά σημαντικότητας.

**Κεφάλαιο 4 ΑΠΟΤΕΛΕΣΜΑΤΙΚΗ
ΑΥΤΟΜΑΤΗ ΑΝΑΚΑΛΥΨΗ “ΚΑΥΤΩΝ”
ΠΡΟΪΟΝΤΩΝ**

ΠΕΡΙΛΗΨΗ

Σε πραγματικές συνθήκες-εφαρμογές το μέχρι τώρα κυρίαρχο μοντέλο της μιας και μοναδικής υποστήριξης, το οποίο θεωρούσε όλα τα προϊόντα της ίδιας φύσης και σημαντικότητας αποδείχτηκε ανεπαρκές. Η ανομοιογένεια των προϊόντων από τη μια μεριά και οι μεγάλες διακυμάνσεις του πλήθους των εμφανίσεων τους από την άλλη επιβάλλουν την χρήση διαφορετικών προσεγγίσεων. Οι τεχνικές που είχαν προταθεί μέχρι τώρα, προσπαθούσαν να αντιμετωπίσουν αυτές τις αδυναμίες αλλά νέα πιο πολύπλοκα προβλήματα ανέκυπταν όπως παράδειγμα ποια προϊόντα είναι πιο σημαντικά από τα άλλα, τι τα ξεχωρίζει, πως μπορούμε να τα ανακαλύψουμε, και τέλος πως θα μπορούσαμε να τα διαχειριστούμε κατάλληλα. Επιπρόσθετα μια κοινή αδυναμία όλων των τεχνικών είναι ότι έχουν μια μεγάλη χρονική υστέρηση στο να ανακαλύπτουν νέες συσχετίσεις και ουσιαστικά δουλεύουν μόνο με υπάρχουσες συσχετίσεις και μοτίβα. Στην παρούσα εργασία προτείνουμε μια μέθοδο που βρίσκει αυτό που καλούμε “καυτά” προϊόντα στην βάση μας, αντιμετωπίζει τα προαναφερθέντα προβλήματα και τελικά αποδεικνύεται αποτελεσματική.

Λέξεις Κλειδιά: *Εξόρυξη Γνώσης; Κανόνες Συσχετίσεων; Καυτά Προϊόντα;*

4.1 Ανακαλύπτοντας Διαφαινόμενες Τάσεις

Η βασική ιδέα πίσω από την εύρεση κανόνων συσχετίσεων είναι το ψάξιμο σε ένα αρκετά μεγάλο όγκο δεδομένων τα οποία θα έχουν συλλεχθεί σε κάποια σχετικά μακρά χρονική περίοδο, και η εφαρμογή διαφόρων τεχνικών προκειμένου να ανακαλύψουμε κάποια λιγότερο ή περισσότερο άγνωστη γνώση μέσα στα δεδομένα. Ένα μεγάλο πρόβλημα των μεθόδων και των προσεγγίσεων που χρησιμοποιούνταν έως τώρα ήταν ότι ανακάλυπταν σχέσεις μεταξύ αντικειμένων οι οποίες προϋπήρχαν για μεγάλο διάστημα αδυνατώντας να ανακαλύψουν τις νέες. Όλες οι προσεγγίσεις έως τώρα ακολουθούσαν παρά συμβάδιζαν με τις πωλήσεις ή γενικότερα τις εμφανίσεις των αντικειμένων. Όμως εμείς κατά βάση χρειαζόμαστε μια προσέγγιση η οποία θα βρίσκει **διαφαινόμενες - ανερχόμενες τάσεις** εν τη γενέσει τους, ταυτόχρονα με τις ήδη υπάρχουσες. Το ζητούμενο αυτό μπορεί να εξηγηθεί καλύτερα στο παρακάτω παράδειγμα.

Έστω ένα προϊόν το οποίο είναι στην αγορά αρκετά χρόνια με μέτριες πωλήσεις και ένα άλλο προϊόν το οποίο μόλις εισήχθη στην αγορά (π.χ. ένα προϊόν το οποίο πωλείται μόλις ένα χρόνο) αλλά εμφανίζει αξιοσημείωτες πωλήσεις. Αν εφαρμόσουμε το κλασσικό στατιστικό μοντέλο των κανόνων συσχετίσεων όπου απλά μετράμε τον αριθμό των εμφανίσεων κάθε προϊόντος και αν αυτό είναι πάνω από ένα προκαθορισμένο - ορισμένο από τον χρήστη όριο θεωρείται σαν συχνό, τότε σίγουρα το νέο προϊόν δεν θα ξεπερνά αυτό το όριο. Ένα προϊόν το οποίο βρίσκεται στην αγορά για τόσο μικρό διάστημα δεν μπορεί πρακτικά ούτε καν να πλησιάσει τις πωλήσεις ενός προϊόντος το οποίο πωλείται για πολύ μεγαλύτερο διάστημα. Έτσι θα πρέπει είτε να περιμένουμε κάποιο διάστημα ώστε το νέο προϊόν μας να συγκεντρώσει αρκετές πωλήσεις, το οποίο μπορεί κάλλιστα να πάρει μήνες, ή να βρούμε έναν τρόπο ώστε να μπορέσουμε αποδοτικά να το λάβουμε υπόψη άμεσα. Η ίδια κατάσταση μπορεί να εμφανιστεί με προϊόντα τα οποία πωλούνταν και παλιότερα αλλά παρουσίαζαν πολύ χαμηλές πωλήσεις και ξαφνικά ξεκίνησαν να παρουσιάζουν ασυνήθιστα υψηλές πωλήσεις για διάφορους λόγους όπως π.χ. επειδή διαφημίζονται πολύ, ξαφνικά έγιναν της μόδας, κάποιοι εξωτερικοί παράγοντες ή καταστάσεις ανέβασαν τις πωλήσεις τους (π.χ. κλιματολογικές συνθήκες) κ.ο.κ. . Η συγκεκριμένη περίπτωση είναι πολύ συνηθισμένη, ιδιαίτερα σε καταστήματα λιανεμπορίου όπου τα προϊόντα που πωλούνται εκεί παρουσιάζουν τέτοιες συμπεριφορές. Σε κάθε περίπτωση τελικά ένα προϊόν το οποίο παρουσιάζει τόσες πολλές πωλήσεις σε τόσο μικρή περίοδο πρέπει να είναι προφανώς πολύ ενδιαφέρον, και μάλιστα τέτοιες “εκρηκτικές” και απρόβλεπτες συμπεριφορές είναι πολύ πιο ενδιαφέρουσες από τις συνηθισμένες. Στην επόμενη ενότητα θα δούμε πως διαχειριζόμαστε αυτά τα προϊόντα.

4.2 “Καυτά” Προϊόντα

Καλούμε τα προϊόντα τα οποία παρουσιάζουν πολύ υψηλές πωλήσεις σε μια συγκεκριμένη χρονική περίοδο σαν “καυτά” προϊόντα. Για κάθε στοιχείο υπολογίζουμε αυτό που ονομάζουμε τον **λόγο ενδιαφέροντος** (**ir – interest ratio**), το οποίο ορίζεται σαν ο αριθμός των πωλήσεων ενός προϊόντος στην τελευταία περίοδο πωλήσεων (δηλ. από την τελευταία φορά που τρέξαμε τον αλγόριθμο μας) δια του μέσου όρου πωλήσεων όλων των προϊόντων στην ίδια περίοδο.

$$ir_i = \frac{sales_i}{sales}$$

Κάθε προϊόν που έχει τον λόγο ενδιαφέροντος μεγαλύτερο ή ίσο από ένα κατώτατο - ορισμένο από τον χρήστη όριο, το οποίο καλείται κατώτατος λόγος ενδιαφέροντος θεωρείται σαν καυτό για την συγκεκριμένη περίοδο. Βέβαια αν ένα προϊόν έχει αριθμό πωλήσεων πάνω από το κατώφλι υποστήριξης τότε αντιμετωπίζεται σαν ένα συχνό προϊόν. Στην ουσία ψάχνουμε για προϊόντα που έχουν πωλήσεις κάτω από το κατώφλι υποστήριξης αλλά ο λόγος ενδιαφέροντος τους είναι πάνω από το κατώτατο λόγο ενδιαφέροντος. Ο χρήστης έχει βέβαια την επιλογή να δίνει σ’ αυτό το λόγο οποιαδήποτε τιμή, ανάλογα με το επιθυμητό αποτέλεσμα.

Μια λογική ερώτηση που ανακύπτει είναι τι συμβαίνει με ένα προϊόν το οποίο ήταν καυτό κατά την προηγούμενη περίοδο αλλά δεν είναι πλέον ούτε καυτό αλλά ούτε και συχνό στην επόμενη. Μια επιλογή θα μπορούσε να ήταν να μεταχειριστούμε αυτά τα προϊόντα σαν σπάνια - μη ενδιαφέροντα στην νέα περίοδο αφού προφανώς δεν είναι πλέον ενδιαφέροντα στους χρήστες. Μια άλλη βέβαια θα μπορούσε να ήταν να δώσουμε σε αυτά τα προϊόντα μια περίοδο χάριτος, και να τα αντιμετωπίσουμε σαν καυτά, προκειμένου να δούμε αν θα επανέλθουν στις υψηλές πωλήσεις. Οποιαδήποτε από τις δύο προαναφερθείσες επιλογές είναι αποδεκτή και εξαρτάται αποκλειστικά από τις ανάγκες του χρήστη. Κάποιος βέβαια θα μπορούσε να ισχυριστεί ότι κακώς θεωρήσαμε τα συγκεκριμένα προϊόντα σαν ενδιαφέροντα, αφού η μετέπειτα τάση τους έδειξε ότι δεν είναι πλέον καυτά. Παρ’ όλ’ αυτά καταφέραμε να εντοπίσουμε άμεσα την περίοδο κατά την οποία έγιναν ενδιαφέροντα, τα λάβαμε υπόψη και προωθήσαμε τις πωλήσεις τους όταν ήταν ενδιαφέροντα και αυτός ήταν ακριβώς ο σκοπός μας. Αν από την άλλη ένα καυτό προϊόν γίνει συχνό στην επόμενη περίοδο τότε καταφέραμε να προβλέψουμε επιτυχώς την μελλοντική του συμπεριφορά και απόδοση αρκετά νωρίς και να το λάβουμε υπόψη εκ των προτέρων χωρίς να χρειάζεται να το περιμένουμε να γίνει συχνό.

Το τελικό ερώτημα είναι πόσο τακτικά θα πρέπει να “τρέχουμε” τον αλγόριθμο μας στα δεδομένα μας. Όπως καταλαβαίνουμε από τους παραπάνω ορισμούς, η περίοδος που τρέχουμε τον αλγόριθμο μας επηρεάζει τόσο την ποσότητα όσο και την φύση των προϊόντων που ανακαλύπτουμε. Σύμφωνα με την εργασία στο [99] προκειμένου να πάρουμε κάποια ασφαλή

συμπεράσματα από τη βάση μας θα πρέπει να περιμένουμε για κάποιο διάστημα προκειμένου να μαζέψουμε αρκετά δεδομένα. Ο χρόνος αυτός μπορεί να ποικίλει από επιχείρηση σε επιχείρηση και από τομέα σε τομέα. Για παράδειγμα μια πολυεθνική εταιρεία που συγκεντρώνει δεδομένα από διαφορετικά καταστήματα σε διάφορες χώρες, ή ένα πολύ δημοφιλές ηλεκτρονικό κατάστημα (π.χ. Amazon ή eBay) ή κάποιοι κλάδοι της βιομηχανίας όπως οι τηλεπικοινωνίες πιθανότατα να χρειάζονταν να τρέχουν κάποιο αλγόριθμο κάθε μια ή δύο ημέρες. Άλλοι όχι τόσο δημοφιλείς πιθανώς να χρειάζονταν μια περίοδο ενός μηνός ή περισσότερο. Έτσι η περίοδος κατά την οποία “τρέχουμε” τον αλγόριθμο μας είναι μια εντελώς υποκειμενική απόφαση η οποία εξαρτάται από την εφαρμογή, την εμπειρία και τις ανάγκες του χρήστη (data miner), και δεν αλλάζει αν χρησιμοποιούσαμε οποιονδήποτε άλλο κλασικό αλγόριθμο. Ως εκ τούτου η αυτόματη επιλογή αυτής της περιόδου δεν κρίνεται ακόμα δυνατή.

4.3 Η προτεινόμενη προσέγγιση

Η κύρια ιδέα της προσεγγίσεως μας είναι να παρακολουθούμε όλα τα προϊόντα για μη φυσιολογική – ασυνήθιστη συμπεριφορά από την τελευταία φορά που τρέξαμε τον αλγόριθμο μας μέχρι την επόμενη. Ξεκινάμε μετρώντας τον αριθμό εμφανίσεων των στοιχείων (Εικόνα 5, γραμμές 4-8). Αφού ο αριθμός εμφανίσεων όλων των στοιχείων έχει βρεθεί, μπορούμε να εντοπίσουμε τρεις τύπους προϊόντων ανάλογα με τον αριθμό εμφανίσεων τους.

- ▮ **Συχνά αντικείμενα (I-items):** Προϊόντα με αριθμό εμφανίσεων μεγαλύτερο από το κατώφλι υποστήριξης.
- ▮ **Καυτά αντικείμενα:** Προϊόντα με αριθμό εμφανίσεων κάτω από το κατώφλι υποστήριξης, αλλά πάνω από το κατώτατο λόγο ενδιαφέροντος.
- ▮ **Σπάνια αντικείμενα:** Προϊόντα με αριθμό εμφανίσεων κάτω τόσο από το κατώφλι υποστήριξης όσο και κάτω από το κατώτατο λόγο ενδιαφέροντος.

Πρώτα βρίσκουμε όλα τα συχνά αντικείμενα, συγκρίνοντας τον αριθμό των εμφανίσεων τους με το κατώφλι υποστήριξης. Ακολούθως βρίσκουμε όλα τα καυτά αντικείμενα παίρνοντας την διαφορά μεταξύ των προηγούμενων και των νέων πωλήσεων και βρίσκοντας τον αριθμό των εμφανίσεων στην παρούσα περίοδο. Τα εναπομείναντα αντικείμενα χαρακτηρίζονται αυτόματα σαν σπάνια (Εικόνα 5, γραμμές 10-15). Το επόμενο βήμα είναι η απόδοση MIS τιμών σε όλα τα αντικείμενα, σύμφωνα με την παρακάτω εξίσωση:

$$MIS(i) = \begin{cases} f(i) & i = hot \\ \min \sup & i = l \arg e \end{cases}$$

```

1) Database=set of transactions;
2) Items=set of items;
3) Transaction=<TID, {x| x ∈ Items }>;
4) foreach transaction  $t \in D$  do begin
5)     foreach item  $x$  in  $t$  do begin
6)          $x.count++$ ;
7)     end
8) end
9) Comment: Categorize all 1-items according to their number of appearances
10) foreach item  $x$  |  $x \in Items$  do
11)     if  $x.count \geq \text{min sup}$  then  $x=\text{large}$ ;
12)     elseif  $x.count \geq \text{ir}$  then  $x=\text{hot}$ ;
13)     else  $x=\text{small}$ ;
14) end
15) end
16) Comment: Assign MIS values to all 1-items
17) foreach 1-item  $x \in \text{large\_or\_hot}$  do
18)     if  $x \in \text{large}$   $MIS(x)=\text{minsup}$ 
19)     else  $MIS(x)=f(x)$ 
20) end
21)  $H_1=\{\text{large and hot 1-items}\}$ ;
22)  $M=\text{sort}(H_1, MIS)$ ;
23)  $L_1=\{\text{sorted large and hot 1-items}\}$ ;
24) for ( $k=2$ ;  $L_{k-1} \neq \emptyset$ ;  $k++$ ) do
25)      $C_k=\text{candidate\_gen}(L_{k-1})$  //create new candidates
26)     foreach transaction  $t \in D$  do begin
27)          $C_t=\text{subset}(C_k, t)$ ;
28)         foreach candidate  $x \in C_t$  do
29)              $x.count++$ ;
30)         end
31)          $L_k=\{x \in C_k \mid x.count \geq MIS(c[1])\}$ 
32)     end
33)  $\text{Answer}=\bigcup_k L_k$ ;

```

Εικόνα 5: Το κυρίως πρόγραμμα του αλγορίθμου μας

Σε κάθε συχνό προϊόν αποδίδεται σαν MIS τιμή η τιμή του κατωφλίου υποστήριξης. Τα καυτά προϊόντα λαμβάνουν σαν MIS τιμή τον αριθμό των εμφανίσεών τους. Τα υπόλοιπα προϊόντα απλά απορρίπτονται (Εικόνα 5, γραμμές 17-20). Έχοντας τώρα τις MIS τιμές για όλα τα αντικείμενα μπορούμε να τα ταξινομήσουμε σε αύξουσα σειρά σύμφωνα με τις τιμές αυτές έτσι ώστε να ικανοποιούν την sorted closure αρχή§ (Εικόνα 5, γραμμή 22). Έτσι βάζουμε πρώτα το καυτό στοιχείο με την χαμηλότερη MIS τιμή, και ακολούθως κάθε καυτό στοιχείο με

§ Για περισσότερες λεπτομέρειες δες την παράγραφο 2.5 ή την δουλειά στο [89].

μεγαλύτερη MIS τιμή. Αφότου όλα τα καυτά στοιχεία έχουν ταξινομηθεί, βάζουμε τα συχνά στοιχεία. Αυτά αποτελούν το σύνολο L_1 , το ταξινομημένο σύνολο των συχνών στοιχείων (Εικόνα 5, γραμμή 23). Όπως μπορούμε εύκολα να καταλάβουμε όλα τα καυτά στοιχεία μπαίνουν πρώτα στην διάταξη σύμφωνα με τις MIS τιμές αφού οι αντίστοιχες MIS τιμές τους είναι πάντα κάτω από το κατώφλι υποστήριξης, και όλα τα υπόλοιπα συχνά στοιχεία ακολουθούν.

Από αυτό το σημείο και έπειτα η προσέγγιση μας πραγματοποιεί τρεις λειτουργίες σε κάθε πέρασμα από τα δεδομένα (Εικόνα 5, γραμμές 24-32). Πρώτα τα υποψήφια αντικείμενα C_k παράγονται χρησιμοποιώντας τα συχνά αντικείμενα που προέκυψαν στο προηγούμενο $(k-1)$ πέρασμα. Κατόπιν η βάση μας σαρώνεται ξανά και η υποστήριξη κάθε αντικειμένου στο C_k βρίσκεται. Τελικά τα συχνά k -αντικείμενα βρίσκονται και χρησιμοποιούνται προκειμένου να σχηματιστεί το L_k το σύνολο των συχνών k -αντικειμένων. Ο αλγόριθμος μας τερματίζει με την απουσία κάποιου υποψήφιου αντικειμένου.

Δεν ψάχνουμε για καυτά αντικείμενα μεγαλύτερης τάξης (δηλ. για k -αντικείμενα, όπου $k \geq 2$) αλλά μόνο για συχνά αφού οποιοδήποτε k -αντικείμενο θα έχει όλα του τα υποσύνολα σαν συχνά ή καυτά. Αυτό σημαίνει ότι οποιοδήποτε k -αντικείμενο μπορεί να έχει ένα μέγιστο αριθμό εμφανίσεων ίσο με τον αριθμό εμφανίσεων του λιγότερου συχνού αντικειμένου που το αποτελεί. Επίσης η MIS τιμή που αποδίδεται σε αυτό (η χαμηλότερη MIS τιμή ανάμεσα σε όλα τα αντικείμενα που το αποτελούν) είναι η χαμηλότερη δυνατή MIS τιμή που μπορεί να λάβει. Έτσι δεν υπάρχει καμία περίπτωση ένα οποιοδήποτε k -αντικείμενο να μην βρεθεί συχνό αλλά να είναι καυτό.

4.4 Πειραματικά Αποτελέσματα

Σ’ αυτή την ενότητα παρουσιάζουμε τα αποτελέσματα των πειραμάτων που πραγματοποιήσαμε προκειμένου να εξετάσουμε την απόδοση της προσέγγισης μας. Οι χρησιμοποιούμενες βάσεις δεδομένων ήταν όλες τεχνητές, και είχαν παραχθεί από την γεννήτρια τεχνητών δεδομένων της IBM.

4.4.1 Συνθετικά Δεδομένα

Η δημιουργία συνθετικών δεδομένων καθώς και η λειτουργία της γεννήτριας επεξηγούνται επαρκώς στο [10] αλλά και στην ενότητα 2.8. Σαν παράδειγμα των βάσεων δεδομένων που δημιουργούνται χρησιμοποιώντας την συγκεκριμένη γεννήτρια, το αρχείο “T15.I4.D100K” αναφέρεται σε μια βάση με μέσο μήκος συναλλαγής τα 15 προϊόντα, ένα μέσο μέγεθος των πιθανώς συχνών προϊόντων 4 και τέλος ένα σύνολο 100 000 συναλλαγών.

Οι βάσεις δεδομένων που δημιουργούνται κατ’ αυτόν τον τρόπο μοιάζουν με αρχεία κειμένου με γραμμές, όπου κάθε γραμμή περιέχει κάποια προϊόντα και αναπαριστά μια συναλλαγή που πραγματοποιήθηκε από ένα πελάτη. Παρ’ όλ’ αυτά, όπως επισημάνθηκε νωρίτερα ο αλγόριθμος μας προσπαθεί να βρει τόσο τα συχνά όσο και τα καυτά προϊόντα από την τελευταία φορά που τρέξαμε τον αλγόριθμο μας. Έτσι αν μια βάση που δημιουργείται με χρήση της συγκεκριμένης γεννήτριας αναπαριστά τις συναλλαγές που πραγματοποιήθηκαν για μια συγκεκριμένη χρονική περίοδο, πρέπει να βρούμε έναν τρόπο να αναπαραστήσουμε τις πωλήσεις στη νέα περίοδο. Ακολουθώντας για την δημιουργία δεδομένων χρησιμοποιούμε μια διαφορετική διαδικασία, η οποία έχει ως ακολούθως:

Κατ’ αρχήν δημιουργούμε μια βάση δεδομένων, χρησιμοποιώντας την γεννήτρια συνθετικών δεδομένων. Υποθέτουμε ότι αυτή η βάση αντιπροσωπεύει τις πωλήσεις όλων των προϊόντων μέχρι τώρα. Προκειμένου να προσομοιώσουμε τις πωλήσεις των προϊόντων στη νέα περίοδο (δηλ. όταν τρέχουμε ξανά τον αλγόριθμο μας), χρησιμοποιούμε μια γεννήτρια τυχαίων αριθμών που αποδίδει ένα θετικό αριθμό σε κάθε προϊόν που αναπαριστά τις νέες του πωλήσεις. Αφού γνωρίζουμε τον αριθμό των πωλήσεων κάθε νέου προϊόντος στις θεωρούμενες προηγούμενες περιόδους, περιορίσαμε τον δυνατό αριθμό πωλήσεων κάθε προϊόντος ανάμεσα στις πωλήσεις του πιο συχνού και του πιο σπάνιου προϊόντος, μειωμένες κατά ένα παράγοντα 10. Με άλλα λόγια εάν το πιο συχνό προϊόν στην αρχική μας βάση είχε 1000 εμφανίσεις και το πιο σπάνιο 100 τότε οποιοδήποτε προϊόν στη νέα περίοδο δεν μπορεί να έχει παραπάνω από 100 εμφανίσεις και λιγότερες από 10. Δημιουργήσαμε μ’ αυτόν τον τρόπο συνολικά 5 βάσεις, οι οποίες φαίνονται στον Πίνακα 4.

Πίνακας 4: Τεχνητές Βάσεις

Όνομα βάσης	T	I	D	Μέγεθος σε MB
T10.I2.D100K	10	2	100 K	4.4
T10.I4.D100K	10	4	100 K	
T20.I2.D100K	20	2	100 K	8.4
T20.I4.D100K	20	4	100 K	
T20.I6.D100K	20	6	100 K	

4.4.2 Χρόνοι Εκτέλεσης

Ο αλγόριθμος μας παρουσιάζει ελάχιστα υψηλότερους χρόνους εκτέλεσης από τον αλγόριθμο MSAPriori. Αυτό οφείλεται κύρια στο γεγονός ότι αντίθετα προς τον αλγόριθμο MSAPriori ο οποίος είχε όλα τα στοιχεία ταξινομημένα σύμφωνα με τις τιμές υποστήριξης τους προτού καν ξεκινήσει, η προσέγγιση μας δεν ταξινομεί τα στοιχεία έως ότου έχει καταμετρήσει τις

εμφανίσεις τους. Έτσι η συγκεκριμένη διαδικασία προσθέτει κάποιο χρόνο στον συνολικό χρόνο εκτέλεσης. Βέβαια αν παραλείψουμε αυτό το βήμα από τον υπολογισμό του συνολικού χρόνου εκτέλεσης προκειμένου να δώσουμε στον αλγόριθμο μας μια δικαιότερη σύγκριση, τότε ο συνολικός χρόνος θα ήταν σχεδόν ο ίδιος με αυτόν που χρειάζεται από άλλους αλγορίθμους. Ένας άλλος παράγοντας ο οποίος αυξάνει τον συνολικό χρόνο εκτέλεσης είναι οι επιπρόσθετες συγκρίσεις που πρέπει να γίνουν αφότου έχουν μετρηθεί όλα τα στοιχεία. Όλοι οι αλγόριθμοι μέχρι τώρα έλεγχαν εάν η υποστήριξη κάθε στοιχείου ήταν πάνω από το κατώφλι υποστήριξης, και ακολούθως το χαρακτήριζαν σαν συχνό ή σπάνιο. Στην περίπτωση μας βέβαια ακόμα και αν ένα στοιχείο έχει αριθμό εμφανίσεων κάτω από το κατώφλι υποστήριξης, πρέπει εντούτοις να ελέγξουμε το λόγο ενδιαφέροντος του. Βέβαια αφού αυτό συμβαίνει μόνο με τα στοιχεία, τα οποία συνήθως κυμαίνονται το πολύ σε μερικές χιλιάδες σε ένα τυπικό κατάστημα λιανεμπορίου, το τελικό κόστος είναι αρκετά μικρό. Τέλος, ο αλγόριθμος μας στην αναζήτηση του για καυτά προϊόντα παράγει ελαφρώς περισσότερα υποψήφια αλλά και τελικά συχνά αντικείμενα από άλλους αλγορίθμους. Οι σχετικοί χρόνοι των δύο αλγορίθμων φαίνονται στην Εικόνα 6.

Εικόνα 6: Σχετική απόδοση του MSAPriori σε σχέση με τον αλγόριθμο μας

4.5 Συμπεράσματα και Μελλοντικές κατευθύνσεις

Καταφέραμε και λύσαμε ένα ιδιαίτερα δύσκολο ανοιχτό πρόβλημα, ονομαστικά την αυτόματη εξεύρεση και διαχείριση των διαφαινόμενων τάσεων σε μια βάση. Έτσι δώσαμε μια διαφορετική διάσταση στην διαδικασία ανακάλυψης γνώσης, όπου συμβαδίζουμε με τις τάσεις

των διαφόρων προϊόντων παρά απλά να ακολουθούμε ουσιαστικά παρελθοντικές συμπεριφορές τους. Η τελική μας πρόταση όχι μόνο έλυσε όλα τα παραπάνω θέματα, αλλά ταυτόχρονα κατόρθωσε να παρουσιάζει ελάχιστα μεγαλύτερους χρόνους εκτέλεσης σε σύγκριση με μεθόδους που δεν κατάφεραν να αντιμετωπίσουν αυτά τα προβλήματα.

**Κεφάλαιο 5 ΔΗΜΙΟΥΡΓΙΑ ΕΝΟΣ
ΑΠΟΤΕΛΕΣΜΑΤΙΚΟΥ ΜΗΧΑΝΙΣΜΟΥ
ΑΠΟΔΟΣΗΣ ΒΑΡΩΝ ΓΙΑ ΤΗΝ ΕΝΣΩΜΑΤΩΣΗ
ΤΟΥ ΜΙΚΡΟΟΙΚΟΝΟΜΙΚΟΥ ΠΑΡΑΓΟΝΤΑ
ΣΤΟΥΣ ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΕΩΝ**

Περίληψη

Οι περισσότεροι αλγόριθμοι και προσεγγίσεις που ασχολούνται με την ανακάλυψη κανόνων συσχετίσεων υπέθεταν ότι όλα τα προϊόντα ήταν της ίδιας φύσεως και σημαντικότητας και χρησιμοποιούσαν ένα ενιαίο κατώφλι υποστήριξης. Ορισμένοι προσπάθησαν να αντιμετωπίσουν την ανομοιογένεια και την ανομοιομορφία τόσο των αντικειμένων όσο και του πλήθους των εμφανίσεων τους. Παρόλα' αυτά καμία προσέγγιση δεν πρότεινε κάποιο συγκεκριμένο τρόπο για να ανακαλύπτουμε και ακολούθως να αποδίδουμε μια κατάλληλη μονάδα μέτρησης της σημαντικότητας σε κάθε αντικείμενο, ούτε και ελάμβανε υπόψη το πλαίσιο μέσα από το οποίο μια διαδικασία ανακάλυψης γνώσης θα πρέπει να εξετάζεται και να υλοποιείται. Στο συγκεκριμένο κεφάλαιο εξετάζουμε την περίπτωση εφαρμογής τεχνικών ανακάλυψης γνώσης κύρια σε επιχειρήσεις λιανεμπορίου και εξετάζουμε τα αντικείμενα της βάσης καθώς και τους παραγόμενους κανόνες συσχετίσεων μέσα από το κατάλληλο μικροοικονομικό πλαίσιο. Έτσι προτείνουμε ένα μηχανισμό απόδοσης βαρών ο οποίος δίνει τις κατάλληλες τιμές υποστήριξης σε όλα τα αντικείμενα, αυτόματα βρίσκει τα πιο σημαντικά και τελικά αποδεικνύεται αποτελεσματικός.

Λέξεις Κλειδιά: Απόδοση Βαρών, Συχνά Αντικείμενα, Περιθώριο Ενδιαφέροντος, Μικροοικονομικός παράγων.

5.1 Εισαγωγή

Οι περισσότερες εργασίες μέχρι στιγμής υπέθεταν ότι όλα τα προϊόντα ήταν της ίδιας σημαντικότητας και βασίζονταν στο μοντέλο που έθετε ένα ενιαίο - μοναδικό κατώφλι υποστήριξης για όλα τα αντικείμενα σε μια βάση. Σε εφαρμογές όμως υπό πραγματικές συνθήκες το μέτρο αυτό αποδεικνύεται ανεπαρκές, αφού η συχνότητα των εμφανίσεων των αντικειμένων σε μια βάση καθώς και τα αντικείμενα αυτά καθ' αυτά κάθε άλλο παρά ομοιόμορφα είναι. Για παράδειγμα υπάρχουν πολλές περιπτώσεις όπου τα σπάνια αντικείμενα είναι πιο ενδιαφέροντα από τα συχνά, ή η φύση και τα χαρακτηριστικά των αντικειμένων είναι πολύ ανομοιόμορφα ώστε να έχουν μια ενιαία αντιμετώπιση, όπως π.χ. loss leader ** προϊόντα, πολύ επικερδή προϊόντα, προϊόντα τα οποία βρίσκονται στη διαδικασία της έρευνας αγοράς κ.α. [77]. Διάφορες προσεγγίσεις και αλγόριθμοι προτάθηκαν προκειμένου να αντιμετωπίσουν το συγκεκριμένο πρόβλημα ([56][73][88][115]) με καλύτερη λύση κατά τη γνώμη μας την εργασία των Liu, Hsu και Ma στο [89]. Η απόδοση αυτής της εργασίας βελτιώθηκε στο [79], αλλά η βασική ιδέα παρέμεινε η ίδια. Παρόλα τα πλεονεκτήματα της, η προσέγγιση των Liu, Hsu και Ma υποφέρει επίσης από σοβαρά μειονεκτήματα όπως οι τιμές υποστήριξης που αποδίδονται σε όλα τα αντικείμενα ή η διαδικασία για την εύρεση και απόδοση αυτών των τιμών. Αυτό που είναι όμως ακόμα πιο σημαντικό είναι ότι τόσο αυτή όσο και όλες οι άλλες προσεγγίσεις δεν καθόριζαν ποια αντικείμενα θα πρέπει να θεωρούνται πιο σημαντικά ούτε και πρότειναν κάποιο τρόπο ώστε να τα ανακαλύπτουμε αυτόματα.

Στην παρούσα εργασία προτείνουμε και εξετάζουμε ένα μηχανισμό απόδοσης βαρών που αποδίδει την κατάλληλη τιμή υποστήριξης σε κάθε αντικείμενο στη βάση μας αυτόματα. Η όλη διαδικασία ακολουθεί μια αυτοκαθοριζόμενη λογική, όπου ο χρήστης (data miner) δεν πρέπει να ορίσει κάτι προκαταβολικά αλλά αφήνει τα αντικείμενα να λάβουν τις δικές τους τιμές υποστήριξης σύμφωνα με μια σειρά παραγόντων όπως οι εμφανίσεις τους στη βάση καθώς και σύμφωνα με τις εμφανίσεις όλων των άλλων αντικειμένων. Η τελική προσέγγιση αποδεικνύεται πιο αποτελεσματική ως προς τις προτάσεις που παράγει.

** Προϊόντα τιμολογημένα ακόμα και κάτω του κόστους προκειμένου να προσελκύσουν πελάτες σε μια επιχείρηση, ακόμα και αν οι πελάτες τελικά δεν αγοράζουν το loss leader προϊόν αλλά άλλα προϊόντα.

5.2 Λαμβάνοντας υπόψη την μικροοικονομική διάσταση των κανόνων συσχετίσεων

5.2.1 Διατύπωση του προβλήματος

Οι αλγόριθμοι εύρεσης κανόνων συσχετίσεων επεξεργάζονται τα δεδομένα με ένα σταθερό κατώφλι υποστήριξης για όλα τα αντικείμενα (ή με διαφορετικές αλλά πάλι σταθερές τιμές MIS για κάθε αντικείμενο στην περίπτωση που χρησιμοποιούμε τον αλγόριθμο MSApriori). Ουσιαστικά λειτουργούν με την υπόθεση ότι η σωστή τιμή του κατωφλίου υποστήριξης είναι ήδη γνωστή; στην πράξη όμως η τιμή αυτή δεν είναι γνωστή και θα μπορούσε μόνο να εκτιμηθεί βασιζόμενοι σε γνώση της περιοχής ή σε προηγούμενη εμπειρία. Επίσης όταν μιλάμε για την “σωστή” τιμή, αυτό κρίνεται εκ των υστέρων με βάση τον αριθμό των συχνών αντικειμένων τα οποία ανακαλύψαμε (δηλ. πολύ λίγα ή πάρα πολλά σε σχέση με το αναμενόμενο), ή με άλλα λόγια μόνο μέσω εντελώς υποκειμενικών και σχετικά υπεραπλουστευμένων κριτηρίων. Ακολουθώντας αν το κατώφλι υποστήριξης αλλάξει, τότε η διαδικασία της ανακάλυψης γνώσης θα πρέπει να επαναληφθεί από την αρχή απαιτώντας τουλάχιστον τον ίδιο αριθμό περασμάτων από την βάση όπως και το προηγούμενο τρέξιμο (αν το νέο κατώφλι είναι χαμηλότερο από το προηγούμενο), και πιο σημαντικό χωρίς να είναι σε θέση να χρησιμοποιήσει γνώση από προηγούμενα τρεξίματα του ίδιου αλγορίθμου. Ειδικότερα δε στην περίπτωση που πρέπει να αποδώσουμε σε κάθε αντικείμενο μια ξεχωριστή τιμή υποστήριξης κατωφλίου, τότε η διαχείριση όλων αυτών των αντικειμένων αποτελεί ένα εξαιρετικά πολύπλοκο πρόβλημα. Άρα πρόκειται για μια επαναληπτική διαδικασία, όπου οι τιμές του κατωφλίου υποστήριξης απαιτούν μικρό-ρυθμίσεις (fine-tuning), οι οποίες μπορούν να επιτευχθούν μόνο μέσω πολλαπλών ολοκληρωμένων τρεξιμάτων ενός αλγορίθμου εύρεσης κανόνων συσχετίσεων.

Όμως διαπερνώντας ολόκληρη την βάση μας πολλές φορές και ψάχνοντας και καταμετρώντας όλους τους δυνατούς έγκυρους συνδυασμούς αντικειμένων μπορεί να μην είναι τόσο πρακτικό όσο αρχικά φαντάζει, ειδικότερα σε εφαρμογές σε πραγματικό περιβάλλον ή για δεδομένα άλλα από αυτά που χρησιμοποιούνται για κλασσική ανάλυση συναλλαγών (π.χ. ιδιαίτερα πεπλεγμένα δεδομένα όπως δεδομένα από απογραφές πληθυσμού ή δεδομένα με πολλαπλά μη δυαδικά χαρακτηριστικά). Για παράδειγμα η βάση Cover Type από την συλλογή του UCI^{††} με μόλις 120 μονοσύνολα, αλλά με τα περισσότερα από αυτά να εμφανίζονται σχεδόν συνέχεια έχει σαν αποτέλεσμα την δημιουργία 15 εκατομμυρίων συχνών αντικειμένων. Χρησιμοποιώντας τον αλγόριθμο Apriori μόνο για να βρούμε τα συχνά αντικείμενα απαιτεί

^{††} <http://kdd.ics.uci.edu/>

παραπάνω από 96 ώρες χρήσης CPU [125]. Ή σε μεγάλες επιχειρήσεις όπως η Wal-Mart, η UPS ή η Amazon με βάσεις δεδομένων της τάξεως των Terabytes, με εκατοντάδες χιλιάδες διαφορετικών αντικειμένων και με αρκετές δεκάδες αντικειμένων σε κάθε συναλλαγή, τότε ακόμα και ένα πέρασμα από ολόκληρη τη βάση απαιτεί τεράστιους πόρους. Έτσι ολόκληρη η διαδικασία απόδοσης τιμών σε όλα τα αντικείμενα σε μια βάση θα πρέπει να είναι όσο το δυνατόν ακριβέστερη προκειμένου να αποφευχθούν άσκοπα τρεξίματα του ίδιου αλγορίθμου πάνω από τα δεδομένα.

Αυτό που είναι όμως ακόμα πιο σημαντικό είναι το γεγονός ότι οι υπάρχουσες τεχνικές για την εύρεση της υποστήριξης ενός αντικειμένου δεν λαμβάνουν υπόψη την εμπορική αξία των αντικειμένων και των συσχετίσεων μεταξύ τους [31], και σαν αποτέλεσμα η υποστήριξη που τους αποδίδεται δεν αντανακλά την πραγματική αξία των αντικειμένων. Αυτό έχει ως αποτέλεσμα την ομογενοποίηση όλων των αντικειμένων σε μια οποιαδήποτε συλλογή αντικειμένων, ή με άλλα λόγια η πώληση ενός κουτιού πολύ ακριβού χαβιαριού αντιμετωπίζεται λίγο πολύ σαν την πώληση ενός πακέτου με τσίχλες. Σύμφωνα με τους Kleinberg, Papadimitriou και Raghavan στο [74], “ένα μοτίβο στα δεδομένα είναι ενδιαφέρον μόνο στο βαθμό που μπορεί να χρησιμοποιηθεί στην διαδικασία της λήψης αποφάσεων της επιχείρησης προκειμένου να αυξήσει τα οφέλη της”. Έτσι η χρησιμότητα των εξαγομένων μοτίβων (όπως κανόνες συσχετίσεων και αλληλεξαρτήσεις) στην διαδικασία της λήψης αποφάσεων μιας επιχείρησης μπορεί να αντιμετωπιστεί μόνο μέσω του μικροοικονομικού πλαισίου του περιβάλλοντος στο οποίο εργαζόμαστε. Αλλά προτού ανακαλύψουμε τα μοτίβα, η ίδια φροντίδα θα πρέπει να ληφθεί επίσης για τα αντικείμενα τα οποία και στην ουσία αποτελούν το βασικό συστατικό των μοτίβων (δηλ. η εμπορική αξία των αντικειμένων θα πρέπει επίσης να ληφθεί υπόψη όταν βρίσκουμε τα συχνά αντικείμενα). Ο τρόπος με τον οποίο οι υπάρχοντες αλγόριθμοι αντιμετωπίζουν τα αντικείμενα σαν απλές στατιστικές πιθανότητες και σαν δυαδικές μεταβλητές είναι μάλλον μια υπεραπλούστευση ενός ιδιαίτερα πολύπλοκου προβλήματος, το οποίο χρειάζεται να αντιμετωπιστεί διαφορετικά.

Με την δουλειά μας αυτή λαμβάνουμε υπόψη όλες τις απαιτήσεις που προαναφέραμε και τις αντιμετωπίζουμε μέσω ενός απλού (αλλά όχι αφελούς) γενικού πλαισίου. Ακολουθώντας προτείνουμε μια εναλλακτική πρόταση για να βρίσκουμε τα συχνά αντικείμενα σε μια βάση, της οποίας το κύριο χαρακτηριστικό είναι ένας αποτελεσματικός μηχανισμός απόδοσης βαρών.

5.2.2 Η προσέγγιση μας

Όπως αναφέρθηκε παραπάνω η προσέγγιση μας εργάζεται με την ίδια λογική όπως οι παραδοσιακοί αλγόριθμοι εύρεσης κανόνων συσχετίσεων, δηλ. βρίσκοντας όλα τα συχνά αντικείμενα χρησιμοποιώντας πολλαπλές τιμές κατωφλιών υποστήριξης (multiple minimum

support values) όπως ο αλγόριθμος MSAPriori και κατόπιν παράγοντας τους αντίστοιχους κανόνες συσχετίσεων. Ο ακρογωνιαίος λίθος της προσέγγισης μας είναι ο μηχανισμός απόδοσης βαρών που χρησιμοποιείται αφού αυτό του επιτρέπει να λάβει υπόψη τις διαφορές ανάμεσα στα αντικείμενα όπως επίσης και τις ιδιαίτερες συνθήκες του περιβάλλοντος στο οποίο εργαζόμαστε. Στο παρόν κεφάλαιο θα ασχοληθούμε μόνο με μια εφαρμογή. Πιο συγκεκριμένα θα προτείνουμε έναν μηχανισμό απόδοσης βαρών ειδικά σχεδιασμένο για την περίπτωση ενός καταστήματος λιανεμπορίου και έτσι θα επικεντρωθούμε εξολοκλήρου στην μικροοικονομική πλευρά των αντικειμένων. Παρόλα' αυτά η φιλοσοφία και η λογική της προσέγγισης μας μπορεί να χρησιμοποιηθεί με ελάχιστες τροποποιήσεις σε όλα τα είδη επιχειρήσεων και εφαρμογών.

Έτσι όταν αναπτύσσαμε το μηχανισμό αυτό, το πρώτο μας μέλημα ήταν να προσδιορίσουμε και να επιλέξουμε τους παράγοντες και τις αντίστοιχες παραμέτρους τους, οι οποίοι είναι και οι πιο σημαντικοί και παρουσιάζουν την μεγαλύτερη επιρροή στο να καθορίσουμε την σπουδαιότητα ενός αντικειμένου. Έτσι καταλήξαμε στους παρακάτω τρεις παράγοντες:

1. Ποια είναι τα καθαρά κέρδη ενός αντικειμένου συγκρινόμενα με τα μέσα καθαρά κέρδη όλων των αντικειμένων στη βάση μας?
2. Ποιες είναι οι συνολικές πωλήσεις ενός αντικειμένου σε σύγκριση με τις μέσες ολικές πωλήσεις όλων των αντικειμένων?
3. Ποιος είναι ο αριθμός των συναλλαγών στις οποίες εμφανίζεται ένα αντικείμενο σε σχέση με τον μέσο αριθμό συναλλαγών που εμφανίζονται όλα τα αντικείμενα.

Στον υπολογισμό κάθε παράγοντα χρησιμοποιήσαμε τις μέσες τιμές και όχι τις ελάχιστες, επειδή σε τέτοιες βάσεις οι τιμές αυτών των παραγόντων παρουσιάζουν πολύ μεγάλες διακυμάνσεις και πιθανώς να παίρναμε τελικά ένα αντικρουόμενο μηχανισμό^{††}. Ο πρώτος παράγοντας είναι αρκετά προφανής αφού ο απλούστερος και πιο ακριβής τρόπος υπολογισμού της σημαντικότητας ενός αντικειμένου σε ένα κατάστημα λιανεμπορίου είναι τα καθαρά κέρδη που παρουσιάζει. Φροντίδα θα πρέπει να ληφθεί με αυτόν τον παράγοντα ώστε να μην χρησιμοποιήσουμε την τιμή πώλησης, αφού αυτή δεν αντιπροσωπεύει πάντα την αληθινή του εμπορική αξία (δηλ. ένα ακριβό προϊόν μπορεί να παρουσιάζει πολύ μικρά καθαρά κέρδη για μια εταιρεία και το αντίστροφο). Ο δεύτερος παράγοντας ουσιαστικά συνδέει τα καθαρά κέρδη ενός προϊόντος με την παρουσία του μέσα στη βάση. Πιο συγκεκριμένα ένα εξόχως επικερδές προϊόν είναι σημαντικό μόνο αν έχει επίσης υψηλές πωλήσεις και ένα όχι τόσο

^{††} Αν κάποιος λάβει υπόψη mottos όπως αυτό του πολυκαταστήματος Harrods (<http://www.harrods.com>), “from a pin to an elephant” γίνεται προφανές γιατί προτιμούμε τις μέσες από τις ελάχιστες τιμές.

επικερδές προϊόν μπορεί να είναι ενδιαφέρον με την προϋπόθεση ότι έχει υψηλές πωλήσεις. Τέλος είναι ο τρίτος παράγοντας που έχει να κάνει ακριβώς με την παρουσία ενός αντικειμένου σε μια βάση. Το γεγονός ότι ένα προϊόν εμφανίζεται πολλές φορές στη βάση είναι πραγματικά σημαντικό αλλά είναι ακόμα πιο σημαντικό αν το δούμε από την πλευρά του αριθμού των συναλλαγών στις οποίες εμφανίζεται. Με άλλα λόγια σε όσες περισσότερες συναλλαγές εμφανίζεται ένα προϊόν τόσο πιο σημαντικό θα πρέπει να θεωρείται, αφού στατιστικά παρουσιάζει περισσότερες πιθανότητες να αγοραστεί ξανά. Προκειμένου να ποσοτικοποιήσουμε τους τρεις αυτούς παράγοντες, ορισμένες παράμετροι θα πρέπει να είναι γνωστές. Πιο συγκεκριμένα πρέπει να έχουμε τις τιμές των παρακάτω παραμέτρων:

1. Σε πόσες συναλλαγές εμφανίζεται ένα προϊόν, το οποίο και συμβολίζουμε με T_i .
2. Ποιος είναι ο μέσος αριθμός συναλλαγών όπου εμφανίζονται όλα τα προϊόντα, το οποίο συμβολίζουμε με $\bar{T}_{total}$.
3. Πόσες φορές εμφανίζεται συνολικά ένα προϊόν, το οποίο συμβολίζουμε με S_i .
4. Ποιος είναι ο μέσος αριθμός εμφανίσεων όλων των προϊόντων, το οποίο συμβολίζουμε με $\bar{S}_{total}$.
5. Ποιο είναι το καθαρό κέρδος ανά προϊόν, το οποίο συμβολίζουμε με N_i .
6. Ποιο είναι το μέσο καθαρό κέρδος όλων των προϊόντων, το οποίο συμβολίζουμε με $\bar{N}_{total}$.
7. Πόσες συναλλαγές υπάρχουν συνολικά, το οποίο το συμβολίζουμε με T_{total} .

Το επόμενο βήμα ήταν να συνδυάσουμε όλους αυτούς τους παράγοντες μαζί με τις αντίστοιχες παραμέτρους σε μια εξίσωση η οποία θα μας έδινε το βάρος του κάθε αντικειμένου. Η λογική μας για να βρούμε αυτή την εξίσωση ήταν απλή. Κάθε προϊόν μπορεί να έχει υποστήριξη το πολύ ίση με: $\maxsup(i) = T_i / \bar{T}_{total}$. Όσο πιο κοντά θέτουμε την υποστήριξη ενός προϊόντος στην μέγιστη υποστήριξη του, τόσο περισσότερες εμφανίσεις χρειάζονται προκειμένου να χαρακτηριστεί σαν συχνό και άρα τόσο πιο ασήμαντο θεωρείται. Στην πράξη αν θέσουμε την υποστήριξη ενός προϊόντος ίση με $\maxsup(i)$ είναι σαν να το αποκλείουμε αφού θα χρειάζεται πλέον να εμφανίζεται σε ένα ποσοστό 100% με οποιοδήποτε άλλο προϊόν (το οποίο είναι πρακτικά πάρα πολύ δύσκολο). Όσο χαμηλότερα θέτουμε την υποστήριξη του σε σχέση με την μέγιστη υποστήριξη $\maxsup(i)$, τόσο λιγότερες εμφανίσεις χρειάζονται προκειμένου να θεωρηθεί συχνό και έτσι τόσο πιο σημαντικό το θεωρούμε. Έτσι αφού πειραματιστήκαμε με αρκετές διαφορετικές μορφές της ζητούμενης εξίσωσης

καταλήξαμε στην παρακάτω μορφή η οποία προσεγγίζει με τον καλύτερο τρόπο την ζητούμενη συμπεριφορά που μόλις περιγράψαμε:

$$wf = \log_A \left(1 + \frac{\bar{T}_{total}}{T_i} + \frac{\bar{S}_{total}}{S_i} + \frac{\bar{N}_{total}}{N_i} \right)$$

Εξίσωση 2

$$\text{όπου} \quad A = \max \left(1 + \frac{\bar{T}_{total}}{T_i} + \frac{\bar{S}_{total}}{S_i} + \frac{\bar{N}_{total}}{N_i} \right)$$

Ο λογάριθμος στην συγκεκριμένη εξίσωση χρησιμοποιήθηκε προκειμένου να αποφύγουμε το φαινόμενο όπου ένα προϊόν με κάποιο παράγοντα μισό από αυτόν κάποιου άλλου να θεωρηθεί δύο φορές πιο σημαντικό (π.χ. ένα προϊόν που έχει τα διπλάσια καθαρά κέρδη από ένα άλλο να θεωρείται δύο φορές πιο σημαντικό και να λαμβάνει το μισό βάρος). Το κυριότερο πλεονέκτημα της συγκεκριμένης εξίσωσης είναι ότι λαμβάνει υπόψη και τους τρεις παράγοντες, και ένα οποιοδήποτε προϊόν θα πρέπει να έχει και τους τρεις αυτούς παράγοντες μικρούς προκειμένου να θεωρηθεί πολύ σημαντικό. Δεν αρκεί δηλαδή ένα προϊόν να εμφανίζεται σε πολλές συναλλαγές, ή να εμφανίζεται πολλές φορές συνολικά, ή να έχει μεγάλα καθαρά κέρδη προκειμένου να πάρει μια αρκετά χαμηλή τιμή υποστήριξης. Είναι ο συνδυασμός και των τριών αυτών παραγόντων που το κάνουν ακόμα πιο σημαντικό. Επίσης στην Εξίσωση 2 κάποιος είναι ελεύθερος να επιλέξει έναν διαφορετικό βαθμό βαρύτητας σε όποιον παράγοντα αποφασίσει. Για παράδειγμα αν μια επιχείρηση ενδιαφέρεται περισσότερο να αυξήσει το μερίδιο αγοράς της (δηλ. τις πωλήσεις τις) παρά τα καθαρά κέρδη της θα μπορούσε να αυξήσει την επίδραση του παράγοντα $\bar{S}_{total}/S_i$ πολλαπλασιάζοντας τον με μια σταθερή θετική ποσότητα κάτω του 1. Επίσης η συγκεκριμένη εξίσωση επιτρέπει την εισαγωγή οποιοδήποτε νέου παράγοντα ο χρήστης θεωρεί σημαντικό. Αφού θα έχουμε βρει το βάρος του κάθε προϊόντος, το τελικό βήμα είναι η απόδοση μιας τιμής υποστήριξης σύμφωνα με τον ακόλουθο τύπο: $\sup(i) = wf * \max \sup(i)$.

5.2.2.1 Ένα παράδειγμα απόδοσης βαρών με την προσέγγισή μας

Ας υποθέσουμε ότι έχουμε μια βάση η οποία περιέχει τις συναλλαγές που εμφανίζονται στον Πίνακα 5, η οποία αποτελείται από 9 προϊόντα και 10 συναλλαγές (parameter T_{total}). Στον Πίνακα 6 βλέπουμε για κάθε προϊόν i τον συνολικό αριθμό των συναλλαγών στις οποίες εμφανίζεται (παράμετρος T_i), τις φορές που εμφανίζεται συνολικά σε όλες τις συναλλαγές (παράμετρος S_i) και το καθαρό κέρδος ανά προϊόν (παράμετρος N_i). Κάθε αντικείμενο στην συγκεκριμένη βάση επιλέχτηκε έτσι ώστε να είναι διαφορετικό από τα άλλα καθώς και προκειμένου να ληφθούν υπόψη όλες οι δυνατές περιπτώσεις. Θέλαμε δηλαδή να έχουμε όλους τους δυνατούς συνδυασμούς των προαναφερθέντων παραγόντων (δηλ. τα καθαρά κέρδη, το

πλήθος των συναλλαγών και τον αριθμό των εμφανίσεων) που χαρακτηρίζουν ένα αντικείμενο. Έτσι με βάση τα χαρακτηριστικά των αντικειμένων μπορούμε να τα ταξινομήσουμε σε τρεις διαφορετικές κατηγορίες. Η πρώτη περιλαμβάνει αντικείμενα τα οποία έχουν ένα μόνο παράγοντα μικρό και τους άλλους δύο μεγάλους (αναφορικά τα αντικείμενα A, B και C). Πιο συγκεκριμένα το αντικείμενο A έχει υψηλά καθαρά κέρδη, το αντικείμενο B εμφανίζεται σε πολλές συναλλαγές ενώ το αντικείμενο C έχει συνολικά πολλές εμφανίσεις. Η δεύτερη κατηγορία περιλαμβάνει τα αντικείμενα αυτά που έχουν και τους τρεις παράγοντες όμοιους. Έτσι το αντικείμενο D έχει και τους τρεις παράγοντες μικρούς, το αντικείμενο E έχει και τους τρεις μεγάλους ενώ το F έχει και τους τρεις σχετικά μέτριους. Η τελευταία κατηγορία αντικειμένων περιλαμβάνει τα αντικείμενα αυτά τα οποία έχουν δύο παράγοντες μικρούς και τον τρίτο μεγάλο. Έτσι το αντικείμενο G εμφανίζεται σε πολλές συναλλαγές όπως επίσης και πολλές φορές συνολικά, το αντικείμενο H εμφανίζεται επίσης σε πολλές συναλλαγές αλλά εμφανίζει υψηλά καθαρά κέρδη και τέλος το αντικείμενο I έχει υψηλά καθαρά κέρδη και επίσης εμφανίζεται πολλές φορές.

Πίνακας 5: Μια βάση συναλλαγών

Αριθμός Συναλλαγής	Αντικείμενο	Αριθμός Συναλλαγής	Αντικείμενο
1	A B C G H I	6	A B D G
2	B C D G H	7	C D F G H
3	B D E F G H	8	B D H I
4	B D F G H	9	B D E F G
5	B D F G H	10	B D G H

Πίνακας 6: Τελικά Αποδιδόμενα Βάρη σε κάθε προϊόν

Αντικείμενο	Συνολικός αριθμός εμφανίσεων Si	Αριθμός συναλλαγών Ti	Καθαρά κέρδη ανά μονάδα Ni	Μέγιστη Υποστήριξη	MIS(i)	Τελικό Βάρος
A	2	2	100	20%	20,00%	10,37%
B	2	9	10	90%	54,00%	61,28%
C	30	3	9	30%	20,00%	11,55%
D	30	9	90	90%	54,00%	2,95%
E	2	2	10	20%	20,00%	20,00%
F	17	5	50	50%	30,00%	7,15%
G	30	9	11	90%	54,00%	16,56%
H	8	8	55	80%	48,00%	13,09%
I	30	2	55	20%	20,00%	3,42%

Με αυτά τα δεδομένα σαν είσοδο αποδίδουμε τιμές υποστήριξης σε κάθε αντικείμενο με τον αλγόριθμο MSAPriori και την προσέγγιση μας. Όπως σημειώσαμε νωρίτερα ο αλγόριθμος MSAPriori δεν λαμβάνει υπόψη τα καθαρά κέρδη των αντικειμένων (ή οποιαδήποτε άλλη παράμετρο που τα διαφοροποιεί) όπως επίσης και τις φορές που εμφανίζεται συνολικά στη

βάση μας προκειμένου να αποδώσει μια τιμή υποστήριξης σε κάθε αντικείμενο, αλλά αντιμετωπίζει όλα τα αντικείμενα σαν όμοιες δυαδικές μεταβλητές. Ο αλγόριθμος μας από την άλλη λαμβάνει υπόψη τόσο τα καθαρά κέρδη κάθε αντικειμένου όπως επίσης και τις φορές που εμφανίζεται κάθε αντικείμενο συνολικά. Οι δύο τελευταίες στήλες του Πίνακα 6 παρουσιάζουν την τιμή υποστήριξης που αποδίδεται σε κάθε αντικείμενο από τον αλγόριθμο μας και από τον αλγόριθμο MSAPriori. Για τον αλγόριθμο MSAPriori θέτουμε την τιμή της ελάχιστης επιτρεπόμενης τιμής υποστήριξης – LS ίση με 0.2 (με άλλα λόγια πρέπει να εμφανίζεται σε τουλάχιστον 2 συναλλαγές από τις 10), και την τιμή της παραμέτρου β ίση με 0.6. Όπως μπορούμε να δούμε και από την Εικόνα 7 οι δύο αλγόριθμοι δεν χρησιμοποιούν την ίδια τιμή υποστήριξης παρά μόνο για ένα αντικείμενο στη βάση μας (το αντικείμενο E) και επίσης μπορούμε να παρατηρήσουμε ότι για τα περισσότερα αντικείμενα οι αποδιδόμενες τιμές υποστήριξης παρουσιάζουν μεγάλες αποκλίσεις.

Οι τιμές υποστήριξης που αποδόθηκαν από την προσέγγισή μας για την πρώτη κατηγορία αντικειμένων κυμαίνονται από 0.385 έως 0.68 της μέγιστης υποστήριξης. Έτσι παρατηρούμε ότι η προσέγγισή μας λαμβάνει υπόψη ότι τα συγκεκριμένα αντικείμενα έχουν ένα παράγοντα μικρό, αλλά προκειμένου να θεωρηθούν πραγματικά σημαντικά και να λάβουν μια πολύ χαμηλή τιμή υποστήριξης τουλάχιστον δύο παράγοντες χρειάζεται να είναι μικροί όπως άλλωστε θα δούμε και στην τελευταία κατηγορία παρακάτω. Για την δεύτερη κατηγορία αντικειμένων οι αποδιδόμενες τιμές κυμαίνονται αρκετά λόγω των τιμών των αντίστοιχων παραγόντων. Το αντικείμενο D το οποίο έχει και τους τρεις παράγοντες μικρούς λαμβάνει μια τιμή υποστήριξης που αντιστοιχεί σε μόλις το 0.033 της μέγιστης τιμής υποστήριξης ενώ το αντικείμενο E από την άλλη που έχει και τους τρεις παράγοντες μεγάλους λαμβάνει σαν τιμή υποστήριξης την μέγιστη τιμή υποστήριξης, αφού είναι σχετικά ασήμαντο. Τέλος το αντικείμενο F έχει και τους τρεις παράγοντες κοντά στις μέσες τιμές, και έτσι η τιμή υποστήριξης που του αποδίδεται είναι σχετικά χαμηλή. Όσον αφορά την τελευταία κατηγορία αντικειμένων τώρα βλέπουμε ότι οι τιμές υποστήριξης που αποδίδονται στα αντικείμενα είναι ακόμα μικρότερες από αυτές που αποδόθηκαν στην πρώτη κατηγορία. Αυτό είναι απόλυτα δικαιολογημένο και ουσιαστικά αναμενόμενο με βάση το γεγονός ότι η προσέγγισή μας επηρεάζεται περισσότερο αν έχει πολλούς παράγοντες με χαμηλές τιμές, και όσο πιο χαμηλές είναι αυτές οι τιμές. Έτσι η συγκεκριμένη κατηγορία όπου τα αντικείμενα έχουν δύο παράγοντες μικρούς σε σύγκριση με την πρώτη όπου τα αντικείμενα είχαν μόνο ένα παράγοντα μικρό, έχουν τιμές υποστήριξης που κυμαίνονται από 0.163 έως 0.184 της μέγιστης τιμής υποστήριξης.

Αν θέλουμε να συγκρίνουμε τον MSAPriori με την προσέγγισή μας θα δούμε ότι σε όλες τις περιπτώσεις ο MSAPriori αγνοεί την πραγματική σημασία των αντικειμένων και αποδίδει αυθαίρετες τιμές υποστήριξης. Ειδικότερα για το αντικείμενο D όπου όπως σημειώσαμε

Εικόνα 7: Οι εκχωρούμενες τιμές υποστήριξης της προσέγγισης μας και του αλγορίθμου MSApriori σε σύγκριση με την μέγιστη τιμή υποστήριξης

νωρίτερα παρουσιάζει πολύ υψηλά κέρδη, εμφανίζεται σχεδόν σε όλες τις συναλλαγές (9 από τις 10) και επίσης εμφανίζεται πολλές φορές συνολικά (30 φορές ή 3 φορές ανά συναλλαγή κατά μέσο όρο) ο MSApriori του αποδίδει μια τιμή υποστήριξης 54% που είναι υπερβολικά υψηλή σε σύγκριση με την δική μας τιμή που είναι μόλις 2.95%. Μόνο στο αντικείμενο B παρατηρούμε την αντίθετη αντιμετώπιση από τις δύο προσεγγίσεις. Πιο συγκεκριμένα η προσέγγιση μας αποδίδει μια υψηλότερη τιμή υποστήριξης από αυτή του MSApriori γιατί ναι μεν το συγκεκριμένο αντικείμενο είναι κάπως σημαντικό (έχει έναν παράγοντα μικρό) αλλά η μέγιστη τιμή υποστήριξης του είναι πολύ μεγάλη (90%) που έχει ως αποτέλεσμα μια υψηλή τιμή υποστήριξης από την προσέγγιση μας. Και οι δύο προσεγγίσεις συμφωνούν στην τιμή υποστήριξης του αντικειμένου E. Παρόλα' αυτά αυτό είναι μάλλον απλά ένα τυχαίο γεγονός αφού ο αλγόριθμος MSApriori αποδίδει την χαμηλότερη δυνατή τιμή υποστήριξης (η οποία είναι ίδια με την μέγιστη τιμή υποστήριξης του συγκεκριμένου αντικειμένου) επειδή έχει τελική υποστήριξη (η τιμή του $f(i)*\beta$) ^{§§} κάτω από το επιτρεπόμενο από το χρήστη ελάχιστο – LS, ενώ η προσέγγιση μας του αποδίδει την μέγιστη τιμή επειδή έχει και τους τρεις παράγοντες μεγάλους και έτσι είναι σχετικά ασήμαντο.

^{§§} Βλέπε ενότητα 2.7.

Εικόνα 8: Οι τιμές υποστήριξης που αποδίδονται από την προσέγγιση μας και τον αλγόριθμο MSAPriori στο αντικείμενο B για κυμαινόμενες τιμές της μεταβλητής β και του LS

Στην Εικόνα 8 τέλος απομονώσαμε ένα συγκεκριμένο αντικείμενο από το σύνολο των δεδομένων μας (το αντικείμενο B) και παρουσιάζουμε την τιμή υποστήριξης που αποδίδεται από την προσέγγιση μας σε σύγκριση με τις τιμές υποστήριξης που αποδίδονται από τον αλγόριθμο MSAPriori μεταβάλλοντας τις τιμές των παραμέτρων που χρησιμοποιεί. Πιο συγκεκριμένα η πρώτη στήλη παρουσιάζει τις τιμές υποστήριξης που προκύπτουν διατηρώντας την παράμετρο β σταθερή (ίση με 0.6) και μεταβάλλοντας την τιμή του LS (από 0% έως 100% με βήμα 5% περίπου), και η δεύτερη τις τιμές υποστήριξης διατηρώντας το LS σταθερό (ίσο με 0.2) και μεταβάλλοντας την τιμή του β (από 0.0476 έως 1) λαμβάνοντας τελικά συνολικά 21 σημεία. Η τιμή υποστήριξης του αλγορίθμου μας ήταν σταθερή (ίση με 61.28%), και φαίνεται στην τρίτη στήλη. Με αυτό το πείραμα μπορούμε να δούμε καθαρά ότι προκειμένου να ανακαλύψουμε την σωστή τιμή υποστήριξης (οποιαδήποτε και αν είναι αυτή, ακόμα και αν κάποιος διαφωνεί με το ότι η δικιά μας τιμή είναι η σωστή ή η καλύτερη) με τον αλγόριθμο MSAPriori, κάποιος θα πρέπει να πραγματοποιήσει αρκετές δοκιμές με διαφορετικές τιμές για την παράμετρο β ή/και για την παράμετρο LS αν εργαζόμαστε με τον αλγόριθμο αυτό ή με την τιμή του κατωφλίου υποστήριξης αν εργαζόμαστε με οποιονδήποτε άλλο αλγόριθμο. Αν

κάποιος λάβει υπόψη ότι υπάρχουν αρκετά μονοσύνολα σε μια βάση και ότι κάθε φορά που ο αλγόριθμος MSAPriori ή οποιοσδήποτε άλλος αλγόριθμος χρησιμοποιείται με μια νέα τιμή του κατωφλίου υποστήριξης πρέπει να κάνει ξανά πολλαπλά περάσματα πάνω από τη βάση χωρίς ωστόσο να λαμβάνει υπόψη τυχόν γνώση από προηγούμενα περάσματα, τότε το κόστος είναι τεράστιο.

5.2.3 Περιθώριο Ενδιαφέροντος

Η προσέγγιση μας μέχρι τώρα θεωρεί όλα τα μονοσύνολα σημαντικά, και δεν αποκλείει κανένα από αυτά. Σημειώσαμε προηγουμένως ότι όσο πιο κοντά θέτουμε την υποστήριξη ενός αντικειμένου στην μέγιστη τιμή υποστήριξης, τόσο πιο ασήμαντο θεωρείται, ενώ αν τη θέσουμε στο 100% της μέγιστης τιμής είναι σαν να το αποκλείουμε από τα επόμενα περάσματα. Παρόλα αυτά τουλάχιστον για το δεύτερο πέρασμα ένας πολύ μεγάλος αριθμός υποψηφίων αντικειμένων θα πρέπει να μετρηθούν (ουσιαστικά όλα τα πιθανά δισύνολα αντικειμένων), παρόλο που στο τέλος αυτού του περάσματος μόνο πολύ λίγα θα αποδειχτούν συχνά. Αν ο αριθμός των μονοσύνολων είναι σχετικά χαμηλός τότε το να μην απαλείψουμε κανένα αντικείμενο δεν δημιουργεί κάποιο μεγάλο πρόβλημα. Αλλά αν ο αριθμός των μονοσύνολων είναι σχετικά υψηλός τότε αυτό οδηγεί σε ένα πολύ μεγάλο πλήθος υποψηφίων δυσυνόλων το οποίο μπορεί να δυσχεράνει εξαιρετικά την όλη διαδικασία^{***}. Στην πράξη μόνο τα αντικείμενα που έχουν χαμηλό βάρος είναι σημαντικά, και θα πρέπει να συμπεριληφθούν στα επόμενα βήματα της προσέγγισης μας. Έτσι προκειμένου να απαλείψουμε κάποια όχι τόσο σημαντικά αντικείμενα δημιουργήσαμε το περιθώριο ενδιαφέροντος. Το περιθώριο ενδιαφέροντος είναι μια μονάδα μέτρησης της σημαντικότητας των αντικειμένων; μια οριζόμενη από το χρήστη τιμή πάνω από την οποία οποιοδήποτε αντικείμενο είναι ασήμαντο. Για παράδειγμα αν θέσουμε το περιθώριο ενδιαφέροντος ίσο με 80%, τότε οποιοδήποτε αντικείμενο έχει τιμή για το w_f πάνω από 80% θεωρείται ασήμαντο και αφαιρείται. Η επιλογή της συγκεκριμένης τιμής εξαρτάται από την εφαρμογή και ο χρήστης (data miner) είναι ελεύθερος να την προσαρμόσει σύμφωνα με τις ανάγκες του. Επίσης μπορεί να ορίσει διαφορετικές τιμές για το περιθώριο ενδιαφέροντος για διαφορετικά αντικείμενα ή ακόμα και για ολόκληρες ομάδες αντικειμένων. Η διαφοροποίηση της συγκεκριμένης τιμής, το οποία ουσιαστικά είναι και το μεγαλύτερο πλεονέκτημα της από μετρικές όπως το κατώφλι υποστήριξης κ.α., είναι ότι υπολογίζεται σε σχέση με το βάρος του κάθε αντικειμένου παρά σαν μια καθολική τιμή η οποία παραβλέπει την ιδιαιτερότητα του κάθε αντικειμένου.

^{***} Σύμφωνα με τους Park, Chen και Yu στο [94] τα πρώτα σύνολα υποψηφίων αντικειμένων, με πιο σημαντικό αυτό των 2-αντικειμένων είναι καθοριστικής σημασίας για την απόδοση ενός αλγορίθμου εύρεσης κανόνων συσχετίσεων.

5.3 Πειραματικά Αποτελέσματα

Ο καλύτερος τρόπος τον οποίο μπορούσαμε να σκεφτούμε προκειμένου να συγκρίνουμε την απόδοση της προσέγγισής μας με όλες τις άλλες προσεγγίσεις σχετικά με την ποιότητα των αποτελεσμάτων της ήταν να ελέγξουμε τα καθαρά κέρδη τα οποία παρήγαγε. Όπως αναφέρθηκε προηγουμένα η χρησιμότητα των κανόνων συσχετίσεων (η οποιασδήποτε άλλης μορφής μοτίβων εξαχθούν) στην διαδικασία λήψεως αποφάσεων μιας επιχείρησης μπορεί να ελεγχθεί λαμβάνοντας υπόψη το μικροοικονομικό περιβάλλον της επιχείρησης ή της εφαρμογής που εργαζόμαστε. Έτσι στην περίπτωση μας δεν υπάρχει καθαρότερος τρόπος ελέγχου για την σημαντικότητα και ορθότητα των προτάσεων που γίνονται από την προσέγγιση μας από τα καθαρά κέρδη που τελικά καταφέρει να δημιουργήσει.

Ακολούθως αποφασίσαμε να δημιουργήσουμε ένα εικονικό κατάστημα, προσβάσιμο μέσω του παγκόσμιου ιστού και να αφήσουμε τους χρήστες να πραγματοποιήσουν κάποιες αγορές. Για το συγκεκριμένο παράδειγμα χρησιμοποιήσαμε ένα σύνολο χρηστών χωρισμένους σε τέσσερις ομάδες. Όλες οι ομάδες αποτελούνταν από ανθρώπους ηλικίας από 18 έως 25 ετών, φοιτητές στο ίδιο πανεπιστήμιο και στο ίδιο τμήμα. Όλοι τους ήταν αρκετά οικείοι με τους ηλεκτρονικούς υπολογιστές και τον παγκόσμιο ιστό, και γενικότερα φιλικόι ως προς τις νέες τεχνολογίες. Αποφασίσαμε να δημιουργήσουμε ένα φανταστικό κατάστημα το οποίο πουλούσε κινητά τηλέφωνα και τα αντίστοιχα αξεσουάρ τους. Τα προϊόντα, τα ονόματα τους και οι τιμές που χρησιμοποιήθηκαν ήταν όλα αληθινά. Ο λόγος που επιλέξαμε την συγκεκριμένη κατηγορία προϊόντων ήταν επειδή τα άτομα που αποτελούσαν το κοινό ελέγχου μας (test audience) είχαν κινητό τηλέφωνο ή/και είχαν επισκεφτεί ένα κατάστημα που τα πουλούσε σε ένα ποσοστό σχεδόν 100%.

Το πείραμα εκτελέστηκε ως εξής. Κατ' αρχήν αφήσαμε το πρώτο σύνολο ανθρώπων να πραγματοποιήσει αγορές έτσι ώστε να συγκεντρώσουμε κάποια δεδομένα. Οι συγκεκριμένες αγορές πραγματοποιήθηκαν χωρίς να κάνουμε καμία πρόταση προϊόντος στους χρηστές. Κατόπιν αφήσαμε τα τρία υπόλοιπα σετ χρηστών να κάνουν κάποιες αγορές χρησιμοποιώντας το κάθε ένα ένα διαφορετικό σύστημα. Πιο συγκεκριμένα το πρώτο σύνολο λάμβανε προτάσεις με βάση τον αλγόριθμο Apriori, το δεύτερο με βάση τον MSApriori και το τρίτο με την προσέγγισή μας. Οι χρήστες δεν γνώριζαν κάτι για το χρησιμοποιούμενο σύστημα, καθώς και για το αντικείμενο του πειράματος μας προκειμένου να κάνουν όλες τις αγορές τους ανεπηρέαστοι. Γνώριζαν μόνο ότι έπρεπε να κάνουν κάποιες αγορές, και ότι είχαν ένα συγκεκριμένο ποσό χρημάτων να ξοδέψουν συνολικά.

Το ζητούμενο ήταν να ελέγξουμε αφού θα είχαμε μαζέψει τα κατάλληλα δεδομένα, ποια λύση θα δημιουργούσε τα περισσότερα καθαρά κέρδη. Τελικά το αποτέλεσμα ήταν ότι η λύση που πρότεινε προϊόντα στους χρήστες χρησιμοποιώντας τον αλγόριθμο MSApriori παρουσίαζε

πρακτικά τα ίδια καθαρά κέρδη όπως ο αλγόριθμος Argiori. Η προσέγγιση μας από την άλλη κατάφερε να αυξήσει τα καθαρά κέρδη κατά ένα ποσοστό σχεδόν της τάξης του 10% και έτσι μπορούμε να συμπεράνουμε ότι το σύστημα μας πράγματι παρέχει καλύτερες προτάσεις στους χρήστες του, αφού ο κύριος σκοπός της αύξησης των καθαρών κερδών επιτεύχθηκε.

5.4 Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Προτείνουμε ένα μηχανισμό απόδοσης βαρών προκειμένου να καταφέρουμε να λάβουμε υπόψη τα ιδιαίτερα χαρακτηριστικά κάθε αντικειμένου και να τα αντιμετωπίσουμε μέσο του ακριβούς περιβάλλοντος όπου μια τέτοια προσέγγιση υλοποιείται. Το αποτέλεσμα ήταν μια προσέγγιση η οποία πράγματι παράγει καλύτερες προβλέψεις στους χρήστες, και κατορθώνει να αυξήσει τα καθαρά κέρδη σε σύγκριση με τις υπάρχουσες προσεγγίσεις. Ίσως το πιο σημαντικό συμπέρασμα της προσέγγισης μας είναι ότι ένα ακριβό ή ένα πολύ δημοφιλές προϊόν δεν παρουσιάζει μεγαλύτερη εμπορική αξία από ένα φθηνό ή ένα όχι και τόσο δημοφιλές. Είναι η εύρεση και ο συνδυασμός όλων των παραμέτρων που επηρεάζουν την σημαντικότητα ενός αντικειμένου που το καθιστούν λιγότερο ή περισσότερο σημαντικό από ένα άλλο στη βάση. Επίσης κατορθώσαμε να διαχειριστούμε ένα πρόβλημα που αγνοούσαν όλες οι προσεγγίσεις έως τώρα, αυτό της απόδοσης των καταλλήλων τιμών υποστήριξης ή γενικότερα οποιοδήποτε άλλου μέτρου σημαντικότητας στα αντικείμενα της βάσης μας και να μην θεωρούμε τις τιμές αυτές σαν ήδη γνωστές.

Κεφάλαιο 6 ΕΝΑΣ ΒΕΛΤΙΩΜΕΝΟΣ
ΑΛΓΟΡΙΘΜΟΣ ΓΙΑ ΤΗΝ ΕΞΟΥΞΗ ΚΑΝΟΝΩΝ
ΣΥΣΧΕΤΙΣΕΩΝ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ
ΠΟΛΛΑΠΛΕΣ ΤΙΜΕΣ ΚΑΤΩΦΛΙΟΥ
ΥΠΟΣΤΗΡΙΞΗΣ

Περίληψη

Σχεδόν όλες οι προσεγγίσεις στον τομέα των κανόνων συσχετίσεων υιοθετούν την χρήση ενός ενιαίου κατωφλίου υποστήριξης, τεχνική που είτε αποκλείει όλα τα σπάνια αντικείμενα ή πάσχει από το πρόβλημα της δημιουργίας και ανάγκης ελέγχου υπερβολικού αριθμού υποψήφιων αντικειμένων. Η καλύτερη εναλλακτική για το συγκεκριμένο πρόβλημα αποτελεί η χρήση πολλαπλών τιμών υποστήριξης για το κάθε αντικείμενο. Στην παρούσα ενότητα προτείνουμε και παρουσιάζουμε έναν νέο αλγόριθμο για την αποτελεσματικότερη και ταχύτερη εξόρυξη κανόνων συσχετίσεων από αντικείμενα, με χρήση διαφορετικών τιμών υποστήριξης. Ο αλγόριθμος μας βασίζει την φιλοσοφία του σε δύο πολύ γνωστούς αλγορίθμους, αναφορικά τον αλγόριθμο DIC [27] (βλέπε επίσης ενότητα 0) και τον αλγόριθμο MSAPriori [89] (βλέπε επίσης ενότητα 2.7). Τα πειράματά μας δείχνουν ότι η νέα αυτή πρόταση είναι όντως καλύτερη από τον αλγόριθμο MSAPriori, αλλά και για ορισμένους τύπους δεδομένων είναι επίσης καλύτερη και από τον αλγόριθμο DIC.

Λέξεις Κλειδιά: *Κανόνες συσχετίσεων, πολλαπλές τιμές υποστήριξης.*

6.1 Εισαγωγή

Όπως είδαμε και στα προηγούμενα κεφάλαια όλα τα αντικείμενα σε μια βάση δεδομένων δεν είναι της ίδιας φύσεως και σημαντικότητας και έτσι απαιτούν ξεχωριστή μεταχείριση. Ιδιαίτερα σε εφαρμογές και δεδομένα όπως αυτά ενός ηλεκτρονικού καταστήματος ή ενός καταστήματος λιανεμπορίου η ανάγκη αυτή είναι ακόμα πιο επιτακτική. Παρόλ' αυτά σχεδόν όλες οι προσεγγίσεις μέχρι στιγμής θεωρούσαν όλα τα αντικείμενα σε μια βάση όμοια, και χρησιμοποιούσαν έτσι μια ενιαία τιμή κατωφλίου υποστήριξης για αυτά. Αυτό είχε σαν αποτέλεσμα, όπως τονίστηκε και από τον Mannila στο [93], αν θέταμε την τιμή αυτή πολύ χαμηλά προκειμένου να βρούμε επίσης τα σπάνια αλλά σημαντικά αντικείμενα να μας επιστρέφονται ένα υπερβολικά μεγάλο πλήθος αντικειμένων εκ των οποίων τα περισσότερα ήταν εντελώς άχρηστα.

Ένας αποτελεσματικός αλγόριθμος για να αποφύγουμε αυτό το πρόβλημα αλλά και ταυτόχρονα να βρούμε όλους τους κανόνες συσχετίσεων είναι όπως είδαμε στην ενότητα 2.7 ο αλγόριθμος MSAPriori [89]. Σύμφωνα με τη συγκεκριμένη προσέγγιση, κάθε αντικείμενο στη βάση μπορεί να έχει την δικιά του τιμή υποστήριξης που καλείται MIS. Μ' αυτόν τον τρόπο είμαστε σε θέση να έχουμε υψηλότερες τιμές υποστήριξης για κανόνες που περιλαμβάνουν μόνο συχνά αντικείμενα και χαμηλότερες τιμές για κανόνες που περιλαμβάνουν επίσης σπάνια αλλά σημαντικά αντικείμενα. Ο MSAPriori όμως, όπως και ο προγονός του ο Apriori, απαιτούσε τόσα περάσματα από τη βάση όσος και ο αριθμός των στοιχείων στο υψηλότερης τάξης υπονήφιο αντικείμενο.

Από την άλλη οι Brin et al. στο [27] παρουσίασαν τον αλγόριθμό DIC, που χρησιμοποιούσε μια ενιαία τιμή κατωφλίου υποστήριξης αλλά μείωνε ταυτόχρονα τον αριθμό των περασμάτων που πραγματοποιούνταν πάνω στη βάση μας, διατηρώντας τον αριθμό των αντικειμένων που θα έπρεπε να ελεγχθούν αρκούντως χαμηλό σε σχέση με τις αντίστοιχες τεχνικές δειγματοληψίας (sampling techniques [118]). Η κύρια ιδέα του αλγορίθμου αυτού ήταν να ξεκινάμε να μετράμε οποιοδήποτε αντικείμενο μόλις αντιληφθούμε ότι μπορεί να είναι συχνό αντί να περιμένουμε μέχρι όλα του τα υποσύνολα να έχουν καταμετρηθεί σε ολόκληρη τη βάση. Δηλαδή παράδειγμα για το αντικείμενο $\langle 1,2 \rangle$, αν τα στοιχεία του 1 και 2 έχουν ήδη αριθμό εμφανίσεων πάνω από το κατώφλι υποστήριξης τότε αντί να περιμένουμε να μετρήσουμε όλες τις εμφανίσεις τους ξεκινάμε να μετράμε άμεσα το αντικείμενο $\langle 1,2 \rangle$. Έτσι ο DIC παρομοιάζεται με ένα τραίνο που τρέχει πάνω από τα δεδομένα με στάσεις που βρίσκονται σε διαστήματα M δεδομένων ή μια από την άλλη (βλέπε Εικόνα 4 στην σελίδα 25). Σε κάθε στάση κάποια νέα αντικείμενα ανεβαίνουν στο τραίνο (ξεκινούν να καταμετρώνται) και κάποια κατεβαίνουν από αυτό (τελειώνει η καταμέτρηση τους). Αυτό που αδυνατεί να

λάβει υπόψη η συγκεκριμένη προσέγγιση είναι ότι όλοι οι επιβάτες δεν είναι ίδιοι. Ορισμένοι είναι επιβάτες πρώτης θέσης και ως εκ τούτου χρειάζονται κάποια ιδιαίτερη προσοχή, ενώ κάποιοι άλλοι ανήκουν στην οικονομική θέση. Έτσι τελικά κάθε αντικείμενο που ανεβαίνει στο τραίνο σε οποιαδήποτε στάση θεωρείται σαν ίδιο με όλα τα υπόλοιπα, ή με άλλα λόγια λαμβάνει το ίδιο κατώφλι υποστήριξης όπως όλα τα υπόλοιπα. Το ερέθισμα έτσι αυτού του κεφαλαίου ήταν να καταφέρουμε να ενσωματώσουμε την φιλοσοφία των πολλαπλών τιμών υποστήριξης που προτάθηκε από τους Liu, Hsu και Ma στο [89], στην διαδικασία που προτάθηκε από τους Brin et al. στο [27] και να καταλήξουμε με έναν ταχύτερο και αποτελεσματικότερο αλγόριθμο για την εξόρυξη κανόνων συσχετίσεων με πολλαπλές τιμές κατωφλίου υποστήριξης.

Το υπόλοιπο του κεφαλαίου αυτού περιλαμβάνει τα εξής. Στην ενότητα 6.2 παρουσιάζεται ο προτεινόμενος αλγόριθμος. Στις ενότητες 6.3 και 6.4 δίνονται αναλυτικά η διάταξη των αντικειμένων κατά την διαδικασία της καταμέτρησης καθώς και η χρησιμοποιούμενη δομή. Τέλος στην ενότητα 6.5 παρουσιάζουμε τα πειραματικά αποτελέσματα από τον έλεγχο του αλγορίθμου μας.

6.2 Ο αλγόριθμος KTM

Ο αλγόριθμος μας, με την ονομασία **KTM**, ουσιαστικά βασίζει την λειτουργία του στον αλγόριθμο DIC ενσωματώνοντας όμως τις αρχές του αλγορίθμου MSApriori. Ας δούμε πως: για την λειτουργία του ο αλγόριθμος μας σημειώνει κάθε αντικείμενο στη βάση με μια διαφορετική κατάσταση. Εκτός από τις τέσσερις καταστάσεις που χρησιμοποιούνταν από τον αλγόριθμο DIC, εισάγουμε δύο ακόμα καταστάσεις για τους σκοπούς μας. Έτσι ο αλγόριθμος μας σημειώνει συνολικά τα αντικείμενα με έξι διαφορετικούς τρόπους οι οποίοι είναι:

- ▣ **Διακεκομμένος Κύκλος (Dashed Circle - DC)** – αντικείμενο το οποίο υποψιαζόμαστε ότι είναι σπάνιο – ένα αντικείμενο το οποίο ακόμα καταμετρούμε και όπου οι εμφανίσεις του είναι κάτω από την MIS τιμή του – επίσης η αρχική κατάσταση όλων των αντικειμένων.
- ▣ **Συμπαγής Κύκλος (Solid Circle - SC)** – επιβεβαιωμένο σπάνιο αντικείμενο – ένα αντικείμενο του οποίου τελειώσαμε την καταμέτρηση αλλά που ο αριθμός εμφανίσεων του είναι κάτω από την δικιά του MIS τιμή όπως επίσης και κάτω από την τιμή του αντικειμένου πριν από αυτό.
- ▣ **Διακεκομμένο Τετράγωνο (Dashed Square - DS)** – αντικείμενο το οποίο υποψιαζόμαστε ότι είναι συχνό – επίσης ένα αντικείμενο το οποίο καταμετρούμε ακόμα

αλλά του οποίου ο αριθμός εμφανίσεων του είναι ήδη υψηλότερος από την MIS τιμή του.

- ▮ **Συμπαγές Τετράγωνο (Solid Square - SS)** – επιβεβαιωμένο συχνό αντικείμενο – ένα αντικείμενο το οποίο τελειώσαμε να καταμετράμε στο σύνολο των συναλλαγών και το οποίο υπερβαίνει την MIS τιμή του.
- ▮ **Διακεκομμένο Διαμάντι (Dashed Diamond - DD)** – αντικείμενο το οποίο επιβεβαιώσαμε ότι οι εμφανίσεις του δεν ξεπερνούν την MIS τιμή του αλλά που είναι πάνω από την MIS τιμή του αντικειμένου πριν από αυτό στην MIS διάταξη, και το οποίο δεν έχει δημιουργήσει όλα τα υποψήφια αντικείμενα που του αναλογούν.
- ▮ **Συμπαγές Διαμάντι (Solid Diamond - SD)** – αντικείμενο το οποίο επιβεβαιώσαμε ότι οι εμφανίσεις του είναι κάτω από την MIS τιμή του αλλά πάνω από την MIS τιμή του αντικειμένου πριν από αυτό στην MIS διάταξη και το οποίο έχει δημιουργήσει όλα τα υποψήφια αντικείμενα που του αναλογούν.

Κάθε μονοσύνολο ξεκινά να καταμετράται με την κατάσταση του ορισμένη σε DC – διακεκομμένο κύκλο, εκτός από το κενό μονοσύνολο το οποίο σημειώνεται απευθείας με συμπαγές τετράγωνο (Εικόνα 9, στιγμιότυπο 1). Μετά από M συναλλαγές, όπου M είναι ο αριθμός των συναλλαγών που πρέπει να διαβαστούν πριν το τραίνο (ο αλγόριθμος μας) πραγματοποιήσει την πρώτη του στάση, ελέγχουμε τον μετρητή του κάθε μονοσύνολου που μετράμε σε σχέση με την MIS τιμή του. Αν ο μετρητής του είναι μεγαλύτερος ή ίσος από την MIS τιμή αυτή τότε η κατάσταση του αλλάζεται σε DS – διακεκομμένο τετράγωνο (Εικόνα 9, στιγμιότυπο 2, στοιχεία 1 και 4). Διαφορετικά παραμένει DC – διακεκομμένος κύκλος και θα ελεγχθεί ξανά μετά από M συναλλαγές. Πριν προχωρήσουμε ξανά με τις επόμενες M συναλλαγές ελέγχουμε ποια δισύνολα θα πρέπει να αρχίσουν να μετρούνται. Έτσι κάθε δισύνολο το οποίο έχει τα στοιχεία του σαν DS – διακεκομμένα τετράγωνα χαρακτηρίζεται σαν DC – διακεκομμένος κύκλος και ξεκινάει να μετράται. Το ίδιο συμβαίνει σε κάθε στάση του αλγόριθμου μας.

Αντίστοιχα στις επόμενες στάσεις εάν ένα άμεσο υπερσύνολο ενός μονοσύνολου με κατάσταση DS – διακεκομμένο τετράγωνο, έχει όλα του τα υποσύνολα σαν συμπαγή ή διακεκομμένα τετράγωνα, τότε κάνουμε την κατάσταση του DC – διακεκομμένο κύκλο και ξεκινάμε να το μετράμε (Εικόνα 9, στιγμιότυπο 2, αντικείμενο <1,4>).

Αν ο μετρητής του παραμείνει χαμηλότερος από τη MIS τιμή του και είναι ένα μονοσύνολο τότε ελέγχουμε τον μετρητή του σε σχέση με την MIS τιμή του μονοσύνολου πριν από αυτό στην MIS διάταξη. Αν είναι επίσης χαμηλότερα από την MIS τιμή του μονοσύνολου πριν από αυτό, τότε αλλάζουμε την κατάσταση του σε SC – συμπαγή κύκλο. Αν από την άλλη ο

μετρητής του είναι ίσος ή μεγαλύτερος από την MIS τιμή του αντικειμένου που προηγείται, τότε αλλάζουμε την κατάστασή του σε DD – διακεκομμένο διαμάντι (Εικόνα 9, στιγμιότυπο 3, αντικείμενο 3), και ξεκινάμε να δημιουργούμε όλα τα δυνατά υποψήφια δισύνολα από αυτό. Αυτή η κατάσταση δεν διατηρείται για πολύ αφού όταν ένα αντικείμενο παράγει όλα τα δυνατά υποψήφια αντικείμενα τότε αλλάζουμε την κατάσταση του σε SD – συμπαγές διαμάντι (Εικόνα 9, στιγμιότυπο 4, αντικείμενο 3). Η μόνη διαφορά ανάμεσα στις δύο καταστάσεις είναι ότι ένα αντικείμενο με κατάσταση SD έχει παράγει όλα τα δυνατά υποψήφια αντικείμενα με βάση αυτό. Ο λόγος για τον οποίο εισαγάγαμε δύο καταστάσεις, διακεκομμένο διαμάντι - DD και συμπαγές διαμάντι – SD αντί για μόνο μια ήταν γιατί θέλαμε να αποτρέψουμε τον αλγόριθμο μας από το να παράγει νέα υποψήφια αντικείμενα σε κάθε στάση του. Ο αλγόριθμος μας τερματίζει με την απουσία κάποιου διακεκομμένου αντικειμένου.

Αυτή η διαδικασία είναι κάπως διαφορετική για τα μονοσύνολα τα οποία έχουν σημειωθεί σαν DD – διακεκομμένα διαμάντια στο ότι δημιουργούμε υπερσύνολα από αυτά χρησιμοποιώντας το μονοσύνολο που είναι σημειωμένο σαν DD και οποιοδήποτε άλλο μονοσύνολο το οποίο είναι σημειωμένο DS – διακεκομμένο τετράγωνο, ή SD – συμπαγές τετράγωνο το οποίο είναι επίσης πριν από αυτό στη διάταξη MIS (Εικόνα 9, στιγμιότυπο 4 όπου δημιουργήσαμε το υποψήφιο αντικείμενο $\langle 1,3 \rangle$ μόνο και όχι και το $\langle 3,4 \rangle$). Επίσης η σειρά κάθε νέου αντικειμένου διατηρείται, βάζοντας στην αρχή το μονοσύνολο με την χαμηλότερη MIS τιμή, και ακολούθως όλα τα υπόλοιπα διατεταγμένα ως προς τις MIS τιμές τους.

Εάν ένα άμεσο υπερσύνολο ενός k -αντικειμένου (όπου $k \geq 2$) το οποίο είναι σημειωμένο σαν DS – διακεκομμένο τετράγωνο έχει όλα τα υποσύνολά του σαν συμπαγή ή διακεκομμένα τετράγωνα, τότε κάνουμε τη κατάσταση του DC – διακεκομμένο κύκλο και ξεκινάμε να το μετράμε. Υπάρχει παρόλα αυτά μια εξαίρεση όταν ένα υποσύνολο ενός αντικειμένου δεν είναι συχνό (δηλ. δεν είναι σημειωμένο σαν DS ή SS), και παρόλα αυτά το υποψήφιο υπερσύνολο δεν μπορεί να απαλειφθεί (στην Εικόνα 9, στο στιγμιότυπο 3 το υπερσύνολο $\langle 1,3,4 \rangle$ δημιουργείται παρόλο που το υποσύνολο $\langle 3,4 \rangle$ δεν είναι συχνό). Αυτή η περίπτωση ανακύπτει όταν το υποσύνολο ενός αντικειμένου δεν περιέχει το πρώτο αντικείμενο του υπερσυνόλου (υπάρχει πάντα ένα τέτοιο υποσύνολο) και δεν είμαστε σίγουροι ότι η MIS τιμή του πρώτου μονοσυνόλου δεν είναι ίδια όπως αυτή του δεύτερου.

Εικόνα 9: Παράδειγμα εκτέλεσης του αλγορίθμου KTM

Όταν ένα αντικείμενο έχει καταμετρηθεί σε όλες τις συναλλαγές ελέγχουμε ξανά τον μετρητή του σε σχέση με την MIS τιμή του. Αν η κατάσταση του ήταν DC – διακεκομμένος κύκλος, και ο μετρητής του είναι τελικά ίσος ή μεγαλύτερος από την MIS τιμή του τότε αλλάζουμε την κατάσταση του σε SS – συμπαγές τετράγωνο. Αν η κατάσταση του ήταν DS – διακεκομμένο τετράγωνο, τότε απλά το αλλάζουμε σε SS – συμπαγές τετράγωνο (Εικόνα 9,

στιγμιότυπο 3, αντικείμενα 1 και 4). Όπως μπορούμε εύκολα να καταλάβουμε μόνο τα μονοσύνολα μπορούν να χαρακτηριστούν σαν DD ή SD, αφού κρατάμε μια διάταξη MIS μόνο για αυτά τα αντικείμενα. Όλα τα υπόλοιπα υψηλότερης τάξης αντικείμενα μπορούν να είναι οποιαδήποτε κατάσταση από DC, DS, SS και SC. Με τον τερματισμό του αλγορίθμου μας καταλήγουμε με ένα πλέγμα αντικειμένων όπως αυτό που φαίνεται στην Εικόνα 10, με το κενό αντικείμενο στο κάτω μέρος και το σύνολο όλων των υπολοίπων αντικειμένων από πάνω. Τα αντικείμενα που δεν έχουν σημειωθεί με κάποια κατάσταση, δηλαδή δεν μετρήθηκαν ποτέ δείχνονται παρόλα αυτά εδώ για να βοηθήσουν στη οπτικοποίηση.

Εικόνα 10: Ένα πλέγμα αντικειμένων

6.3 Διάταξη των Αντικειμένων

Η διάταξη των αντικειμένων στη δομή μας παίζει ένα καθοριστικό ρόλο και στους δύο αλγορίθμους, στον καθένα για διαφορετικούς σκοπούς. Στην παράγραφο αυτή δείχνουμε ότι η διάταξη την οποία εμείς εφαρμόζουμε εξυπηρετεί και τους δύο αλγορίθμους και έχει επιπλέον ως αποτέλεσμα την δημιουργία ενός πιο αποτελεσματικού αλγορίθμου.

Οι Liu, Hsu και Ma ([89]) πρότειναν την διάταξη των αντικειμένων σύμφωνα με τις MIS τιμές τους σε αύξουσα σειρά προκειμένου να ικανοποιούν αυτό που ονόμαζαν sorted closure property (δες επίσης την παράγραφο 2.7). Οι Brin et al. ([27]) από την άλλη χρησιμοποιούσαν ένα ενιαίο κατώφλι υποστήριξης που ικανοποιούσε την subset closure property (δες επίσης την

παράγραφο 2.3 ή την δουλειά των Agrawal και Srikant [8]). Στην εργασία τους ανέφεραν το ζήτημα του πώς θα έπρεπε τα αντικείμενα να διατάσσονται στην δομή που χρησιμοποιείται για την καταμέτρηση τους. Έτσι σύμφωνα με τα λεγόμενα τους, έχοντας τα αντικείμενα που εμφανίζονται σε πολλά αντικείμενα μεγαλύτερης τάξης να είναι τελευταία στην διάταξη των αντικειμένων (οπότε θα μένουν λίγα αντικείμενα μετά από αυτά) και τα αντικείμενα που εμφανίζονται σε λίγα αντικείμενα μεγαλύτερης τάξης πρώτα στη διάταξη, επιτυγχάνει αποτελεσματικότερο μέτρημα. Οι συγκεκριμένοι ερευνητές προσπάθησαν να εφαρμόσουν μια τεχνική επαναδιάταξης, της οποίας όμως τα αποτελέσματα ήταν απογοητευτικά, αφού η επαναδιάταξη έπαιζε έναν αμελητέο ρόλο στη τελική απόδοση, ενώ πολλές φορές είχε το ακριβώς αντίθετο αποτέλεσμα, διατάσσοντας τα αντικείμενα ακριβώς αντίστροφα. Αυτό που προσπάθησαν δηλαδή, ήταν μετά τις πρώτες M συναλλαγές να βάλουν τα στοιχεία στο δέντρο σύμφωνα με τις εμφανίσεις τους σε αυτές τις συναλλαγές. Έτσι οι μελλοντικές συναλλαγές θα καταμετρώνται σύμφωνα με αυτή τη σειρά. Το συγκεκριμένο πρόβλημα ήταν περισσότερο εμφανές σε βάσεις οι οποίες είχαν τα δεδομένα υπερβολικά ανομοιόμορφα κατανεμημένα (skewed), και τοπικές αλλαγές ή ιδιαίτερα χαρακτηριστικά στα δεδομένα μπορούν να έχουν εντελώς απρόβλεπτα αποτελέσματα. Η διάταξη που προτείνουμε εμείς στον αλγόριθμο μας πρώτα απ' όλα ικανοποιεί την sorted closure property που προτάθηκε από τους Liu, Hsu και Ma ([89]). Επιπρόσθετα αφού σύμφωνα με αυτή τη διάταξη τα πιο συχνά αντικείμενα αποθηκεύονται πρώτα στην δομή μας και τα λιγότερα συχνά τελευταία, επιτυγχάνουμε πιο αποτελεσματικό μέτρημα, αποφεύγουμε το κόστος που συνεπάγεται η επαναδιάταξη που πρότειναν οι Brin et al. ([27]), και τελικά όπως θα δούμε και στην τελευταία ενότητα για εξαιρετικά ανομοιόμορφα κατανεμημένες βάσεις ο αλγόριθμος μας αποδεικνύεται αποτελεσματικότερος από τον DIC.

6.4 Η Δομή Δεδομένων

Η δομή δεδομένων που χρησιμοποιείται από τον αλγόριθμο μας είναι ακριβώς όπως το hash tree που χρησιμοποιεί ο DIC με μια μικρή αλλά πολύ σημαντική διαφοροποίηση. Σε κάθε κόμβο του δέντρου αποθηκεύουμε επίσης την αντίστοιχη MIS τιμή του συγκεκριμένου κόμβου. Αυτό σημαίνει ότι αφού δεν έχουμε ένα κοινό κατώφλι υποστήριξης για κάθε αντικείμενο στο δέντρο, κάθε αντικείμενο πρέπει να έχει τη δική του τιμή υποστήριξης. Οι MIS τιμές για τα μονοσύνολα αποδίδονται χειροκίνητα από το χρήστη, και από το σημείο αυτό και μετά ο αλγόριθμος ο ίδιος αποδίδει τις MIS τιμές σε κάθε κόμβο αυτόματα. Το βήμα αυτό δίνεται παρακάτω:

1. *for every itemset $c = \langle c.item_1, c.item_2, \dots, c.item_k \rangle$ that we begin counting*
2. *select $c.item_1$*
3. $c.MIS = c.item_1.MIS$

Βασικά για κάθε αντικείμενο το οποίο γίνεται υποψήφιο (δηλ. χαρακτηρίζεται σαν DC) και ξεκινά να καταμετράται παίρνουμε το πρώτο στοιχείο, το οποίο έχει και την χαμηλότερη MIS τιμή από όλα τα άλλα, και αποδίδουμε στο αντικείμενο αυτή την τιμή.

Στην Εικόνα 11 παρουσιάζουμε τα σύμβολα που χρησιμοποιούνται για τα διάφορα αντικείμενα καθώς και για τις καταστάσεις τους, τις μεταβάσεις ανάμεσα στις καταστάσεις αυτές και τέλος πότε θα πρέπει να λαμβάνουν χώρα αυτές οι μεταβάσεις. Όπως μπορούμε να δούμε και στην Εικόνα 11 η διαχείριση των μεταβάσεων ανάμεσα στις διάφορες καταστάσεις (από μη καταμετρημένες σε καταμετρημένες και από σπάνιες σε συχνές) και η ανίχνευση πότε θα πρέπει αυτές οι μεταβάσεις να λαμβάνουν χώρα είναι ένα πολύπλοκο έργο.

Εικόνα 11: Οι μεταβάσεις ανάμεσα στις διαφορετικές καταστάσεις

6.5 Σύγκριση Απόδοσης

Σ' αυτή την ενότητα παρουσιάζουμε τα πειράματα που πραγματοποιήσαμε για να ελέγξουμε την αποτελεσματικότητα της πρότασης μας. Έτσι αρχικά συγκρίναμε τον αλγόριθμο μας σε σχέση με τον αλγόριθμο Apriori για διάφορες τιμές χρησιμοποιούμενων μεταβλητών. Ακολούθως συγκρίναμε τον αλγόριθμο μας με τον αλγόριθμο DIC σε ανομοιομορφα κατανομημένα δεδομένα προκειμένου να καταδείξουμε την βελτίωση που επιτυγχάνουμε σε αυτού του τύπου δεδομένα.

6.5.1 Πειράματα με συνθετικά δεδομένα

Προκειμένου να πραγματοποιήσουμε τα πειράματα μας με συνθετικά δεδομένα χρησιμοποιήσαμε την γεννήτρια συνθετικών δεδομένων††. Όπως διατυπώθηκε και από τους Liu, Hsu και Ma στο [89] στην σύγκριση του MSAPriori με τον αλγόριθμο Apriori, όσον αναφορά τα τεχνητά δεδομένα ο χρόνος που απαιτείται από τον MSAPriori είναι πρακτικά ίδιος με αυτόν που απαιτείται από τον αλγόριθμο Apriori αφού το διάβασμα της βάσης κυριαρχεί στην όλη διαδικασία. Έτσι συγκρίναμε τον αλγόριθμο μας απευθείας με τον αλγόριθμο Apriori. Ο αριθμός των υποψήφιων και των συχνών αντικειμένων είναι ακριβώς ίδιος και για τους δύο αλγόριθμους και έτσι δεν δείχνεται εδώ.

Προκειμένου να αποδώσουμε MIS τιμές στα δεδομένα μας χρησιμοποιήσαμε την ίδια μέθοδο όπως αυτή των Liu, Hsu και Ma στο [89], και πιο συγκεκριμένα χρησιμοποιήσαμε την ακόλουθη εξίσωση:

$$MIS(i) = \begin{cases} M(i) & M(i) > LS \\ LS & otherwise \end{cases} \quad where \quad \begin{cases} M(i) = \beta f(i) \\ \beta = 1/\alpha \end{cases}$$

Όπου $f(i)$ είναι η ακριβής συχνότητα εμφανίσεων ενός αντικειμένου i , LS είναι μια οριζόμενη από το χρήστη κατώτερη επιτρεπόμενη τιμή υποστήριξης και β είναι μια παράμετρος που ελέγχει το πως οι MIS τιμές των αντικειμένων θα πρέπει να σχετίζονται με τις συχνότητές τους.

Για τα πειράματα μας δημιουργήσαμε ένα σύνολο βάσεων προκειμένου να δοκιμάσουμε αλγόριθμο μας. Όλες οι βάσεις είχαν παρόμοια συμπεριφορά, και έτσι παρουσιάζουμε μόνο μια. Στην παρούσα έκδοση του ο αλγόριθμος μας δεν υποστηρίζει την αυτόματη απόδοση MIS τιμών, και έτσι τα πάντα έπρεπε να γίνουν χειροκίνητα. Έτσι δοκιμάσαμε τον αλγόριθμο μας με βάσεις οι οποίες αποτελούνταν από μερικές δεκάδες διαφορετικών αντικειμένων. Η βάση η οποία δείχνεται εδώ δημιουργήθηκε χρησιμοποιώντας κατά μέσο όρο 10 στοιχεία ανά συναλλαγή, 50 διαφορετικά στοιχεία και 100.000 συναλλαγές. Το βήμα του αλγόριθμου μας καθορίστηκε στο 20% της βάσης (δηλ. 20.000 συναλλαγές).

Στην Εικόνα 12 δείχνουμε την σύγκριση ανάμεσα στους χρόνους εκτέλεσης των δύο αλγορίθμων σαν ποσοστό. Όπως μπορούμε να δούμε ο αλγόριθμος μας καταφέρνει να μειώσει τους χρόνους εκτέλεσης σημαντικά. Έτσι για $LS=1\%$ ο αλγόριθμος μας τρέχει σχεδόν 70% ταχύτερα (για $\alpha=1$), και για $LS=2\%$ σχεδόν 60% ταχύτερα (για $\alpha=1$).

†††Βλέπε παράγραφο 2.8

Εικόνα 12: Σύγκριση των χρόνων εκτέλεσης σε ποσοστά

6.5.2 Ανομοιόμορφα Κατανομημένα - Skewed Δεδομένα

Ο αλγόριθμος DIC όπως επίσης και ο αλγόριθμος μας είναι και οι δύο ευεπηρεάστοι στην ανομοιομορφία των δεδομένων. Αφού δεν υπάρχει κάποια γνωστή γεννήτρια για την δημιουργία ανομοιόμορφα κατανομημένων συνθετικών βάσεων δεδομένων, τουλάχιστον κάποια που να γνωρίζουμε, αποφασίσαμε να δημιουργήσουμε κάποια δικά μας τέτοια δεδομένα. Ξεκινήσαμε με μια βάση η οποία περιείχε 5 διαφορετικά αντικείμενα και 100.000 συναλλαγές και αυξήσαμε τον αριθμό των συναλλαγών βαθμιαία μέχρι τα 9.

Οι βάσεις που δημιουργήσαμε ήταν τόσο ανομοιόμορφες που ορισμένα αντικείμενα εμφανίζονταν τόσο σπάνια όσο μόλις 100 φορές σε μια βάση με 100.000 συναλλαγές, όλες σε συνεχόμενες συναλλαγές και τις περισσότερες φορές στο τέλος ολόκληρης της βάσης. Αυτό σήμαινε ότι ορισμένες φορές και οι δύο αλγόριθμοι έπρεπε να περιμένουν μέχρι το τέλος της βάσης προτού να ξεκινήσουν να μετρούν ορισμένα αντικείμενα, χάνοντας έτσι το πλεονέκτημα που προσφερόταν και από τους δύο αλγόριθμους και το οποίο δεν ήταν άλλο από το να ξεκινούν να μετρούν υψηλότερης τάξης αντικείμενα αρκετά νωρίτερα. Αυτό εξηγεί και τους ελαφρώς υψηλούς χρόνους εκτέλεσης που χρειαζόνταν και οι δύο αλγόριθμοι. Σ' αυτή τη σειρά πειραμάτων δεν συμπεριλάβαμε τον αλγόριθμο MSApriori καθώς θα απαιτούσε ακόμα μεγαλύτερους χρόνους εκτέλεσης. Σ' αυτά τα πειράματα, όπως μπορούμε να δούμε και από την Εικόνα 13, ο αλγόριθμος KTM είχε καλύτερη απόδοση από τον αλγόριθμο DIC σε όλες τις περιπτώσεις, μειώνοντας τον χρόνο εκτέλεσης από 2% σε μια βάση με 5 αντικείμενα σε σχεδόν 10% στη βάση με τα 9 αντικείμενα επιβεβαιώνοντας αυτό που είχαμε επισημάνει νωρίτερα.

Εικόνα 13: % Βελτίωση του χρόνου εκτέλεσης

6.6 Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Στο μέλλον θα θέλαμε να πραγματοποιήσουμε πιο εκτεταμένα πειράματα προκειμένου να ελέγξουμε την συμπεριφορά του αλγορίθμου μας. Θα θέλαμε να πειραματιστούμε με περισσότερα χαρακτηριστικά όπως το μήκος του βήματος, το κατώφλι βεβαιότητας (minimum confidence) καθώς και με περισσότερα μονοσύνολα. Θα θέλαμε επίσης να τον δοκιμάσουμε σε πραγματικά δεδομένα.

**Κεφάλαιο 7 ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ
ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΤΗΝ
ΥΠΟΣΤΗΡΙΞΗ ΑΝΑΚΑΛΥΨΗΣ ΓΝΩΣΗΣ**

Περίληψη

Η κλασσική προσέγγιση του αλγορίθμου Apriori αλλά και όλων των απογόνων του, η οποία συνίστατο αρχικά στην εύρεση όλων των συχνών αντικειμένων και μετέπειτα στη παραγωγή όλων των κανόνων συσχετίσεων δούλεψε καλά για συγκεκριμένες κατηγορίες δεδομένων. Παρόλα' αυτά για πολλούς άλλους τύπους δεδομένων αυτή η προσέγγιση παρουσιάζει μειωμένη απόδοση και αποδεικνύεται μάλλον ανεπαρκής.

Διαφωνούμε ότι πρέπει να ελέγξουμε όλους τους δυνατούς συνδυασμούς υποψηφίων αντικειμένων αλλά μάλλον να αφήσουμε την βάση μας να αποκαλύψει μόνη της τα μυστικά της καθώς οι χρήστες θα την χρησιμοποιούν. Προτείνουμε ένα σύστημα το οποίο δεν ψάχνει απλά όλη τη βάση, αλλά αντίθετα λειτουργεί σαν μια μηχανή αναζήτησης ειδικά σχεδιασμένη για να κάνει προτάσεις στους χρήστες χρησιμοποιώντας τεχνικές δανεισμένες από την ανάκτηση πληροφορίας.

***Λέξεις Κλειδιά:** Ανακάλυψη γνώσης, Προτάσεις προϊόντων, Δημιουργία ευρετηρίων, Ερωτήματα Δυναδικά-Διάταξης.*

7.1 Εισαγωγή

Ένα μεγάλο μειονέκτημα των παραδοσιακών τεχνικών εξόρυξης δεδομένων και ιδιαίτερα των αλγορίθμων εύρεσης κανόνων συσχετίσεων έχει να κάνει με τη τεχνική που χρησιμοποιούν για να αποκλείσουν κάποια αντικείμενα (pruning technique) από την διαδικασία της καταμέτρησης. Σύμφωνα με αυτή την τεχνική αν ένα αντικείμενο βρεθεί σπάνιο τότε όλα τα υψηλότερης τάξης αντικείμενα που συμπεριλαμβάνουν αυτό το αντικείμενο είναι καταδικασμένα να είναι επίσης σπάνια. Έτσι απλά ξεχνάμε οτιδήποτε έχει να κάνει με το συγκεκριμένο αντικείμενο. Αλλά τι συμβαίνει αν ένας χρήστης ξεκινήσει μια συναλλαγή ή βάλει σχετικά νωρίς στο καλάθι αγορών του ένα σπάνιο αντικείμενο; Τότε ο συγκεκριμένος χρήστης με βάση την τρέχουσα συναλλαγή του αποκλείεται να λάβει άλλες προτάσεις αγορών. Οτιδήποτε αυτός/αυτή αγοράσει από τη στιγμή που επέλεξε ένα σπάνιο αντικείμενο θα είναι επίσης σπάνιο. Στην πράξη όμως θα θέλαμε σίγουρα να προτείνουμε στους χρήστες έστω και ένα όχι και τόσο συχνό αντικείμενο, ακόμα και αν με βάση τους παραδοσιακούς αλγορίθμους δεν ξεπερνά το κατώφλι υποστήριξης, από το να μην προτείνουμε απολύτως τίποτα. Η προσέγγιση που δημιουργήσαμε και αναλύουμε στο παρόν κεφάλαιο ξεπερνά αυτό το πρόβλημα καθώς δεν έχει κάποια μετρική για να αποκλείει αντικείμενα τα οποία δεν εμφανίζονται αρκετά συχνά. Αυτό που κάνει είναι να λαμβάνει δεδομένα – είσοδο από τους ίδιους τους χρήστες και να δημιουργεί τις δικές του προτάσεις άμεσα, αγνοώντας για οποιαδήποτε τεχνική αποκλεισμού (pruning technique).

Από την άλλη το μοντέλο των δύο φάσεων που χρησιμοποιούν όλοι οι κλασικοί αλγόριθμοι είναι σωστό αν το εξετάσουμε με βάση αυστηρά στατιστικά κριτήρια αλλά αδυνατεί να λάβει υπόψη και να αντιμετωπίσει τη σημασιολογία των εμφανίσεων των αντικειμένων στις συναλλαγές καθώς και τις διαφοροποιήσεις στη φύση τους. Με άλλα λόγια είναι άλλο πράγμα αν ένα αντικείμενο εμφανίζεται δέκα φορές στην ίδια συναλλαγή και άλλο αν εμφανίζεται μόνο μια ή δύο φορές, καθώς επίσης δεν είναι το ίδιο αν ένα πολύ σημαντικό αντικείμενο εμφανίζεται σε μια συναλλαγή και αν ένα όχι τόσο σημαντικό εμφανίζεται στην ίδια συναλλαγή. Σαν αποτέλεσμα, εκτός από το πλήθος των εμφανίσεων των αντικειμένων κάποιες άλλες πολύ σημαντικές παράμετροι πρέπει να λαμβάνονται επίσης υπόψη. Όλες αυτές οι απαιτήσεις και οι παράμετροι λαμβάνονται υπόψη με τον καλύτερο τρόπο και αντιμετωπίζονται από ένα πλήρες σύστημα το οποίο και παρουσιάζουμε παρακάτω.

Το υπόλοιπο του παρόντος κεφαλαίου είναι οργανωμένο ως εξής: Ξεκινάμε δίνοντας κάποιες σχετικές εργασίες καθώς και μια γενική περιγραφή του προτεινόμενου συστήματος στις ενότητες 7.2 και 7.3 αντίστοιχα. Στην ενότητα 7.4 δίνουμε μια σύντομη επισκόπηση της πρώτης λειτουργίας του συστήματος μας, σαν ένας κλασικός αλγόριθμος κανόνων

συσχετίσεων. Ακολούθως στην ενότητα 7.5 παρουσιάζουμε με λεπτομέρεια την δεύτερη και κύρια λειτουργία του συστήματος μας, σαν μια μηχανή που κάνει προτάσεις στους χρήστες με χρήση είτε δυαδικών είτε ερωτημάτων διάταξης και δίνουμε όλες τις προσφερόμενες υπηρεσίες. Στην ενότητα 7.6 δίνουμε κάποια αποτελέσματα από τις πειραματικές μεθόδους που χρησιμοποιήσαμε και καταλήγουμε με την ενότητα 7.7 όπου παρουσιάζουμε τα ευρήματα μας καθώς και μελλοντικές κατευθύνσεις.

7.2 Σχετικές Εργασίες

Το σύστημα το οποίο προτείνουμε στο παρόν κεφάλαιο βασίζεται σε μεθόδους και τεχνικές από την ανάκτηση πληροφορίας προκειμένου να υποστηρίξει λειτουργίες ανακάλυψης γνώσης. Σύμφωνα με την βιβλιογραφία κάποιο παρόμοιο σύστημα δεν έχει ξαναπροταθεί. Οι περισσότερες τεχνικές και λειτουργίες που προτείνονται εδώ είναι εντελώς καινοτόμες ακόμα και για την κλασσική εξόρυξη δεδομένων. Μόνο ορισμένες λειτουργίες παρουσιάζουν κάποιες ομοιότητες με τα collaborative filtering συστήματα.

Τα Collaborative filtering συστήματα προσπαθούν να προσφέρουν προσωποποιημένες προτάσεις αντικειμένων (συνήθως σε άρθρα ειδήσεων, ταινίες και μουσική) βασιζόμενα στην πληροφορία σχετικά με τις ομοιότητες ανάμεσα στις προτιμήσεις των χρηστών. Ορισμένα από τα πιο γνωστά collaborative filtering συστήματα μέχρι τώρα είναι το Tapestry [54], το GroupLens [76][104], το Ringo [111] και ο Video Recommender [61] (για μια πιο εμπεριστατωμένη και λεπτομερή επισκόπηση τέτοιων συστημάτων ο αναγνώστης μπορεί να συμβουλευτεί το [105]). Μια διαδικασία collaborative filtering μπορεί να χωριστεί σε τρεις υπό-διαδικασίες, αναφορικά στην αναπαράσταση, τη δημιουργία κλικών και την παροχή προτάσεων στους χρήστες. Στην διαδικασία της αναπαράστασης προσπαθούμε να μοντελοποιήσουμε τα αντικείμενα που σχετίζονται με τον κάθε χρήστη (δηλ. στην πράξη να δημιουργήσουμε ένα προφίλ για κάθε υπάρχοντα χρήστη). Κατά τη δημιουργία της κλίμακας προσπαθούμε να βρούμε χρήστες που ταιριάζουν καλύτερα με τον κάθε νέο χρήστη. Τέλος η διαδικασία των προτάσεων ασχολείται με το να προτείνει στους νέους χρήστες τα πιο σχετικά αντικείμενα με βάση τις προτιμήσεις των παλιών χρηστών. Συνήθως τα collaborative filtering συστήματα κατηγοριοποιούνται σύμφωνα με το πως συλλέγεται το feedback από τους χρήστες, σε έμμεσης (implicit) και άμεσης (explicit) ψηφοφορίας. Κατά την άμεση ψηφοφορία το σύστημα ζητάει από τους χρήστες να βαθμολογήσουν ή να αξιολογήσουν με βάση κάποια κλίμακα αυτό που μόλις τους παρουσιάστηκε, ενώ κατά την έμμεση ψηφοφορία το σύστημα προσπαθεί να παρατηρήσει ή να ανακαλύψει κάποια πληροφορία σχετικά με τις προτιμήσεις των χρηστών βασιζόμενο σε διάφορους εξωγενείς – τρίτους παράγοντες. Η άμεση ψηφοφορία είναι πολύ πιο ακριβής από την έμμεση αφού ζητάει την συνειδητή αξιολόγηση των χρηστών

ενώ η έμμεση βασίζεται σε παράγοντες οι οποίοι μπορεί να μην είναι και τόσο ακριβείς. Από την άλλη το γεγονός ότι η άμεση ψηφοφορία απαιτεί συνεχή αξιολόγηση για οτιδήποτε μπορεί να έχει δει ένας χρήστης, μπορεί να γίνει πολύ ενοχλητικό και να έχει διάφορες ανεπιθύμητες “παρενέργειες”. Αντί της άμεσης-έμμεσης κατηγοριοποίησης προτιμούμε μια εναλλακτική, η οποία βασίζεται στο αν το σύστημα χρησιμοποιεί κάποιο μοντέλο ή τη μνήμη του υπολογιστή. Τα συστήματα που βασίζονται στη μνήμη χρησιμοποιούν κάθε φορά ολόκληρη την διαθέσιμη βάση προκειμένου να κάνουν προβλέψεις ενώ τα συστήματα που βασίζονται σε κάποιο μοντέλο χρησιμοποιούν τη βάση μόνο για να κατασκευάσουν το μοντέλο και κατόπιν χρησιμοποιούν μόνο το μοντέλο. Σε ένα μοντέλο που βασίζεται στη μνήμη οι προβλέψεις σχετικά με τις προτιμήσεις των χρηστών γίνονται χρησιμοποιώντας διάφορες μετρικές όπως ο συντελεστής συσχέτισης Pearson ή η μετρική του ημίτονου. Διάφορες βελτιώσεις έχουν γίνει στο μοντέλο της μνήμης οι πιο γνωστές εκ των οποίων είναι η default voting, η case amplification [24] και η inverse user frequency [108]. Οι τεχνικές που βασίζονται σε κάποιο μοντέλο από την άλλη χρησιμοποιούν τεχνικές όπως τα Bayesian δίκτυα ή το SVD - Singular Value Decomposition προκειμένου να κατασκευάσουν το μοντέλο το οποίο κατόπιν θα χρησιμοποιηθεί για την δημιουργία των προβλέψεων. Και τα δύο συστήματα παρουσιάζουν μικρότερη ακρίβεια στις προβλέψεις τους και έχουν τα σχετικά μειονεκτήματα και πλεονεκτήματα τους. Έτσι τα συστήματα μνήμης καταφέρνουν να είναι πάντα ενημερωμένα και να ανταποκρίνονται άμεσα σε οποιεσδήποτε αλλαγές στις προτιμήσεις των χρηστών με το κόστος από την άλλη να χρειάζονται τεράστιους πόρους σε μνήμη και υπολογιστική ισχύ, και έτσι πολλές φορές καταλήγοντας σε ένα μάλλον δυσκίνητο και αργό σύστημα. Τα συστήματα μοντέλων από την άλλη έχουν μια πιο εγγυημένη συμπεριφορά ανεξάρτητα από την χρησιμοποιούμενη βάση αλλά έχουν μια πολύ μεγάλη καθυστέρηση στο να λάβουν υπόψη όποιες αλλαγές συμβαίνουν στις προτιμήσεις των χρηστών. Τέλος είναι τα συστήματα αυτά που έχουν παρόμοια συμπεριφορά με τα collaborative filtering συστήματα αλλά χρησιμοποιούν αρκετά διαφορετικές προσεγγίσεις προκειμένου να επιτύχουν το σκοπό τους (βλέπε [2][16][34][35][58]). Για μια πιο λεπτομερή μελέτη και παρουσίαση collaborative filtering συστημάτων οι αναγνώστες μπορούν να ανατρέξουν στα [1][17][58].

Όπως σημειώσαμε και προηγουμένως ορισμένες λειτουργίες του συστήματος μας προσομοιάζουν ένα collaborative filtering σύστημα και πιο συγκεκριμένα ένα σύστημα βασισμένο στη μνήμη, με χρήση έμμεσης ψηφοφορίας. Παρόλ’ αυτά αυτές οι λειτουργίες του δικού μας συστήματος παρουσιάζουν ουσιώδεις διαφορές σε σχέση με όλες τις προσεγγίσεις collaborative filtering που προτάθηκαν μέχρι τώρα κυρίως λόγω του τρόπου με τον οποίο προσπαθούν να ανακαλύψουν τα αντικείμενα που θα αποτελέσουν την καλύτερη δυνατή πρόταση καθώς επίσης και λόγω της λογικής που χρησιμοποιούν. Έτσι κατ’ αρχήν αντίθετα με τα collaborative filtering συστήματα όπου το σύστημα δημιουργεί προφίλ χρηστών με βάση τα

υπάρχοντα δεδομένα και κατόπιν προσπαθεί να ταιριάζει τα προφίλ των νέων χρηστών με αυτά των παλιών, το σύστημα μας προσπαθεί να ταιριάζει οποιαδήποτε παλιά συναλλαγή με μια νέα. Με αυτό τον τρόπο το πρόβλημα των συνεχώς αυξανόμενων προφίλ χρηστών τα οποία και πολλές φορές δεν αντικατοπτρίζουν τις τρέχουσες τάσεις των χρηστών αποφεύγονται. Επίσης λίγη έως καθόλου δουλειά έχει γίνει σε συστήματα και περιβάλλοντα τα οποία πωλούν μια μεγάλη γκάμα διαφορετικών προϊόντων όπως τα καταστήματα λιανεμπορίου, αφού τα collaborative filtering συστήματα που έχουν δημιουργηθεί μέχρι σήμερα επικεντρώνονταν κύρια σε συστήματα με ομογενή αντικείμενα. Η δουλειά μας επίσης λαμβάνει υπόψη τις ειδικές μεταβλητές (στην περίπτωση μας κυρίως τις μικροοικονομικές) του περιβάλλοντος στο οποίο το σύστημα μας εφαρμόζεται. Τέλος κάνοντας χρήση καλά μελετημένων μεθόδων δανεισμένων από την Ανάκτηση Πληροφορίας, το σύστημα μας καταφέρνει να ξεπεράσει μεγάλα προβλήματα που εμφανίζονται στα συστήματα collaborative filtering όπως κλιμάκωση (scalability), προβλήματα απόδοσης και μνήμης που παρουσιάζονται κύρια στα συστήματα μνήμης όπως επίσης και την σύμφυτη στατική φύση των συστημάτων μοντέλων.

7.3 Το προτεινόμενο σύστημα

Το σύστημα το οποίο προτείνουμε παίρνει σαν είσοδο δεδομένα από τη βάση μας, δημιουργεί το αντίστοιχο **ευρετήριο (index)** και ακολούθως είναι έτοιμο να το χρησιμοποιήσει με διάφορους τρόπους. Το συγκεκριμένο σύστημα είναι μια ολοκληρωμένη λύση η οποία αποτελείται ουσιαστικά από δύο τμήματα (Εικόνα 14). Το πρώτο, στο οποίο θα αναφερόμαστε και σαν **batch module**, λειτουργεί ακριβώς όπως οι αλγόριθμοι εύρεσης κανόνων συσχέτισεων που δουλεύουν βήμα-βήμα (π.χ. ο Apriori), βρίσκοντας δηλαδή όλα τα συχνά αντικείμενα και παράγοντας όλους τους πιθανούς κανόνες. Το δεύτερο, στο οποίο θα αναφερόμαστε και σαν το **διαδραστικό τμήμα (interactive module)**, έχει μια περισσότερο αλληλεπιδραστική και on-line συμπεριφορά (βλέπε ενότητα 7.5). Και τα δύο τμήματα έχουν σαν βάση τους ένα ευρετήριο και λειτουργούν χρησιμοποιώντας ερωτήματα (δυαδικά ή ερωτήματα διάταξης), τεχνικές ευρέως χρησιμοποιούμενες στην ανάκτηση πληροφορίας. Ειδικά το δεύτερο τμήμα προσφέρει υπηρεσίες οι οποίες αποκλίνουν από το κλασσικό σενάριο που χρησιμοποιείται από όλες τις άλλες προσεγγίσεις. Πιο συγκεκριμένα αντί να διαβάσει όλα τα αντικείμενα και τους δυνατούς συνδυασμούς τους και να τα αξιολογεί με τον ένα ή τον άλλο τρόπο, κάνει προτάσεις με βάση μια προσέγγιση παρόμοια με αυτή που χρησιμοποιείται όταν θέτουμε ερωτήματα στις μηχανές αναζήτησης. Στην πράξη προσπαθεί να απαντήσει το παρακάτω ερώτημα: Ας υποθέσουμε ότι έχουμε μια βάση που περιέχει παλιότερες συναλλαγές, και ένας πελάτης ο οποίος πραγματοποιεί μια νέα συναλλαγή. Ποιά/ες συναλλαγή/ες καλύτερα προσεγγίζουν την συναλλαγή του συγκεκριμένου πελάτη, με βάση τα αντικείμενα που έχει ήδη επιλέξει;

Εικόνα 14: Η λειτουργία του συστήματος μας

7.4 Χρησιμοποιώντας την προσέγγισή μας σαν ένα κλασσικό αλγόριθμο Εξόρυξης Δεδομένων (Batch Module) INDEX DATA

Η **batch module** του συστήματος μας λειτουργεί όπως όλοι οι κλασσικοί αλγόριθμοι εύρεσης κανόνων συσχετίσεων, βρίσκοντας δηλαδή όλα τα συχνά αντικείμενα και δίνοντας στο χρήστη όλους τους παραγόμενους κανόνες συσχετίσεων. Σ' αυτή την ενότητα δίνουμε μόνο μια σύντομη παρουσίαση του πως λειτουργεί η συγκεκριμένη module, αφού ο κύριος σκοπός μας δεν ήταν να δώσουμε απλά μια ακόμα υλοποίηση ενός κλασσικού αλγόριθμου εξόρυξης δεδομένων που θα παρήγαγε ακριβώς τα ίδια αποτελέσματα, αλλά να δώσουμε μια νέα προοπτική στο πρόβλημα της εύρεσης κανόνων συσχετίσεων και των προσεγγίσεων που προτείνουν αντικείμενα στους χρήστες (δηλ. την διαδραστική module που παρουσιάζεται στην ενότητα 7.5).

Ας υποθέσουμε ότι έχουμε την βάση η οποία παρουσιάζεται στον Πίνακας 7. Πρώτ' απ' όλα η βάση διαβάζεται και ένα ευρετήριο, με χρήση ανεστραμμένων αρχείων, δημιουργείται. Κατόπιν το ευρετήριο αυτό συμπίεζεται και αποθηκεύεται στην κύρια μνήμη για γρήγορη πρόσβαση. Αυτό το βήμα είναι ισοδύναμο με το πρώτο βήμα των παραδοσιακών αλγορίθμων, όπου η βάση διαβάζεται μια φορά προκειμένου να βρούμε τα υποψήφια στοιχεία και να καταμετρήσουμε τις εμφανίσεις τους. Τα υποψήφια στοιχεία έχουν ήδη βρεθεί και απο την προσέγγισή μας, αφού το ευρετήριο αποθηκεύει κάθε ξεχωριστό στοιχείο μαζί με την

SYSTEM

Batch
moduleFULL SET OF
DISCOVERED
RULES

πληροφορία σε πόσες και ποιες συναλλαγές εμφανίζεται. Μετά από το πρώτο αυτό πέρασμα, η βάση δεν πρέπει να ψαχτεί ξανά. Κάθε μετέπειτα πέρασμα από τη βάση αντικαθίσταται με ένα ψάξιμο του ευρετηρίου, το οποίο αποθηκεύεται στην κύρια μνήμη (Εικόνα 15). Ας υποθέσουμε τώρα ότι ο χρήστης θέτει το κατώφλι υποστήριξης στο 2, δηλαδή ένα αντικείμενο θα πρέπει να εμφανίζεται σε τουλάχιστον δύο συναλλαγές για να θεωρηθεί συχνό. Μπορούμε να βρούμε άμεσα το σύνολο των συχνών στοιχείων L1 επιλέγοντας από το συμπιεσμένο ευρετήριο αυτά τα οποία έχουν αριθμό εμφανίσεων τουλάχιστον δύο. Προκειμένου να δημιουργήσουμε το σύνολο των υποψήφιων δισυνόλων συνδυάζουμε τα συχνά στοιχεία του L1 και δημιουργούμε το C2. Η υποστήριξη κάθε δισυνόλου στο C2 βρίσκεται θέτοντας διαδοχικά ερωτήματα στο ευρετήριο. Για παράδειγμα για το αντικείμενο AB θέτουμε το εξής ερώτημα: A AND B. Βρίσκουμε την τομή των λιστών των αντικειμένων A και B αντίστοιχα και η υποστήριξη (οι κοινές συναλλαγές των δύο λιστών) επιστρέφεται. Αυτό γίνεται συνέχεια μέχρις ότου για κάθε αντικείμενο στο C2 έχει γίνει το αντίστοιχο ερώτημα και έχει βρεθεί η υποστήριξη του, δημιουργώντας έτσι το L2 – το σύνολο που περιέχει όλα τα συχνά 2 αντικείμενα. Συνδυάζοντας τα συχνά 2-αντικείμενα δημιουργούμε το C3 – το σύνολο των υποψήφιων 3-αντικειμένων κοκ.

Πίνακας 7: Ένα παράδειγμα βάσης που χρησιμοποιούμε για datamining

DATABASE D	
TID	ITEMS
100	A C D
101	B C E
102	A B C E
103	B E

Ένα πλεονέκτημα της προσέγγισης μας είναι ότι η δημιουργία του ευρετηρίου είναι εντελώς αποσυνδεδεμένη από οποιοδήποτε στατιστικό μέτρο. Το ευρετήριο δημιουργείται ανεξάρτητα από το κατώφλι υποστήριξης ή οποιοδήποτε άλλο μέτρο, και δεν αλλάζει αν μεταβάλλουμε την τιμή του. Έτσι αν δεν είμαστε ικανοποιημένοι με το αποτέλεσμα, για παράδειγμα πάρα πολλά ή πολύ λίγα συχνά αντικείμενα δημιουργήθηκαν, μπορούμε σε οποιοδήποτε βήμα και σε οποιαδήποτε στιγμή να επανεκκινήσουμε την όλη διαδικασία. Το κόστος είναι πάρα πολύ μικρότερο σε σύγκριση με τις άλλες μεθόδους αφού τα πάντα είναι στην κύρια μνήμη και η βάση μας δεν πρέπει να ξαναδιαβαστεί. Αυτό που πρέπει να κάνουμε είναι απλά να ξαναθέσουμε τα ερωτήματα μας στο ευρετήριο βήμα-βήμα χρησιμοποιώντας το νέο κατώφλι. Χρησιμοποιώντας έναν κλασσικό αλγόριθμο, όπως για παράδειγμα τον Apriori [10], θα έπρεπε να περάσουμε ολόκληρη τη βάση πολλές φορές ξανά.

Εικόνα 15: Χρήση του συστήματος μας σαν ένας κλασσικός αλγόριθμος που δουλεύει σε βήματα

7.5 Διαδραστικό Module

Η βάση του **διαδραστικού module** είναι και πάλι το ίδιο ευρετήριο όπως στο batch module, το οποίο όμως χρησιμοποιούμε με διαφορετικό τρόπο. Το συγκεκριμένο module βασίζεται είτε σε δυαδικά ερωτήματα είτε σε ερωτήματα διάταξης, και όλες οι υπηρεσίες εκτός από μια έχουν σαν σκοπό να προτείνουν στους χρήστες τα αντικείμενα που είναι πιο πιθανό να αγοράσουν, με βάση τα αντικείμενα που βρίσκονται ήδη στο καλάθι τους. Δίνουμε πρώτα τις υπηρεσίες που παρέχονται από τα δυαδικά ερωτήματα και κατόπιν παρουσιάζουμε αυτές που παρέχονται από τα ερωτήματα διάταξης.

7.5.1 Υπηρεσίες που βασίζονται σε Δυναμικά Ερωτήματα

7.5.1.1 Εύρεση όλων των αντικειμένων που περιέχουν ένα συγκεκριμένο στοιχείο/α

Η πρώτη υπηρεσία που προτείνει αντικείμενα στους χρήστες λειτουργεί προσπαθώντας να βρει όλους τους συνδυασμούς ενός συγκεκριμένου αντικειμένου. Ας υποθέσουμε ότι έχουμε μια επιχείρηση που πουλάει προϊόντα μέσω κάποιου ηλεκτρονικού καταστήματος. Ένας πελάτης περιηγείται ανάμεσα στα προϊόντα και σε κάποιο σημείο τοποθετεί το προϊόν B στο καλάθι του. Επιθυμούμε να βρούμε όλους τους δυνατούς συνδυασμούς του B με όλα τα υπόλοιπα προϊόντα και να του παρουσιάσουμε τα πιο συχνά. Αν υπάρχουν N διαφορετικά αντικείμενα στη βάση μας, τότε μπορεί να υπάρχουν το πολύ N-1 συνδυασμοί του B με όλα τα άλλα αντικείμενα, τα οποία βρίσκουμε και του τα παρουσιάζουμε σε φθίνουσα σειρά συχνότητας εμφάνισης. Ας υποθέσουμε ότι ο ίδιος πελάτης βάζει τώρα το προϊόν F στο καλάθι του. Τώρα βρίσκουμε όλους τους δυνατούς συνδυασμούς του αντικειμένου BF με όλα τα υπόλοιπα αντικείμενα, τα οποία είναι το πολύ N-2. Όσο ο πελάτης συνεχίζει να βάζει προϊόντα στο καλάθι του συνεχίζουμε να βρίσκουμε όλους τους νέους συνδυασμούς. Στην Εικόνα 16 δείχνουμε όλους τους συνδυασμούς που θα πρέπει να πραγματοποιηθούν σε κάθε βήμα για την σειρά αγορών B – F – D – C, και για το σύνολο των προϊόντων (A, B, C, D, E, και F). Φυσικά η διαδικασία που προτάθηκε παραπάνω έχει ένα προφανές μειονέκτημα. Αν ο αριθμός των διαφορετικών στοιχείων είναι πάρα πολύ μεγάλος (για παράδειγμα χιλιάδες ή δεκάδες χιλιάδες) τότε όπως μπορούμε εύκολα να καταλάβουμε ο αριθμός των συνδυασμών που θα πρέπει να πραγματοποιηθούν κάθε φορά που ένα προϊόν εισέρχεται στο καλάθι αγορών μας είναι πολύ μεγάλος, και έτσι ο χρόνος και οι πόροι που απαιτούνται για να διαχειριστούμε κάθε αντικείμενο είναι υπερβολικός, κάνοντας την όλη διαδικασία μάλλον σχετικά δυσεφάρμοστη. Έτσι η διαδικασία που περιγράψαμε παραπάνω είναι ιδανική για ένα μικρό αριθμό διαφορετικών στοιχείων. Παρακάτω προτείνουμε μια τεχνική που αντιμετωπίζει αυτό το μειονέκτημα σε ένα μεγάλο βαθμό και βελτιώνει σημαντικά την απόδοση.

Αντί να διαβάζουμε όλα τα αντικείμενα και να τα συνδυάζουμε με αυτό ή αυτά που έχουμε ήδη διαλέξει, σταματάμε σε ένα σημείο όπου έχουμε βρει τα πιο συχνά. Ας υποθέσουμε ξανά ότι ένας πελάτης επιλέγει ένα αντικείμενο. Τότε ξεκινάμε να συνδυάζουμε το συγκεκριμένο αντικείμενο με τα αντικείμενα τα οποία είναι και τα πιο συχνά αφού αυτά παρουσιάζουν τις περισσότερες πιθανότητες να περιέχουν το συγκεκριμένο αντικείμενο, και κατόπιν με όλα τα άλλα αντικείμενα. Προκειμένου να το επιτύχουμε αυτό κρατάμε μια ακόμα λίστα με όλα τα αντικείμενα, εκτός από αυτή που χρησιμοποιούμε σαν λεξικό, όπου όλα τα αντικείμενα διατάσσονται σύμφωνα με τις συχνότητες εμφανίσεών τους σε φθίνουσα σειρά (εναλλακτικά το ίδιο το λεξικό θα μπορούσε να αποθηκευτεί σε διατεταγμένη μορφή). Συνήθως ένας

χρήστης ενδιαφέρεται για τρεις ή το πολύ τέσσερις προτάσεις προϊόντων. Τα πιο δημοφιλή ηλεκτρονικά καταστήματα που προτείνουν προϊόντα στους χρήστες τους (π.χ. το Amazon [66], το Egghead [68], το eBay [67] κ.α.) επίσης σταματάνε στις τέσσερις προτάσεις προϊόντων. Αν παρουσιάσουμε σε ένα χρήστη περισσότερα από τέσσερα προϊόντα σαν προτάσεις τότε αυτό γίνεται πληροφοριακή συμφόρηση και περισσότερο μπερδεύει παρά βοηθά. Φυσικά ο εκάστοτε χρήστης έχει την δυνατότητα να προσαρμόσει το σύστημα του στις ανάγκες του, και να καθορίσει τον αριθμό των αντικειμένων που θα προτείνονται στους πελάτες. Για τους σκοπούς μας περιορίσαμε τον συγκεκριμένο αριθμό επίσης στο τέσσερα. Έτσι ξεκινάμε συνδυάζοντας αντικείμενα και σταματάμε όταν έχουμε συλλέξει τα τέσσερα πιο συχνά. Το ερώτημα τώρα είναι πως μπορούμε να γνωρίζουμε ότι έχουμε βρει τα πιο συχνά. Αν βρούμε τέσσερα αντικείμενα με υποστήριξη ίση με αυτή του αντικειμένου μας και βεβαιότητα 100% τότε είμαστε σίγουροι ότι έχουμε βρει τουλάχιστον μερικά από τα πιο σημαντικά και δεν είναι ανάγκη να βρούμε και άλλα. Μπορεί να υπάρχουν και άλλα συχνά αντικείμενα αλλά μπορούν να έχουν υποστήριξη και βεβαιότητα το πολύ ίση με αυτή των αντικειμένων που βρήκαμε μέχρι τώρα και έτσι δεν θα προσθέσουν κάποια νέα πληροφορία στις ήδη υπάρχουσες προτάσεις αντικειμένων.

Εικόνα 16: Οι συνδυασμοί για την σειρά B – F – D – C.

Στην περίπτωση τώρα που δεν έχουμε συλλέξει τέσσερα τέτοια προϊόντα αυτό που κάνουμε είναι να ψάχνουμε κάθε φορά την υποστήριξη του επόμενου αντικειμένου στη λίστα σε σχέση με την υποστήριξη του τελευταίου αντικειμένου που δημιουργήσαμε συνδυάζοντας τα αντικείμενα που βρίσκονται ήδη στο καλάθι μας με τα εναπομείναντα. Αν η συγκεκριμένη υποστήριξη είναι μεγαλύτερη από του επομένου και ως εκ τούτου και από αυτή όλων των επόμενων αντικειμένων τότε έχουμε βρει όλα τα συχνότερα αντικείμενα που ψάχναμε και μπορούμε να σταματήσουμε σε αυτό το σημείο. Έτσι η παρακάτω εξίσωση πρέπει να είναι αληθής προκειμένου ο αλγόριθμος μας να σταματήσει:

$$\text{support}(I_n, I_k) > \text{support}(I_n) \text{ Εξίσωση 3}$$

όπου I_x είναι τα αντικείμενα ήδη στο καλάθι μας, I_k είναι το αντικείμενο στο διατεταγμένο λεξικό μας το οποίο εξετάζουμε αυτή τη στιγμή και I_n είναι το επόμενο αντικείμενο στο λεξικό. Ο αλγόριθμος για να βρίσκουμε τα k πιο συχνά αντικείμενα ανάμεσα σε όλα τα αντικείμενα δίνεται στην Εικόνα 17.

1.	Algorithm Find_k_TopItemsets(itemset(X), k)
2.	N=number of 1-itemsets
3.	begin
4.	while $i \leq k$
5.	if $itemset(i) \notin X$
6.	$top_list(i) = itemset(X_I);$
7.	$i = i + 1;$
8.	end;
9.	end;
10.	sort (top_list);
11.	$minsup = sup[itemset(k) \text{ in } top_list]$
12.	m =i-1;
13.	while $minsup < sup[itemset(m+1)]$
14.	if $itemset(i) \notin X$
15.	Query the index [itemset(X) AND itemset(m)]
16.	if $support[itemset(X_m)] > minsup$
17.	insert itemset(X_m) into top_LIST;
18.	sort top_LIST;
19.	end;
20.	end;
21.	m =m+1;
22.	end;
23.	end;

Εικόνα 17: Βρίσκοντας τα k πιο συχνά αντικείμενα

Ο αλγόριθμος Find_k_TopItemsets ξεκινάει συνδυάζοντας τα αντικείμενα που βρίσκονται ήδη στο καλάθι μας με τα πρώτα k αντικείμενα στην διατεταγμένη λίστα με βάση την συχνότητα, τα οποία επίσης δεν περιλαμβάνονται επίσης ήδη στο καλάθι μας. Τα αντικείμενα που δημιουργούνται από αυτούς τους συνδυασμούς αποθηκεύονται σε ένα πίνακα, τον οποίο διατάσσουμε σύμφωνα με την υποστήριξη τους σε φθίνουσα σειρά. Ακολουθώντας προχωράμε με τα επόμενα αντικείμενα στη διατεταγμένη λίστα. Η υποστήριξη κάθε νέου αντικειμένου που δημιουργείται βρίσκεται και συγκρίνεται με αυτή του τελευταίου αντικειμένου στον πίνακα. Αν βρεθεί μικρότερη τότε το τελευταίο αντικείμενο αντικαθίσταται από ένα νέο και ο πίνακας επαναδιατάσσεται. Ο αλγόριθμος σταματάει όταν το τελευταίο αντικείμενο που εισήλθε στη

λίστα μας βρεθεί να έχει υποστήριξη μεγαλύτερη από αυτή του επόμενου αντικειμένου στην διατεταγμένη λίστα σύμφωνα με την συχνότητα. Η λειτουργία του αλγορίθμου αυτού μπορεί να φανεί καλύτερα στο παρακάτω παράδειγμα:

Παράδειγμα 1: Ας υποθέσουμε ότι έχουμε την παρακάτω λίστα αντικειμένων διατεταγμένη σε φθίνουσα σειρά εμφανίσεων (Πίνακας 8) και ένας πελάτης τοποθετεί το προϊόν D στο καλάθι του.

Πίνακας 8: Μια διατεταγμένη λίστα αντικειμένων

Αντικείμενο	Υποστήριξη
W	10
B	8
C	6
A	5
D	4
G	3
Z	2

Ξεκινάμε συνδυάζοντας το στοιχείο D με το στοιχείο W, μετά με το στοιχείο B κ.ο.κ. Ας υποθέσουμε τώρα ότι τα αντικείμενα DW και DB έχουν και τα δύο υποστήριξη και βεβαιότητα 100%, το αντικείμενο DC έχει υποστήριξη 0, και το στοιχείο A εμφανίζεται με το στοιχείο D σε τέσσερις συναλλαγές. Τότε δεν υπάρχει ανάγκη να ψάξουμε και να συνδυάσουμε άλλα αντικείμενα αφού δεν υπάρχει περίπτωση τα υπόλοιπα αντικείμενα να είναι πιο συχνά (το αντικείμενο G όπως και όλα τα υπόλοιπα αντικείμενα μπορούν να εμφανιστούν το πολύ σε τόσες συναλλαγές όσες είναι και ο αριθμός των εμφανίσεων τους) και ήδη έχουμε συλλέξει τα τρία πιο συχνά αντικείμενα.

Παρά την παραπάνω βελτίωση η συγκεκριμένη προσέγγιση δεν έχει δυστυχώς εγγυημένη απόδοση και συμπεριφορά. Μπορεί να χρειαστούμε έτσι και αλλιώς να ελέγξουμε όλα τα στοιχεία στη λίστα, το οποίο θα χρειαστεί τόσους συνδυασμούς όσους και ο αριθμός των διαφορετικών στοιχείων μέχρις ότου να καταλήξουμε στα πιο συχνά⁺⁺⁺. Παρακάτω παρουσιάζουμε μια εναλλακτική προσέγγιση η οποία αντιμετωπίζει αυτή την πιθανότητα και έχει επίσης εγγυημένη απόδοση.

⁺⁺⁺ Βέβαια κάτι τέτοιο είναι πρακτικά εξαιρετικά δύσκολο να συμβεί αλλά θεωρητικά υπάρχει αυτή η πιθανότητα.

7.5.1.2 Προτείνοντας αντικείμενα με βάση παρόμοιες συναλλαγές

Η δεύτερη προσφερόμενη υπηρεσία, προσπαθεί επίσης να κάνει προτάσεις στους πελάτες αλλά χρησιμοποιεί μια διαφορετική προσέγγιση για την αξιολόγηση των αντικειμένων που είναι πιο πιθανό να αγοραστούν. Ας υποθέσουμε ξανά ότι ένας πελάτης επιλέγει το αντικείμενο B. Αυτό δίνεται σαν ένα δυαδικό ερώτημα στο σύστημα μας το οποίο επιστρέφει όλες τις συναλλαγές που αποτελούν την ανεστραμμένη λίστα του συγκεκριμένου αντικειμένου. Όλες οι συναλλαγές διαβάζονται και ακολούθως μπορούμε να ακολουθήσουμε δύο διαφορετικές τεχνικές με αυτές.

Μια θα ήταν αν μετρούσαμε τις εμφανίσεις κάθε αντικειμένου στις συγκεκριμένες συναλλαγές και ακολούθως προτείναμε αυτές με τις περισσότερες εμφανίσεις. Για παράδειγμα ας υποθέσουμε ότι το στοιχείο B εμφανίζεται στις συναλλαγές 4, 7, 10 και 16, οι οποίες φαίνονται στον Πίνακα 9.

Πίνακας 9: Τα αντικείμενα που περιέχονται στις συναλλαγές 4, 7, 10 και 16

TID	Itemsets
4	B, C, E
7	A, B, E, F
10	A, B, E
16	B, D

Όπως μπορούμε να δούμε το αντικείμενο E είναι το συχνότερα εμφανιζόμενο αντικείμενο σε όλες τις συναλλαγές που περιέχουν επίσης το B (εμφανίζεται σε τρεις συναλλαγές) και έτσι προτείνεται πρώτο στο χρήστη. Κατόπιν το αντικείμενο A είναι το δεύτερο συχνότερα εμφανιζόμενο αντικείμενο (εμφανίζεται δύο φορές) και έτσι προτείνεται στον χρήστη αμέσως μετά το αντικείμενο E. Τα υπόλοιπα αντικείμενα (τα αντικείμενα C, F και D) εμφανίζονται όλα μόλις μια φορά και θα μπορούσαν να προταθούν με οποιαδήποτε σειρά.

Προτείνοντας πρώτα τα αντικείμενα που εμφανίζονται τις περισσότερες φορές είναι λογικά και στατιστικά σωστό αφού αυτά τα αντικείμενα παρουσιάζουν τις μεγαλύτερες πιθανότητες να αγοραστούν τελικά από ένα πελάτη. Αλλά τι συμβαίνει όταν όλα ή ακόμα και ορισμένα από τα αντικείμενα έχουν τον ίδιο αριθμό εμφανίσεων όπως συνέβη προηγουμένως; Το να τα προτείναμε σύμφωνα με τον κωδικό αριθμό τους (bar code) ή σύμφωνα με το ονομά τους θα ήταν η πιο απλοϊκή και εύκολη λύση, αλλά εμείς επιθυμούμε να κάνουμε την καλύτερη δυνατή πρόταση. Αφού όλα τα αντικείμενα εμφανίζονται σε ακριβώς τον ίδιο αριθμό συναλλαγών όλα παρουσιάζουν τις ίδιες πιθανότητες να αγοραστούν. Έτσι πρέπει να προτείνουμε τα στοιχεία με τη σειρά που τα θεωρούμε πιο σημαντικά για εμάς. Θα μπορούσαμε να προτείνουμε στον πελάτη πρώτα τα στοιχεία που αποφέρουν μεγαλύτερο κέρδος ή γενικότερα αυτά που έχουν τη

μεγαλύτερη βαρύτητα. Φυσικά αυτό προϋποθέτει ότι όλα τα αντικείμενα θα έχουν λάβει κάποιο βάρος ανάλογα με το πόσο σημαντικά θεωρούνται.

Μια άλλη δυνατότητα θα ήταν αν χρησιμοποιούσαμε αυτό που ορίζουμε σαν συντελεστή συσχέτισης. Ο συντελεστής συσχέτισης είναι ένα μέτρο που μας δείχνει πόσο συσχετισμένα είναι κάποια αντικείμενα, και ορίζεται ως εξής:

Ορισμός 1: Συντελεστής Συσχέτισης - (Correlation Coefficient). Έστω $I = \{i_1, i_2, \dots, i_m\}$ το σύνολο όλων των αντικειμένων στη βάση μας. Ας υποθέσουμε ότι έχουμε ένα σύνολο αντικειμένων $T = \{i_1, i_2, \dots, i_k\}$ τα οποία αποτελούν τα αντικείμενα που αγόρασε ένας πελάτης μέχρι τώρα έτσι ώστε $T \subseteq I$, και $\text{sup}(T)$ είναι η υποστήριξη του συγκεκριμένου αντικειμένου. Ο συντελεστής συσχέτισης ανάμεσα στο αντικείμενο T και κάθε υποψήφιο αντικείμενο i_c το οποίο είναι επίσης παρών στις συναλλαγές όπου το T είναι παρών ορίζεται σαν ο λόγος ανάμεσα στην υποστήριξη του αντικειμένου T δια της υποστήριξης του αντικειμένου που εξετάζουμε σε ολόκληρη την βάση (το οποίο δείχνουμε σαν $\text{sup}_{total}(i_c)$):

$$cc(i_1, i_2, \dots, i_k \mapsto i_c) = \frac{\text{sup}(i_1, i_2, \dots, i_k)}{\text{sup}_{total}(i_c)} \quad \text{Εξίσωση 4}$$

Όσο υψηλότερη είναι η τιμή του συγκεκριμένου συντελεστή τόσο μεγαλύτερη είναι η σχέση ανάμεσα στα θεωρούμενα αντικείμενα. Ας δούμε πως χρησιμοποιείται αυτός ο συντελεστής μέσα από ένα παράδειγμα.

Παράδειγμα 2: Ας υποθέσουμε ότι έχουμε μια βάση η οποία αποτελείται από 100 συναλλαγές και ένας πελάτης διαλέγει το αντικείμενο E το οποίο εμφανίζεται στις ακόλουθες 4 συναλλαγές (Πίνακας 10). Ελέγχοντας τώρα αυτές τις συναλλαγές βρίσκουμε ότι περιέχουν επίσης τα αντικείμενα A , B , F και G , όπου το στοιχείο F εμφανίζεται 3 φορές και τα υπόλοιπα 2 δύο φορές το καθένα. Το στοιχείο F θα προταθεί πρώτο αφού εμφανίζεται τις περισσότερες φορές. Μένει τώρα να προτείνουμε τα υπόλοιπα αντικείμενα.

Πίνακας 10: Οι συναλλαγές που περιέχουν το στοιχείο E

TID	Itemsets
3	A, B, E, F
9	B, E, F
77	A, E, F, G
106	E, G

Ας υποθέσουμε ότι οι συνολικές υποστηρίξεις για αυτά τα αντικείμενα είναι:

$$\text{sup}_{\text{total}}(A)=10, \text{sup}_{\text{total}}(B)=50, \text{sup}_{\text{total}}(G)=3$$

η συνολική υποστήριξη για κάθε στοιχείο είναι άμεσα διαθέσιμη, αφού είτε στο λεξικό ή στην αντίστοιχη ανεστραμμένη λίστα κάθε στοιχείου αποθηκεύουμε επίσης τον αριθμό των εμφανίσεων του. Έτσι ο συντελεστής συσχέτισης για κάθε αντικείμενο θα είναι:

$$cc(E \mapsto A) = \frac{\text{sup}(E)}{\text{sup}_{\text{total}}(A)} = \frac{4}{10} = 0.4,$$

$$cc(E \mapsto B) = 0.08 \text{ και } cc(E \mapsto G) = 1.33.$$

Βλέπουμε ότι παρά το γεγονός ότι όλα τα αντικείμενα μοιράζονται τον ίδιο αριθμό εμφανίσεων το στοιχείο G είναι αυτό που σχετίζεται πιο στενά με το στοιχείο E αφού όλες οι εμφανίσεις του λαμβάνουν χώρα με το στοιχείο E, και έτσι θα πρέπει να προταθεί πρώτο. Το στοιχείο B από την άλλη εμφανίζεται τις ίδιες φορές με το στοιχείο E, αλλά επίσης εμφανίζεται στις μισές συναλλαγές σε ολόκληρη την βάση μας και έτσι δεν είναι τόσο στενά σχετισμένο με το στοιχείο E. Έτσι προτείνεται τελευταίο.

Η ίδια διαδικασία ακολουθείται κάθε φορά που ένα νέο αντικείμενο προστίθεται στο καλάθι ενός καταναλωτή, αλλά με μειούμενο αριθμό κοινών συναλλαγών (βλέπε ενότητα 3.2). οι προτάσεις σταματούν όταν δεν υπάρχουν άλλες κοινές συναλλαγές για τον συγκεκριμένο συνδυασμό αντικειμένων.

7.5.1.3 Χρήση Δυαδικών Ερωτημάτων συμπληρωματικά προς τους Κλασσικούς Αλγόριθμους

Μια άλλη πιθανή χρήση της συγκεκριμένης προσέγγισης θα ήταν συμπληρωματικά προς τους κλασσικούς αλγόριθμους εύρεσης κανόνων συσχέτισεων (π.χ. Apriori [10], DHP [98], DIC [27] κ.α.), παρά για να κάνει προτάσεις στους χρήστες όπως προηγούμενα. Πιο συγκεκριμένα θα μπορούσαμε να τρέξουμε κάποιο αλγόριθμο εύρεσης κανόνων, να πάρουμε τους παραγόμενους κανόνες και ακολούθως να πραγματοποιήσουμε μια περαιτέρω εξέταση χρησιμοποιώντας δυαδικά ερωτήματα.

Έτσι το προτεινόμενο σύστημα μας είναι σε θέση να δώσει απαντήσεις σε οποιαδήποτε τύπο ερωτήματος που περιλαμβάνει οποιονδήποτε από τους τρεις κύριους δυαδικούς τελεστές AND – OR – NOT. Για παράδειγμα μπορεί να δώσει απαντήσεις σε ερωτήματα όπως πόσοι χρήστες αγόρασαν το προϊόν A και το προϊόν B, ή πόσοι αγόρασαν το προϊόν C ή το προϊόν D. Επιπρόσθετα το σύστημα μας είναι σε θέση να θέτει και να επιλύει οποιαδήποτε ερώτημα περιλαμβάνει συνδυασμούς οποιασδήποτε μορφής των τριών κύριων τελεστών. Για παράδειγμα πόσοι αγόρασαν το προϊόν C και A αλλά όχι το προϊόν B. Αυτό δίνει απόλυτη ελευθερία στον χρήστη, αφού του επιτρέπει να παίρνει συγκεκριμένα αποτελέσματα για συγκεκριμένα αντικείμενα με έναν τρόπο που είναι πιο φυσικός και πιο απλός προς αυτόν, και μοιάζει με μια μηχανή αναζήτησης. Τέλος η χρήση ενός ευρετηρίου του επιτρέπει την επίλυση όλων των μορφών ερωτημάτων πρακτικά άμεσα και πιο σημαντικά με ελάχιστους πόρους.

7.5.2 Υπηρεσίες που βασίζονται σε Ερωτήματα Διάταξης

Ξεκινάμε δίνοντας την λογική της δεύτερης λύσης μας η οποία βασίζεται επίσης σε ένα ευρετήριο αλλά σε ερωτήματα διάταξης αντί σε δυαδικά ερωτήματα όπως προηγούμενως, και συνεχίζουμε παρουσιάζοντας την λύση αυτή καθ' εαυτή. Τέλος παρουσιάζουμε ορισμένα θέματα σχετικά με μια καλύτερη υλοποίηση του ευρετηρίου για την συγκεκριμένη λύση.

7.5.2.1 Σημαντικότητα των Αντικειμένων

Οι περισσότερες προσεγγίσεις μέχρι στιγμής στηρίζονταν στην μετρική της υποστήριξης για την εύρεση όλων των συχνών αντικειμένων, και υπέθεταν όλα τα αντικείμενα είναι της ίδιας δυαδικής φύσης, δηλ. είτε εμφανίζονται είτε όχι σε μια συναλλαγή. Δεν έδιναν οποιοδήποτε άλλο μέτρο σημαντικότητας στα αντικείμενα ούτε λάμβαναν υπόψη τον αριθμό των εμφανίσεων του ίδιου αντικειμένου σε μια συναλλαγή. Εμείς διαφωνούμε ότι μόνο ο αριθμός των συνολικών εμφανίσεων ενός αντικειμένου πρέπει να είναι το μοναδικό κριτήριο της σημαντικότητας του. Αν λάβουμε υπόψη για παράδειγμα την περίπτωση ενός supermarket, οι πωλήσεις ενός αντικειμένου μας λένε μόνο μέρος της αλήθειας για το προϊόν αυτό καθ' εαυτό. Το υπόλοιπο μέρος έχει να κάνει με το συνολικό κέρδος που μας αποφέρει όπως επίσης και με τον αριθμό των φορών που εμφανίζεται σε κάθε συναλλαγή. Προϊόντα όπως οι τσίχλες εμφανίζονται σχεδόν σε κάθε συναλλαγή αλλά αν κοιτάξουμε τα συνολικά τους κέρδη θα δούμε ότι δεν είναι και τόσο μεγάλα. Από την άλλη προϊόντα όπως τα κρασιά δεν πωλούνται τόσο συχνά αλλά τα περιθώρια κέρδους τους είναι πολύ μεγαλύτερα, και έτσι τα συνολικά κέρδη που αποφέρουν είναι πολλές φορές μεγαλύτερα. Επίσης ορισμένα προϊόντα (π.χ. οι μπύρες) γενικά αγοράζονται σε μεγάλες ποσότητες σε μια συναλλαγή και όχι σε τακτά χρονικά διαστήματα και σε μικρότερες ποσότητες όπως για παράδειγμα το γάλα. Λαμβάνοντας υπόψη μόνο αν τα στοιχεία εμφανίζονται ή όχι σε μια συναλλαγή και αδιαφορώντας για την

πραγματική τους αξία - σημαντικότητα, οποιαδήποτε και αν είναι αυτή, όπως επίσης για τις ακριβείς ποσότητες τους μπορεί να μας οδηγήσει σε λανθασμένα συμπεράσματα και προτάσεις.

Διάφορες εργασίες προσπάθησαν να αντιμετωπίσουν τα προβλήματα που συνεπάγεται η χρήση ενός ενιαίου κατωφλίου υποστήριξης για όλα τα αντικείμενα σε μια βάση. Μια ιδιαίτερα επιτυχημένη δουλειά ήταν και αυτή στο [89], όπου σε κάθε αντικείμενο αποδιδόταν μια διαφορετική τιμή υποστήριξης προκειμένου να αντικατροπτίσει τις διαφορές στη φύση και τις μεγάλες διακυμάνσεις στις συχνότητες εμφανίσεων των στοιχείων στη βάση. Έτσι αν ένα αντικείμενο θεωρείται πιο σημαντικό από τα υπόλοιπα τότε του αποδίδουμε μια χαμηλότερη τιμή υποστήριξης. Αυτό θα συμβαίνει με όλα τα ενδιαφέροντα/σημαντικά αντικείμενα. Παρόλα αυτά και αυτή η προσέγγιση δεν λάμβανε υπόψη τον αριθμό των φορών όπου ένα αντικείμενο εμφανιζόταν σε μια συναλλαγή αλλά τα θεωρούσε σαν δυαδικής φύσεως (είτε εμφανίζονται είτε όχι σε μια συναλλαγή) και το πιο σημαντικό είχε έναν μάλλον αυθαίρετο και διφορούμενο τρόπο για να αποδίδει υποστηρίξεις (multiple minimum supports – MIS) στα αντικείμενα. Στο [32] προτάθηκε μια εναλλακτική στην ευρέως χρησιμοποιούμενη μετρική της υποστήριξης, η οποία λάμβανε υπόψη την ιδιαίτερη σημασία που ενδεχομένως είχε κάποιο στοιχείο. Σύμφωνα με αυτή την προσέγγιση σε κάθε αντικείμενο αποδιδόταν ένα διαφορετικό βάρος το οποίο αναπαριστούσε την σημασία του. Ο πολλαπλασιασμός αυτού του βάρους με την υποστήριξη του προϊόντος μας έδινε την μετρική weighted support. Αν η weighted support ενός στοιχείου ήταν πάνω από ένα όριο το οποίο καθόριζε ο χρήστης (το οποίο ονομάζεται weighted minimum support – wminsup), τότε αυτό το αντικείμενο θεωρείτο συχνό. Η συγκεκριμένη προσέγγιση αντιμετώπιζε επιτυχώς το πρόβλημα της απόδοσης κάποιου μέτρου σημαντικότητας σε κάθε αντικείμενο, αλλά υπέφερε από διάφορα άλλα μειονεκτήματα όπως η διαδικασία δημιουργίας και περικοπής υποψηφίων αντικείμενων η οποία αναπόφευκτα είχε μια μεγάλη επίδραση στην τελική απόδοση. Επίσης ούτε και αυτή η προσέγγιση δεν λάμβανε υπόψη το πλήθος των εμφανίσεων ενός στοιχείου σε μια συναλλαγή.

Με την προσέγγισή μας καταφέρνουμε να λάβουμε υπόψη τόσο την σημασία ενός αντικειμένου όπως αυτή εκφράζεται με την μορφή κάποιου βάρους, όσο και τον αριθμό των εμφανίσεων των αντικείμενων σε κάθε συναλλαγή και να παρουσιάσουμε στους χρήστες ένα σύστημα που λειτουργεί σαν μια μηχανή αναζήτησης. Αυτό που είναι όμως πιο σημαντικό είναι πως με την προσέγγισή μας αυτές οι παράμετροι λαμβάνονται υπόψη τόσο με τις υπάρχουσες συναλλαγές όσο και για οποιαδήποτε νέα συναλλαγή πραγματοποιήσει κάποιος χρήστης. Η πρότασή μας είναι ιδιαίτερα κατάλληλη όταν δουλεύουμε με βάρη στα αντικείμενα ή με αντικείμενα με περισσότερα από ένα χαρακτηριστικά.

7.5.2.2H προσέγγιση μας

Η προσέγγιση μας, η οποία δουλεύει επίσης σε τρεις φάσεις, μοιάζει όπως είπαμε και προηγούμενα με το μοντέλο του ανυσματικού χώρου που χρησιμοποιείται στην ανάκτηση πληροφορίας. Έτσι, κατ' αρχήν ένα βάρος αποδίδεται σε κάθε στοιχείο, όλες οι υπάρχουσες συναλλαγές διαβάζονται από την βάση μας και δημιουργείται ένα ευρετήριο. Κατόπιν κάθε νέα συναλλαγή μεταφράζεται σαν ένα ερώτημα διάταξης και τροφοδοτείται στο σύστημα μας. Τέλος το σύστημα βρίσκει τις πιο σχετικές συναλλαγές σε οποιαδήποτε νέα συναλλαγή και οι πελάτες λαμβάνουν κάποιες προτάσεις (προϊόντων – αντικειμένων) για να διαλέξουν. Ας δούμε πως δουλεύει:

7.5.2.2.1 Απόδοση Βαρών στις Υπάρχουσες συναλλαγές

Ας υποθέσουμε ότι έχουμε ένα σύνολο συναλλαγών $T = \{t_1, t_2, t_3, \dots, t_n\}$, τα οποία και αποτελούν τη βάση μας και ένα σύνολο στοιχείων $I = \{i_1, i_2, i_3, \dots, i_n\}$. Κάθε συναλλαγή t είναι ένα σύνολο στοιχείων τέτοιο ώστε $T \subseteq I$. Αναπαριστούμε αυτά τα στοιχεία σαν ένα Ευκλείδειο χώρο m -διαστάσεων όπου κάθε διάσταση αντιστοιχεί σε ένα στοιχείο. Ακολουθώντας κάθε συναλλαγή αναπαριστάται σαν ένας άξονας σε αυτό το χώρο. Η συντεταγμένη μιας συναλλαγής t στην κατεύθυνση ενός στοιχείου i εξαρτάται από δύο ποσότητες:

Βάρος στοιχείου: Το βάρος αυτό προσπαθεί να αναπαραστήσει την σημαντικότητα που δίνει ένας χρήστης σε ένα στοιχείο i . Ένα σημαντικό στοιχείο θα απολαμβάνει ένα μεγάλο βάρος ενώ ένα ασήμαντο ένα μικρότερο βάρος. Η εύρεση αυτού του βάρους εξαρτάται από την εφαρμογή καθώς και σε άλλες παραμέτρους όπως η επιθυμητή έξοδος ή η φύση των δεδομένων μας. Στις ελάχιστες εργασίες που απέδιδαν βάρη στα στοιχεία [32], τόσο η εύρεση τους όσο και η απόδοση τους ήταν τελείως αυθαίρετη ή δεν αντιμετωπίζονταν καθόλου. Δεν υπάρχει κάποιος γενικός τύπος ή μέθοδος για να βρίσκουμε αυτές τις τιμές, ενώ η χρήση τεχνικών που χρησιμοποιούνταν σε άλλες περιπτώσεις μπορούν να μας οδηγήσουν σε λανθασμένα συμπεράσματα. Για παράδειγμα αν δανειστούμε και χρησιμοποιήσουμε την αρχή του Zipf που χρησιμοποιείται στις βάσεις αρχείων κειμένων, στην περίπτωση θα ήταν εντελώς λανθασμένο μας αφού σε καμία περίπτωση δεν θα πρέπει ένα πολύ συχνό αντικείμενο να θεωρείται ασήμαντο.

Το βάρος αυτό έχει μια παρόμοια λειτουργία όπως ο παράγοντας IDF στο TFIDF σχήμα, ή με άλλα λόγια να μειώνει την επιρροή ασήμαντων στοιχείων, ή να αυξάνει αυτή των σημαντικών στοιχείων. Αν σκεφτούμε και πάλι την περίπτωση ενός supermarket όπου η πρώτη μας προτεραιότητα είναι τα συνολικά κέρδη που έχουμε από κάθε προϊόν θα μπορούσαμε να χρησιμοποιήσουμε τον παρακάτω τύπο:

$$w_i = \ln \left(1.718 + \frac{f_i \cdot PM_i}{\max_k TP_k} \right)$$

Όπου $\max TP_k$ είναι τα μέγιστα συνολικά κέρδη που έχουμε από ένα προϊόν στη βάση μας, f_i είναι ο αριθμός εμφανίσεων ενός προϊόντος i και PM είναι το περιθώριο κέρδους ενός προϊόντος i . Στην πράξη αυτό σημαίνει ότι μια εμφάνιση ενός πολύ κερδοφόρου στοιχείου αξίζει πολύ περισσότερο από τις πούμε 10 εμφανίσεις ενός όχι και τόσο κερδοφόρου.

Intra Transaction Itemset Frequency: Το intra transaction itemset frequency στοιχείο $itf_{t,i}$ στην ουσία ελέγχει πως θα αξιολογούνται οι εμφανίσεις των αντικειμένων μέσα στις ίδιες τις συναλλαγές. Αυτός ο αριθμός έχει παρόμοια λειτουργία όπως ο παράγοντας TF του TFIDF σχήματος. Όπως είπαμε προηγουμένως με την προσέγγιση μας θέλουμε να λαμβάνουμε υπόψη των αριθμό των εμφανίσεων ενός στοιχείου σε κάθε συναλλαγή. Μια απλή λύση θα ήταν αν θέταμε αυτή τη μεταβλητή ίση με τον αριθμό των φορών που εμφανίζεται το συγκεκριμένο στοιχείο σε μια συναλλαγή ($itf_{t,i} = f_{t,i}$). Παρόλα αυτά αυτό έχει το προφανές μειονέκτημα ότι συναλλαγές με πάρα πολλές εμφανίσεις ενός πολύ σημαντικού στοιχείου ή γενικότερα πολύ μεγάλες συναλλαγές θα βαθμολογούνται πάντα πρώτες. Αυτό δεν είναι απαραίτητα εντελώς κακό, αλλά επιθυμούμε να μειώσουμε αυτό το φαινόμενο. Στην πράξη θέλουμε να λαμβάνουμε υπόψη τις ποσότητες των στοιχείων που αγοράζονται σε μια συναλλαγή, αλλά ταυτόχρονα να μην ευνοούμε τις εμφανίσεις αυτές πάρα πολύ σε σχέση με τα υπόλοιπα στοιχεία. Για αυτό το λόγο θα χρησιμοποιήσουμε την ακόλουθη εξίσωση, όπου η πρώτη εμφάνιση ενός στοιχείου μετράει πολύ περισσότερο από τις επόμενες.

$$itf_{t,i} = \left(K + (1 - K) \frac{f_{t,i}}{\max_c f_{t,c}} \right) \quad \text{Εξίσωση 5}$$

Η μεταβλητή K είναι ένας ρυθμιστικός παράγοντας που ελέγχει την ισορροπία ανάμεσα στην πρώτη και στις επόμενες εμφανίσεις ενός στοιχείου σε μια συναλλαγή, με βέλτιστες τιμές ανάμεσα στο 0.3 και στο 0.5 [52]. Ο παράγοντας $\max_c f_{t,c}$ είναι η μέγιστη συχνότητα εμφανίσεων ενός οποιουδήποτε στοιχείου σε μια συναλλαγή t . Έτσι το διάνυσμα μιας συναλλαγής t στην κατεύθυνση ενός στοιχείου i υπολογίζεται ως εξής:

$$w_{t,i} = itf_{t,i} \cdot w_i$$

Η διαδικασία που περιγράψαμε παραπάνω είναι στην ουσία η δημιουργία ενός ευρετηρίου από μια συλλογή δεδομένων (στην περίπτωση μας οι συναλλαγές μας) αλλά αποδίδοντας βάρος σε όλα τα στοιχεία. Το ευρετήριο αυτό μπορεί να χρησιμοποιηθεί για την πραγματοποίηση διαφόρων τύπων ερωτημάτων, όπως θα δούμε και στην επόμενη ενότητα.

7.5.2.2 Αξιολόγηση Νέων Συναλλαγών

Μέχρι στιγμής καταφέραμε να αναπαραστήσουμε όλες τις συναλλαγές που ήδη υπάρχουν στη βάση μας σαν διανύσματα σε έναν πολυδιάστατο χώρο. Το ερώτημα τώρα είναι πως θα μπορούσαμε να χρησιμοποιήσουμε αυτή την πληροφορία και τι συμβαίνει με κάθε νέα συναλλαγή. Όταν ένας πελάτης αγοράζει ένα προϊόν ή γενικότερα όταν ένας χρήστης μια συλλογής αντικειμένων επικεντρώνει το ενδιαφέρον του σε ένα αντικείμενο μέσα στη συλλογή, τότε αυτό δίνεται σαν ένα ερώτημα στο σύστημα μας. Ας υποθέσουμε για παράδειγμα ότι έχουμε ξανά την περίπτωση ενός σουπερμάρκετ, και ένας πελάτης τοποθετεί ένα προϊόν στο καλάθι του. Τότε θα ενδιαφερόμαστε για τη συναλλαγή ή τις συναλλαγές που καλύτερα προσεγγίζουν τις επόμενες αγορές του συγκεκριμένου πελάτη. Φυσικά είναι σχετικά δύσκολο να προβλέψουμε αυτό από μόλις ένα προϊόν αλλά καθώς ο ίδιος πελάτης θα συνεχίσει να τοποθετεί περισσότερα προϊόντα στο καλάθι του οι απαντήσεις θα γίνονται συνεχώς όλο και πιο ακριβείς. Κάθε ερώτημα (δηλ. κάθε νέα συναλλαγή) μεταφράζεται σαν μια συναλλαγή και μετατρέπεται σε ένα διάνυσμα στον ίδιο χώρο. Το διάνυσμα ενός ερωτήματος q στην κατεύθυνση ενός στοιχείου i υπολογίζεται ως εξής:

$$w_{q,i} = itf_{q,i} \cdot w_i$$

Το $itf_{q,i}$ στοιχείο δίνεται από την ίδια εξίσωση όπως το $itf_{i,i}$ στοιχείο παραπάνω (Εξίσωση 5). Χρησιμοποιώντας αυτή την εξίσωση καταφέρνουμε να λάβουμε υπόψη τον αριθμό των εμφανίσεων κάθε αντικειμένου σε μια νέα συναλλαγή όπως επίσης και το βάρος του, με τον ίδιο τρόπο που συνέβαινε για τις συναλλαγές ήδη στη βάση μας, και έτσι καταφέρνουμε να βρίσκουμε με μεγαλύτερη ακρίβεια τις πιο σχετικές συναλλαγές.

Το τελευταίο βήμα είναι να μετρήσουμε την εγγύτητα ανάμεσα στα διανύσματα $\vec{q}$ και $\vec{d}$ για όλα τα $t \in T$. Διάφορες τεχνικές και ευρεστικές έχουν προταθεί προκειμένου να δώσουν απαντήσεις σε ερωτήματα διάταξης. Στην περίπτωση μας θα χρησιμοποιήσουμε την ευρέως χρησιμοποιούμενη μετρική του ημίτονου η οποία δίνεται από τον παρακάτω τύπο:

$$\cos(Q, D_d) = \frac{1}{w_d w_q} \sum_{t \in Q \cap D_d} (1 + \log_e f_{d,t}) \cdot w_{d,t}$$

Σύμφωνα με την μετρική αυτή προσπαθούμε να βρούμε το ημίτονο της γωνίας ανάμεσα στο διάνυσμα του ερωτήματος $\vec{q}$ και το διάνυσμα $\vec{d}$ κάθε σχετικής συναλλαγής. Όσο μεγαλύτερη είναι αυτή η τιμή τόσο μεγαλύτερη είναι και η ομοιότητα. Τα κλασσικά διανυσματικά μοντέλα, που χρησιμοποιούνται για αρχεία κειμένων σταματάνε σε αυτό το σημείο, και διατάσσουν όλες τις απαντήσεις σύμφωνα με το similarity score που έλαβαν και τα παρουσιάζουν στους

χρήστες. Παρόλα αυτά όπως θα δούμε στην ενότητα 7.5.2.2.3 παρακάτω, στην περίπτωση μας εισάγουμε ένα επιπλέον βήμα.

7.5.2.2.3 Φάση Προτάσεων

Σε μια βάση με κείμενα, όταν οι χρήστες έθεταν ένα ερώτημα τους παρουσιάζονταν τα πιο σχετικά κείμενα για να διαλέξουν. Στην περίπτωση μας όμως, οι χρήστες δεν ενδιαφέρονται για προτάσεις συναλλαγών (όπως γινόταν σε μια συλλογή με κείμενα), αλλά μάλλον για προτάσεις αντικειμένων. Στην πράξη μπορεί να ψάχνουμε για την πιο σχετική συναλλαγή αλλά τελικά θέλουμε να προτείνουμε στους χρήστες τα αντικείμενα που είναι πιθανότερο να αγοράσουν. Προκειμένου να το επιτύχουμε αυτό μπορούμε να ακολουθήσουμε δύο εναλλακτικές στρατηγικές.

Ας υποθέσουμε ότι ένας χρήστης έχει τοποθετήσει ορισμένα αντικείμενα στο καλάθι του, και το σύστημα μας έχει καταλήξει σε μερικές συναλλαγές οι οποίες είναι οι πιο σχετικές για το συγκεκριμένο χρήστη μέχρι τώρα. Η πρώτη επιλογή θα ήταν αν προτείναμε όλα τα στοιχεία από την πιο σημαντική συναλλαγή δηλ. αυτή που τελικά κατατάχθηκε πρώτη. Σ' αυτή την περίπτωση όλα τα στοιχεία μέσα σ' αυτή τη συναλλαγή θα ταξινομούνταν επίσης σύμφωνα με το γινόμενο του βάρους τους με τον αριθμό των εμφανίσεων τους στη συναλλαγή και θα εμφανίζονταν στο χρήστη σε φθίνουσα σειρά σπουδαιότητας. Για παράδειγμα αν η πιο σχετική συναλλαγή σε κάποια νέα συναλλαγή περιείχε τα ακόλουθα στοιχεία (κάθε δυάδα περιέχει πρώτα το στοιχείο και κατόπιν το πλήθος των εμφανίσεων του στην συγκεκριμένη συναλλαγή) $\langle (1,3), (4,2), (7,1), (9,6) \rangle$ και τα βάρη^{§§§} αυτών των στοιχείων είναι αντίστοιχα $w_1=0.3$, $w_4=0.5$, $w_7=0.8$ και $w_9=0.1$ τότε ο πολλαπλασιασμός του βάρους κάθε στοιχείου με τον αριθμό των εμφανίσεων του θα μας έδινε (Πίνακας 11).

Πίνακας 11: Βρίσκοντας τα πιο σημαντικά αντικείμενα σε μια συναλλαγή

Αντικείμενο	Πλήθος Εμφανίσεων	Ειδικό Βάρος	Τελικό Βάρος
1	3	0.3	0.9
4	4	0.5	2
7	1	0.8	0.8
9	6	0.1	0.6

Άρα σύμφωνα με τα τελικά βάρη η σειρά προτάσεων θα είναι: **[4, 1, 7, 9]**.

Η δεύτερη στρατηγική θα ήταν αν αξιολογούσαμε επίσης ορισμένες από τις μετέπειτα συναλλαγές. Για παράδειγμα ας υποθέσουμε ότι ελέγχαμε επίσης τις δύο ή τις τρεις επόμενες

^{§§§} Τα βάρη αυτά μπορούν να προκύψουν από μια διαδικασία σαν αυτή που περιγράψαμε στο Κεφάλαιο 5

συναλλαγές και από κάθε μια από αυτές παρουσιάζαμε στους χρήστες το ένα ή τα δύο πιο σημαντικά στοιχεία. Ακολούθως, κάθε φορά που ένας πελάτης τοποθετεί περισσότερα προϊόντα (στοιχεία) στο καλάθι του τότε αυτό τίθεται σαν ένα ακόμα ερώτημα διάταξης στο σύστημα μας το οποίο τώρα αποτελείται απ' όλα τα στοιχεία. Οι υψηλότερες συναλλαγές επιστρέφονται κοκ. μέχρι ο πελάτης μας να ολοκληρώσει τις αγορές του.

Και οι δύο στρατηγικές θα μπορούσαν να χρησιμοποιηθούν παράλληλα με μια τεχνική γνώστη και ως Rocchio's relevance feedback [29];[30], τεχνική ευρέως χρησιμοποιούμενη κύρια στην ανάκτηση πληροφορίας, προκειμένου να μας παρέχει καλύτερες και ακριβέστερες προβλέψεις. Σύμφωνα με την συγκεκριμένη τεχνική οι απαντήσεις σε ένα ερώτημα κάποιου χρήστη παρουσιάζονται μαζί με μια απλή φόρμα βαθμολόγησης και ο χρήστης καθορίζει ποια από τα επιστρεφόμενα κείμενα είναι χρήσιμα - σχετικά. Το σύστημα κατόπιν επαναδιαμορφώνει το ερώτημα και το τροφοδοτεί πίσω στο σύστημα μας, προσθέτοντας στο αρχικό διάνυσμα του ερωτήματος $\vec{q}$ ένα σταθμισμένο άθροισμα όλων των διανυσμάτων που αντιστοιχούν στα σχετικά κείμενα D_+ και αφαιρώντας ένα σταθμισμένο άθροισμα των μη σχετικών κειμένων D_- σύμφωνα με τον παρακάτω τύπο:

$$\vec{q}' = \alpha \vec{q} + \beta \sum_{D_+} \vec{d} - \gamma \sum_{D_-} \vec{d}$$

Οι τιμές των σταθερών α , β και γ συνήθως τίθενται ίσες με 8, 16, και 4 αντίστοιχα αφού αυτές οι τιμές θεωρήθηκαν λογικές επίσης για το TREC 2****. Φυσικά το να απαιτούμε οι χρήστες να αξιολογούν χειροκίνητα κάθε τι που τους προτείνεται θα δημιουργούσε τα ίδια προβλήματα όπως αυτά που εμφανίζονται στα συστήματα collaborative filtering που χρησιμοποιούν άμεση ψηφοφορία (βλέπε ενότητα 7.2). Έτσι αντί να ρωτάμε τους χρήστες να παρέχουν χειροκίνητα όλες αυτές τις παραμέτρους θα μπορούσαμε να χρησιμοποιήσουμε μια μέθοδο με την ονομασία pseudo-relevance feedback - PRF. Με την PRF ουσιαστικά υποθέτουμε ότι ένας αριθμός των υψηλότερα καταταγμένων κειμένων είναι σχετικά (δηλ. το D_+) και θέτουμε το γ ίσο με το 0 (δηλαδή το D_- , τα άσχετα κείμενα δεν χρησιμοποιούνται εκτός και αν έχουμε αυτή την πληροφορία). Κατ' αυτόν τον τρόπο το σύστημα λαμβάνει το feedback του αλλά δίχως να ενοχλεί ή να γίνεται αντιληπτό από τους χρήστες. Παρά το γεγονός ότι και οι δύο αυτές βελτιώσεις θα συντελούσαν σε πιο ακριβείς προτάσεις, όπως θα δούμε αργότερα (ενότητα 7.6.5) έχουν συγκεκριμένα μειονεκτήματα που τα καθιστούν μάλλον μη πρακτικά στην περίπτωση μας.

**** Συνέδριο σχετικό με την ανάκτηση πληροφορίας από αρχεία κειμένων (βλέπε επίσης στο <http://trec.nist.gov/>)

7.5.2.3 Δημιουργία Ευρετηρίου ειδικά για Ερωτήματα Διάταξης

Αν αποφασίσουμε να χρησιμοποιήσουμε το σύστημα μας μόνο για την δεύτερη προσέγγιση και όχι για την πρώτη (δηλ. μόνο για να κάνουμε ερωτήματα διάταξης και όχι για δυαδικά ερωτήματα) τότε θα έπρεπε να χρησιμοποιήσουμε μια εναλλακτική δομή ευρετηρίου, ειδικά υλοποιημένη για αυτόν τον σκοπό [100]. Ένα ευρετήριο που κατασκευάζεται χρησιμοποιώντας την κλασσική τεχνική των ανεστραμμένων αρχείων αποθηκεύει τη λίστα κάθε αντικειμένου διατεταγμένη σύμφωνα με τον αριθμό της συναλλαγής, και με κάθε ζευγάρι να αναπαριστά τον αριθμό της συναλλαγής t και την αντίστοιχη συχνότητα μέσα στην συναλλαγή $f_{d,t}$. Ας υποθέσουμε ότι έχουμε ένα αντικείμενο το οποίο εμφανίζεται σε 4 συναλλαγές και πιο συγκεκριμένα δυο φορές στη συναλλαγή 100, τέσσερις στην συναλλαγή 105, δέκα στην συναλλαγή 130 και δύο στην συναλλαγή 199. Η λίστα αυτού του αντικειμένου θα αποθηκεύεται ως εξής:

<4; (100,2), (105,4), (130,10), (199,2)>

Προκειμένου τώρα να επιταχύνουμε την όλη διαδικασία της κατάταξης διατάσουμε κάθε λίστα σε φθίνουσα σειρά με βάση την τιμή του $f_{d,t}$. Έτσι παραπάνω λίστα μπορεί να γραφτεί ως εξής:

<4; (130, 10), (105, 4), (100, 2), (199, 2)>

Με αυτόν τον τρόπο οι συναλλαγές με τις μεγαλύτερες τιμές $f_{d,t}$ μπορούν να βρεθούν πολύ γρήγορα. Παρόλα αυτά η παραπάνω λίστα θα πρέπει να αναπαρισταθεί με κάποιον άλλο τρόπο, επειδή αφού οι δείκτες των αριθμών των συναλλαγών δεν είναι πλέον διατεταγμένες σε αύξουσα σειρά, δεν μπορούμε πλέον να επεξεργαστούμε τις λίστες παίρνοντας τα διάκενα ανάμεσα στις εγγραφές (d-gaps). Προκειμένου να το επιτύχουμε αυτό μπορούμε να σπάσουμε κάθε λίστα σε κομμάτια μέσα στα οποία οι τιμές $f_{d,t}$ είναι ίδιες. Κατ' αυτόν τον τρόπο μπορούμε να παίρνουμε τα διάκενα μέσα σε κάθε κομμάτι. Έτσι τελικά η παραπάνω λίστα μπορεί να γραφτεί ως εξής:

<4; (10,1 : 130), (4,1 : 105), (2,2 : 100,199)>

Κάθε κομμάτι παίρνει σαν πρόθεμα την κοινή $f_{d,t}$ τιμή που έχουν όλοι οι δείκτες μέσα στο συγκεκριμένο κομμάτι, ακολουθούμενο από ένα μετρητή που μας δείχνει πόσοι δείκτες υπάρχουν στο συγκεκριμένο κομμάτι. Τέλος αποθηκεύουμε σε κάθε κομμάτι τους ακριβείς αριθμούς των συναλλαγών που έχουν κοινή τιμή $f_{d,t}$. Στην παραπάνω λίστα το πρώτο και το δεύτερο κομμάτι περιέχουν ένα δείκτη, και το τρίτο δύο. Το μέγεθος του τελικού ευρετηρίου το οποίο είναι διατεταγμένο σύμφωνα με τις συχνότητες είναι μόλις ελάχιστα μεγαλύτερο από αυτό που δημιουργείται με τον παραδοσιακό τρόπο [101].

7.6 Αξιολόγηση

Δοκιμάσαμε τις προσεγγίσεις μας σε δύο διαφορετικές κατευθύνσεις, αναφορικά ποιοτικά και ποσοτικά. Πιο συγκεκριμένα θέλαμε να ελέγξουμε τα αποτελέσματα που παρήγαν (δηλ. τον ποιοτικό παράγοντα) όπως επίσης και την συνολική τους απόδοση (δηλ. τον ποσοτικό παράγοντα). Η απόδοση της προσέγγισης μας ως προς τον απαιτούμενο χρόνο αλλά και ως προς το χώρο δοκιμάστηκε χρησιμοποιώντας συνθετικά δεδομένα από την γεννήτρια συνθετικών δεδομένων. Η ποιότητα των αποτελεσμάτων καθώς επίσης και η χρησιμότητα της προσέγγισης μας από την άλλη επιβεβαιώθηκε χρησιμοποιώντας πραγματικά δεδομένα.

7.6.1 Συνθετικά Δεδομένα

Προκειμένου να δοκιμάσουμε την απόδοση της προσέγγισης μας χρησιμοποιήσαμε παρόμοιες βάσεις με τις συνθετικές βάσεις που χρησιμοποιήθηκαν στο [10] και στις περισσότερες μετέπειτα εργασίες. Τα δεδομένα που δημιουργούνται από αυτή τη γεννήτρια θεωρούνται ότι προσομοιάζουν επιτυχώς την αγοραστική συμπεριφορά πελατών σε ένα περιβάλλον καταστήματος λιανεμπορίου. Αξιολογήσαμε το σύστημα μας με είκοσι διαφορετικές βάσεις, οκτώ προκειμένου να δοκιμάσουμε την συμπεριφορά του για μεταβαλλόμενο μήκος συναλλαγών και μέσο μέγεθος πιθανών συχνών αντικειμένων και τα υπόλοιπα κρατώντας όλες τις άλλες παραμέτρους σταθερές και αυξάνοντας τον αριθμό των συναλλαγών. Για τα πρώτα οκτώ θέσαμε: $|D|=100K$, $N=1000$ και $|L|=2000$. Οι τιμές για το $|T|$ τέθηκαν σε 5, 10 και 20. Το μέσο μέγεθος των μέγιστων πιθανών συχνών αντικειμένων ήταν 2, 4, 6 και 8. Στη διαδικασία δημιουργίας συναλλαγών εισαγάγαμε ένα επιπλέον βήμα προκειμένου να αναπαραστήσουμε και να λάβουμε κατόπιν υπόψη τον αριθμό των φορών που εμφανίζεται το κάθε αντικείμενο σε μια συναλλαγή. Πιο συγκεκριμένα χρησιμοποιήσαμε μια γεννήτρια τυχαίων αριθμών η οποία και καθόριζε τον αριθμό των φορών που κάθε προϊόν εμφανίζεται σε κάθε συναλλαγή. Περιορίσαμε τους πιθανούς αριθμούς που παράγονται από την γεννήτρια μας σε 20, αφού είναι σχεδόν αδύνατο ένας πελάτης να αγοράσει περισσότερες από 20 φορές το ίδιο προϊόν σε μια και μόνη συναλλαγή. Οι τελικές βάσεις είχαν το διπλάσιο μέγεθος σε σύγκριση με αυτά που είχαν παραχθεί στο [10]. Στον Πίνακα 12 δείχνουμε τις ρυθμίσεις των παραμέτρων για τις χρησιμοποιηθείσες βάσεις.

Πίνακας 12: Ρυθμίσεις παραμέτρων για τις χρησιμοποιηθείσες βάσεις

Όνομα	T	I	Μέγεθος σε MB
T5.I2.D100K	5	2	4.8
T5.I4.D100K	5	4	
T10.I2.D100K	10	2	8.8
T10.I4.D100K	10	4	
T10.I6.D100K	10	6	
T20.I2.D100K	20	2	16.8
T20.I4.D100K	20	4	
T20.I6.D100K	20	6	
T20.I8.D100K	20	8	

Για την δημιουργία των βαρών χρησιμοποιήσαμε μια παρόμοια τεχνική με αυτή που χρησιμοποιήθηκε στο [32]. Πιο συγκεκριμένα σε κάθε αντικείμενο αποδίδεται ένα βάρος σύμφωνα με μια εκθετική κατανομή (Εικόνα 18) βασισμένη στην συχνότητα εμφανίσεων του στη βάση.

Εικόνα 18: Κατανομή βαρών σε μια βάση

7.6.2 Χρονικές και Χωρικές Απαιτήσεις για την Δημιουργία του Ευρετηρίου

Η βάση διαβάζεται μια φορά και δημιουργείται ένα ευρετήριο βασισμένο στα ανεστραμμένα αρχεία. Προκειμένου να συμπίεσουμε το ευρετήριο χρησιμοποιήσαμε το local Bernoulli μοντέλο [126] με τον κώδικα Golomb [55] για να αναπαραστήσουμε τα κενά ανάμεσα στις διαδοχικές συναλλαγές (στην περίπτωση μας σημειώνονται σαν TIDs). Τέλος κάθε ανεστραμμένη λίστα υπάρχει σαν πρόθεμα ένας ακέραιος f_i που μας επιτρέπει τον υπολογισμό της παραμέτρου b για τον Golomb. Οι ακέραιοι αυτοί (f_i) αναπαριστώνται χρησιμοποιώντας κωδικοποίηση γ . Τα παραγόμενα ευρετήρια είχαν ένα μέγεθος που αντιστοιχούσε περίπου στο 20% του αρχικού αρχείου και έτσι λαμβάνοντας υπόψη και τη μεγάλη διαθεσιμότητα σε μνήμη, το ευρετήριο μπορεί να κρατηθεί και να είναι προσβάσιμο από την κύρια μνήμη σχετικά εύκολα. Γενικά αφού το ζητούμενο μας δεν ήταν ο βαθμός συμπίεσης του ευρετηρίου που θα επιτυγχάναμε αλλά το πόσο γρήγορα θα γινόταν τελικά η δημιουργία και η συμπίεση του προτιμήσαμε μια λιγότερο πολύπλοκη μέθοδο. Φυσικά υπάρχει μια αντίστροφη σχέση ανάμεσα στην πολυπλοκότητα της μεθόδου και το τελικό μέγεθος του ανεστραμμένου αρχείου αλλά αφού ήδη επιτύχαμε σχετικά αρκετά μικρά μεγέθη για τα αρχεία μας δεν βρήκαμε κίνητρο να επενδύσουμε σε περισσότερο πολύπλοκες μεθόδους.

Στην Εικόνα 19 βλέπουμε το χρόνο που απαιτείται για να δημιουργήσουμε και να συμπίεσουμε τις βάσεις που φαίνονται στον Πίνακα 12, όλες αποτελούμενες από 100.000 συναλλαγές. Στην Εικόνα 20 βλέπουμε το απαιτούμενο χρόνο για την δημιουργία και συμπίεση του ευρετηρίου για τρεις βάσεις για αυξανόμενο αριθμό συναλλαγών (ξεκινώντας από 100.000 συναλλαγές και φθάνοντας έως και τις 1.000.000). οι βάσεις αυτές είχαν μεγέθη που κυμαίνονταν από 15,4MB η μικρότερη (T5.I2.D250K) έως 224MB (T20.I6.D1000K) η μεγαλύτερη.

Εικόνα 19: Απαιτούμενοι χρόνοι για την δημιουργία και συμπίεση ευρετηρίου

Εικόνα 20: Απαιτούμενος χρόνος για την δημιουργία και συμπίεση του ευρετηρίου για αυξανόμενο αριθμό συναλλαγών

7.6.3 Χρόνοι Απόκρισης

Αφού επιθυμούμε το σύστημα μας να έχει παρόμοια λειτουργία με αυτή μιας μηχανής αναζήτησης και να έχει μια on-line συμπεριφορά, ο χρόνος που απαιτείται να επεξεργαστούμε

κάθε ερώτημα θα πρέπει να είναι όσο το δυνατόν μικρότερος. Οι χρήστες δεν μπορούν να περιμένουν πάρα πολύ προκειμένου να πάρουν μια πρόταση. Το σύστημα μας έχει ουσιαστικά την ίδια απόδοση όπως οποιαδήποτε μηχανή αναζήτησης που υπάρχει στον παγκόσμιο ιστό και έτσι η απάντηση σε οποιοδήποτε ερώτημα είναι πρακτικά άμεση ή δίνεται σε επιτρεπτά χρονικά όρια.

7.6.4 Αξιολόγηση Ποιότητας

Η ποιοτική αξιολόγηση της προσέγγισης μας αποδείχτηκε ένα πολύ περίπλοκο εγχείρημα αφού καμία γνωστή τεχνική από τη βιβλιογραφία δεν μπορούσε να χρησιμοποιηθεί στην περίπτωση μας. Πιο συγκεκριμένα μια επιλογή θα ήταν να προσπαθούσαμε να την αξιολογήσουμε σε σχέση με ορισμένες παραδοσιακές προσεγγίσεις (π.χ. Apriori, DIC, DHP κτλ.) σε όρους απαιτούμενου χρόνου, πόρων και παραγομένου αποτελέσματος. Παρόλα αυτά αυτό απορρίφθηκε αμέσως αφού θα συγκρίναμε δυο εντελώς διαφορετικές λογικές με διαφορετικές λειτουργίες και αποτελέσματα. Η δεύτερη επιλογή που σκεφτήκαμε ήταν η αξιολόγηση χρησιμοποιώντας τεχνικές από την ανάκτηση πληροφορίας όπως την ανάκληση (recall) και την ακρίβεια (precision). Παρόλα αυτά αυτό απορρίφθηκε επίσης αφού στην περίπτωση μας παρουσίαζε ανυπέρβλητα προβλήματα.

Η τελευταία σκέψη ήταν η χρήση ενός κοινού ελέγχου (test audience) που θα αξιολογούσε τις προτάσεις που θα έκανε το σύστημα μας. Εντούτοις όπως αναφέραμε και προηγούμενα αυτό αναπόφευκτα θα εμπεριείχε και ένα μεγάλο βαθμό υποκειμενικότητας αφού τα αντικείμενα δεν παρέχουν κάποιο καθαρό μέτρο του τι είναι σχετικό και τι όχι. Επίσης αφού η εξόρυξη δεδομένων είναι εξ ορισμού η ανακάλυψη κρυμμένης γνώσης θα ήταν παράλογο να ζητάμε από τους χρήστες να αξιολογήσουν τις παραγόμενες προτάσεις. Πιθανώς να μπορούσαν να κρίνουν και να αξιολογήσουν τις προφανείς απαντήσεις αλλά θα ήταν σίγουρα αβοήθητοι με τις σίγουρα πιο σημαντικές μη αναμενόμενες απαντήσεις. Είναι σχετικά εύκολο να κρίνεις μια πρόταση που βάζει μαζί γάλα και δημητριακά, αλλά πως μπορεί πραγματικά να αποφασίσεις αν μια πρόταση που τοποθετεί μαζί πάνες και μπύρες είναι σωστή ή όχι;

Η απάντηση τελικά δόθηκε χρησιμοποιώντας πραγματικά ένα κοινό ελέγχου, το οποίο παρόλα αυτά είχε μια εντελώς διαφορετική λειτουργία. Πιο συγκεκριμένα χρησιμοποιήσαμε ένα κοινό ελέγχου απλά και μόνο για να κάνουμε κάποιες συναλλαγές παρά για να τις αξιολογήσει. Για το συγκεκριμένο πείραμα χρησιμοποιήσαμε τρεις ομάδες ατόμων, η κάθε μια αποτελούμενη από 50 άτομα με τον ίδιο αριθμό γυναικών και ανδρών. Και οι τρεις ομάδες αποτελούνταν από άτομα ηλικία 18 έως 22 ετών όπου όλοι σπούδαζαν στο ίδιο πανεπιστήμιο και στο ίδιο τμήμα. Δεν βάλαμε κάποιους άλλους περιορισμούς όταν σχηματίζαμε τις συγκεκριμένες ομάδες όπως για παράδειγμα την οικονομική κατάσταση, αφού θέλαμε να είναι

όσο το δυνατόν πιο αντιπροσωπευτικό του κοινού ενός πραγματικού καταστήματος. Όλα τα άτομα ήταν εξοικειωμένοι με την τεχνολογία με σχετική άνεση με τους υπολογιστές και το Ίντερνετ. Αποφασίσαμε να φτιάξουμε ένα εικονικό κατάστημα το οποίο πουλούσε κινητά τηλέφωνα καθώς και διάφορα αξεσουάρ τους. Τα προϊόντα, τα ονόματα τους και οι τιμές που χρησιμοποιήθηκαν στο κατάστημα μας ήταν όλα αληθινά. Ο λόγος για τον οποίο επιλέξαμε το συγκεκριμένο προϊόν ήταν επειδή τα άτομα στο κοινό ελέγχου είχαν κινητό τηλέφωνο και/ή είχαν επισκεφτεί κάποιο κατάστημα που τα πουλούσε σε ένα ποσοστό σχεδόν 100%.

Ακολουθώντας το πείραμα μας πραγματοποιήθηκε ως εξής. Επιτρέψαμε και στις τρεις ομάδες ατόμων να πραγματοποιήσουν κάποιες αγορές έτσι ώστε να συγκεντρώσουμε αρκετά δεδομένα. Οι συγκεκριμένες αγορές πραγματοποιήθηκαν χωρίς να κάνουμε κάποιες προτάσεις στους χρήστες. Κατόπιν η πρώτη ομάδα ατόμων έκανε αγορές λαμβάνοντας προτάσεις χρησιμοποιώντας κάποιον κλασσικό αλγόριθμο (π.χ. τον *Argiori*) η δεύτερη με την προσέγγιση μας αλλά με βάση τα δυαδικά ερωτήματα και η τρίτη πάλι με την προσέγγιση μας αλλά με βάση τα ερωτήματα διάταξης. Οι χρήστες δεν ήξεραν τίποτα για το χρησιμοποιούμενο σύστημα, όπως επίσης και για το αντικείμενο του πειράματός μας έτσι ώστε να μπορούσαν να κάνουν όλες τις αγορές τους ανεπηρέαστοι. Ήξεραν μόνο ότι έπρεπε να κάνουν κάποιες αγορές, και ότι είχαν ένα συγκεκριμένο ποσό να δαπανήσουν κάθε φορά αλλά και συνολικά. Το ποσό των χρημάτων που κάθε χρήστης είχε στη διάθεση του να ξοδέψει δεν προέκυψε αυθαίρετα αλλά αφού είχαμε προηγούμενα ρωτήσει τους χρήστες τι ποσό χρημάτων θα μπορούσαν να διαθέσουν ή είχαν διαθέσει στο κοντινό παρελθόν για την αγορά τέτοιων προϊόντων. Από τις απαντήσεις τους υπολογίσαμε ένα μέσο ποσό χρημάτων το οποίο και παρείχαμε στους χρήστες προκειμένου να το ξοδέψουν. Φυσικά μπορούσαν να ξοδέψουν μόνο μέρος των χρημάτων και όχι όλο, καθώς επίσης τους δινόταν η ευχέρεια να χαλάσουν και λίγο παραπάνω. Οι χρήστες ήταν ελεύθεροι να περιηγηθούν σε όλα τα διαθέσιμα προϊόντα ελεύθερα με την ησυχία τους, επισκεπτόμενοι κάθε σελίδα προϊόντος όσες φορές επιθυμούσαν, προτού αποφασίσουν τελικά να αγοράσουν κάτι.

Το ζητούμενο ήταν να ελέγξουμε αφότου θα είχαμε συγκεντρώσει κάποιο αριθμό πωλήσεων, ποια λύση θα παρήγαγε τα μεγαλύτερα καθαρά κέρδη. Το αποτέλεσμα ήταν ότι η λύση που πρότεινε προϊόντα στους χρήστες χρησιμοποιώντας τα δυαδικά ερωτήματα αύξησε τα καθαρά κέρδη κατά ένα ποσοστό σχεδόν 10%, ενώ η λύση που πρότεινε προϊόντα στους χρήστες χρησιμοποιώντας τα ερωτήματα διάταξης τα αύξησε κατά ένα ποσοστό λίγο μεγαλύτερο του 15%. Έτσι μπορούμε να συμπεράνουμε ότι το σύστημα μας πραγματικά παρέχει καλύτερες προτάσεις στους χρήστες, αφού το κυρίως ζητούμενο της αύξησης των καθαρών κερδών επιτεύχθηκε.

7.6.5 Συζήτηση πάνω στην Αξιολόγηση

Στην διαδικασία της αξιολόγησης αποφασίσαμε να μην χρησιμοποιήσουμε την relevance feedback ή την pseudo-relevance feedback βελτιστοποίηση όχι επειδή δεν είναι αποτελεσματικές στο να συλλαμβάνουν τις προτιμήσεις των χρηστών αλλά επειδή όπως σημειώσαμε προηγούμενα παρουσιάζουν συγκεκριμένα ελαττώματα τα οποία τις καθιστούν σχετικά δύσχρηστες. (ακόμα και στις μηχανές αναζήτησης η relevance feedback είναι ένα πολύ σπάνια εμφανιζόμενο χαρακτηριστικό).

Έτσι πρώτα απ' όλα οι χρήστες συνήθως δεν επιθυμούν να ενοχλούνται από συστήματα feedback, αλλά απλά θέλουν να πάρουν τις σωστές απαντήσεις και μόλις τις πάρουν επικεντρώνονται στο να τις χρησιμοποιήσουν παρά να τις αξιολογήσουν. Ακόμα και αν η απάντηση ή οι απαντήσεις που επιστράφηκαν δεν ήταν οι αναμενόμενες είτε είναι πολύ ενοχλημένοι ή πολύ ανυπόμονοι να βρουν τη σωστή που δεν επιθυμούν να ξανασπαταλήσουν άλλο χρόνο παρέχοντας στο σύστημα feedback [34]. Φυσικά αυτό το πρόβλημα θα μπορούσε να ξεπεραστεί με την pseudo relevance feedback – PRF μέθοδο, αλλά οποιαδήποτε από τις δύο αυτές μεθόδους και αν χρησιμοποιούσαμε θα δημιουργούσαμε ένα άλλο μεγαλύτερο πρόβλημα. Ένα σύστημα που χρησιμοποιεί οποιαδήποτε μορφή relevance feedback, προσθέτει αρκετά στην πολυπλοκότητα του δίνοντας καλύτερες απαντήσεις αλλά ρίχνοντας κατά πολύ την απόδοση [34]. Και στην συγκεκριμένη εφαρμογή δεν μπορούμε να θυσιάσουμε την απόδοση για χάρη της ποιότητας, αφού οι χρήστες δεν ενδιαφέρονται τόσο πολύ στο να λαμβάνουν καλύτερες απαντήσεις όπως συνέβαινε σε μια μηχανή αναζήτησης ή στις βάσεις κειμένων.

Στις εμπορικές εφαρμογές η χρήση τέτοιων μεθόδων πιθανότατα να έγειρε σοβαρούς ενδοιασμούς σχετικά με την προστασία προσωπικών δεδομένων και την ιδιωτικότητα από την πλευρά των καταναλωτών. Αντίθετα με τις αναζητήσεις σε μια μηχανή αναζήτησης στο διαδίκτυο ή σε μια δημόσια βιβλιοθήκη όπου οι χρήστες αισθάνονται πολύ περισσότερο “ανώνυμοι”, κατά την επίσκεψη τους σε ένα κατάστημα λιανεμπορίου αισθάνονται πιο ευάλωτοι και έχουν σοβαρές επιφυλάξεις σχετικά με τα προσωπικά δεδομένα τους. Έτσι παρόλο το γεγονός ότι η χρήση δεδομένων για διαδικασίες εξόρυξης δεν είναι παράνομη, συνεπάγεται παρόλα αυτά και έναν υψηλό βαθμό ρίσκου για την φήμη της εταιρείας. Με άλλα λόγια αν θέλαμε να το θέσουμε όπως μας το εξέφρασαν οι μάνατζερ και οι υπεύθυνοι πληροφορικής διαφόρων επιχειρήσεων λιανεμπορίου: “*μπορεί να εφαρμόζουμε τεχνικές εξόρυξης δεδομένων αλλά δεν υπάρχει κανένας λόγος να το υπενθυμίζουμε στους πελάτες μας όλη την ώρα*”. Τέλος όπως σημειώσαμε στην ενότητα 7.6.4 η εξόρυξη δεδομένων είναι εξ ορισμού η ανακάλυψη άγνωστης, υπονοούμενης και πιθανώς χρήσιμης γνώσης. Έτσι αντίθετα με τις σελίδες στο διαδίκτυο ή τα κείμενα που επιστρέφει ένα κλασσικό σύστημα ανάκτησης

πληροφορίας, τα προϊόντα ή γενικότερα τα αντικείμενα δεν παρέχουν κάποιο καθαρό μέτρο για να από φασίσουμε τι είναι σχετικό και τι όχι. Είναι παράλογο να ζητάς από ένα χρήστη που μόλις αγόρασε μπύρες και το σύστημα του προτείνει πάνες να αξιολογήσει πόσο σχετική ήταν αυτή η πρόταση.

Όσο αναφορά την τελική μορφή της τεχνικής αξιολόγησης που χρησιμοποιήθηκε (δηλ. τα καθαρά κέρδη που επιτύγχανε) πέραν του γεγονότος ότι ήταν η καλύτερη λύση που μπορέσαμε να σκεφτούμε ικανοποιούσε και μια άλλη επίσης σημαντική παράμετρο. Σύμφωνα με τους Kleinberg, Papadimitriou και Raghavan στο [74], ένα μοτίβο στα δεδομένα είναι ενδιαφέρον μόνο στο βαθμό που μπορεί να χρησιμοποιείται στην διαδικασία λήψεως αποφάσεων της επιχείρησης προκειμένου να βελτιώσει την αποτελεσματικότητα της. Έτσι η αποτελεσματικότητα των εξαγομένων μοτίβων (όπως οι κανόνες συσχετίσεων) στην διαδικασία λήψεως αποφάσεων μιας επιχείρησης μπορεί να εξεταστεί μόνο μέσα από το μικροοικονομικό της πρίσμα. Κατά τη γνώμη μας δεν υφίσταται κάποιος καλύτερος τρόπος για να το επιβεβαιώσουμε αυτό, από τα καθαρά κέρδη που επιτυγχάνουμε από την κάθε λύση. Ο τρόπος με τον οποίο οι υπάρχοντες αλγόριθμοι αντιμετωπίζουν τα αντικείμενα και τα εξαγόμενα μοτίβα γενικότερα σαν απλές στατιστικές πιθανότητες και σαν δυαδικές μεταβλητές είναι μάλλον μια υπεραπλούστευση ενός πολύπλοκου προβλήματος που εμείς καταφέραμε τελικά να το διαχειριστούμε διαφορετικά.

Καθώς προσπαθούσαμε να καθορίσουμε το καθαρό κέρδος το οποίο θα δημιουργούσε το κάθε προϊόν βρεθήκαμε αντιμέτωποι με μια σειρά ερωτημάτων. Αρχικά η εύρεση της τιμής πώλησης κάθε προϊόντος ήταν μια διαδικασία η οποία θα μπορούσε εύκολα να αντιμετωπιστεί με μια απλή επίσκεψη σε κάποιο κατάστημα (φυσικό ή ηλεκτρονικό) που πουλάει τέτοια προϊόντα. Παρόλα αυτά η τιμή πώλησης ενός προϊόντος είναι κάτι εντελώς διαφορετικό από τα καθαρά κέρδη που αποκομίζουμε. Επίσης όπως σημειώσαμε προηγουμένως ένα ακριβό προϊόν δεν είναι πάντα ένα κερδοφόρο προϊόν και ένα φθηνό προϊόν μπορεί να είναι επικερδές. Έτσι έπρεπε να κάνουμε κάποιες υποθέσεις. Κατ' αρχήν υποθέσαμε ότι οι επιχειρήσεις αγοράζουν τα συγκεκριμένα προϊόντα κατά μέσο όρο 20% φθηνότερα απ' ότι τα πωλούν. Επίσης υποθέσαμε ένα 4% κατά μέσο όρο κόστους το οποίο περιλάμβανε κόστη επεξεργασίας προϊόντων (product handling costs) και διατήρησης αποθεμάτων (inventory costs). Τα κόστη επεξεργασίας προϊόντων αναφέρονται στα κόστη που αφορούν στην φυσική επεξεργασία των προϊόντων και τα κόστη αποθεμάτων περιλαμβάνουν τα οικονομικά κόστη για την διατήρηση αποθεμάτων και την προμήθεια νέων. Έτσι καταλήξαμε σε ένα καθαρό κέρδος της τάξεως του 16% επί της τιμής πώλησης τους κάθε προϊόντος.

Ανακαλύψαμε τέλος ότι οι περισσότεροι πελάτες πραγματοποιούσαν τις αγορές τους σε μια συνεδρία (δηλ. ξόδευαν όλα τα διαθέσιμα χρήματα τους με μια φορά). Επίσης παρατηρήσαμε

ότι σχεδόν όλα τα άτομα ακολουθούσαν το ίδιο μοτίβο στις αγορές τους διαλέγοντας πρώτα το κυρίως προϊόν σαν κάτι πολύ ακριβό και τα υπόλοιπα προϊόντα όσο το δυνατόν φθηνότερα.

7.7 Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Στις περισσότερες περιπτώσεις η χρήση μιας τεχνικής από ένα άλλο επιστημονικό τομέα, οδηγεί σε μη βέλτιστες λύσεις. Στην περίπτωση μας όμως, οι τεχνικές που δανειστήκαμε από τον τομέα της ανάκτησης πληροφορίας όχι μόνον μας οδήγησαν σε μια γενικά καλή λύση αλλά κατόρθωσε να υπερκεράσει ορισμένα προβλήματα που παρουσίαζε όταν χρησιμοποιούταν παραδοσιακά σε αρχεία κειμένου. Οι κυριότεροι λόγοι για αυτό είναι: στις συλλογές κειμένων και ειδικά για τα ερωτήματα διάταξης ένα σοβαρό πρόβλημα παρουσιαζόταν με την φύση αυτών καθ' εαυτών των λέξεων. Πιο συγκεκριμένα αντιμετωπίζαμε προβλήματα όπως πολυσημία (polysemy), συνωνυμία (synonymy) κ.ο.κ. που καθιστούσαν δύσκολο τόσο να κάνουμε το κατάλληλο ερώτημα όσο και να επιστρέψουμε τις πιο σχετικές απαντήσεις. Στην περίπτωση μας όμως αυτά τα προβλήματα δεν υφίστανται μιας και η αναγνώριση των αντικειμένων είναι 100% ακριβής (π.χ. με βάση τον μοναδικό τους κωδικό – bar code). Έτσι όταν θέτουμε ένα ερώτημα σχετικά με ένα συγκεκριμένο αντικείμενο είμαστε σίγουροι ότι αναφερόμαστε στο συγκεκριμένο αντικείμενο. Οι απαντήσεις σε οποιοδήποτε ερώτημα είναι έτσι πολύ πιο ουδέτερες και ανεπηρέαστες από τα χρησιμοποιούμενα ερωτήματα, ακόμα και όταν έχουμε πολύ λιγότερους όρους (αντικείμενα). Επίσης ο αριθμός των στοιχείων σε μια βάση με συναλλαγές λιανεμπορίου είναι πολύ μικρότερος από το πλήθος των διαφορετικών λέξεων σε μια τυπική βάση κειμένων. Λιγότερα στοιχεία σημαίνει λιγότεροι συνδυασμοί αντικειμένων και γενικότερα λιγότερη πολυπλοκότητα.

Ένα άλλο πρόβλημα που επιλύεται είναι όταν περισσότερα του ενός άτομα θέλουν να έχουν πρόσβαση και να επεξεργάζονται τα συγκεντρωμένα δεδομένα, ειδικότερα αν δεν βρίσκονται στην ίδια τοποθεσία. Μια λύση θα ήταν να αναπαράγουμε τα δεδομένα όπου αυτό ήταν αναγκαίο και να αφήναμε κάθε χρήστη να τα επεξεργαστεί με τις δικές του παραμέτρους στην τοποθεσία του. Όμως τα μεγέθη των βάσεων στις μέρες μας όπως επίσης και νομικοί περιορισμοί καθιστούν τις οποιεσδήποτε σκέψεις για την διανομή ή έστω την πρόσβαση στις βάσεις αυτές απαγορευτική. Από την άλλη αν κάθε ενδιαφερόμενος χρήστης τρέχει έναν κλασσικό αλγόριθμο με τις δικές του παραμέτρους στην αρχική βάση τότε αναπόφευκτα θα προσθέταμε ένα μεγάλο βάρος στη βάση μας και θα επιδεινώναμε κατ' αυτόν τον τρόπο δραματικά την απόδοση. Το μέγεθος του ευρετηρίου που δημιουργείται από τη βάση μας όσο και η ευκολία χρησιμοποίησης του καθιστούν δυνατό είτε να το διανείμουμε ή να δώσουμε πρόσβαση σε πολλούς χρήστες που επιθυμούν να επεξεργαστούν τη βάση, χωρίς να χρειάζεται να έχουν πρόσβαση στη βάση αυτή καθ' εαυτή.

Η διαχείριση των ενημερώσεων είναι μια άλλη επίσης πολύ σημαντική και δύσκολη διαδικασία. Σύμφωνα με το [98] *“είναι απαραίτητο να συλλέξουμε έναν επαρκή αριθμό δεδομένων συναλλαγών (για παράδειγμα των τελευταίων 30 ημερών) προτού να μπορέσουμε να καταλήξουμε σε λογικά συμπεράσματα από αυτά”*. Φυσικά αυτή η περίοδος εξαρτάται από διάφορες παραμέτρους όπως το μέγεθος και ο τομέας όπου μια επιχείρηση ή ένας οργανισμός δραστηριοποιείται, όπως επίσης και στη φύση της υλοποιούμενης εφαρμογής. Για παράδειγμα σε τομείς όπως οι τηλεπικοινωνίες ή σε επιχειρήσεις όπως η Wal-Mart όπου παράγονται και συλλέγονται τεράστιες ποσότητες δεδομένων καθημερινά, ή σε εφαρμογές όπως η ανακάλυψη απατών σε πιστωτικές κάρτες όπου η εύρεση πιθανής απάτης θα πρέπει να γίνει όσο το δυνατόν ταχύτερα, μπορεί να είναι απαραίτητο να επεξεργαζόμαστε τα δεδομένα μας πάρα πολύ συχνά (πιθανώς και σε ημερήσια βάση). Χρησιμοποιώντας οποιονδήποτε από τους κλασσικούς αλγορίθμους εξόρυξης δεδομένων θα έπρεπε να τρέχουμε πάνω από όλα τα δεδομένα, παλαιά και νέα από την αρχή. Αν κάποιος λάβει υπόψη ότι πιθανότατα να χρειαζόμαστε πολλαπλά τρεξίματα του ίδιου αλγορίθμου προκειμένου να συντονίσουμε τις απαιτούμενες παραμέτρους (το κατώφλι υποστήριξης - *minsupport* και το κατώφλι βεβαιότητας - *minconfidence*) προκειμένου να πάρουμε το επιθυμητό αποτέλεσμα, τότε οι πόροι που απαιτούνται είναι υπερβολικοί. Ορισμένες λύσεις έχουν δοθεί έτσι ώστε να μην χρειάζεται να επεξεργαζόμαστε όλα τα δεδομένα από την αρχή αλλά μόνο τα νέα, όπως παράδειγμα η δουλειά στο [40], η γενίκευση στο [41] ή η πιο πρόσφατη δουλειά στο [129]. Η δική μας λύση για το συγκεκριμένο πρόβλημα είναι πολύ αποτελεσματική και ταυτόχρονα απλή και υλοποιείται χρησιμοποιώντας είτε ένα *stop press file* για την διαχείριση των ερωτημάτων ή την ακόμα πιο αποτελεσματική εναλλακτική του [132]. Στο μέλλον θα θέλαμε να ελέγξουμε την συμπεριφορά της προσέγγισης μας σε σχέση με τις προηγούμενες προσεγγίσεις.

Το σύστημα το οποίο προτείνουμε υποδεικνύει αντικείμενα στους χρήστες του, λαμβάνοντας υπόψη πολύτιμες πληροφορίες που κανένα άλλο σύστημα δεν λάμβανε. Η τελική λύση είναι πολύ αποτελεσματική, απαιτώντας ελάχιστους πόρους για τις λειτουργίες της. Μια άλλη πιθανότητα θα ήταν η πρόταση και αξιολόγηση μιας διαφορετικής μετρικής συνάφειας, εκτός της μετρικής του συνημίτονου, ειδικά υλοποιημένου για την περίπτωση των δεδομένων συναλλαγών. Ενδιαφέρουσα θα ήταν επίσης η επιβεβαίωση με πειράματα αν η απόδοση της προσέγγισης μας σαν ένας κλασσικός αλγόριθμος εξόρυξης δεδομένων (δηλ. το *batch module* στην ενότητα 7.4) είναι καλύτερο από τους προηγούμενους αλγορίθμους. Τέλος θα θέλαμε να εξετάσουμε πιο προσεκτικά την διαδικασία ανάθεσης βαρών σε όλα τα αντικείμενα στη βάση.

Κεφάλαιο 8 ΑΥΤΟΜΑΤΗ ΑΝΑΚΑΛΥΨΗ
ΤΟΠΙΚΑ ΣΥΧΝΩΝ ΠΡΟΪΟΝΤΩΝ ΣΕ
ΣΥΝΔΥΑΣΜΟ ΜΕ ΠΟΛΥ ΣΥΧΝΑ ΠΡΟΪΟΝΤΑ

Περίληψη

Πολλές εναλλακτικές έχουν προταθεί για την ανακάλυψη κανόνων συσχετίσεων που περιλαμβάνουν τα σπάνια αλλά “ενδιαφέροντα” αντικείμενα σε μια βάση που υπάρχουν επίσης πολύ συχνά αντικείμενα. Παρόλα αυτά, όλες οι προσεγγίσεις μέχρι στιγμής υπέθεταν ότι γνωρίζαμε ποια είναι τα ενδιαφέροντα αντικείμενα, όπως επίσης ποια είναι η κατάλληλη τιμή υποστήριξης γι’ αυτά. Καμία προσέγγιση δεν πρότεινε κάποιο τρόπο για την αυτόματη εύρεση τέτοιων αντικειμένων. Στην παρόν κεφάλαιο εισάγουμε την έννοια των τοπικά συχνών αντικειμένων και υποστηρίζουμε την ύπαρξη τους σαν η μεγαλύτερη και πιο συχνά εμφανιζόμενη κατηγορία σπάνιων αλλά ενδιαφερόντων αντικειμένων ειδικότερα σε εφαρμογές λιανεμπορίου. Ακολουθώντας προτείνουμε δύο αλγορίθμους για την εύρεση και διαχείριση τέτοιων αντικειμένων. Η κύρια ιδέα είναι να χωρίσουμε την βάση μας σε διαμερίσεις σύμφωνα με τις ανάγκες του προβλήματος και να ψάχνουμε και για αντικείμενα που είναι συχνά σε σχέση με κάποια διαμέριση, πέρα από αυτά που είναι συχνά σε ολόκληρη τη βάση. Η προσέγγισή μας αποδεικνύεται πολύ αποτελεσματική και αποδοτική σε σχέση με τους παραδοσιακούς αλγορίθμους τόσο σε συνθετικά όσο και σε πραγματικά δεδομένα.

Λέξεις Κλειδιά: *Εξόρυξη Δεδομένων, Κανόνες Συσχετίσεων, Τοπικά Συχνά Αντικείμενα, Διαμερίσεις Βάσεων*

8.1 Εισαγωγή

Το πρόβλημα της εύρεσης κανόνων συσχετίσεων από τα δεδομένα αναφέρεται συχνά σαν το πρόβλημα των “καλαθιών αγορών” (market basket problem). Σύμφωνα με την διατύπωση αυτού του προβλήματος δίνεται ένα σύνολο αντικειμένων και μια μεγάλη συλλογή συναλλαγών, όπου τα αντικείμενα εμφανίζονται μαζί σε σύνολα (καλάθια). Το συγκεκριμένο πρόβλημα πήρε αρχικά το όνομα του από την εφαρμογή του σε σουπερμάρκετ, όπου τα προϊόντα είναι τα αντικείμενα και τα καλάθια είναι οι αγορές που πραγματοποιούνται από τους πελάτες. Το μοντέλο αυτό υπέθετε ότι όλα τα σημαντικά αντικείμενα έχουν παρόμοιες συχνότητες εμφανίσεων, και είναι της ίδιας φύσης. Παρόλα αυτά σε πραγματικές εφαρμογές και ειδικότερα σε δεδομένα σουπερμάρκετ, ορισμένα αντικείμενα εμφανίζονται πρακτικά σε κάθε συναλλαγή ενώ άλλα πολύ σπάνια. Συνήθως οι χρήστες ενδιαφέρονται για τους κανόνες αυτούς που περιλαμβάνουν τα συχνά αντικείμενα αλλά αρκετές φορές εξίσου ενδιαφέροντα για αυτούς είναι και τα σπάνια αντικείμενα. Οι λόγοι είναι διάφοροι, όπως για παράδειγμα επειδή τα προϊόντα αυτά μπορούν να θεωρηθούν σαν loss leaders προϊόντα ή επειδή είναι στην διαδικασία του ελέγχου αγοράς, ή επειδή έχουν πολύ μεγάλα περιθώρια κέρδους κτλ.

Για παράδειγμα οι πωλήσεις αντηλιακών λοσιόν στα σουπερμάρκετ κατά τους καλοκαιρινούς μήνες είναι σίγουρα σημαντικές (για να μην αναφέρουμε τα περιθώρια κέρδους τους), αλλά συγκρινόμενα με τις πωλήσεις προϊόντων που πωλούνται ολόκληρο το χρόνο και σχεδόν σε καθημερινή βάση (π.χ. γαλακτοκομικά) είναι αμελητέες. Έτσι θα πρέπει είτε να ξέρουμε εκ των προτέρων ποια σπάνια αντικείμενα είναι ενδιαφέροντα, ή πρέπει να βρούμε κάποιο τρόπο να τα ανακαλύψουμε. Φυσικά ακόμα και εάν γνωρίζαμε ποια αντικείμενα είναι ενδιαφέροντα για εμάς θα έπρεπε να μαντέψουμε την σωστή MIS τιμή για αυτά, να τρέξουμε κάποιο αλγόριθμο, να ελέγξουμε τα αποτελέσματα και αν δεν είμαστε ευχαριστημένοι να επαναλάβουμε ολόκληρη την διαδικασία με καινούργιες MIS τιμές από την αρχή ξανά και ξανά.

8.1.1 Ενδιαφέροντα Αντικείμενα

Σαν ενδιαφέρον μπορεί να θεωρηθεί κάθε αντικείμενο το οποίο θα πρέπει να λαμβάνεται υπόψη όταν δημιουργούμε τους κανόνες συσχετίσεων, ακόμα και αν δεν ικανοποιεί το κατώφλι υποστήριξης ή οποιαδήποτε άλλο μέτρο που χρησιμοποιείται για την εύρεση όλων των συχνών αντικειμένων. Παραδείγματα τέτοιων κριτηρίων για τα ενδιαφέροντα αντικείμενα θα μπορούσαν να είναι το περιθώριο κέρδους, ο χρόνος ζωής, η αντοχή κ.α. Δυστυχώς για όλα τα αντικείμενα δεν υπάρχει ακόμα κάποιος αυτόματος τρόπος ώστε να ανακαλύπτουμε ποια θα πρέπει να θεωρούνται ως σημαντικά. Ο χρήστης θα πρέπει να ξέρει εκ των προτέρων ποια

αντικείμενα έχουν ιδιαίτερο ενδιαφέρον όπως επίσης ποια είναι η κατάλληλη τιμή υποστήριξης για αυτά. Εμείς ασχοληθήκαμε με μια συγκεκριμένη κατηγορία ενδιαφερόντων αντικειμένων, αναφορικά με αυτά τα οποία είναι πολύ συχνά αν θεωρηθούν σε σχέση σε ένα συγκεκριμένο κομμάτι της βάσης μας, αλλά ο συνολικός αριθμός εμφανίσεων τους δεν καταφέρνει να ξεπεράσει το κατώφλι υποστήριξης σε ολόκληρη τη βάση. Αυτή η περίπτωση είναι αρκετά κοινή και έχει εξαιρετικό ενδιαφέρον ειδικά σε μεγάλα σούπερ μάρκετ ή σε καταστήματα λιανεμπορίου όπου ορισμένα προϊόντα πωλούνται καθ' όλη τη διάρκεια του χρόνου ενώ άλλα μόνο σε εποχική βάση. Τέτοια προϊόντα μπορούν να φτάσουν τιμές υποστήριξης πολύ πάνω από 60% σε συγκεκριμένες περιόδους και μάλιστα με σημαντικά κέρδη, αλλά το υπόλοιπο του έτους δεν βρίσκονται καν σε κάποιο ράφι και σαν αποτέλεσμα η συνολική τους υποστήριξη να γίνεται ιδιαίτερα χαμηλή^{††††}. Έτσι αν κάποιος θέλει να παρουσιάσει μια ολοκληρωμένη άποψη όλων των κανόνων που αφορούν μια βάση, η οποία να αντανάκλα την πραγματικότητα, αυτά τα αντικείμενα θα πρέπει σε κάθε περίπτωση να ληφθούν υπόψη. Αυτή η περίπτωση ανακύπτει συχνά και σε πολλές άλλες περιπτώσεις, όπως για παράδειγμα με αντικείμενα που αγοράζονται σε συγκεκριμένες γεωγραφικές περιοχές μόνο, ή λόγω ειδικών περιστάσεων κοκ.

Παρόλα αυτά, ακόμα και αν αυτά τα αντικείμενα ταυτοποιηθούν κάποιος μπορεί να αναρωτηθεί ποια είναι η κατάλληλη τιμή υποστήριξης για ένα αντικείμενο, ή πως μπορεί να βρεθεί αυτή η τιμή. Μια απλοϊκή μέθοδος θα ήταν αν χρησιμοποιούσαμε έναν άλλο αλγόριθμο, όπως π.χ. τον Apriori, μόνο για το πρώτο βήμα προκειμένου να βρούμε τον αριθμό εμφανίσεων κάθε στοιχείου. Αυτό θα γινόταν πολύ γρήγορα και τότε θα γνωρίζαμε την υποστήριξη κάθε στοιχείου, και θα μπορούσαμε να διαλέξουμε την κατάλληλη τιμή υποστήριξης για τους σκοπούς μας. Παρόλα αυτά αυτή η προσέγγιση έχει τα εξής πρακτικά μειονεκτήματα.

1. Ο αλγόριθμος Apriori, ή οποιοσδήποτε άλλος αλγόριθμος θα έπρεπε να υλοποιηθεί ή να αποκτηθεί προκειμένου να χρησιμοποιηθεί για να βρει τις τιμές υποστήριξης.
2. Θα πρέπει να διαβάσουμε όλες τις τιμές υποστήριξης όλων των στοιχείων προκειμένου να βρούμε αυτά που μας ενδιαφέρουν.
3. Θα πρέπει να τρέξουμε ξανά μέσα απ' όλα τα στοιχεία, να αποδώσουμε MIS τιμές στο καθένα χειροκίνητα και μετά να ξανά-ξεκινήσουμε ολόκληρη την διαδικασία από την αρχή.
4. Δεν θα μπορούσαμε τελικά να πούμε αν κάποιο αντικείμενο το οποίο βρέθηκε σπάνιο σε ολόκληρη την βάση έγινε συχνό σε κάποιο κομμάτι αυτής.

^{††††} Από την άλλη βέβαια τα συγκεκριμένα αντικείμενα τις περισσότερες φορές παρ' όλες τις σχετικά χαμηλές πωλήσεις που παρουσιάζουν δεν μπορούν να αποκλειστούν.

Ειδικότερα το τελευταίο αυτό μειονέκτημα καθιστά αυτή την τεχνική ακατάλληλη για τους σκοπούς μας. Εν κατακλείδι οι υπάρχουσες προσεγγίσεις και αλγόριθμοι δεν προτείνουν ένα αποτελεσματικό τρόπο για την ανακάλυψη και διαχείριση αυτών των αντικειμένων. Σε αυτό το κεφάλαιο προτείνουμε μια προσέγγιση που αυτόματα βρίσκει τα αντικείμενα που είναι συχνά σε σχέση με ένα συγκεκριμένο κομμάτι της βάσης μας όπως επίσης και αντικείμενα που είναι συχνά σε σχέση με ολόκληρη τη βάση, τους αποδίδει τις σωστές MIS τιμές και τελικά παράγει όλους τους κανόνες.

Το συγκεκριμένο κεφάλαιο είναι οργανωμένο ως εξής. Στην ενότητα 8.2 παρουσιάζουμε την έννοια των τοπικά συχνών αντικειμένων. Στις ενότητες 8.3 και 8.4 παρουσιάζουμε δύο αλγορίθμους που βρίσκουν τα συχνά αντικείμενα ενώ λαμβάνουν υπόψη τα τοπικά συχνά αντικείμενα. Τέλος στην ενότητα 8.5 παρουσιάζουμε ορισμένα πειραματικά αποτελέσματα, και κλείνουμε με την ενότητα 8.6 όπου δίνουμε τα συμπεράσματα και μελλοντικές κατευθύνσεις.

8.2 Τοπικά Συχνά Αντικείμενα

Ουσιώδης για την προσέγγισή μας είναι η διαφοροποίηση ανάμεσα στα αντικείμενα που είναι συχνά και σε αυτά που θα πρέπει να θεωρούνται συχνά σε σχέση με ένα συγκεκριμένο κομμάτι της βάσης μας. Έτσι προσπαθούμε να δώσουμε κατ' αρχήν έναν ορισμό του τι ορίζουμε σαν Τοπικά Συχνά Αντικείμενα. Ένα **Τοπικά Συχνό - Locally Frequent (LF)** αντικείμενο είναι ένα αντικείμενο το οποίο είναι πολύ συχνό σε σχέση με ένα συγκεκριμένο κομμάτι της βάσης μας (π.χ. για πωλήσεις που πραγματοποιήθηκαν σε μια συγκεκριμένη χρονική περίοδο, σε ένα συγκεκριμένο τμήμα της επιχείρησης κοκ.) αλλά όχι σε σχέση με ολόκληρη τη βάση (οι συνολικές του εμφανίσεις είναι κάτω από το κατώφλι υποστήριξης). Αυτό μας αφήνει μόνο το πρόβλημα του πως μπορούμε να ξεχωρίσουμε ανάμεσα στα τοπικά συχνά αντικείμενα και σε αυτά που είναι απλώς σπάνια. Ένα αντικείμενο καλείται τοπικά συχνό αν ο λόγος τοπικής συχνότητας του είναι μεγαλύτερος από ένα καθορισμένο όριο. Ορίζουμε σαν λόγο τοπικής συχνότητας (local frequency ratio – LFR) τον λόγο του αριθμού των εμφανίσεων ενός αντικειμένου σε μια συγκεκριμένη διαμέριση της βάσης προς αυτή την διαμέριση.

$$\text{LFR} = \text{Itemset interval count} / \text{Interval}$$

Αν αυτό ο λόγος αυτός είναι πάνω από το όριο που έχουν θέσει οι χρήστες, τότε αυτό το αντικείμενο θεωρείται σαν τοπικά συχνό. Ο συγκεκριμένος λόγος εξαρτάται άμεσα από το διάστημα που επιλέγουμε να μετρήσουμε την υποστήριξη των αντικειμένων (η επιλογή αυτών των διαμερίσεων εξηγείται στην ενότητα 8.2.1). Σε μια βάση μπορούν να υπάρχουν τοπικά συχνά αντικείμενα διαφόρων τάξεων (π.χ. μονοσύνολα, δισύνολα, τρι-σύνολα κοκ.), και ο χρήστης μπορεί να ορίσει την τάξη των αντικειμένων τα οποία και τον ενδιαφέρουν. Αν αποφασίσουμε να ψάξουμε και να συμπεριλάβουμε όλα τα υψηλότερης τάξης αντικείμενα θα

χρησιμοποιούμε περισσότερους πόρους και προσπάθεια και θα δημιουργήσουμε ένα σημαντικό ποσό επιπρόσθετων κανόνων οι οποίοι πιθανώς να μην είναι σημαντικοί. Η αναζήτηση τοπικά συχνών αντικειμένων όλων των τάξεων είναι αυτό που ονομάζουμε πλήρες ψάξιμο. Τέλος όπως θα δούμε στη συνέχεια μόνο τα τοπικά συχνά στοιχεία - μονοσύνολα συνεισφέρουν στην δημιουργία περισσότερων υποψήφια αντικειμένων υψηλότερης τάξης. Όλα τα υπόλοιπα τοπικά συχνά αντικείμενα υψηλότερης τάξης απλά αναφέρονται στον χρήστη, και δεν παράγουν περισσότερα υποψήφια αντικείμενα ούτε και αλλάζουν την δικιά τους MIS τιμή ή αυτή άλλων αντικειμένων.

8.2.1 Εύρεση του πλήθους και του μεγέθους των διαμερίσεων

Η εύρεση του πλήθους και του μεγέθους των διαμερίσεων είναι μια απόφαση η οποία βασίζεται κύρια στη φύση της εξεταζόμενης περίπτωσης και στην εμπειρία και τις ανάγκες του data miner παρά σε συγκεκριμένα κριτήρια όπως γινόταν για παράδειγμα στο [109] όπου το πλήθος των διαμερίσεων επιλεγόταν με τέτοιο τρόπο ώστε να μειώσει το disk I/O, και περιοριζόταν από την διαθέσιμη κύρια μνήμη. Παρακάτω παραθέτουμε ορισμένες παρατηρήσεις για το πως μπορούμε να καταλήξουμε σ' αυτή την απόφαση.

1. Η φύση των δεδομένων: τα δεδομένα τα οποία συλλέγονται έρχονται από την ίδια επιχείρηση, από διαφορετικούς τομείς της ίδιας επιχείρησης, από διαφορετικά καταστήματα στην ίδια πόλη, περιοχή ή χώρα κ.ο.κ.
2. Το επιθυμητό αποτέλεσμα: ψάχνουμε για εποχιακά προϊόντα; Ψάχνουμε για προϊόντα που αγοράζονται σε ειδικές περιπτώσεις; Ψάχνουμε για προϊόντα που αγοράζονται σε συγκεκριμένα τμήματα ή υποκαταστήματα;

Αν για παράδειγμα έχουμε μια βάση η οποία αποτελείται από συναλλαγές που πραγματοποιήθηκαν κατά την διάρκεια ενός έτους, συνήθως ενδιαφερόμαστε για πωλήσεις οι οποίες πραγματοποιούνται σε μια συγκεκριμένη εποχή, ή σε συγκεκριμένους μήνες. Κατόπιν διαμερίζουμε την βάση μας σε εποχές και επεξεργαζόμαστε κάθε μια ξεχωριστά. Αν από την άλλη συλλέγουμε δεδομένα από διαφορετικά υποκαταστήματα μιας επιχείρησης (π.χ. μια μεγάλη αλυσίδα supermarket, μια τράπεζα, μια πολυεθνική κοκ.), θα επεξεργαζόμασταν κάθε υποκατάστημα ξεχωριστά. Αυτές ακριβώς οι απαιτήσεις μας περιορίζουν από το να αυτοματοποιήσουμε την όλη διαδικασία επιλογής διαμερίσεων και έτσι το πλήθος και το μέγεθος των διαμερίσεων που θα πρέπει να γίνουν εξαρτώνται σε μεγάλο βαθμό από τη φύση του οργανισμού ή γενικότερα τη φύση του προβλήματος μας και επηρεάζουν τόσο την ποσότητα όσο και την ποιότητα των παραγομένων συχνών αντικειμένων.

8.3 Ο αλγόριθμος LFI

Ο αλγόριθμος LFI (Locally Frequent Itemsets) βρίσκει αυτόματα όλα τα συχνά και τοπικά συχνά αντικείμενα, τους αποδίδει τις κατάλληλες MIS τιμές και βρίσκει όλους τους κανόνες συσχετίσεων χωρίς ανθρώπινη παρέμβαση. Ο χρήστης θέτει μόνο το επιθυμητό κατώφλι υποστήριξης (minsup) και κατώφλι τοπικής συχνότητας - minimum local frequency (minLFR) όπως επίσης και το πλήθος και το μέγεθος των διαμερίσεων κατά την έναρξη του αλγορίθμου. Η διαδικασία που χρησιμοποιείται για την απόδοση τιμών MIS στα στοιχεία μοιάζει με αυτή που χρησιμοποιείται και στο [89], αλλά ξεπερνάει ορισμένα μεγάλα μειονεκτήματα και ανεπάρκειες που είχε η συγκεκριμένη μέθοδος (όπως ο μάλλον αυθαίρετος τρόπος απόδοσης MIS τιμών). Έτσι τελικά αυτό που κάνουμε είναι η απόδοση των MIS τιμών σύμφωνα με τον παρακάτω εξίσωση:

$$MIS(i) = \begin{cases} \min sup & , i = l \arg e \\ \min ratio & , i = locally_frequent \end{cases} \quad \text{Εξίσωση 6}$$

Αν κάποιο στοιχείο βρεθεί συχνό τότε η MIS τιμή του είναι η τιμή του κατωφλίου υποστήριξης. Αν βρεθεί Τοπικά Συχνό σε μια ή περισσότερες διαμερίσεις, τότε υπολογίζουμε το λόγο των εμφανίσεων που είχε σε κάθε διαμέριση προς τον συνολικό αριθμό των συναλλαγών σε ολόκληρη τη βάση και αποδίδουμε σαν MIS τιμή τον χαμηλότερο λόγο σε όλες τις διαμερίσεις (Εικόνα 23). Όλα τα υπόλοιπα στοιχεία είναι απλά σπάνια και έτσι απορρίπτονται. Ο χρήστης έχει την δυνατότητα να ορίσει για μέχρι ποιας τάξης τοπικά συχνά αντικείμενα ενδιαφέρεται. Ας δούμε όμως πως λειτουργεί ο αλγόριθμος μας αναλυτικότερα:

Κατ' αρχήν ο χρήστης χωρίζει την βάση του σε ορισμένες λογικές, μη-επικαλυπτόμενες διαμερίσεις, και ξεκινά να επεξεργάζεται κάθε διαμέριση ξεχωριστά (Εικόνα 21, γραμμή 1). Ξεκινούμε μετρώντας τον αριθμό των εμφανίσεων όλων των αντικειμένων στην πρώτη διαμέριση. Όταν τελειώσουμε με την συγκεκριμένη διαμέριση ελέγχουμε αν ο αριθμός των εμφανίσεων κάθε αντικειμένου είναι μεγαλύτερος ή ίσος με το κατώφλι υποστήριξης (Εικόνα 23, γραμμές 8-10). Για κάθε αντικείμενο για το οποίο αυτό είναι αληθές το σημειώνουμε σαν συχνό και η MIS τιμή που του αποδίδουμε είναι ίση με το κατώφλι υποστήριξης. Επίσης δεν χρειάζεται πλέον να ελέγξουμε αν το συγκεκριμένο αντικείμενο είναι τοπικά συχνό σε αυτή ή στις επόμενες διαμερίσεις. Αν από την άλλη ο αριθμός εμφανίσεων ενός αντικειμένου είναι μικρότερος από το κατώφλι υποστήριξης τότε υπολογίζουμε το LFR του και το συγκρίνουμε με το minLFR (Εικόνα 23, γραμμή 11). Αν είναι μεγαλύτερο ή ίσο από την τιμή αυτή, τότε το σημειώνουμε σαν τοπικά συχνό (Locally Frequent – LF Εικόνα 23, γραμμή 15). Η MIS τιμή που αποδίδεται στο συγκεκριμένο αντικείμενο είναι ίση προς τον αριθμό των εμφανίσεων στη

συγκεκριμένη διαμέριση δια του αριθμού των συναλλαγών σε ολόκληρη τη βάση (Εικόνα 23, γραμμή 16).

```

1.      P=partition_database(D)
2.      K=number of transactions
3.      n=number of partitions
4.      m=number of distinct 1-itemsets
5.      count_1-itemsets
6.      S=sort(m,MIS)
7.      L1={large 1-itemsets};
8.      for (k=2;Lk-1!=∅ 0;k++) do
9.          Ck=candidate_gen(Lk-1)
10.         forall transactions t ∈ D do
11.             Ct=subset(Ck,t);
12.             forall candidates c ∈ Ct do
13.                 c.count++;
14.             end
15.             Lk={c ∈ Ck | c.count ≥ MIS(c[1])}
16.         end
17.     Answer=∪k Lk;

```

Εικόνα 21: Ο Αλγόριθμος LFI

Κατόπιν επεξεργαζόμαστε τις επόμενες διαμερίσεις. Για κάθε αντικείμενο που βρίσκεται σε μια προηγούμενη διαμέριση συχνό (δηλ. το πλήθος των εμφανίσεων του είναι ήδη ίσο ή μεγαλύτερο από το κατώφλι υποστήριξης) απλά συνεχίζουμε να το μετράμε προκειμένου να βρούμε την τελική του υποστήριξη (Εικόνα 23, γραμμές 8-10). Κάθε αντικείμενο το οποίο βρίσκεται σε μια προηγούμενη διαμέριση τοπικά συχνό (Locally Frequent - LF), ελέγχεται ξανά προκειμένου να δούμε αν είναι τοπικά συχνό σε μια άλλη διαμέριση. Αν βρούμε ότι είναι τοπικά συχνό και σε μια άλλη διαμέριση τότε ελέγχουμε εάν το πλήθος των εμφανίσεων του στην τρέχουσα διαμέριση δια του συνολικού αριθμού των συναλλαγών σε ολόκληρη τη βάση μας είναι μικρότερο ή ίσο από την MIS τιμή που του έχει αποδοθεί σε μια προηγούμενη διαμέριση όπου βρέθηκε τοπικά συχνό (Εικόνα 23, γραμμή 12). Αν κάτι τέτοιο συμβαίνει τότε αυτή είναι η νέα τιμή MIS που του αποδίδεται (Εικόνα 23, γραμμή 13). Αν όχι, τότε η MIS τιμή παραμένει η ίδια και συνεχίζουμε να το καταμετρούμε προκειμένου να δούμε αν οι

εμφανίσεις του είναι μεγαλύτερες ή ίσες από το κατώφλι υποστήριξης (δηλ. αν γίνεται τελικά συχνό). Μετά την ολοκλήρωση ενός περάσματος από τα δεδομένα, μπορούμε να βρούμε τρεις διαφορετικούς τύπους στοιχείων, κατηγοριοποιημένα σύμφωνα με το πλήθος των εμφανίσεων τους:

- ▮ **Συχνά αντικείμενα.** Αντικείμενα για τα οποία ο αριθμός των εμφανίσεων τους είναι μεγαλύτερος από το κατώφλι υποστήριξης.
- ▮ **Τοπικά Συχνά Αντικείμενα.** Αντικείμενα τα οποία ευρέθησαν συχνά σε μια διαμέριση της βάσης μας (το LFR τους υπερέβαινε το $\min LFR$), αλλά ο συνολικός αριθμός εμφανίσεων τους είναι κάτω από το κατώφλι υποστήριξης.
- ▮ **Σπάνια Αντικείμενα.** Αντικείμενα για τα οποία ο αριθμός των εμφανίσεων τους είναι κάτω από το κατώφλι υποστήριξης και το LFR τους δεν ξεπερνούσε το όριο του χρήστη σε καμία διαμέριση.

```

1.  procedure candidate-gen
2.  Join step
3.  insert into  $C_k$ 
4.  select  $p.item_1, p.item_2, \dots, p.item_{k-1}, q.item_{k-1}$ 
5.      from  $L_{k-1} p, L_{k-1} q$ 
6.      where  $p.item_1 = q.item_1, \dots, p.item_{k-2} = p.item_{k-2}, p.item_{k-1} < q.item_{k-1}$ 
7.  Prune step
8.  for each itemset  $c \in C_k$  do
9.      for each  $(k-1)$ -subset  $s$  of  $c$  do
10.         if ( $c[1] \in s$ ) or ( $MIS(c[2]) = MIS(c[1])$ ) then
11.            if ( $s \notin L_{k-1}$ ) then delete  $c$  from  $C_k$ 
12.  end procedure

```

Εικόνα 22: Η διαδικασία candidate-gen

```

1.  procedure count_1-itemsets
2.      K=number of transactions
3.      n=number of partitions
4.      m=number of distinct 1-itemsets
5.      for j=1 to n do
6.          for i=1 to m do
7.              read_in_partition( $p_j \in P$ )
8.              if i.count  $\geq$  minsup & i.type=large
9.                  i.count= i.count1+...+i.countj
10.                 i.mis=minsup; i.type=large
11.             elseif i.mis  $\geq$  minLFR
12.                 if i.type=locally_frequent & i.mis > i.countj/K
13.                     i.mis= i.countj/K
14.                 else
15.                     i.type=locally_frequent
16.                     i.mis= i.countj/K
17.             end
18.         end procedure

```

Εικόνα 23: Η διαδικασία count_1-itemsets

Το επόμενο βήμα το οποίο θα πρέπει να γίνει είναι η διάταξη των στοιχείων σύμφωνα με τις MIS τιμές τους προκειμένου να ικανοποιούν την sorted closure ιδιότητα, αφού και πάλι χρησιμοποιούμε πολλαπλές τιμές υποστήριξης για τα διαφορετικά στοιχεία. Αυτή είναι μια απλή διαδικασία η οποία τοποθετεί πρώτο το τοπικά συχνό στοιχείο με την χαμηλότερη MIS τιμή, ακολουθούμενο από όλα τα υπόλοιπα τοπικά συχνά στοιχεία σε αύξουσα σειρά υποστήριξης και τέλος από όλα τα άλλα συχνά στοιχεία (Εικόνα 21, γραμμή 6). Αυτά τα διατεταγμένα στοιχεία αποτελούν τα συχνά μονοσύνολα (1-itemsets), τα οποία συνδυάζονται προκειμένου να σχηματίσουν τα υποψήφια δισύνολα (2-itemsets). Αν αποφασίσουμε να μην ψάξουμε για τοπικά συχνά αντικείμενα μεγαλύτερης τάξης τότε από το σημείο αυτό κ έπειτα ο αλγόριθμος μας επεξεργάζεται τα δεδομένα μας συνολικά και όχι σε διαμερίσεις όπως συνέβαινε για τα μονοσύνολα. Έτσι για παράδειγμα στο τέλος του δεύτερου περάσματος από

τα δεδομένα μας κάθε υποψήφιο δισύνολο ελέγχεται για το εάν έχει αριθμό εμφανίσεων πάνω από το κατώφλι υποστήριξης ή όχι και ακολούθως χαρακτηρίζεται σαν συχνό ή σπάνιο.

Αν από την άλλη υποθέσουμε ότι ο χρήστης θέλει να βρει επίσης τα τοπικά συχνά αντικείμενα μεγαλύτερης τάξης (k -αντικείμενα, όπου $k \geq 2$) τότε η διαδικασία που ακολουθείται είναι ελαφρώς διαφορετική. Ξανά επεξεργαζόμαστε τη βάση μας σε διαμερίσεις όπως και προηγούμενα, ξεκινώντας με τα δισύνολα. Στο τέλος κάθε διαμερίσης ελέγχουμε τις εμφανίσεις κάθε δισυνόλου σε σχέση με την MIS τιμή του και σε σχέση με το LFR του. Αν ένα δισύνολο βρεθεί τοπικά συχνό τότε απλά το χαρακτηρίζουμε σαν τοπικά συχνό όπως ακριβώς συνέβαινε με τα μονοσύνολα. Αν όχι συνεχίζουμε να το μετρούμε στις επόμενες διαμερίσεις. Παρόλ' αυτά αντίθετα με τα μονοσύνολα οι MIS τιμές των τοπικά συχνών δισυνόλων παραμένουν ίδιες (όπως η αρχική τιμή) και στο τέλος όλων των διαμερίσεων αν το συγκεκριμένο αντικείμενο παρέμεινε τοπικά συχνό (ο τελικός αριθμός εμφανίσεων του δεν ξεπέρασε το κατώφλι υποστήριξης), απλά αναφέρεται σαν τοπικά συχνό και δεν συνεισφέρει στην δημιουργία υποψηφίων τρισυνόλων. Ο αλγόριθμος μας καθώς και όλες οι διαδικασίες που τον αποτελούν δίνονται στις εικόνες 18, 19 και 20.

8.4 Ο Αλγόριθμος P-DIC

Ο αλγόριθμος που περιγράφηκε προηγούμενα ανακαλύπτει όλα τα συχνά και τοπικά συχνά αντικείμενα αλλά η απόδοση του θα μπορούσε να βελτιωθεί κάνοντας χρήση μιας διαφορετικής τεχνικής. Αυτό που προτείνουμε είναι η χρήση μιας τροποποιημένης έκδοσης του αλγορίθμου Dynamic Itemset Counting – DIC [27], με την ονομασία Partition DIC (P-DIC), αφού ολοκληρωθεί το πρώτο βήμα, δηλαδή αφότου όλα τα συχνά και τοπικά συχνά μονοσύνολα βρεθούν, και έχουν διαταχθεί σύμφωνα με τις MIS τιμές τους. Ιδανικά θα θέλαμε να χρησιμοποιήσουμε τον τροποποιημένο αλγόριθμο DIC από την αρχή αλλά η φύση του προβλήματος δεν μας επιτρέπει να κάνουμε κάτι τέτοιο, αφού πρέπει να γνωρίζουμε τις MIS τιμές όλων των μονοσυνόλων και να τα έχουμε διατάξει σύμφωνα με αυτές τις τιμές από την αρχή. Παρόλα αυτά αφού τα μονοσύνολα σε τέτοιες εφαρμογές κυμαίνονται συνήθως από εκατοντάδες μέχρι μερικές χιλιάδες, και ο χρόνος που χρειάζεται για να τα επεξεργαστούμε είναι γενικά μικρός δεν υπάρχει μεγάλη απώλεια αν χρησιμοποιήσουμε τον P-DIC σε ένα μετέπειτα στάδιο. Στην πράξη ο αριθμός των υποψηφίων δισυνόλων και ιδιαίτερα των τρισυνόλων είναι αυτά που κύρια επηρεάζουν την συνολική απόδοση, αφού το πλήθος τους είναι συγκριτικά πολύ μεγάλο [98]. Ο αλγόριθμος P-DIC έχει επίσης την δυνατότητα να ορίζουμε μέχρι ποια τάξη τοπικά συχνών αντικειμένων ενδιαφερόμαστε να βρούμε.

Έτσι, κατ' αρχήν βρίσκονται όλα τα συχνά και τοπικά συχνά μονοσύνολα και τους αποδίδεται μια MIS τιμή χρησιμοποιώντας την διαδικασία που εξηγήθηκε στην προηγούμενη

προσέγγιση για να βρούμε όλα τα μονοσύνολα. Κατόπιν κάθε αντικείμενο στην βάση μας χαρακτηρίζεται με μια διαφορετική κατάσταση. Εκτός από τις τέσσερις καταστάσεις που χρησιμοποιούνταν επίσης από τον αλγόριθμο DIC, εισάγουμε δύο επιπρόσθετες καταστάσεις για τους σκοπούς μας. Έτσι ο αλγόριθμος μας χαρακτηρίζει τα αντικείμενα με έξι διαφορετικούς τρόπους, οι οποίοι και είναι:

- ▣ **Διακεκομμένος Κύκλος (Dashed Circle - DC)** – πιθανολογούμενο σπάνιο αντικείμενο – ένα αντικείμενο το οποίο ακόμα μετράμε και οι εμφανίσεις του είναι λιγότερες από την MIS τιμή του – επίσης η αρχική κατάσταση όλων των αντικειμένων.
- ▣ **Συμπαγής κύκλος (Solid Circle - SC)** – επιβεβαιωμένο σπάνιο αντικείμενο – ένα αντικείμενο το οποίο τελειώσαμε το μέτρημα του αλλά οι εμφανίσεις του ήταν κάτω από την δικιά του MIS τιμή και δεν βρέθηκε τοπικά συχνό σε κάποια διαμέριση.
- ▣ **Διακεκομμένο Τετράγωνο (Dashed Square - DS)** – πιθανολογούμενο συχνό αντικείμενο – ένα αντικείμενο το οποίο ακόμα μετράμε, και οι εμφανίσεις του ήδη ξεπερνούν την MIS τιμή του.
- ▣ **Συμπαγές Τετράγωνο (Solid Square - SS)** – βεβαιωμένο συχνό αντικείμενο – ένα αντικείμενο το οποίο τελειώσαμε το μέτρημα του σε όλη τη βάση και που ξεπερνά την MIS τιμή του.
- ▣ **Διακεκομμένο Τρίγωνο (Dashed Triangle - DT)** – ένα αντικείμενο που βρέθηκε τοπικά συχνό σε κάποια διαμέριση και το οποίο ακόμα μετράμε προκειμένου να δούμε αν οι συνολικές εμφανίσεις του είναι μεγαλύτερες από την MIS τιμή του.
- ▣ **Διακεκομμένο Τρίγωνο (Solid Triangle - ST)** – ένα αντικείμενο το οποίο τελειώσαμε να μετράμε σε ολόκληρη τη βάση, και βρέθηκε τοπικά συχνό σε κάποια διαμέριση αλλά όπου οι συνολικές του εμφανίσεις ήταν κάτω από την MIS τιμή του.

Τα μονοσύνολα που βρίσκονται είτε συχνά είτε τοπικά συχνά χαρακτηρίζονται όλα SS – συμπαγή τετράγωνα, με όλα τα υπόλοιπα σπάνια μονοσύνολα να χαρακτηρίζονται σαν SC – συμπαγής κύκλος στο τέλος του πρώτου περάσματος (Εικόνα 24, στιγμιότυπο 1, στοιχεία 1-2-3-4). Όλα τα υπόλοιπα αντικείμενα ψηλότερης τάξης μπορούν να είναι οποιαδήποτε από τις έξι καταστάσεις. Ο αλγόριθμος P-DIC κατόπιν παίρνει σαν είσοδο όλα τα μονοσύνολα που είναι χαρακτηρισμένα σαν SS στο προηγούμενο πέρασμα, και παράγει όλα τα υποψήφια δισύνολα.

Εικόνα 24: Το “τρέξιμο” του αλγορίθμου P-DIC

Αν ένα άμεσο υπερσύνολο ενός μονοσύνολου το οποίο είναι χαρακτηρισμένο σαν DS – διακεκομμένο τετράγωνο, έχει όλα τα υποσύνολα συμπαγή τετράγωνα, τότε κάνουμε την κατάσταση του DC – διακεκομμένο κύκλο και ξεκινάμε να το μετράμε (Εικόνα 24, στιγμιότυπο 1, αντικείμενα 13 – 14 – 34). Κατόπιν ο αλγόριθμος ξεκινά να επεξεργάζεται όλα τα αντικείμενα υψηλότερης τάξης. Μετά από M, όπου M είναι το μέγεθος της πρώτης διαμέρισης, ελέγχουμε το μετρητή κάθε αντικειμένου που μετράμε σε σχέση με την MIS τιμή του. Αν ο μετρητής ενός αντικειμένου είναι μεγαλύτερος ή ίσος προς την MIS τιμή του τότε η κατάσταση του αλλάζει σε DS – διακεκομμένο τετράγωνο (Εικόνα 24, στιγμιότυπο 2, αντικείμενο 13). Αν όχι τότε υπολογίζουμε το LFR του και το συγκρίνουμε με το οριζόμενο από το χρήστη ελάχιστο LFR. Αν είναι μεγαλύτερο από το ελάχιστο LFR τότε αλλάζουμε την κατάσταση του σε DT – διακεκομμένο τρίγωνο και προχωρούμε με τις επόμενες M συναλλαγές (Εικόνα 24, στιγμιότυπο 2, αντικείμενο 14).

Αν ένα άμεσο υπερσύνολο ενός k-συνόλου (όπου $k \geq 2$) που είναι χαρακτηρισμένο DS – διακεκομμένο τετράγωνο έχει όλα τα υποσύνολα του σαν συμπαγή ή διακεκομμένα τετράγωνα, τότε κάνουμε την κατάσταση του DC – διακεκομμένο κύκλο και ξεκινάμε να το μετράμε. Υπάρχει παρόλα αυτά μια εξαίρεση όπου ένα υποσύνολο ενός αντικειμένου δεν είναι συχνό

(δηλ. δεν είναι χαρακτηρισμένο DS ή SS), και παρόλα αυτά το υποψήφιο υπερσύνολο δεν μπορεί να περικοπεί. Αυτή η περίπτωση εμφανίζεται όταν το υποσύνολο που δεν περιέχει το πρώτο στοιχείο δεν είναι συχνό και δεν είμαστε σίγουροι αν η υποστήριξη αυτού του υποσυνόλου είναι μικρότερη από την υποστήριξη του πρώτου στοιχείου. Στην Εικόνα 24, στο στιγμιότυπο 3 το υπερσύνολο $\langle 1,3,4 \rangle$ δημιουργείται παρόλο που το υποσύνολο $\langle 3,4 \rangle$ δεν είναι συχνό, επειδή δεν είμαστε σίγουροι ότι η υποστήριξη του είναι επίσης πάνω από την MIS τιμή του στοιχείου 1 ή η MIS τιμή του στοιχείου 1 είναι η ίδια όπως αυτή του στοιχείου 3 (για περισσότερες λεπτομέρειες δείτε και το [89]). Αντίθετα με τον αλγόριθμο LFI, ο αλγόριθμος P-DIC επεξεργάζεται την βάση σε διαμερίσεις (αφού αυτό είναι και το μεγάλο του πλεονέκτημα) ανεξάρτητα αν ο χρήστης ψάχνει μόνο για τοπικά συχνά μονοσύνολα. Αυτό που αλλάζει στον P-DIC αν δεν ψάχνουμε για τοπικά συχνά αντικείμενα υψηλότερης τάξης είναι ότι ψάχνουμε μόνο για συχνά αντικείμενα, και τα αντικείμενα δεν μπορούν να χαρακτηριστούν σαν τρίγωνα.

Όταν ένα αντικείμενο έχει καταμετρηθεί σε όλες τις συναλλαγές τότε ελέγχουμε ξανά τον μετρητή του σε σχέση με την MIS τιμή του. Αν η κατάσταση του ήταν DC – διακεκομμένος κύκλος, και ο μετρητής του είναι τελικά μεγαλύτερος ή ίσος προς την MIS τιμή του τότε αλλάζουμε την κατάσταση του σε SS – συμπαγή κύκλο. Αν η κατάσταση του ήταν DS – διακεκομμένο τετράγωνο, τότε απλά την αλλάζουμε σε SS – συμπαγές τετράγωνο (Εικόνα 24, στιγμιότυπο 4, αντικείμενα 13-14). Στην Εικόνα 25 βλέπουμε τις μεταβάσεις ανάμεσα στις διαφορετικές καταστάσεις. Ο αλγόριθμος μας τερματίζει με την απουσία κάποιου διακεκομμένου αντικειμένου. Κάθε υποψήφιο δισύνολο, λαμβάνει σαν MIS τιμή την μικρότερη MIS τιμή ανάμεσα στα δύο μονοσύνολα που το σχηματίζουν. Επίσης η σειρά κάθε νέου αντικειμένου διατηρείται, βάζοντας πρώτα το μονοσύνολο που έχει την μικρότερη MIS τιμή, και κατόπιν όλα τα υπόλοιπα στοιχεία διατεταγμένα σε αύξουσα σειρά σύμφωνα με τις MIS τιμές τους.

Εικόνα 25: Οι αλλαγές καταστάσεων για το αλγόριθμο P-DIC

8.5 Πειραματικά Αποτελέσματα

Σε αυτή την ενότητα περιγράφουμε τα πειράματα που πραγματοποιήσαμε προκειμένου να δοκιμάσουμε την προσέγγιση μας, καθώς και τα αποτελέσματά τους. Τα πειράματα χωρίζονται σε τρία διαφορετικά τμήματα. Στο πρώτο τμήμα συγκρίνουμε τους αλγορίθμους μας μεταξύ τους για διάφορες τιμές των παραμέτρων που χρησιμοποιούν. Στο επόμενο τμήμα αξιολογούμε την απόδοση των δύο αλγορίθμων μας σε σχέση με ορισμένους παραδοσιακούς αλγορίθμους εύρεσης κανόνων συσχετίσεων και στο τρίτο και τελευταίο τμήμα επιχειρούμε μια ποιοτική αξιολόγηση της ορθότητας και της χρησιμότητας των ευρημάτων μας. Το μέτρο σύγκρισης της απόδοσης για τα δύο πρώτα τμήματα ήταν οι χρόνοι εκτέλεσης των αλγορίθμων για διάφορες βάσεις δεδομένων.

Τα πρώτα δύο τμήματα των πειραμάτων μας πραγματοποιήθηκαν σε συνθετικές βάσεις δεδομένων που δημιουργήθηκαν χρησιμοποιώντας την γεννήτρια συνθετικών δεδομένων καθώς και διάφορες άλλες πραγματικές βάσεις. Η συγκεκριμένη γεννήτρια έχει επεξηγηθεί και τεκμηριωθεί επαρκώς στο [10], όπως επίσης σε πολλές μετέπειτα εργασίες. Θεωρείται το στάνταρ σημείο αναφοράς για τις περισσότερες συγκρίσεις στον τομέα της εύρεσης κανόνων συσχετίσεων. Τα κυριότερα χαρακτηριστικά των χρησιμοποιούμενων βάσεων φαίνονται στον Πίνακα 13, ενώ οι βάσεις αυτές καθ' εαυτές όπως και λεπτομερής περιγραφή τους μπορούν να βρεθούν στο <http://fimi.cs.helsinki.fi/data/>. Στο τρίτο και τελευταίο τμήμα των πειραμάτων μας χρησιμοποιήσαμε πραγματικά δεδομένα. Όλα τα πειράματα πραγματοποιήθηκαν χρησιμοποιώντας το χειρότερο δυνατό σενάριο της πλήρους αναζήτησης (δηλ. ο χρήστης έψαχνε για τοπικά συχνά αντικείμενα όλων των τάξεων). Για τα πειράματα μας υλοποιήσαμε τους αλγόριθμους MSAPriori, LFI και P-DIC σε C++, ενώ χρησιμοποιήσαμε την υλοποίηση του Ferenc Bodon [22] για τον αλγόριθμο Apriori.

8.5.1 LFI εναντίον P-DIC

Σε αυτή την πρώτη σειρά πειραμάτων ελέγξαμε την απόδοση των δύο αλγορίθμων μας μεταξύ τους σε σχέση με τον απαιτούμενο χρόνο για τις παραμέτρους που χρησιμοποιούν και οι δύο. Πιο συγκεκριμένα συγκρίναμε τον αλγόριθμο LFI ενάντια στον P-DIC για διαφορετικές τιμές του κατωφλίου υποστήριξης, του ελάχιστου LFR και του μεγέθους των διαμερίσεων για διάφορες βάσεις δεδομένων.

8.5.1.1 Επίδραση του Μεγέθους της Διαμέρισης

Όπως σημειώσαμε προηγούμενα (ενότητα 8.2.1) ο χρήστης δεν έχει την δυνατότητα να μεταβάλλει το μέγεθος και το πλήθος των διαμερίσεων που χρησιμοποιούνται, αφού η επιλογή τους σε μεγάλο βαθμό επηρεάζει την φύση και την ποσότητα των συχνών αντικειμένων.

Παρόλα αυτά, για πειραματικούς σκοπούς μόνο εξετάσαμε την απόδοση των αλγορίθμων μας για μεταβαλλόμενα μεγέθη των διαμερίσεων που χρησιμοποιούνται για την μέτρηση μας.

Για το συγκεκριμένο πείραμα διατηρήσαμε την ίδια τιμή για το κατώφλι υποστήριξης (2%) και ξεκινήσαμε με ένα συγκεκριμένο LFR (3%) σε ένα μέγεθος διαμέρισης ίσο προς 50.000 συναλλαγές. Ακολουθώντας μειώσαμε το μέγεθος των διαμερίσεων μαζί με το ελάχιστο LFR (διατηρώντας έτσι τον λόγο τους σταθερό) προκειμένου να διατηρήσουμε τον αριθμό των υποψηφίων και των συχνών αντικειμένων όσο το δυνατόν σταθερά. Ταυτόχρονα διατηρήσαμε το κατώφλι υποστήριξης ίδιο. Στην Εικόνα 26 βλέπουμε την επίδραση στην απόδοση της μεταβολής του μεγέθους της διαμέρισης.

Βλέπουμε ότι η απόδοση του αλγορίθμου P-DIC σε σχέση με τον LFI βελτιώνεται όσο τα μεγέθη των διαμερίσεων μικραίνουν αλλά μόνο μέχρι κάποια συγκεκριμένη τιμή (10.000 συναλλαγές), κάτω απ' όπου ο P-DIC ξεκινάει χειροτερεύει. Αυτό οφείλεται στο γεγονός ότι πράγματι τα υποψήφια αντικείμενα δημιουργούνται και μετρώνται νωρίτερα, αλλά κάτω από τις 10.000 συναλλαγές αυτό το όφελος αντισταθμίζεται από τις πάρα πολλές συγκρίσεις που έπρεπε να πραγματοποιούνται σε κάθε σταμάτημα. Παρόλα αυτά ακόμα και σε αυτά τα μεγέθη διαμερίσεων η απόδοση του P-DIC ακόμα παραμένει πολύ καλύτερη από αυτή του LFI.

Εικόνα 26: Η επίδραση της αλλαγής του μεγέθους των διαμερίσεων στην απόδοση

8.5.1.2 Επίδραση του Κατωφλίου Υποστήριξης

Σε αυτή τη σειρά πειραμάτων το LFR διατηρήθηκε σταθερό (3%) όπως επίσης και το μέγεθος της κάθε διαμέρισης (20.000 συναλλαγές) και μεταβάλλαμε μόνο το κατώφλι υποστήριξης. Πάλι η σχετική απόδοση των δύο αλγορίθμων παρουσιάζει διακυμάνσεις καθώς το κατώφλι

υποστήριξης λαμβάνει διαφορετικές τιμές αλλά σε όλες τις περιπτώσεις ο P-DIC παραμένει καλύτερος από τον LFI (Εικόνα 27).

Εικόνα 27: Η επίδραση της μεταβολής του κατώφλιου υποστήριξης στην απόδοση

8.5.1.3 Επίδραση του Local Frequency Ratio

Για τα συγκεκριμένα πειράματα διατηρήσαμε το κατώφλι υποστήριξης σταθερό (5%) όπως επίσης και το μέγεθος της διαμέρισης (25.000 συναλλαγές) και μεταβάλλαμε το ελάχιστο Local Frequency Ratio – LFR. Χρησιμοποιήσαμε μια σχετικά υψηλή τιμή για το κατώφλι υποστήριξης προκειμένου να επιτρέψουμε το κατώφλι LFR να επηρεάσει την απόδοση και των δύο αλγορίθμων σε μεγαλύτερο βαθμό, αφού ο αριθμός των εμφανίσεων των στοιχείων σε αυτές τις βάσεις δεν παρουσιάζει μεγάλες διακυμάνσεις. Όπως μπορούμε να δούμε και από την Εικόνα 28, ο αλγόριθμος P-DIC αποδίδει καλύτερα από τον αλγόριθμο LFI σε όλες τις περιπτώσεις, χωρίς ωστόσο να εμφανίζουν μεγάλες διαφορές.

Εικόνα 28: η επίδραση της μεταβολής του Local Frequency Ratio στην απόδοση

8.5.2 Οι αλγόριθμοι LFI και P-DIC σε σχέση με τους παραδοσιακούς αλγορίθμους

Πέρα από τον έλεγχο των αλγορίθμων μας μεταξύ τους θέλαμε να ελέγξουμε την απόδοση τους σε σχέση με κάποιους παραδοσιακούς αλγορίθμους για την εύρεση κανόνων συσχετίσεων. Όπως είδαμε νωρίτερα ο P-DIC αποδίδει καλύτερα από τον LFI σε όλες τις περιπτώσεις. Έτσι σε αυτά τα πειράματα ελέγξαμε την χειρότερη περίπτωση (δηλ. τον αλγόριθμο LFI) σε σχέση με τους αλγορίθμους Apriori [10], και MSApriori [89]. Όπως συζητήσαμε νωρίτερα (δες ενότητα 8.1.1) οι παραδοσιακοί αλγόριθμοι δεν λαμβάνουν υπόψη τους τα τοπικά συχνά αντικείμενα, αλλά κάτω από συγκεκριμένες συνθήκες μπορούν να “εξαναγκαστούν” να το κάνουν. Ας δούμε πως μέσω ενός παραδείγματος. Ας υποθέσουμε ότι έχουμε μια βάση την οποία χωρίζουμε σε δύο ίσες διαμερίσεις η κάθε μια 50% της αρχικής βάσης, και θέτουμε την ελάχιστη τιμή του LFR ίση με 20% και το κατώφλι υποστήριξης ίσο με 30%. Αν θέλαμε τώρα ο αλγόριθμος Apriori να έβρισκε τα τοπικά συχνά αντικείμενα που βρίσκονται επίσης από την προσέγγιση μας τότε θα έπρεπε να θέσουμε το κατώφλι υποστήριξης του ίσο με 10%. Η τιμή του κατωφλίου υποστήριξης που χρησιμοποιείται από τον αλγόριθμο Apriori υπολογίζεται χρησιμοποιώντας την ελάχιστη τιμή LFR και το μέγεθος της μικρότερης διαμέρισης των αλγορίθμων μας. Έτσι όσο χαμηλότερη είναι η ελάχιστη τιμή του LFR και όσο χαμηλότερο είναι το μέγεθος της μικρότερης διαμέρισης σε σύγκριση με το μέγεθος ολόκληρης της βάσης μας τόσο χαμηλότερη είναι και η τιμή του κατωφλίου υποστήριξης που θα πρέπει να χρησιμοποιηθεί. Το πρόβλημα είναι ότι τις περισσότερες φορές η τιμή του κατωφλίου υποστήριξης είναι τόσο χαμηλή που δημιουργεί ανυπέρβλητες

παρενέργειες στους αλγόριθμους που την χρησιμοποιούν. Πιο συγκεκριμένα όταν ο Apriori χρησιμοποιεί ένα τόσο χαμηλό ενιαίο κατώφλι υποστήριξης κατ' αρχήν βρίσκονται σαν συχνά όχι μόνο τα αντικείμενα που είναι τοπικά συχνά σε μια διαμέριση αλλά και κάθε άλλο αντικείμενο με τιμή υποστήριξης σε ολόκληρη τη βάση πάνω από το κατώφλι υποστήριξης. Αυτό σημαίνει ότι πολύ περισσότερα υποψήφια αντικείμενα θα πρέπει να δημιουργηθούν και να καταμετρηθούν και πολλά περισσότερα συχνά αντικείμενα θα βρεθούν, το οποίο με τη σειρά του σημαίνει μεγάλες απαιτήσεις σε χρόνο και μνήμη. Αυτό επιβεβαιώνεται επίσης και από το γεγονός ότι για ορισμένες βάσεις ο Apriori δεν κατάφερε καν να τελειώσει (στην πράξη οι χρόνοι εκτέλεσης ήταν τόσο μεγάλοι ώστε αναγκάζομαστε να τον σταματήσουμε). Αυτό που είναι όμως ακόμα πιο σημαντικό είναι ότι δεν υπάρχει κάποιος τρόπος για να ξεχωρίσουμε τα τοπικά συχνά αντικείμενα από τα αντικείμενα αυτά που είχαν απλά περισσότερες εμφανίσεις από το κατώφλι υποστήριξης σε ολόκληρη τη βάση. Τέλος είναι το γεγονός ότι ο αριθμός των συχνών αντικειμένων που βρέθηκαν ήταν τόσο μεγάλος που θα έπρεπε να κάνουμε εφαρμόσουμε τεχνικές εξόρυξης δεδομένων στους ίδιους τους κανόνες για να βρούμε αυτούς που πραγματικά θα μας ενδιαφέρουν. Όσον αφορά τον αλγόριθμο MSApriori η απόδοση MIS τιμών σε κάθε μονοσύνολο γίνεται με χρήση των ίδιων εξισώσεων (βλέπε Κεφάλαιο 2, παράγραφο 2.7) όπως αυτοί που χρησιμοποιούνταν στο [89].

Προκειμένου τώρα ο αλγόριθμος MSApriori να βρίσκει επίσης όλα τα τοπικά συχνά αντικείμενα θα έπρεπε να θέσουμε την ελάχιστη επιτρεπόμενη τιμή υποστήριξης (δηλ. την τιμή του LS) ίση προς 10%. Μ' αυτόν τον τρόπο κάθε στοιχείο με αριθμό εμφανίσεων σε ολόκληρη τη βάση (και κατά συνέπεια σε οποιαδήποτε διαμέριση της βάσης) πάνω από την ελάχιστη τιμή υποστήριξης θα ανακαλυπτόταν και θα αναφερόταν σαν συχνό. Και πάλι ισχύουν τα ίδια προβλήματα όπως με τον αλγόριθμο Apriori, δηλαδή λόγω της πολύ χαμηλής τιμής της μεταβλητής LS ο αλγόριθμος MSApriori δεν μπορεί να τερματίσει για ορισμένες βάσεις.

Πίνακας 13: Χαρακτηριστικά των Βάσεων μας

	Πλήθος Συναλλαγών	Πλήθος Αντικείμενων	Μέσο μήκος Συναλλαγής	Μέγεθος Βάσης (MB)
T40I10D100K	100.000	942	40	15
kosarak	990.002	41.270	8,1	32
connect	67.557	129	43	9
pumsb	49.046	2.113	74	16
accidents	340.183	468	45	35
BMS-WebView-1_mb	59.602	497	2,5	0,8
BMS-WebView-2_mb	77.512	3.340	5	2

Για τα συγκεκριμένα πειράματα χρησιμοποιήσαμε όλες τις βάσεις του Πίνακα 13, τις οποίες χωρίσαμε όλες σε τρεις ίδιες διαμερίσεις η κάθε μια με μέγεθος 37%, 33% και 30% της αρχικής βάσης. Ακολουθώντας αρχίσαμε με ένα συγκεκριμένο κατώφλι υποστήριξης και μια ελάχιστη τιμή LFR τα οποία και χαμηλώσαμε αμφότερα, δημιουργώντας έξι ζεύγη τιμών συνολικά. Η τιμή του LS για τον αλγόριθμο MSArriori όπως επίσης και η τιμή του κατωφλίου υποστήριξης για τον αλγόριθμο Arriori υπολογίστηκε χρησιμοποιώντας σαν βάση την ελάχιστη τιμή LFR από τους αλγορίθμους μας (όλες οι παράμετροι για τα συγκεκριμένα πειράματα δίνονται στον Πίνακα 14). Τέλος για τον αλγόριθμο MSArriori χρησιμοποιήσαμε τιμή για την παράμετρο β ίση προς 0,125. Στις Εικόνα 29 έως Εικόνα 32 βλέπουμε τη σχετική απόδοση του αλγορίθμου LFI σε σχέση με τον Arriori και τον MSArriori για τις διάφορες βάσεις.

Όπως αναμέναμε οι δύο αλγόριθμοι μας αποδίδουν καλύτερα από τους δύο παραδοσιακούς αλγορίθμους σε όλες τις περιπτώσεις. Ειδικότερα για τις βάσεις connect, pumsb και accidents τόσο ο Arriori όσο και ο MSArriori δεν μπορούσαν καν να τελειώσουν με τόσο χαμηλές τιμές υποστήριξης. Το πλήθος των υποψήφιων αντικειμένων που πρέπει να καταμετρηθούν όσο και ο τελικός αριθμός των συχνών αντικειμένων που βρίσκονταν από τους κλασσικούς

αλγορίθμους υπερβαίνουν κατά πολύ αυτά των αλγορίθμων μας και αυτό έχει μια άμεση επίδραση στον απαιτούμενο χρόνο και μνήμη.

Εικόνα 29: Οι σχετικοί χρόνοι εκτέλεσης για την βάση T40110D100K

Εικόνα 30: Οι σχετικοί χρόνοι εκτέλεσης για την βάση Korasak

Εικόνα 31: Οι σχετικοί χρόνοι εκτέλεσης για την βάση BMS-WebView-1_mb

Εικόνα 32: Οι σχετικοί χρόνοι εκτέλεσης για την βάση BMS-WebView-2_mb

8.5.3 Ποιοτική Αξιολόγηση

Η εξέταση της ποιοτικής αξίας των προτάσεων μας και η επαλήθευση της σημασίας τους αποδείχτηκε ένα αρκετά πολύπλοκο έργο. Σ' αυτό το τμήμα του πειράματος μας αφού το

κυρίως ζητούμενο μας δεν ήταν η απόδοση των αλγορίθμων μας (δηλ. οι χρόνοι εκτέλεσης τους ή γενικότερα οι απαιτούμενοι πόροι) θα έπρεπε να ελέγξουμε τα αντικείμενα που επιστρέφονται σαν συχνά από την προσέγγιση μας. Η πρώτη μας σκέψη ήταν να προσπαθήσουμε να την συγκρίνουμε με ορισμένους παραδοσιακούς αλγορίθμους για την εύρεση κανόνων συσχετίσεων (π.χ. τον Apriori, τον MSApriori κ.α.), αλλά αυτό απορρίφθηκε άμεσα. Ο λόγος ήταν γιατί απλούστατα όπως αναφέραμε και προηγούμενα θέτοντας το κατώφλι υποστήριξης πολύ χαμηλά στον αλγόριθμο Apriori (ή την τιμή του LS για τον αλγόριθμο MSApriori) ήταν ουσιαστικά σαν να τον καταδικάζαμε. Ο συνολικός αριθμός των συχνών αντικειμένων ήταν τόσο μεγάλος και ακόμα σημαντικότερο, η σχέση των συχνών αυτών αντικειμένων με το επιθυμητό αποτέλεσμα ήταν τόσο απογοητευτική που το τελικό σύνολο των αντικειμένων που ανακαλύψαμε δεν είχαν τελικά πρακτική χρήση. Για παράδειγμα στα προηγούμενα πειράματα στη βάση T40I10D100K, τρέχοντας τον Apriori με ένα κατώφλι υποστήριξης της τάξης του 0.01 και τους αλγορίθμους μας με μια τιμή κατωφλίου υποστήριξης 0.04 και ένα ελάχιστο LFR 0.033 είχε σαν αποτέλεσμα την εύρεση 65.228 συχνών αντικειμένων, το οποίο ήταν σχεδόν 14 φορές περισσότερο από τους αλγορίθμους μας. Για τις υπόλοιπες βάσεις οι διαφορές αυτές ήταν ακόμα μεγαλύτερες.

Μια άλλη πιθανή επιλογή θα ήταν η χρήση μετρικών που χρησιμοποιούνται σε άλλους τομείς για την αξιολόγηση της αποτελεσματικότητας μιας μεθόδου, όπως οι ευρέως γνωστές μετρικές της ανάκλησης και της ακρίβειας που χρησιμοποιούνται στην ανάκτηση πληροφορίας. Πιο συγκεκριμένα η ακρίβεια μιας μεθόδου ανάκτησης πληροφορίας ορίζεται σαν ο λόγος του αριθμού των σχετικών κειμένων (η γενικότερα απαντήσεων) που ανακτώνται προς τον συνολικό αριθμό των ανακτημένων κειμένων, ενώ η ανάκληση ορίζεται σαν ο λόγος του αριθμού των σχετικών κειμένων που ανακτώνται προς τον συνολικό αριθμό κειμένων σε ολόκληρη τη συλλογή μας (ανακτημένων και μη).

$$\text{Precision} = \frac{\text{αριθμός_ανάκτημενων_που_είναι_σχετικά}}{\text{συνολικός_αριθμός_ανακτημένων}}$$

$$\text{Recall} = \frac{\text{αριθμός_σχετικών_που_ανακτήθηκαν}}{\text{συνολικός_αριθμός_σχετικών}}$$

Στην περίπτωση μας, τα κείμενα είναι στην ουσία τα αντικείμενα στη βάση μας και ένα ερώτημα δεν είναι τίποτα άλλο από το τρέξιμο του αλγορίθμου μας με ένα συγκεκριμένο κατώφλι υποστήριξης και τοπικής συχνότητας το οποίο με την σειρά του επιστρέφει όλα τα συχνά και τοπικά συχνά αντικείμενα (θα αναφερόμαστε στο παρών κεφάλαιο στα τοπικά συχνά και στα συχνά αντικείμενα της βάσης μας σαν σχετικά από εδώ και στο εξής). Το

πρόβλημα είναι τώρα σε σχέση με ποια αντικείμενα θα συγκρίνουμε τα αντικείμενα που βρήκαμε σαν σχετικά με την προσέγγιση μας και πως θα ανακαλύψουμε αυτά τα αντικείμενα; Σύμφωνα με τους Witten, Moffat και Bell στο [126] η μέθοδος της ανάκλησης (recall) και της ακρίβειας (precision) για την αξιολόγηση μιας μετρικής ομοιότητας (similarity measure) προϋποθέτει ότι υπάρχει ήδη ένα σύνολο έτοιμων κειμένων και ερωτήσεων και ότι ένα πλήρες σύνολο αποφάσεων – κριτηρίων σχετικότητας για το ποια κείμενα είναι σχετικά σε κάθε ερώτημα είναι διαθέσιμα. Αυτή είναι μια φρουδα απαίτηση, ειδικά αν το πείραμα μας εκτελείται σε παραδείγματα και βάσεις από πραγματικά δεδομένα και όχι τεχνητά. Πρακτικά για το πείραμα μας αυτό σημαίνει ότι θα πρέπει να ξέρουμε εκ των προτέρων όλα τα αντικείμενα που θα πρέπει να θεωρούνται σαν σημαντικά στην συγκεκριμένη βάση προκειμένου να είμαστε σε θέση να τα συγκρίνουμε με αυτά που βρίσκουμε σαν σχετικά με την προσέγγιση μας. Αυτή η προϋπόθεση όμως, ειδικά στην περίπτωση μας παρουσιάζει ακόμα μεγαλύτερα προβλήματα. Έτσι, κατ' αρχήν αφού τέτοιες έτοιμες βάσεις, τουλάχιστον αυτές όπου τα τοπικά συχνά αντικείμενα είναι γνωστά και ειδικά πραγματικές βάσεις δεν υπάρχουν, θα έπρεπε να βρεθούν ή να δημιουργηθούν. Αλλά ιδιαίτερα σε επιχειρήσεις λιανεμπορίου και σε εμπορικές εφαρμογές υπάρχουν κάποια ειδικά προβλήματα όπως ανταγωνισμός ή προβλήματα ιδιωτικότητας που δεν υπάρχουν σε άλλες εφαρμογές όπως στην ανάκτηση πληροφορίας, οι οποίες περιπλέκουν περαιτέρω την απόκτηση και παραπέρα χρήση τέτοιων βάσεων. Δεύτερον μια λογική ερώτηση η οποία ανακύπτει είναι ποιος θα αποφασίσει ποια είναι τα σχετικά αντικείμενα σε μια βάση και ποια κριτήρια θα πρέπει να χρησιμοποιεί κάποιος κατά την εύρεση τους. Στην περίπτωση που είχαμε μια συλλογή κειμένων και θέταμε ένα ερώτημα τότε οποιοσδήποτε με απλές γραμματικές γνώσεις και μια μέτρια γνώση του θέματος των κειμένων αυτών θα μπορούσε να διαβάσει τις απαντήσεις και να αποφασίσει αν είναι σχετικές ή όχι. Στην περίπτωση μας τα πράγματα δεν είναι τόσο απλά, αφού τα αντικείμενα δεν παρέχουν κάποιο άμεσο τρόπο προκειμένου να αποφασίσουμε τι είναι σχετικό και τι όχι. Επίσης αφού η εξόρυξη δεδομένων είναι εξ ορισμού η ανακάλυψη κρυμμένης γνώσης θα ήταν σχετικά παράλογο να ζητάμε από τους χρήστες να αξιολογήσουν την σχέση και την σημασία των αντικειμένων που επεστράφησαν. Πιθανώς να μπορούσαν να κρίνουν και να αξιολογήσουν τις προφανείς απαντήσεις αλλά θα ήταν σίγουρα αβοήθητοι με τις ομολογουμένως πιο πολύτιμες απροσδόκητες απαντήσεις. Παράδειγμα είναι σχετικά εύκολο να κρίνεις μια πρόταση η οποία τοποθετεί μαζί γάλα και δημητριακά, αλλά πως μπορεί κάποιος να αποφασίσει αν μια πρόταση η οποία βάζει μαζί μύρρες και παιδικές πάνες είναι σωστή ή όχι Έτσι θα πρέπει να λάβουν μέρος κάποιοι σχετικοί χρήστες - κριτές, με βαθιά γνώση και εμπειρία σχετικά με την φύση της εταιρείας απ' όπου θα ληφθούν τα δεδομένα όπως επίσης των δεδομένων αυτών καθ' εαυτών και τέλος του ίδιου του θέματος της εξόρυξης δεδομένων. Τέλος πολύ σημαντικό είναι και το γεγονός ότι ένα ερώτημα στην περίπτωση μας είναι στην

πράξη το τρέξιμο του αλγορίθμου μας με κάποια τιμή κατωφλίου υποστήριξης και ελάχιστου LFR, το οποίο με τη σειρά του σημαίνει ότι επιστρέφονται όλα τα συχνά και τοπικά συχνά αντικείμενα (συμπεριλαμβανομένων των συνδυασμών τους) σε αντίθεση με ένα σύστημα ανάκτησης πληροφορίας όπου σε οποιοδήποτε ερώτημα ένα μικρό πλήθος απαντήσεων επιστρεφόταν και αξιολογούταν. Τα επιστρεφόμενα αντικείμενα στην περίπτωση μας μπορούν κάλλιστα να κυμαίνονται από χιλιάδες μέχρι δεκάδες χιλιάδες και έτσι το να ψάξουμε και να αξιολογήσουμε ένα τόσο μεγάλο πλήθος αντικειμένων αν είναι σχετικά ή όχι είναι πρακτικά αδύνατο. Ακόμα και αν αυτά τα προβλήματα έχουν αντιμετωπιστεί, ο υπολογισμός της ακρίβειας και της ανάκλησης στην περίπτωση μας είναι επίσης διαφορετικός. Σε μια διαδικασία διάταξης που χρησιμοποιείται σε ένα κλασσικό σύστημα ανάκτησης πληροφορίας, εκτός από τις απαντήσεις που επιστρέφονται λαμβάνουμε υπόψη επίσης την σειρά κατά την οποία αυτές οι απαντήσεις επιστρέφονται. Έτσι με βάση αυτή την σειρά ο χρήστης δημιουργεί την καμπύλη ανάκλησης-ακρίβειας. Στην περίπτωση μας τα αντικείμενα έχουν μια περισσότερο δυαδική υπόσταση (δηλ. είτε είναι τοπικά συχνά ή συχνά αντικείμενα οπότε και επιστρέφονται ή είναι σπάνια και αγνοούνται) και δεν υπάρχει κάποια σειρά με την οποία τα αντικείμενα επιστρέφονται. Έτσι μπορούμε να καταλήξουμε μόνο σε συγκεκριμένες τιμές ανάκλησης και ακρίβειας σε δοσμένες τιμές κατωφλίου υποστήριξης και LFR παρά να κατασκευάσουμε καμπύλη ανάκλησης-ακρίβειας.

Λαμβάνοντας υπόψη όλες αυτές τις απαιτήσεις-περιορισμούς οδηγηθήκαμε στο παρακάτω πείραμα. Συλλέξαμε ένα σύνολο αντικειμένων συνδυάζοντας δεδομένα από ορισμένες εταιρείες λιανεμπορίου καθώς και προσαρμοσμένες συνθετικές βάσεις από διάφορα προβλήματα εξόρυξης δεδομένων (όπως αυτή που χρησιμοποιούταν στο [80]). Η προκύπτουσα βάση μελετήθηκε από ένα σύνολο ατόμων τους οποίους θεωρήσαμε σαν ειδικούς (δηλ. ο υπεύθυνος πληροφορικών συστημάτων, ο διευθυντής μάρκετινγκ και ο υπεύθυνος πωλήσεων ορισμένων επιχειρήσεων καθώς επίσης και ορισμένοι ερευνητές του τομέα της εξόρυξης δεδομένων). Η βάση αναπαριστούσε τις ετήσιες πωλήσεις ενός συνόλου 50 προϊόντων από διάφορες κατηγορίες, φθάνοντας συνολικά τις 50.000 συναλλαγές με κάθε συναλλαγή να αποτελείται κατά μέσο όρο από 13-14 προϊόντα και το επιθυμητό αποτέλεσμα (δηλ. τα συχνά και τα τοπικά συχνά αντικείμενα) ήταν γνωστό και προαποφασισμένο από τους ειδικούς μας. Πειραματιστήκαμε με διάφορες τιμές κατωφλίου υποστήριξης και ελάχιστου LFR προσπαθώντας να επιτύχουμε την υψηλότερη τιμή ανάκλησης (δηλ. να βρούμε όσο το δυνατόν περισσότερα σχετικά αντικείμενα) αλλά ταυτόχρονα να κρατήσουμε την τιμή της ακρίβειας μέχρι ενός ορίου (δηλ. να μην δημιουργήσουμε πάρα πολλά “σκουπίδια”). Η καλύτερη τιμή ανάκλησης και ακρίβειας που καταφέραμε να επιτύχουμε για ένα συγκεκριμένο συνδυασμό τιμών κατωφλίου υποστήριξης και LFR ήταν 92% και 86% αντίστοιχα.

Όπως αναμέναμε η τιμή της ακρίβειας της προσέγγισης μας ήταν εξαιρετικά υψηλή ενώ η τιμή της ακρίβειας είναι κάπως χαμηλότερη. Αυτό δικαιολογείται και από το γεγονός ότι στην αναζήτηση μας για τοπικά συχνά αντικείμενα ανακαλύψαμε επίσης ορισμένα ψευδή (δηλ. αντικείμενα τα οποία τελικά δεν ήταν ενδιαφέροντα). Αν είχαμε χρησιμοποιήσει έναν παραδοσιακό αλγόριθμο αλλά στον οποίο θα είχαμε θέσει το κατώφλι υποστήριξης πολύ χαμηλά (ίσο με το minLFR) τότε θα είχαμε επιτύχει σχεδόν τις ίδιες τιμές ανάκλησης αλλά εντελώς απογοητευτικές – απαγορευτικές τιμές ακρίβειας (αφού θα είχαμε επιστρέψει κάθε αντικείμενο με αριθμό εμφανίσεων πάνω από το κατώφλι υποστήριξης και όχι μόνο τα τοπικά συχνά αντικείμενα).

Όταν ελέγχαμε τα αντικείμενα που δεν βρέθηκαν από την προσέγγιση μας σαν σχετικά βρεθήκαμε απέναντι σε ορισμένα πολύ ενδιαφέροντα ευρήματα. Ανάμεσα στα αντικείμενα που δεν βρέθηκαν σαν συχνά ή τοπικά συχνά από την προσέγγιση μας υπήρχαν ορισμένα τα οποία πωλούνταν μόνο σε συγκεκριμένα καταστήματα της επιχείρησης, και επίσης είχαν εποχική συμπεριφορά. Αυτά τα αντικείμενα θεωρούνταν σαν σχετικά από τους ειδικούς μας, αφού παρά το γεγονός ότι δεν είχαν αρκετές πωλήσεις συνολικά αλλά ούτε σε σχέση με το συγκεκριμένο κατάστημα είχαν αρκετές πωλήσεις για μια συγκεκριμένη περίοδο στο συγκεκριμένο κατάστημα. Προκειμένου να βρούμε επίσης αυτά τα αντικείμενα κάποιος θα έπρεπε να διαμερίσει την αρχική βάση στα καταστήματα απ' όπου τα δεδομένα προέρχονταν και να εφαρμόσει μια περαιτέρω διαμέριση σε περιόδους προκειμένου να ανακαλύψει αυτά τα αντικείμενα. Μια πολύ απλούστερη λύση θα ήταν να ελέγξουμε με την προσέγγιση μας μόνο τα δεδομένα από το συγκεκριμένο κατάστημα. Ένα ακόμα πιο ενδιαφέρον εύρημα ήταν ότι υπήρχαν κάποια προϊόντα που είχαν μια πολύ συμπτυκνωμένη και εκρηκτική (bursty) συμπεριφορά ως προς τις πωλήσεις τους. Πιο συγκεκριμένα υπήρχαν προϊόντα τα οποία παρουσίαζαν το μεγαλύτερο ποσοστό πωλήσεων κατά τη διάρκεια μιας ή δύο εβδομάδων σε ολόκληρο το χρόνο και επίσης αυτές οι πωλήσεις ήταν πολλές τάξεις μεγέθους μεγαλύτερες από τις μέσες πωλήσεις ολόκληρου του έτους, όπως για παράδειγμα προϊόντα που παράγονται και καταναλώνονται κύρια σε συγκεκριμένες ημερομηνίες, όπως εθνικές εορτές ή σε συγκεκριμένες μικρές περιόδους μιας χρονιάς (π.χ. χαρτικά και βιβλία στις πρώτες ημέρες της σχολικής χρονιάς). Παρόλ' αυτά οι πωλήσεις αυτών των προϊόντων ήταν κάτω από το minLFR για τις συγκεκριμένες διαμερίσεις και έτσι η προσέγγιση μας δεν μπορούσε να τα ανακαλύψει.

8.6 Συμπεράσματα

Εισαγάγαμε την έννοια των τοπικά συχνών αντικειμένων σε μια βάση, τα οποία είναι ιδιαιτέρως εμφανή σε εμπορικές εφαρμογές. Σε μια μικρή έρευνα που διεξαγάγαμε και περιελάμβανε τους υπευθύνους μάρκετινγκ ορισμένων εκ των μεγαλύτερων επιχειρήσεων

λιανεμπορίου στην Ελλάδα, όλοι συμφώνησαν στην ύπαρξη αυτής της κατηγορίας αντικειμένων και όλοι βαθμολόγησαν αυτή την κατηγορία σαν την πιο ενδιαφέρουσα τόσο όσο αναφορά την συχνότητα εμφάνισης της όσο και από άποψης οικονομικού ενδιαφέροντος. Υπάρχουν και άλλες κατηγορίες ενδιαφερόντων αντικειμένων αλλά δεν παρουσιάζουν τόσο μεγάλο ενδιαφέρον. Επίσης δεν έχει μέχρι στιγμής προταθεί κάποιος ευρέως γνωστός τρόπος για την ανακάλυψη όλων των ενδιαφερόντων αντικειμένων αυτόματα, χωρίς την ανάγκη ανθρώπινης παρέμβασης. Είναι πολύ αμφισβητήσιμο τι είναι ενδιαφέρον και τι όχι και είναι μια απόφαση που περιλαμβάνει μεγάλο βαθμό υποκειμενικότητας. Η διαδικασία της πρότασης του τι είναι ενδιαφέρον και τι όχι, καθώς και να ανακαλύπτουμε όλα τα ενδιαφέροντα αντικείμενα αυτόματα παραμένει ένα ανοικτό ζήτημα. Ακολουθώς προτείναμε δύο αλγορίθμους για την εύρεση των τοπικά συχνών αντικειμένων σε μια βάση όπου υπάρχουν επίσης συχνά αντικείμενα οι οποίοι αποδείχτηκαν τόσο αποδοτικοί όσο και αποτελεσματικοί στην πράξη.

Πίνακας 14: Οι παράμετροι που χρησιμοποιούνται για τα πειράματα με τον LFI και τον P-DIC σε σχέση με τους παραδοσιακούς αλγορίθμους

ΟΝΟΜΑ ΒΑΣΗΣ	Αλγόριθμος	Παράμετρος	Διάστημα 1	Διάστημα 2	Διάστημα 3	Διάστημα 4	Διάστημα 5	Διάστημα 6
1 T40I10D100K	Apriori/MSApriori	Minsup/LS	0,050000	0,020000	0,010000	0,008500	0,007000	0,006000
	LFI & P-DIC	minLFR	0,166667	0,066667	0,033333	0,028333	0,023333	0,020000
		minsup	0,200000	0,080000	0,040000	0,034000	0,028000	0,024000
2 kosarak	Apriori/MSApriori	Minsup/LS	0,050000	0,003000	0,002000	0,001000	0,000900	0,000850
	LFI & P-DIC	minLFR	0,166667	0,010000	0,006667	0,003333	0,003000	0,002833
		minsup	0,200000	0,012000	0,008000	0,004000	0,003600	0,003400
3 connect	Apriori/MSApriori	Minsup/LS	0,282178	0,267326	0,255445	0,243564	0,225742	0,216831
	LFI & P-DIC	minLFR	0,940594	0,891089	0,851485	0,811881	0,752475	0,722772
		minsup	0,950000	0,900000	0,860000	0,820000	0,760000	0,730000
4 pumsb	Apriori/MSApriori	Minsup/LS	0,282178	0,267326	0,252475	0,237623	0,225742	0,219801
	LFI & P-DIC	minLFR	0,940594	0,891089	0,841584	0,792079	0,752475	0,732673
		minsup	0,950000	0,900000	0,850000	0,800000	0,760000	0,740000
5 accidents	Apriori/MSApriori	Minsup/LS	0,2376238	0,2227723	0,2079208	0,1930693	0,1871287	0,1782178
	LFI & P-DIC	minLFR	0,7920792	0,7425743	0,6930693	0,6435644	0,6237624	0,5940594
		minsup	0,8	0,75	0,7	0,65	0,63	0,6
6 BMS-WebView-1_mb	Apriori/MSApriori	Minsup/LS	0,000700	0,000640	0,000620	0,000600	0,000580	0,000560
	LFI & P-DIC	minLFR	0,002333	0,002133	0,002067	0,002000	0,001933	0,001867
		minsup	0,002800	0,002560	0,002480	0,002400	0,002320	0,002240
7 BMS-WebView-2_mb	Apriori/MSApriori	Minsup/LS	0,001000	0,000800	0,000400	0,000300	0,000180	0,000100
	LFI & P-DIC	minLFR	0,003333	0,002667	0,001333	0,001000	0,000600	0,000333
		minsup	0,004000	0,003200	0,001600	0,001200	0,000720	0,000400

Κεφάλαιο 9 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ ΤΗΣ ΕΡΕΥΝΑΣ

9.1 Εισαγωγή

Με το παρόν κεφάλαιο ουσιαστικά κλείνουμε την παρούσα διατριβή δίνοντας πιθανές μελλοντικές κατευθύνσεις της έρευνας σαν αποτέλεσμα της εργασίας μας. Έτσι παρουσιάζουμε καταρχήν σύντομα την διαχρονική εξέλιξη στην έρευνα για την ανακάλυψη γνώσης από την εισαγωγή της, όπως αυτή προκύπτει από την διεθνή βιβλιογραφία και τις δημοσιευμένες εργασίες πάνω στο θέμα. Ακολούθως δίνουμε πιθανές μελλοντικές κατευθύνσεις όπου θα μπορούσε κάποιος να κινηθεί ερευνητικά. Οι κατευθύνσεις αυτές έχουν χωριστεί σε τρία διαφορετικά κομμάτια, αναφορικά αυτό στο οποίο όπου θα μπορούσε να κινηθεί κάποιος σαν προέκταση των μέχρι στιγμής εργασιών μας, αυτό στο οποίο προτείνονται κάποιες εντελώς νέες προσεγγίσεις άσχετες με τα μέχρι τώρα αποτελέσματα μας, και τέλος αυτό που έχει να κάνει με τις τάσεις πάνω στην έρευνα της ανακάλυψης γνώσης όπως αυτή προκύπτει από την διεθνή βιβλιογραφία και τις σύγχρονες πρακτικές. Κλείνουμε παρουσιάζοντας ένα πολύ ενδιαφέρον θέμα, γενικότερου προβληματισμού σχετικά με την ανακάλυψη γνώσης και το πώς αυτή μπορεί να επηρεάσει αλλά και να επηρεαστεί από την προσπάθεια για διαφύλαξη της ιδιωτικότητας και της ασφάλειας των δεδομένων των ανθρώπων γενικότερα.

9.2 Η έρευνα στην Ανακάλυψη Γνώσης διαχρονικά

Δεν είναι εύκολο κάποιος να καταγράψει και να παρουσιάσει το πώς μπορεί να κινήθηκε ερευνητικά μια συγκεκριμένη επιστημονική περιοχή, καθώς με τον καιρό εμφανίζονται εντός της περιοχής κάποια θέματα τα οποία αποκτούν λιγότερο ή περισσότερο ενδιαφέρον, έχουμε αλληλεπιδράσεις και αλληλεπικαλύψεις με άλλες περιοχές και τέλος διαφορετικές ερευνητικές προτεραιότητες ως προς τη θεωρία ή την πράξη. Ένας έμεσος τρόπος για να πάρουμε μια γενική εικόνα της, είναι να εξετάσουμε τις εργασίες που παρήχθησαν όλα αυτά τα χρόνια. Κάτι ανάλογο έκαναν και οι Cios και Kurgan το 2005 στο [42] σχετικά με την ανακάλυψη γνώσης, με τη βοήθεια της πλατφορμής Axiom[®]††††. Η συγκεκριμένη πλατφόρμα παρέχει πρόσβαση στις βάσεις INSPEC, Compendex[®], PageOne[™] και στην Derwent World Patents Index. Έτσι δίνοντας κάποιες λέξεις κλειδιά και ένα συγκεκριμένο χρονικό εύρος συγκέντρωσαν τις εργασίες που παρήχθησαν πάνω στην ανακάλυψη γνώσης, και οι οποίες και παρουσιάζονται στις Εικόνα 33, Εικόνα 34 και Εικόνα 35.

Όπως ειπώθηκε και νωρίτερα η εξόρυξη δεδομένων αποτελεί τα τελευταία χρόνια ουσιαστικά ένα αυτόνομο πεδίο έρευνας. Όμως πριν αυτονομηθεί για πολλούς θεωρείτο ότι αποτελούσε ένα παρακλάδι της μηχανικής μάθησης. Αντίστοιχα και η ανακάλυψη γνώσης

†††† AXIOM, <http://axiom.iop.org/>

(Knowledge Discovery in Databases) θεωρείται από ορισμένους ερευνητές σαν κάτι ξεχωριστό από την εξόρυξη δεδομένων (Data Mining). Ουσιαστικά είναι σχετικά δύσκολο να ξεχωρίσει κάποιος εντελώς αυτά τα πεδία και να χαράξει επακριβώς κάποια σαφή όρια ανάμεσα τους. Έστω και δύσκολα όμως σύμφωνα με μια τέτοια θεώρηση αλλά και με βάση τις εργασίες που έχουν δημοσιευτεί όλα αυτά τα χρόνια προκύπτει το διάγραμμα της Εικόνας 33. Βλέπουμε έτσι ότι η μηχανική μάθηση παρουσίαζε διαχρονικά μια σταθερά ανοδική πορεία η οποία όμως από το 1999 και έπειτα ξεκίνησε να εξασθενεί. Η ανακάλυψη γνώσης αυτόνομα αλλά και η εξόρυξη δεδομένων σαν συνδυασμός της μηχανικής μάθησης ακολουθούν μια παρόμοια ανοδική πορεία από το 1995 και έπειτα, με την ανακάλυψη γνώσης να συγκεντρώνει ελαφρά μεγαλύτερο ενδιαφέρον. Η εξέλιξη τέλος της έρευνας εξόρυξης δεδομένων ήταν εκρηκτική και από τις μόλις 20 εργασίες που είχαμε το 1994 φτάσαμε σχεδόν στις 1400 το 2000.

Εικόνα 33: Η εξέλιξη της ανακάλυψης γνώσης σε σχέση με την μηχανική μάθηση [42]

Με την ίδια τεχνική προκύπτει το διάγραμμα της Εικόνας 34, όπου απεικονίζονται τα κυριότερα ζητήματα που επικράτησαν τα χρόνια αυτά στην έρευνα της ανακάλυψης γνώσης, και που είναι η έρευνα για χρήση της XML στην ανακάλυψη γνώσης, η χρήση OLAP τεχνικών, η έρευνα για τους κανόνες συσχετίσεων και τέλος η ενσωμάτωση των συστημάτων αποθήκευσης και διαχείρισης βάσεων δεδομένων με τις εφαρμογές ανακάλυψης γνώσης. Όπως εύκολα φαίνεται και στην εικόνα 34, ορισμένες τεχνικές όπως η XML ή οι κανόνες συσχετίσεων εξακολουθούν να έχουν αυξανόμενο ενδιαφέρον (με τους κανόνες συσχετίσεων να έχουν τεράστια αύξηση ειδικότερα από το 1994 και έπειτα), ενώ οι υπόλοιπες δύο να εξασθενούν. Πιθανότερη αιτία είναι μάλλον ότι κάποια ζητήματα που παρουσιάστηκαν να επιλυθηκαν σχετικά νωρίς, δεδομένου ότι τόσο οι OLAP τεχνικές όσο και οι Data Warehouses είχαν μελετηθεί αρκετά κατά το παρελθόν.

Εικόνα 34: Ζητήματα κατά την έρευνα ανακάλυψης γνώσης [42]

Μια ταξινόμηση των εργασιών πάνω στον τομέα της ανακάλυψης γνώσης θα μπορούσε να ήταν σε εργασίες πάνω σε θεωρητικά ζητήματα, σε πρακτικές εφαρμογές και τέλος σε ότι έχει να κάνει με την έρευνα στους αλγόριθμους. Παρατηρώντας και πάλι την κατανομή των δημοσιευμένων εργασιών χρονικά (Εικόνα 35), βλέπουμε κατ' αρχήν την διαχρονικά ανοδική πορεία τόσο σ' ότι αφορά τις εργασίες για πρακτικές εφαρμογές όσο και σε αλγοριθμικά ζητήματα. Μάλιστα οι αλγόριθμοι έχουν ξεπεράσει αρκετά τις αντίστοιχες εργασίες σε πρακτικά θέματα. Η περίπτωση των αλγορίθμων είναι απόλυτα φυσιολογική μιας και η φύση της ανακάλυψης γνώσης είναι τόσο απαιτητική αλλά και οι όγκοι των εξεταζόμενων δεδομένων τόσο μεγάλοι, όπου το θέμα της δημιουργίας αποτελεσματικών και κλιμακούμενων αλγορίθμων είναι εξαιρετικά κρίσιμο. Ακόμα και στις μέρες μας παραμένει ίσως το κύριο μέλημα των ερευνητών, ιδιαίτερα με την διαρκώς αυξανόμενη συγκέντρωση δεδομένων. Σημείο σταθμός θα πρέπει να θεωρηθεί κατά τη γνώμη μου το 1994 και η εργασία των Agrawal και Srikant [10] με την πρόταση του αλγορίθμου Apriori, έπειτα απ' όπου βλέπουμε μια κατακόρυφη αύξηση των σχετικών εργασιών. Ο αλγόριθμος αυτός αποτέλεσε ουσιαστικά την πρώτη αποτελεσματική λύση στο πρόβλημα της εύρεσης κανόνων συσχετίσεων και μέχρι σήμερα αποτελεί σημείο αναφοράς για όλες σχεδόν τις εργασίες πάνω στο θέμα^{§§§§}. Στην περίπτωση των πρακτικών εφαρμογών το γεγονός των τόσο πολλών εργασιών αποδίδεται στο μεγάλο ενδιαφέρον που δόθηκε από οργανισμούς και επιχειρήσεις πάνω στην ανάπτυξη τέτοιων τεχνικών με την διάθεση πολλών κονδυλίων, καθώς αναγνωρίστηκαν τα μεγάλα πιθανά οφέλη που θα μπορούσαν να αποκομισθούν. Αυτό από την άλλη είχε και σαν έμμεση

^{§§§§} Σύμφωνα με στοιχεία του citeseer (www.citeseer.ist.psu.edu) υπάρχουν σχεδόν 900 αναφορές στην συγκεκριμένη εργασία.

παρενέργεια την στροφή πολλών ερευνητών από την μελέτη θεωρητικών προβλημάτων, ιδιαίτερα από το 1999 και έπειτα. Μπορεί βέβαια κάποια θεωρητικά θέματα να λύθηκαν με τον καιρό αλλά η ελάττωση αυτή δεν μπορεί σε καμία περίπτωση να σημαίνει ότι δεν υπάρχουν νέες προκλήσεις και ανοιχτά προβλήματα. Πιο σημαντικό βέβαια είναι και το γεγονός ότι για να δημιουργηθούν καλές εφαρμογές απαιτείται ένα σταθερό θεωρητικό υπόβαθρο.

Εικόνα 35: Τάσεις στην έρευνα ανακάλυψης γνώσης [42]

9.3 Μελλοντικές κατευθύνσεις της Έρευνας

Σ' ότι αναφορά τις μελλοντικές κατευθύνσεις τις έρευνας, τις έχουμε χωρίσει και τις παρουσιάζουμε σε τρία ξεχωριστά τμήματα. Κατ' αρχήν παρουσιάζουμε πιθανές προεκτάσεις των υπάρχουσων εργασιών μας αλλά και νέα προβλήματα που τυχόν έθεσαν. Κατόπιν παρουσιάζουμε δύο εντελώς νέες ερευνητικές προτάσεις και κλείνουμε δίνοντας μια επισκόπηση σχετικά με το ποιες τάσεις επικρατούν παγκοσμίως στις μέρες μας στην περιοχή σύμφωνα με τη διεθνή βιβλιογραφία.

9.3.1 Προεκτάσεις της υπάρχουσας εργασίας μας

Οι εργασίες που πραγματοποιήσαμε κατά την έρευνα μας, και τις οποίες παρουσιάσαμε στα προηγούμενα κεφάλαια της παρούσας διατριβής, μπορεί να έλυσαν κάποια προβλήματα της περιοχής αλλά ταυτόχρονα μπορεί να δημιούργησαν κάποια νέα ή να άφησαν περιθώρια πιθανών προεκτάσεων. Κατά τη γνώμη μας τα σημαντικότερα από αυτά είναι:

Στο Κεφάλαιο 5 (Δημιουργία ενός αποτελεσματικού μηχανισμού απόδοσης βαρών για την ενσωμάτωση του μικροοικονομικού παράγοντα στους κανόνες συσχετίσεων), ενδιαφέρον θα είχε η εφαρμογή της διαδικασίας απόδοσης βαρών και ο έλεγχος της συμπεριφοράς της και σε άλλες περιοχές εκτός από τα δεδομένα λιανεμπορίου.

Στο Κεφάλαιο 6 (Ένας βελτιωμένος αλγόριθμος για την εξόρυξη κανόνων συσχετίσεων χρησιμοποιώντας πολλαπλές τιμές κατωφλίου υποστήριξης) θα μπορούσαμε να πειραματιστούμε με περισσότερα χαρακτηριστικά του αλγορίθμου μας όπως το μήκος του βήματος, το κατώφλι βεβαιότητας (minimum confidence) καθώς και με περισσότερα μονοσύνολα. Θα θέλαμε επίσης να τον δοκιμάσουμε σε πραγματικά δεδομένα.

Η διαχείριση των ενημερώσεων σε μια διαδικασία ανακάλυψης γνώσης είναι μια πολύ σημαντική και δύσκολη διαδικασία προκειμένου να αποφύγουμε τη συνεχή επεξεργασία τεράστιων ποσοτήτων δεδομένων κάθε φορά που πραγματοποιείται μια αλλαγή. Έτσι σύμφωνα με την εργασία που παρουσιάσαμε στο Κεφάλαιο 7 (Χρήση τεχνικών ανάκτησης πληροφορίας για την υποστήριξη ανακάλυψης γνώσης) θα μπορούσαμε να χρησιμοποιήσουμε και να πειραματιστούμε είτε με ένα stop press file για την διαχείριση των ερωτημάτων ή την ακόμα πιο αποτελεσματική εναλλακτική του [132]. Μια άλλη πιθανότητα θα ήταν η πρόταση και αξιολόγηση μιας διαφορετικής μετρικής συνάφειας, εκτός της μετρικής του συνημίτονου, ειδικά υλοποιημένου για την περίπτωση των δεδομένων συναλλαγών. Ενδιαφέρον θα ήταν επίσης η επιβεβαίωση με πειράματα αν η απόδοση της προσέγγισης μας σαν ένας κλασσικός αλγόριθμος εξόρυξης δεδομένων (δηλ. το batch module στην ενότητα 7.4) είναι καλύτερο από τους προηγούμενους παραδοσιακούς αλγορίθμους. Ανοιχτό παραμένει επίσης, άσχετα απ' όλα τα εμπόδια που φαινομενικά μπορεί να εμφανίζονται, η χρήση και αξιολόγηση relevance feedback τεχνικών. Τέλος θα θέλαμε να εξετάσουμε πιο προσεκτικά την διαδικασία ανάθεσης βαρών σε όλα τα αντικείμενα στη βάση.

Είναι πολύ αμφισβητήσιμο τι είναι ενδιαφέρον και τι όχι και είναι μια απόφαση που περιλαμβάνει μεγάλο βαθμό υποκειμενικότητας. Η διαδικασία της πρότασης του τι είναι ενδιαφέρον και τι όχι, καθώς και να ανακαλύπτουμε όλα τα ενδιαφέροντα αντικείμενα αυτόματα παραμένει ένα ανοικτό ζήτημα. Έτσι θα είχε εξαιρετικό ενδιαφέρον, και όχι απαραίτητα μόνο από την πλευρά της ανακάλυψης γνώσης αλλά σε συνδυασμό και με άλλους τομείς όπως το μάρκετινγκ ή η ψυχολογία η εξέταση της διαδικασίας της πρότασης του τι είναι ενδιαφέρον και τι όχι. Τέλος σε συνδυασμό με αυτό δεν έχει μέχρι στιγμής προταθεί κάποιος ευρέως γνωστός τρόπος για την ανακάλυψη όλων των ενδιαφερόντων αντικειμένων αυτόματα, χωρίς την ανάγκη ανθρώπινης παρέμβασης.

9.3.2 Νέες εργασίες

9.3.2.1 Ενσωμάτωση της αρνητικής επίδρασης κάποιων αντικειμένων

Ενώ οι θετικοί κανόνες συσχετίσεων διαχειρίζονται την ύπαρξη αντικειμένων μέσα στις συναλλαγές, οι αρνητικοί, τουλάχιστον όπως είχαν τυπικά οριστεί στη βιβλιογραφία ([26][110][121][117][13][14]) ασχολούνται με την απουσία τους. Για παράδειγμα ένας θετικός κανόνας συσχετίσεων θα λάμβανε υπόψη τις συναλλαγές που περιείχαν μαζί μύρρες και τσιπς

και θα παρήγαγε όλους τους σχετικούς κανόνες συσχετίσεων. Μια προσέγγιση που θα ασχολείτο με αρνητικούς κανόνες συσχετίσεων από την άλλη θα ελάμβανε υπόψη και κανόνες όπως: αυτοί που αγοράζουν μπύρες αγοράζουν επίσης τσιπς αλλά όχι ξηρούς καρπούς ($M\acute{\pi}\upsilon\rho\alpha \rightarrow T\sigma\iota\pi\varsigma \wedge \neg \Xi\eta\rho\acute{o}\iota \text{ \textit{καρποί}}$), ή αυτοί που αγοράζουν μπύρες αλλά όχι κρασί αγοράζουν επίσης τσιπς ($M\acute{\pi}\upsilon\rho\alpha \wedge \neg K\rho\alpha\sigma\acute{\iota} \rightarrow T\sigma\iota\pi\varsigma$). Οι μετρικές που χρησιμοποιούνται για την εύρεση των συχνών αντικειμένων και την παραγωγή των αρνητικών κανόνων συσχετίσεων είναι στις περισσότερες περιπτώσεις και πάλι η υποστήριξη και η βεβαιότητα. Ένας γενικευμένος κανόνας συσχετίσεων μπορεί να περιλαμβάνει διάφορα θετικά και αρνητικά αντικείμενα (δηλ. αντικείμενα τα οποία εμφανίζονται σε μια συναλλαγή καθώς και αυτά που είναι απών) είτε στο αρχικό είτε στο απορρέον κομμάτι ενός κανόνα. Ένα παράδειγμα ενός τέτοιου κανόνα θα μπορούσε να ήταν: $A \wedge \neg C \wedge \neg F \wedge W \rightarrow B \wedge \neg D$. Παρόλ' αυτά το προφανές ανυπέρβλητο πρόβλημα που δημιουργείται από μια τέτοια θεώρηση είναι ο αριθμός των πιθανών αντικειμένων τα οποία θα έπρεπε να δημιουργηθούν αλλά και να καταμετρηθούν (κάθε θετική συσχέτιση μήκους k παράγει $O(k)$ αρνητικές συσχετίσεις, κάνοντας τον χώρο στον οποίο θα πρέπει αναζητήσουμε αντικείμενα εκθετικά μεγάλο) καθώς επίσης και ο αριθμός των τελικών κανόνων, αφού πέρα από τα υπάρχοντα αντικείμενα θα έπρεπε να λάβουμε υπόψη και όλα αυτά που δεν εμφανίζονται καθώς και τους συνδυασμούς τους.

Όμως το πιο σημαντικό κατά τη γνώμη μας είναι το γεγονός ότι η απουσία ενός αντικειμένου από μια συναλλαγή δεν υποδηλώνει και άμεσα αρνητική συσχέτιση μεταξύ τους αλλά θα μπορούσε να αποδοθεί σε διάφορους παράγοντες όπως τύχη ή απλή σύμπτωση. Αυτή ακριβώς η παρατήρηση δημιουργεί τεράστιο ερευνητικό ενδιαφέρον. Τα υπάρχοντα ηλεκτρονικά καταστήματα (e-shops) και τα συστήματα που διαθέτουν καταγράφουν μόνο τα αντικείμενα τα οποία τελικά παρέμειναν σε μια συναλλαγή. Πιο συγκεκριμένα αποθηκεύουν για κάθε συναλλαγή ένα μοναδικό κωδικό συναλλαγής (γνωστός και σαν TID), την ημερομηνία και ορισμένες φορές και την ακριβή ώρα της συναλλαγής και τέλος το χαρακτηριστικό αριθμό (υπό την μορφή bar codes) των αντικειμένων που αγοράστηκαν. Έτσι δεν καταγράφουν κάτι σχετικά με τα αντικείμενα αυτά που μπορεί να εισήλθαν σε ένα καλάθι αγορών αλλά για κάποιο λόγο απομακρύνθηκαν κατά τις επόμενες αγορές από τους πελάτες. Τα αντικείμενα αυτά όμως περιέχουν πολύτιμη γνώση η οποία δυστυχώς αγνοείται από τις υπάρχουσες προσεγγίσεις και συστήματα, καθώς τα αντικείμενα δύνανται να έχουν πέρα από την θετική και αρνητική επιρροή το ένα στο άλλο (π.χ. η παρουσία ενός αντικειμένου επηρεάζει αρνητικά την επιλογή ή την παρουσία ενός άλλου). Για παράδειγμα ας υποθέσουμε ότι ο πελάτης ενός ηλεκτρονικού καταστήματος έχει ήδη επιλέξει και τοποθετήσει τα αντικείμενα A, D, E και F στο καλάθι αγορών του. Το σύστημα τώρα με βάση τα αντικείμενα

αυτά αποφασίζει να προτείνει το αντικείμενο I στον συγκεκριμένο πελάτη, ο οποίος όμως κατόπιν αποφασίζει να απομακρύνει το αντικείμενο D από το καλάθι. Αυτή είναι μια κατ' αρχήν ένδειξη ότι τα αντικείμενα I και D είναι πράγματι αρνητικά συσχετισμένα, ή ότι ίσως υπάρχει κάποια “σύγκρουση” με ένα ή περισσότερα αντικείμενα τα οποία βρίσκονται ήδη στο συγκεκριμένο καλάθι αγορών. Έτσι όταν εμείς μιλάμε για αρνητικούς κανόνες συσχετίσεων μιλάμε για κανόνες που μοντελοποιούν τις περιπτώσεις εκείνες όπου η αφαίρεση ενός αντικειμένου πιθανώς να σημαίνει αρνητική επιρροή σε κάποιο άλλο. Η πρόταση μας, παρά το γεγονός ότι επίσης προσπαθούμε να ανακαλύψουμε την εμφάνιση αρνητικών συσχετίσεων ανάμεσα στα αντικείμενα μιας βάσης δεδομένων, παρεκκλίνει αλλά ταυτόχρονα κατά μια έννοια επεκτείνει αυτές τις εργασίες. Έτσι αντί να λαμβάνουμε υπόψη κανόνες οι οποίοι προκύπτουν από συναλλαγές όπου η απουσία ενός αντικειμένου υποδηλώνει μια αρνητική σχέση θεωρούμε τους κανόνες αυτούς όπου η σύγκρουση ή η ασυμβατότητα ανάμεσα στα αντικείμενα καταγράφεται υπό την μορφή της απομάκρυνσης ενός αντικειμένου.

9.3.2.2 Χρήση της ακριβής σειράς εμφάνισης των αντικειμένων στη βάση

Όλες οι γνωστές προσεγγίσεις σε όλες τις περιοχές της ανακάλυψης γνώσης θεωρούσαν την σειρά των αντικειμένων μέσα στις συναλλαγές σαν ασήμαντη και έτσι τα διέτασσαν με οποιονδήποτε τρόπο θεωρούσαν καταλληλότερο για τις ανάγκες τους. Αυτή η θεώρηση όμως δεν βασίζονταν σε κάποια συγκεκριμένη και τεκμηριωμένη θεωρητική βάση αλλά είχε σαν σκοπό να εξυπηρετήσει κάποιες πρακτικές ανάγκες (π.χ. σύμφωνα με τους Agrawal και Srikant στο [10] καθώς και σχεδόν σε όλες τις μετέπειτα εργασίες θεωρείται ότι τα αντικείμενα μέσα σε κάθε συναλλαγή είναι λεξικογραφικά διατεταγμένα έτσι ώστε ο σχηματισμός του συνόλου των υπονήφων αντικειμένων καθώς και η καταμέτρηση τους να μπορεί να γίνει αποτελεσματικά). Με βάση αυτή τη θεώρηση, οι παραδοσιακοί αλγόριθμοι στον τομέα των κανόνων συσχετίσεων λειτουργούν στηριζόμενοι στην σχετικά απλοϊκή παραδοχή ότι κάθε αντικείμενο μέσα σε μια συναλλαγή επηρεάζει και όλα τα υπόλοιπα αντικείμενα στην ίδια συναλλαγή. Όμως παρόλο που τα αντικείμενα μπορεί να εμφανίζονται μαζί σε διάφορες συναλλαγές, αυτό δεν σημαίνει απαραίτητα ότι επηρεάζουν το ένα το άλλο. Αυτό μπορεί να γίνει καλύτερα κατανοητό μέσα από ένα παράδειγμα. Έστω ότι έχουμε μια συναλλαγή η οποία περιέχει τα ακόλουθα αντικείμενα, αποθηκευμένα με την ακριβή σειρά με την οποία αγοράστηκαν:

[V, G, A, S, C]

Τότε σύμφωνα με τις υπάρχουσες προσεγγίσεις το αντικείμενο V θεωρείται ότι επηρέασε την αγορά ή γενικότερα την εμφάνιση όλων των άλλων αντικειμένων στην ίδια συναλλαγή (δηλ. των αντικειμένων G, A, S και C). Τα αντικείμενα G, A, S, C επίσης θεωρούνται ότι

επηρέασαν την αγορά όλων των υπολοίπων αντικειμένων στην ίδια συναλλαγή. Παρόλ' αυτά το λογικό ερώτημα που προκύπτει από αυτή την υπόθεση είναι πως μπορεί ένα γεγονός που έλαβε χώρα στο μέλλον να επηρεάζει ένα γεγονός στο παρελθόν. Πιο συγκεκριμένα πως μπορούμε να ισχυριστούμε ότι η επιλογή του αντικειμένου A επηρεάστηκε από την επιλογή του αντικειμένου S ή του C αφού όταν επιλέχτηκε το A τα αντικείμενα S και C δεν βρίσκονταν καν στο καλάθι μας. Είναι έτσι λογικό να υποθέτουμε ότι υπάρχει η πιθανότητα ένα αντικείμενο που αγοράστηκε τώρα να επηρεάσει τις μετέπειτα επιλογές μας (π.χ. στο συγκεκριμένο παράδειγμα το αντικείμενο A να επηρεάσει την επιλογή των S και C) αλλά πως μπορούμε να ισχυριστούμε το αντίθετο? Αυτό το σενάριο είναι ιδιαίτερα εμφανές σε ηλεκτρονικά καταστήματα όπου οι αγορές πραγματοποιούνται με μια “αλυσιδωτή” λογική (δηλ. το ένα αντικείμενο μετά το άλλο και συνήθως χωρίς να υπάρχει κάποια προκαθορισμένη λίστα αγορών). Έτσι κατά τη γνώμη μας κάθε αντικείμενο μέσα σε μια συναλλαγή επηρεάζει μόνο τα επόμενα αντικείμενα στην ίδια συναλλαγή.

Για να λάβουμε υπόψη και να χρησιμοποιήσουμε όλη αυτή την πληροφορία πρέπει να έχουμε στη διάθεση μας την ακριβή σειρά αγοράς των προϊόντων, κάτι σχετικά εύκολο να πραγματοποιηθεί ακόμα και με τα υπάρχοντα συστήματα. Τέλος η συγκεκριμένη θεώρηση μπορεί να συνδυαστεί με την ύπαρξη αρνητικών συσχετίσεων που αναφέρθηκε προηγούμενα, αλλά και με την σημαντικότητα που θα πρέπει να έχει το κάθε αντικείμενο προκειμένου να καταλήξουμε σε ακόμα πιο ακριβή αποτελέσματα. Έτσι ουσιαστικά απομένει να συνδυαστούν όλες αυτές οι παράμετροι σε ένα συνολικό μοντέλο και να χρησιμοποιηθούν οι κατάλληλες τεχνικές ανακάλυψης γνώσης.

9.3.3 Άλλες κατευθύνσεις με βάση τη διεθνή βιβλιογραφία

Έκτος από τις κατευθύνσεις που προτείνουμε εμείς με βάση την δική μας εργασία αλλά και με βάση τα αποτελέσματά μας, σε γενικές γραμμές οι μελλοντικές κατευθύνσεις που προκύπτουν σύμφωνα με τα διεθνή δεδομένα είναι:

- ▮ Ενοποίηση δεδομένων.** Πέρα από την προσπάθεια για καταγραφή όλο και νέων πληροφοριών, η σύγχρονη τάση σχετικά με τα δεδομένα για εφαρμογές ανακάλυψης γνώσης έχει να κάνει με τον συνδυασμό υπαρχόντων δεδομένων από πολλές διαφορετικές πηγές. Αυτό δημιουργεί νέες προκλήσεις μιας και τα δεδομένα αυτά θα πρέπει να έρθουν σε μια ενιαία μορφή προκειμένου να γίνει εφικτή η επεξεργασία τους. Τεχνολογίες όπως η XML, η οποία ήδη χρησιμοποιείται ευρέως για την μεταφορά και διαμοίραση πληροφοριών, αλλά και νέες θα πρέπει να δημιουργηθούν προκειμένου να διευκολύνουν την διαδικασία αυτή.

- ▮ **Κλιμάκωση αλγορίθμων ανακάλυψης γνώσης.** Το συγκεκριμένο κομμάτι αποτελεί διαχρονικά ένα εξαιρετικά απαιτητικό ζήτημα στην ανακάλυψη γνώσης. Όχι μόνο έχουμε υπερβολικά μεγάλους όγκους δεδομένων για επεξεργασία, αλλά το μέγεθος τους αυξάνει συνεχώς. Ιδιαίτερα με την προσπάθεια για ενοποίηση όλο και περισσότερων πηγών δεδομένων το συγκεκριμένο ζήτημα θα γίνει όλο και πιο φλέγον.
- ▮ **Ενσωμάτωση εφαρμογών ανακάλυψης γνώσης στα συστήματα βάσεων δεδομένων.** Η ανακάλυψη γνώσης δεν αποτελεί πλέον μια εξεζητημένη - ειδική εφαρμογή αλλά θεωρείται σαν κάτι σχετικά κοινό. Εφαρμογές ανακάλυψης γνώσης, έστω και μικρής κλίμακας, βρίσκονται ήδη ενσωματωμένες σε ορισμένα συστήματα διαχείρισης και ανάπτυξης βάσεων δεδομένων. Έτσι μελλοντικά τα συστήματα βάσεων δεδομένων θα προσπαθήσουν να συμπεριλάβουν όσο το δυνατόν περισσότερες τέτοιες εφαρμογές.
- ▮ **Δημιουργία και Τυποποίηση μιας ενιαίας γλώσσας Ανακάλυψης Γνώσης.** Καθώς σταδιακά εφαρμογές ανακάλυψης γνώσης θα γίνουν ένα αναπόσπαστο κομμάτι των βάσεων δεδομένων η δημιουργία και τυποποίηση μιας ενιαίας γλώσσας ανακάλυψης γνώσης είναι πλέον αναγκαία. Μια τέτοια γλώσσα θα προωθήσει περαιτέρω τη χρήση τεχνικών ανακάλυψης γνώσης, θα βελτιώσει την ανταλλαγή δεδομένων και πληροφοριών ανάμεσα στα διάφορα συστήματα και θα απλοποιήσει την διαδικασία ανακάλυψης γνώσης. Παράδειγμα τέτοιας προσπάθειας αποτελεί και η Microsoft με την MS OLE DB for Data Mining*****.
- ▮ **Web Mining.** Δεδομένου του τεράστιου όγκου πληροφοριών που υπάρχει διαθέσιμο στο διαδίκτυο, την ελεύθερη πρόσβαση που υπάρχει σ' αυτό αλλά και του ρόλου που διαδραματίζει πλέον ο παγκόσμιος ιστός σε όλους τους τομείς της ζωής μας, η ανάπτυξη και εφαρμογή τεχνικών ανακάλυψης γνώσης στο συγκεκριμένο χώρο θα συνεχίσει να συγκεντρώνει όλο και περισσότερο ενδιαφέρον. Παράδειγμα τέτοιων εφαρμογών αποτελούν η ανακάλυψη γνώσης σε διαδικτυακά περιεχόμενα (web content mining), σε συνήθειες χρηστών (Weblog/usage mining), αλλά και ειδικές εφαρμογές ανακάλυψης γνώσης [91]. Για περισσότερες πληροφορίες στο θέμα ο αναγνώστης μπορεί να συμβουλευτεί το [114]
- ▮ **Ασφάλεια Δεδομένων και διαφύλαξη της ιδιωτικότητας.** Φαντάζει τουλάχιστον σαν παράδοξο από τη μια να προσπαθούμε να καταγράψουμε, να συνδυάσουμε και να επεξεργαστούμε όσο το δυνατόν περισσότερα δεδομένα και από την άλλη να προσπαθούμε να διαφυλάξουμε τα δεδομένα αυτά αλλά και τα όποια συμπεράσματα

μπορεί να προκύψουν. Το θέμα της διαφύλαξης της ιδιωτικότητας των χρηστών από τη μια αλλά και της περαιτέρω ανάπτυξης των εφαρμογών ανακάλυψης γνώσης από την άλλη χωρίς όμως το ένα να επηρεάζει αρνητικά το άλλο θα αποτελέσει ένα ιδιαίτερα ακανθώδες θέμα τόσο ερευνητικά όσο και νομικά – φιλοσοφικά. Μια σύντομη περιγραφή του θέματος επιχειρείται και στην ενότητά 9.4.

▮ **Βιοπληροφορική – ανάλυση βιολογικών δεδομένων.** Ένας απλοϊκός ορισμός της βιοπληροφορικής θα μπορούσε να είναι η εφαρμογή υπολογιστών και πληροφορικής για την ανάλυση βιολογικών δεδομένων. Η βιοπληροφορική και γενικότερα η ανάλυση βιολογικών δεδομένων (π.χ. σειρές DNA διαφόρων οργανισμών) αποτελούν αυτή τη στιγμή ίσως ένα από τα πιο ενεργά πεδία της έρευνας με τεράστιο παγκόσμιο ενδιαφέρον. Η φύση των δεδομένων αυτών αλλά και η πολυπλοκότητα την οποία παρουσιάζουν καθιστούν την ανάπτυξη και χρήση νέων αποδοτικών τεχνικών ανακάλυψης γνώσης επιτακτική. Βέβαια γεννιούνται και τεράστια ερωτήματα ως προς την χρήση αυτών των δεδομένων αλλά και των πιθανών συμπερασμάτων που προκύπτουν (όπως π.χ. κατά την περίπτωση όπου κάποιος θα παρουσιάζει προδιάθεση για ασθένειες).

▮ **Ανακάλυψη γνώσης και πόλεμος κατά της τρομοκρατίας – εγκληματικότητας.** Αποτελεί την πιο σύγχρονη τάση στην έρευνα ανακάλυψης γνώσης, και είναι επακόλουθο της συνεχιζόμενης τρομολαγνίας που ακολούθησε τις επιθέσεις της 11^{ης} Σεπτεμβρίου στις Η.Π.Α. Ήδη μάλιστα από το 2004 υπάρχει και σχετικό workshop^{††††}. Ο αποτελεσματικός εντοπισμός και ανακάλυψη τρομοκρατικών, εγκληματικών και εν γένει πιθανών παράνομων δραστηριοτήτων απαιτεί την ανάλυση τεράστιων ποσοτήτων δεδομένων, τα οποία συν τοις άλλης είναι εξαιρετικά πολύπλοκα και πεπλεγμένα. Ταυτόχρονα οι αναζητούμενες περιπτώσεις είναι ουσιαστικά τόσο σπάνιες και μοναδικές που περιπλέκουν ακόμα περισσότερο την κατάσταση. Αντίστοιχα με την βιοπληροφορική και αυτός ο τομέας ήδη έχει προξενήσει μεγάλες αντιδράσεις ως προς την διαφύλαξη της ιδιωτικότητας των πολιτών.

***** <http://www.microsoft.com/downloads/details.aspx?FamilyID=01005f92-dba1-4fa4-8ba0-af6a19d30217&DisplayLang=en>

†††† <http://www.siam.org/meetings/sdm04/>

9.4 Ανακάλυψη Γνώσης και προστασία προσωπικών δεδομένων

Η ανακάλυψη γνώσης και η ιδιωτικότητα αποτελούν για πολλούς δύο έννοιες εντελώς αντιφατικές και αλληλοσυγκρουόμενες. Από την μια προσπαθούμε να συνδυάσουμε όσο το δυνατόν περισσότερα δεδομένα από διαφορετικές πηγές, και να εισχωρήσουμε όσο το δυνατόν πιο βαθιά γίνεται σε αυτά ανακαλύπτοντας πολυδιάστατη κρυμμένη γνώση, και από την άλλη να διαφυλάξουμε τα δεδομένα μας αλλά και τα συμπεράσματα που μπορεί να προκύψουν από την επεξεργασία αυτών με κάθε δυνατό τρόπο. Στις μέρες μας παράγουμε ένα τεράστιο όγκο δεδομένων σαν ένα υποπροϊόν των καθημερινών μας συναλλαγών, των επισκέψεων μας σε διαδικτυακούς τόπους, των αλληλεπιδράσεων μας με διάφορους οργανισμούς και κυβερνητικές υπηρεσίες κ.α.. Έτσι δεν αυξάνει μόνο ο αριθμός των “εγγράφων” που δημιουργούμε στις διάφορες βάσεις που κρατούν στοιχεία για εμάς, αλλά και το μέγεθος των δεδομένων που αποθηκεύονται σε κάθε εγγραφή. Έτσι πλέον η ανησυχία όλο και περισσότερων ανθρώπων στις μέρες μας είναι κοινή: τι μπορούν να αποκαλύψουν τα δεδομένα μας σε κάποιον τρίτο; Οι άνθρωποι ανησυχούν για τα ιατρικά, τα οικονομικά, τα επαγγελματικά αλλά ακόμη και για τα δημογραφικά τους δεδομένα. Νιώθουν τουλάχιστον άβολα γνωρίζοντας ότι οι άλλοι μπορεί να ξέρουν τι βιβλία διαβάζουν, τι ταινίες παρακολουθούν, σε ποιες ομάδες ή πολιτικές παρατάξεις ανήκουν, πότε και που ταξιδεύουν και γενικότερα τις προτιμήσεις και τις συνήθειες τους σε διάφορους τομείς. Φοβούνται ότι η αποκάλυψη και διάθεση τόσων πολλών στοιχείων τους μπορεί να οδηγήσει σε αύξηση των “οχλήσεων” από εταιρείες μάρκετινγκ (υπό την μορφή τηλεφωνικών κλήσεων για προώθηση προϊόντων ή spam mail), προβλήματα κατά την αναζήτηση ή αλλαγή εργασίας (σύμφωνα με την προδιάθεση τους σε κάποιες ασθένειες ή με βάση το εργασιακό τους ιστορικό), κοινωνικό στιγματισμό (σύμφωνα με τις σεξουαλικές ή θρησκευτικές πεποιθήσεις τους), αλλά ακόμα και περιορισμό ελευθεριών και λανθασμένη προσοχή από τις διωκτικές αρχές (αν το προφίλ τους ταιριάζει με αυτό ενός πιθανού εγκληματία ή τρομοκράτη).

Νομικά το όλο θέμα περιπλέκεται ακόμα περισσότερο, μιας και είτε υπάρχουν χώρες οι οποίες αγνοούν εντελώς (ηθελήμενα ή όχι) το όλο ζήτημα ή χώρες οι οποίες έχουν προχωρήσει πάρα πολύ στον συγκεκριμένο τομέα. Ως αποτέλεσμα αυτής της ανομοιομορφίας προκαλούνται προβλήματα τόσο μέσα στην εκάστοτε χώρα όσο και κατά τις συναλλαγές με άλλες χώρες. Έτσι σήμερα υπάρχουν ήδη σε ισχύ μια σειρά οδηγιών, νόμων και πρακτικών από υπηρεσίες των Η.Π.Α., της Ευρωπαϊκής Ένωσης (Ε.Ε.), της Ιαπωνίας αλλά και του Καναδά αλλά παράλληλα υπάρχουν και χώρες οι οποίες θεωρούνται παράδεισοι δεδομένων (σε αντιστοιχία με τους φορολογικούς παραδείσους) και όπου τα πράγματα είναι εντελώς άναρχα. Μάλιστα αρκετές e-επιχειρήσεις είχαν προχωρήσει ή ακόμα επιχειρούν να μεταφέρουν την

βάση δεδομένων τους σε τέτοιες χώρες ώστε να αποφύγουν τον έλεγχο και να χρησιμοποιήσουν τα δεδομένα τους κατά το δοκούν. Σε κάθε περίπτωση όμως τόσο η υπερπροστασία των δεδομένων αλλά όσο και η πλήρης αποδέσμευση τους από οποιονδήποτε έλεγχο μπορούν να δράσουν σαν μια τροχοπέδη στην ανάπτυξη και την εφαρμογή τεχνικών ανακάλυψης γνώσης.

Έτσι για παράδειγμα στις Η.Π.Α., όπου παρόλο που γενικότερα υιοθετείται μια πολύ πιο φιλελεύθερη αντιμετώπιση στο θέμα της ιδιωτικότητας αφήνοντας στην ουσία την αγορά ελεύθερη να διαμορφώσει τους κανόνες της και να αυτορυθμιστεί έχουν εμφανιστεί αρκετές περιπτώσεις καταπάτησης των δεδομένων των πολιτών και πρόκληση εντελώς αντίθετων αποτελεσμάτων. Χαρακτηριστική περίπτωση αποτελεί μια διαμάχη που έλαβε χώρα το 2003 και η οποία έφτασε μέχρι την γερουσία σχετικά με το πρόγραμμα Total Information Awareness⁺⁺⁺⁺ του υπουργείου άμυνας, όπου παρ' ολίγον τελικά να απαγορευτεί κάθε εφαρμογή και χρήση τεχνικών ανακάλυψης γνώσης. Αντίστοιχη ήταν και η περίπτωση της Ιαπωνίας το 2002 στην οποία μάλιστα πραγματοποιήθηκαν και μαζικές διαδηλώσεις στους δρόμους για το θέμα [116]. Οι περισσότερες εφαρμογές ανακάλυψης γνώσης λειτουργούν με την παραδοχή ότι όλα τα δεδομένα βρίσκονται διαθέσιμα σε μια κεντρική βάση. Αυτό προφανώς δημιουργεί ένα τεράστιο πρόβλημα ασφάλειας μιας και η παραβίαση αυτής της βάσης σημαίνει αυτόματα έκθεση όλων των δεδομένων. Έτσι παρ' όλο που οι άνθρωποι μπορεί να εμπιστεύονται ορισμένες οντότητες (π.χ. οργανισμούς ή επιχειρήσεις) με κάποια από τα δεδομένα τους, δεν εμπιστεύονται κάποια η οποία θα έχει πρόσβαση και θα χειρίζεται όλα τους τα δεδομένα μαζί. Αυτό ήταν ακριβώς και το πρόβλημα στην περίπτωση της Ιαπωνίας, όπου η κυβέρνηση δεν συνέλλεγε νέες πληροφορίες και δεδομένα απλά συγκέντρωνε τα υπάρχοντα δεδομένα από διαφορετικές πηγές σε μια κεντρική βάση.

Η Ευρωπαϊκή Ένωση (Ε.Ε.) από την άλλη διατηρεί μια πιο ενεργή στάση στο θέμα. Έτσι από τον Οκτώβριο του 1995 κιόλας, σχετική οδηγία^{§§§§} της Ευρωπαϊκής Ένωσης για την προστασία των δεδομένων απαγορεύει την μετακίνηση προσωπικών δεδομένων σε χώρες οι οποίες δεν έχουν θεσμοθετημένους και σε ισχύ επαρκείς νόμους για την προστασία της ιδιωτικότητας. Φυσικά από την άλλη διάφοροι επιχειρηματικοί όμιλοι και τομείς της βιομηχανίας και του εμπορίου διαμαρτύρονται λέγοντας ότι οι εθελοντικοί έλεγχοι που ισχύουν τουλάχιστον για τις Η.Π.Α. είναι αρκετοί. Μάλιστα ένα γκρουπ 10 εταιριών με την ονομασία Global Privacy Alliance (GPA), μεταξύ των οποίων ανήκουν και η IBM, η Oracle και η VeriSign, πιστεύουν ότι η Ευρωπαϊκή Ένωση έχει δώσει υπερβολικό βάρος στην προστασία

++++ <http://www.epic.org/privacy/profiling/tia/>

§§§§ Οδηγία 95/46/EC της 24^{ης} Οκτωβρίου 1995 για την προστασία των πολιτών σχετικά με την επεξεργασία και ελεύθερη διακίνηση προσωπικών δεδομένων.

της ιδιωτικότητας των πολιτών της, και όχι αρκετό για την διασφάλιση της ελεύθερης διακίνησης των δεδομένων ανάμεσα στις επιχειρήσεις. Πιθανώς αυτού του είδους οι πιέσεις να οδήγησαν την Ε.Ε. να αναθεωρήσει την στάση της στο θέμα, και να χαλαρώσει λίγο την ισχύουσα νομοθεσία. Έτσι σύμφωνα με ένα πολύ πρόσφατο έγγραφο εργασίας^{*****} (working paper) της Ευρωπαϊκής Ένωσης, ακόμα και αν δεν υφίσταται σχετικό νομοθετικό πλαίσιο σε μια χώρα στην οποία επιθυμείται να γίνει μεταφορά και επεξεργασία δεδομένων η μεταφορά αυτή μπορεί εντούτοις να πραγματοποιηθεί αν ο ενδιαφερόμενος προσκομίσει επαρκή στοιχεία ότι ικανοποιείται η ασφάλεια ή αν τα υποκείμενα των δεδομένων (δηλ. οι πολίτες) δώσουν την συγκατάθεση τους.

Σε κάθε περίπτωση ο καταναλωτής ή γενικότερα ο πολίτης βρίσκεται εκτεθειμένος, μη γνωρίζοντας τι ακριβώς δεδομένα συλλέγονται για αυτόν, πότε, και πως χρησιμοποιούνται. Σύμφωνα με το [33], οι καταναλωτές θα πρέπει να ενημερώνονται πως θα χρησιμοποιηθούν τα δεδομένα που θα συλλεχθούν και αν θα αποκαλυφθούν σε τρίτους. Η δεύτερη παράμετρος αυτής της απαίτησης μπορεί να ικανοποιηθεί σχετικά εύκολα. Σ' ότι αναφορά όμως την πρώτη, του πως δηλαδή θα χρησιμοποιηθούν τα συλλεγμένα δεδομένα, κάτι τέτοιο έρχεται σε αντίθεση με την φύση της ανακάλυψης γνώσης. Αφού η ανακάλυψη γνώσης στηρίζεται στην εξαγωγή άγνωστων μοτίβων από μια βάση δεδομένων, δεν είναι δυνατόν να γνωρίζουμε εκ των προτέρων ποια δεδομένα θα είναι τελικά χρήσιμα ή τι γνώση θα αποκαλυφθεί. Έτσι, ο εξ αρχής προσδιορισμός του τελικού σκοπού και του αποτελέσματος μιας διαδικασίας ανακάλυψης γνώσης και ο περιορισμός κάποιου στη χρήση μόνο συγκεκριμένων δεδομένων και μεθοδολογιών - αλγορίθμων έρχεται σε πλήρη αντίθεση με την φιλοσοφία της ανακάλυψης γνώσης. Πρακτικά οι πολίτες θα πρέπει να ενημερώνονται πως θα χρησιμοποιηθούν τα δεδομένα που θα συλλεχθούν για αυτούς, αν θα δοθούν σε τρίτους και τέλος όποτε και αν γίνει οποιαδήποτε αλλαγή χρήσης τους. Έτσι οι πολίτες θα πρέπει να έχουν τουλάχιστον τρεις επιλογές σε σχέση με τα προσωπικά τους δεδομένα, άσχετα αν τελικά δεν καταλαβαίνουν ή δεν δώσουν τελικά τη δέουσα σημασία σ' αυτές τις επιλογές:

1. Να μην επιτρέπεται καμία επεξεργασία των δεδομένων τους
2. Να επιτρέπεται η επεξεργασία μόνο για χρήση μέσα στην ίδια την επιχείρηση
3. Να επιτρέπεται η επεξεργασία τόσο για εσωτερική χρήση όσο και προς τρίτους..

Ερευνητικά η διασφάλιση της ιδιωτικότητας σε εφαρμογές ανακάλυψης γνώσης αποτελεί μια πολύ νέα ερευνητική κατεύθυνση τόσο στον τομέα της ανακάλυψης γνώσης όσο και στις στατιστικές βάσεις (statistical databases) [3], όπου ουσιαστικά αναλύονται αλγόριθμοι

^{*****} Έγγραφο εργασίας SEC/2006/ 95 της 20^{ης} Ιανουαρίου 2006.

ανακάλυψης γνώσης για τις παρενέργειες που ενδέχεται να έχουν στην ιδιωτικότητα των δεδομένων αλλά και πως μπορούν να διαφυλάξουν αυτή.

Το κυρίαρχο ζήτημα στην διασφάλιση της ιδιωτικότητας κατά την εφαρμογή τεχνικών ανακάλυψης γνώσης έχει δύο πτυχές. Πρώτον, ευαίσθητα πρωτογενή δεδομένα όπως ονόματα, διευθύνσεις και άλλα παρόμοια, θα πρέπει να μετατραπούν ή να αποκοπούν από την αρχική βάση προκειμένου ο τελικός παραλήπτης των δεδομένων αυτών να μην είναι σε θέση να θέσει την ιδιωτικότητα οποιουδήποτε σε κίνδυνο. Δεύτερον, ευαίσθητη γνώση που μπορεί να προκύψει από μια βάση δεδομένων με τη χρήση τεχνικών ανακάλυψης γνώσης, θα πρέπει επίσης να αποκλείεται, επειδή ακόμα και αυτή η γνώση μπορεί να θέσει σε κίνδυνο την ιδιωτικότητα κάποιου ατόμου. Έτσι η κύρια φροντίδα είναι να αναπτύξουμε αλγορίθμους προκειμένου να μετατρέπουμε τα αρχικά δεδομένα κατά κάποιο τρόπο, έτσι ώστε τα ευαίσθητα και προσωπικά δεδομένα και γνώση να παραμένουν ιδιωτικά ακόμα και μετά την διαδικασία ανακάλυψης γνώσης. Έχουν προταθεί και εφαρμοστεί πολλές προσεγγίσεις για την διασφάλιση της ιδιωτικότητας κατά την ανακάλυψη γνώσης. Αυτές μπορούν να ταξινομηθούν στις ακόλουθες κατηγορίες:

- ▮ Κατανομή δεδομένων (data distribution)
- ▮ Τροποποίηση δεδομένων (data modification)
- ▮ Αλγόριθμοι ανακάλυψης γνώσης (data mining algorithm)
- ▮ Απόκρυψη δεδομένων ή κανόνων (data or rule hiding)
- ▮ Διαφύλαξη της ιδιωτικότητας (privacy preservation)

Δεν θα αναλύσουμε περαιτέρω την λειτουργία της κάθε κατηγορίας επειδή ξεφεύγει από τους σκοπούς της παρούσας διατριβής αλλά ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στο [122] για περισσότερες πληροφορίες.

Εν κατακλείδι θα πρέπει να σημειώσουμε ότι παρόλα τα μεγάλα οφέλη που μπορεί να μας προσφέρει η ανακάλυψη γνώσης σε καμία περίπτωση δεν μπορούν να αντισταθμίσουν την ανάγκη και το δικαίωμα των πολιτών για την προστασία των δεδομένων τους. Η όποια διεργασία ανακάλυψης γνώσης θα πρέπει να θεσμοθετείται και να γίνεται με διαφανείς διαδικασίες. Παρότι αντιφατικές έννοιες, η ανακάλυψη γνώσης και η διαφύλαξη της ιδιωτικότητας μπορούν εντούτοις να συνυπάρξουν αρμονικά.

ΑΝΑΦΟΡΕΣ

- [1] K. Aas, *A Survey on Personalized Information Filtering Systems for the World Wide Web*, Report no. 922, Norwegian Computing Center, December 1997.
- [2] D. Achlioptas, and F. McSherry, “Fast Computation of Low Rank Matrix Approximations,” in *Proceedings of ACM STOC*, 2001, pp. 611-618.
- [3] N. Adam, and J. C. Wortmann, “Security-Control Methods for Statistical Databases: A Comparison Study,” *ACM Computing Surveys*, vol. 21, no. 4, pp. 515–556, Dec. 1989.
- [4] Jean-Marc Adamo. *Data Mining for Association Rules and Sequential Patterns*. Springer, New York, 2001.
- [5] C. C. Aggarwal, and P. S. Yu, “A New Approach to Online Generation of Association Rules,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 4, pp. 527-540, 2001.
- [6] C. C. Aggarwal, and P. S. Yu, “A new framework for itemset generation,” in *Proceedings of the ACM PODS International Conference on Principles of Database Systems*, 1998, pp. 18-24.
- [7] C. C. Aggarwal, and P. S. Yu, “Online generation of association rules,” in *Proceedings of 14th International Conference on Data Engineering – ICDE*, 1998, pp. 402-411.
- [8] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207-216.
- [9] R. Agrawal, and J. C. Shafer, *Parallel Mining of Association Rules, Design, Implementation and Experience*, IBM Research Report, 1996.
- [10] R. Agrawal, and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” in *Proceedings of the 20th International Conference on Very Large Data Bases*, 1994, pp. 487-499.
- [11] R. Agrawal, and S. Srikant, “Mining Sequential Patterns,” in *Proceedings of 11th International Conference Data Engineering ICDE*, 1995, pp.3-14.
- [12] R. Agrawal, and R. Srikant, “Privacy-preserving data mining,” in *Proceedings of the ACM SIGMOD Conference on Management of Data*, 2000, pp. 439-450.
- [13] M.-L. Antonie, and O. R. Zaiane, “Mining Positive and Negative Association Rules: An Approach for Confined Rules,” in *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD’04)*, 2004, pp. 27-38.
- [14] B. Arunasalam, S. Chawla, and P. Sun, “Striking Two Birds With One Stone: Simultaneous Mining of Positive and Negative Spatial Patterns,” in *Proceedings of the 5th SIAM International Conference on Data Mining (SDM’05)*, 2005, pp. 173-183.

-
- [15] Y. Aumann, and Y. Lindell, “Statistical theory for quantitative association rules,” *Journal of Intelligent Information Systems*, vol. 20, no. 3, pp. 255-283, 2003
 - [16] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia, “Spectral analysis of data,” in Proceedings of the 33rd Symposium on Theory of Computing STOC, 2001, pp. 619-626.
 - [17] R. Baeza Yates, and B. Ribeiro Neto, *Modern Information Retrieval*, Addison Wesley, 1999.
 - [18] B. Barber, and H. J. Hamilton, “Extracting share frequent itemsets with infrequent subsets,” *Data Mining and Knowledge Discovery*, vol. 7, no., pp. 153-185, 2003.
 - [19] R. J. Bayardo, and R. Agrawal, “Mining the Most Interesting Rules,” in Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, 1999, pp. 145-154.
 - [20] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, “The R*-tree: An efficient and robust access method for points and rectangles,” in Proceedings of ACM-SIGMOD International Conference Management of Data, 1990, pp. 322-331.
 - [21] M. J.A. Berry, and G. Linoff. *Data mining Techniques for Marketing, Sales and Customer Support*. John Willey & Sons Inc., 1996.
 - [22] F. Bodon, “A fast APRIORI implementation,” in Workshop on Frequent Itemset Mining Implementations (FIMI'03), 2003.
 - [23] F. Bodon, I.N Kouris, C.H. Makris, and A.K. Tsakalidis, “Automatic Discovery of Locally Frequent Itemsets in the Presence of Highly Frequent Itemsets,” *Intelligent Data Analysis Journal*, vol. 9, no. 1, pp. 83-104, 2005.
 - [24] J. Breese, D. Heckerman, and C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” in Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998, pp. 43-52.
 - [25] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, “Using association rules for product assortment decisions: a case study,” in Proceedings of the International Conference on Knowledge Discovery and Data Mining, 1999, pp. 254-260.
 - [26] S. Brin, R. Motwani, and C. Silverstein, “Beyond Market Baskets: Generalizing association rules to correlations,” in Proceedings of ACM SIGMOD International Conference on Management of Data, 1997, pp.265-276.
 - [27] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, “Dynamic itemset counting and implication rules for market basket data,” in Proceedings of ACM SIGMOD International Conference on Management of Data, 1997, pp. 255-264.
 - [28] S. Brin and L. Page, “Dynamic Data Mining: Exploring Large Rule Spaces by Sampling,” <http://www-db.stanford.edu/~sergey/ddm.ps>.
-

-
- [29] C. Buckley, G. Salton, and J. Allan, "The Effect of Adding Relevance Information in a Relevance Feedback Environment," in Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 1994, pp. 292-300.
 - [30] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic Query Expansion Using SMART:TREC3," in Proceedings of the 3rd Text REtrieval Conference (TREC3) NIST, 1994, pp.69-80.
 - [31] P. Cabena, P. Hadjinian, R. Stadler, J. Verhess, and A. Zanasi, *Discovering data mining: from concept to implementation*. NJ: Prentice Hall, 1997.
 - [32] C. H. Cai, Ada Wai-Chee Fu, C. H. Cheng, and W. W. Kwong, "Mining association rules with weighted items," in Proceedings of 1998 International Database Engineering and Applications Symposium (IDEAS'98), 1998, pp. 68-77,.
 - [33] A. Cavoukian, "Data Mining: Staking a Claim on Your Privacy," 2001, Report on Information Privacy in the Private Sector Report,
http://www.legis.gov.bc.ca/cmt/36thparl/priv_ps/Reports/report010320.htm
 - [34] S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan Kauffman, 2001.
 - [35] S. Chakrabarti, B. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Spectral filtering for resource discovery," in Proceedings of the ACM SIGIR Workshop on Hypertext Analysis, 1998, pp. 13-21.
 - [36] P. Cheeseman, and J. Stutz, "Bayesian classification (AutoClass): Theory and results," in *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, AAAI/MIT Press, 1996, pp. 153-180.
 - [37] M.S. Chen, J. Han, and P.S. Yu, "Data Mining: An overview from Database Perspective," IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866--883, December 1996.
 - [38] X. Chen, I. Petrounias, and H. Heathfield, "Discovering temporal association rules in temporal databases," in Proceedings of International Workshop on Issues and Applications of Database Technology (IADT), 1998, pp. 312-319.
 - [39] D. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu, "A Fast Distributed Algorithm for Mining Association Rules," in Proceedings of the 4th International Conference on Parallel and Distributed Information System (PDIS), 1996, pp.31-43.
 - [40] D. Cheung, J. Han, V. Ng, and C.Y. Wong, "Maintenance of discovered association rules in large databases: An incremental updating technique," In Proceedings of International Conference on Data Engineering, 1996, pp. 106-114.
-

-
- [41] W. L. Cheung, S.D. Lee, and Benjamin Kao, "A general incremental technique for maintaining discovered association rules," in Proceedings of the 5th International Conference On Database Systems For Advanced Applications (DASFAA), 1997, pp. 185-194.
 - [42] K.J. Cios, and L. Kurgan, "Trends in Data Mining and Knowledge Discovery," in *Knowledge Discovery in Advanced Information Systems*, N.R. Pal, L.C. Jain, and N. Teoderesku, editors, Springer, 2002.
 - [43] C. Clifton, and D. Marks, "Security and privacy implications of data mining," in Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1996, pp. 15-19.
 - [44] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. MIT Press, 1990.
 - [45] R. Coster, and M. Svenson, "Inverted File Search Algorithms for Collaborative Filtering," in Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 246-252.
 - [46] T. Cover, and P. Hart, "Nearest Neighbor Pattern Classification," in IEEE Transactions on Information Theory, pp. 21-27, 1967.
 - [47] N. Cristianini, and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.
 - [48] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 217-228.
 - [49] U.M Fayyad, G. Piatetsky-Shapiro, P. Smuth and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996.
 - [50] D. Fisher, "Improving inference through conceptual clustering," in Proceedings of AAAI Conference, 1987, pp. 461-465.
 - [51] D. Fisher, "Optimization and simplification of hierarchical clusterings," in Proceedings of 1st International Conference on Knowledge Discovery and Data Mining, 1995, pp. 118-123.
 - [52] W. Frakes, and R. Baeza-Yates, *Information Retrieval: data structures and algorithms*. Prentice Hall, Englewood Cliffs, New Jersey, 1992.
 - [53] J. Friedman, "Flexible Metric Nearest Neighbor Classification," Technical Report, Department of Statistics, Stanford University, 1994.
 - [54] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61-70, Dec. 1992.
 - [55] S.W. Golomb, "Run_Length Encodings," *IEEE Transactions Information Theory*, vol. 12, no. 3, pp. 399-401, July 1966.
-

-
- [56] J. Han, and Y. Fu, "Discovery of multiple-level association rules from large databases," in Proceedings of 21st International Conference on Very Large Databases, 1995, pp. 420-431.
 - [57] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.
 - [58] U. Hanani, B. Shpira, and P. Shoval, "Information Filtering: Overview of Issues, Research and Systems," *User Modeling and User-Adapted Interaction*, vol. 11, pp. 203-259, 2001.
 - [59] T. Hastie, and R. Tibshirani, "Discriminant Adaptive Nearest Neighbor Classification," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 6, pp. 607-615, 1996.
 - [60] C. Hider, "Online Association Rule Mining," in Proceedings of ACM SIGMOD International Conference on Management of Data, 1997, pp.171-182.
 - [61] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, "Recommending and Evaluating Choices in a Virtual Community of Use," in Proceedings of CHI, pp. 194-201.
 - [62] J. D. Holt, and S. M. Chung, "Mining association rules using inverted hashing and pruning," *Information Processing Letters*, vol. 83, no. 4, pp. 211-220, 2002.
 - [63] J. D. Holt, and S. M. Chung, "Mining of Association Rules in Text Databases Using Inverted Hashing and Pruning," in Proceedings of International Conference on Data Warehousing and Knowledge Discovery – DaWak, 2000, pp. 290-300.
 - [64] J. D. Holt, and S. M. Chung, "Multipass Algorithms for Mining Association Rules in Text Databases," *Knowledge Information Systems*, vol. 3, no. 2, pp.168-183, 2001.
 - [65] J. D. Holt, and S. M. Chung, "Efficient Mining of Association Rules in Text Databases," in CIKM, 1999, pp. 234-242
 - [66] <http://www.amazon.com>
 - [67] <http://www.ebay.com>
 - [68] <http://www.egghead.com>
 - [69] A.K. Jain, M.N. Murty, and P.J Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, September 1999.
 - [70] A.K. Jain, and R.C Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.
 - [71] T. Joachims, "Text Categorization with Support Vector Machines," in Proceedings of European Conference on Machine Learning, 1998.
 - [72] L. Kaufman, and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series In Probability And Mathematical Statistics, John Wiley and Sons, Inc., 1989.
-

- [73] W. Kim, *Introduction to Object-Oriented Databases*. The MIT Press, Cambridge, Massachusetts, 1990.
- [74] J. Kleinberg, C. Papadimitriou, and P. Raghavan, A microeconomic view of data mining. *In Data Mining and Knowledge Discovery Journal*, vol. 2, no. 4, pp. 311-324, 1998.
- [75] M. Klemettinen, H. Mannila, and H. Toivonen, "Rule discovery in telecommunication alarm data," *Journal of Network and Systems Management*, vol. 7, no. 4, pp. 395-423, Dec. 1999.
- [76] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," *Communications of the ACM*, vol. 40, no. 3, pp. 77-87, 1997.
- [77] P. Kotler, *Marketing Management*. Tenth Edition, Prentice Hall, Upper Saddle River, New Jersey, 2000.
- [78] I.N. Kouris, C. H. Makris, and A. K. Tsakalidis, "On-Line Generation of Association Rules using Inverted File Indexing and Compression," in Proceedings of IADIS International WWW/Internet Conference, 2002, pp. 29-37.
- [79] I. N. Kouris, C. H. Makris, and A. K. Tsakalidis, "An Improved Algorithm for Mining Association Rules using multiple support values," in Proceedings of FLAIRS International Conference, May 2003, pp. 309-314.
- [80] I.N Kouris, C.H. Makris, and A.K. Tsakalidis, "A Spatiotemporal View of Transactional Data for Data Mining," in 9th WSEAS International Conference on COMPUTERS, Special Session on Algorithmic Techniques for Handling Spatiotemporal and Environmental Data, Vouliagmeni.
- [81] I.N Kouris, C.H. Makris and A.K. Tsakalidis, "Using Information Retrieval techniques for supporting data mining," *In Data & Knowledge Engineering*, vol. 52, no. 3, pp. 353-383, March 2005.
- [82] I.N Kouris, C.H. Makris and A.K. Tsakalidis, "Efficient automatic discovery of 'hot' itemsets," *Information Processing Letters*, vol. 90, no. 2, pp. 65-72, April 2004.
- [83] I.N Kouris, C.H. Makris and A.K. Tsakalidis, "A Spatiotemporal View of Transactional Data for Data Mining," in *WSEAS Transactions on Information Science and Applications*, vol. 2, no. 8, August 2005, pp. 1179-1184.
- [84] I.N Kouris, E. Theodoridis and A.K. Tsakalidis, "Spatial Indexing Structures," in *WSEAS Transactions on Information Science and Applications*, vol. 2, no. 7, August 2005, pp. 971-976.
- [85] I.N Kouris, C.H. Makris, and A.K. Tsakalidis, "Assessing the Microeconomic Facet of Association Rules via an Efficient Weighting Scheme," in *Proceedings of ICKM*, 2004.

-
- [86] I.N Kouris, C.H. Makris and A.K. Tsakalidis, “On-line Generation of Association Rules Using Inverted File Indexing and Compression,” in *Proceedings of IADIS International Conference*, 2002.
 - [87] I.N Kouris, C.H. Makris, and A.K. Tsakalidis, *On-line Generation of Association Rules Using Inverted File Indexing and Compression*, Technical Report no. 2002/07/02. Computer Technology Institute, 2002.
 - [88] W. Lee, S. J. Stolfo and K. W. Mok, “Mining audit data to built intrusion detection models,” in *Proceedings of ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1998, pp. 66-72.
 - [89] B. Liu, W. Hsu, and Y. Ma, „Mining Association Rules with Multiple Minimum Supports,” in *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 337-341
 - [90] B. Liu, W. Hsu, and Y. Ma, “Pruning and summarizing the discovered associations,” in *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 125-134.
 - [91] B. Liu, Y. Ma, and P. S. Yu, “Discovering unexpected information from your competitors’ web sites,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2001, pp. 144-153.
 - [92] Y. Ma, B. Liu, C. K. Wong, P. S. Yu, and S. M Lee, “Targeting the Right Students Using Data Mining,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2000, pp. 457-464.
 - [93] H. Mannila, “Database methods for Data Mining,” in *Proceedings of ACM SIGKDD Conference on Knowledge Discovery & Data Mining tutorial*, 1998.
 - [94] T. Mitchel, *Machine Learning*. McGraw-Hill, 1997.
 - [95] R.T. Ng, “Spatial Data Mining: Discovering Knowledge of Clusters from Maps,” in *Proceedings of ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1996, pp. .
 - [96] R.T. Ng, and J. Han, “Efficient and effective clustering methods for spatial data mining,” in *Proceedings of 12th International Conference on Very Large Data Bases*, 1994, pp. 144-155.
 - [97] R. Ng, and J. Han, “Efficient and effective clustering method for spatial data mining,” in *Proceedings of International Conference Very Large Data Bases*, 1994, pp. 144-155.
 - [98] J.-S. Park, M.-S. Chen, and P.S. Yu, “An effective hash based algorithm for mining association rules,” in *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1995, pp. 175-186.
-

-
- [99] J. -S. Park, M.-S. Chen and P. S. Yu, "Efficient Parallel and Data Mining for Association Rules," in *Proceedings of the 4th International Conference on Information and Knowledge Management*, 1995, pp. 31-36.
 - [100] M. Persin, "Document filtering for fast ranking," in *Proceedings of ACM-SIGIR International Conference on Research and Development in Information Retrieval*, 1994, pp. 339-348.
 - [101] M. Persin, J. Zobel, and R. Sacks-Davis, "Filtered document retrieval with frequency-sorted indexes," *Journal of the American Society for Information Science*, vol. 47, no. 10, pp. 749-764, May 1996.
 - [102] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in *Knowledge Discovery in Databases*, 1991, pp. 229-248.
 - [103] N. Raymond, and H. Jiawei, "Efficient and Effective Clustering Methods for Spatial Data Mining," in *Proceedings of the 20th VLDB Conference*, 1994, pp. .
 - [104] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," in *Proceedings of CSCW*, 1994, pp. 175-186.
 - [105] P. Resnick, and H.R. Varian, "Recommender Systems," *Special issue of Communications of the ACM*, vol. 40, no. 3, pp. 56-58, 1997.
 - [106] J.F. Roddick, K. Hornsby, and M. Spiliopoulou, "YABTSSTDMR - yet another bibliography of temporal, spatial and spatio-temporal data mining research," in *SIGKDD Temporal Data Mining Workshop*, 2001, pp. 167-175.
 - [107] J.F. Roddick, K. Hornsby, and M. Spiliopoulou, "A Survey of Temporal Knowledge Discovery Paradigms and Methods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no.4, pp. 750-767, 2002.
 - [108] G. Salton, and M. McGill, *Introduction to modern Information Retrieval*. McGraw-Hill, New York, 1983..
 - [109] A. Savasere, E. Omiecinski, and S. Navathe, "An efficient algorithm for mining association rules in large databases," in *Proceedings of the 21st VLDB Conference*, 1995, pp. 432-443.
 - [110] A. Savasere, E. Omiecinski, and S. B. Navathe, "Mining for strong negative associations in a large database of customer transactions," in *Proceedings of 14th International Conference on Data Engineering (ICDE '98)*, 1998, pp. 494-502.
 - [111] U. Shardanand, and P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," in *Proceedings of CHI*, 1995, pp. 210-217.
 - [112] C. Silverstein, S. Brin, R. Motwani, and J. D. Ullman, "Scalable techniques for mining causal structures," *Data Mining and Knowledge Discovery*, vol. 4, no3, pp. 163-192, 2000.
-

-
- [113] L. Singh, P. Scheuermann, and B. Chen, “Generating Association Rules from Semistructured Documents using an extended Concept Hierarchy,” in *Proceedings of the 6th International Conference on Information and Knowledge Management*, 1997, pp. 193-200.
 - [114] M. Spiliopoulou, “The laborious way from data mining to web mining,” *International Journal of Comp. Sys., Sci. & Eng., Special Issue on “Semantics of the Web”*, vol. 14 , pp. 113-126, Mar. 1999.
 - [115] R. Srikant, and R Agrawal, “Mining Generalized Association Rules,” in Proceedings of 21st International Conference on Very Large Databases, 1995, pp. 407-419.
 - [116] D. Struck, “Don’t Store My Data, Japanese Tell Government,” August 25, 2002, Int’l Herald Tribune, p. 1.
 - [117] D. Thiruvady, and G. Webb, “Mining Negative Rules using GRD,” in *Proceedings of the 8th Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD’04)*, 2004, pp. 161-165.
 - [118] H. Toivonen, “Sampling large databases for finding association rules,” in Proceedings of the 22nd International Conference on Very Large Databases, 1996, pp. 134-145.
 - [119] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila, “Pruning and grouping of discovered association rules,” in Workshop Notes of the ECML Workshop on Statistics Machine Learning and Knowledge Discovery in Databases, 1995, pp. 47-52.
 - [120] S. Ye, and J.A. Keane, “Mining association rules in temporal databases,” in Proceedings of International Conference on Systems, Man and Cybernetics, 1998, pp. 2803-2808.
 - [121] X. Yuan, B. Buckles, Z. Yuan, and J. Zhang, “Mining negative association rules,” in *Proceedings of the 7th International Symposium on Computers and Communications (ISCC’02)*, 2002, pp. 623-629
 - [122] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, “State-of-the-art in privacy preserving data mining,” *SIGMOD Record*, vol. 33, no. 1, pp 50-57, 2004.
 - [123] K. Wang, Y. He, and J. Han, “Pushing support constraints into frequent itemset mining,” in Proceedings of 25th International Conference on Very Large Databases, 1999, pp. 363-374.
 - [124] K. Wang, Y. He and J. Han, “Mining Frequent Itemsets Using Support Constraints,” in Proceedings of 26th International Conference on Very Large Data Bases, 2000, pp. 43-52.
 - [125] G. I. Webb, “Efficient search for association rules,” in Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pp. 99-107.
 - [126] I. Witten, A. Moffat, and T. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco, second edition, 1999.
-

- [127] M. J. Zaki, and K. Gouda, *Fast vertical mining using diffset*,” Technical Report 01-1, Department of Computer Science, Rensselaer Polytechnic Institute, 2001.
- [128] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: an efficient data clustering method for very large databases,” in Proceedings of ACM-SIGMOD International Conference on Management of Data 1996, pp. 103-114.
- [129] S. Zhang, C. Zhang and X. Yan, “PostMining: Maintenance of Association Rules by Weighting,” *Information Systems*, vol. 28, no. 7, pp. 691-707, October 2003.
- [130] G. K. Zipf, *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Boston, MA, 1949.
- [131] J. Zobel, and A. Moffat, “Exploring the similarity space,” in ACM SIGIR FORUM, 32(1):18--34, 1998.
- [132] J. Zobel, A. Moffat, and R. Sacks-Davis, “Storage management for files of dynamic records,” in Proceedings of 4th Australian Database Conference, 1993, pp. 26-38.

ΕΥΡΕΤΗΡΙΑ

ΠΡΟΛΟΓΟΣ	3
ΚΕΦΑΛΑΙΟ 1 ΕΙΣΑΓΩΓΗ	1
1.1 ΕΙΣΑΓΩΓΗ.....	2
1.2 ΤΙ ΕΙΝΑΙ ΤΟ DATA MINING.....	4
1.3 ΣΕ ΠΟΙΑ ΕΙΔΗ ΒΑΣΕΩΝ – ΔΕΔΟΜΕΝΩΝ ΜΠΟΡΟΥΜΕ ΝΑ ΚΑΝΟΥΜΕ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ	6
1.3.1 ΣΧΕΣΙΑΚΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ.....	6
1.3.2 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΣΥΝΑΛΛΑΓΩΝ	6
1.3.3 ΧΩΡΙΚΕΣ ΚΑΙ ΧΡΟΝΙΚΕΣ ΒΑΣΕΙΣ	7
1.3.4 ΒΑΣΕΙΣ ΚΕΙΜΕΝΩΝ ΚΑΙ ΠΟΛΥΜΕΣΙΚΕΣ ΒΑΣΕΙΣ	8
1.3.5 Ο ΠΑΓΚΟΣΜΙΟΣ ΙΣΤΟΣ	8
1.4 ΤΙ ΕΙΔΗ ΤΕΧΝΙΚΩΝ ΜΠΟΡΟΥΝ ΝΑ ΕΦΑΡΜΟΣΤΟΥΝ	8
1.4.1 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΕΩΝ.....	9
1.4.2 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ	10
1.4.3 ΣΥΣΤΑΔΟΠΟΙΗΣΗ	10
1.5 ΑΠΑΙΤΗΣΕΙΣ ΚΑΙ ΠΡΟΚΛΗΣΕΙΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ.....	11
1.6 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΤΕΧΝΙΚΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ	14
ΚΕΦΑΛΑΙΟ 2 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΕΩΝ	16
2.1 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΕΩΝ	18
2.2 ΕΥΡΕΣΗ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΕΩΝ.....	19
2.3 Ο ΑΛΓΟΡΙΘΜΟΣ APRIORI	21
2.3.1 ΔΗΜΙΟΥΡΓΙΑ ΥΠΟΨΗΦΙΩΝ ΑΝΤΙΚΕΙΜΕΝΩΝ.....	22
2.4 Ο ΑΛΓΟΡΙΘΜΟΣ DYNAMIC ITEMSET COUNTING - DIC	24
2.5 Ο ΑΛΓΟΡΙΘΜΟΣ DHP (DIRECT HASHING AND PRUNING).....	25
2.6 Ο ΑΛΓΟΡΙΘΜΟΣ PARTITION.....	27
2.7 Ο ΑΛΓΟΡΙΘΜΟΣ MSAPRIORI	28
2.8 Η ΓΕΝΝΗΤΡΙΑ ΣΥΝΘΕΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ	31
ΚΕΦΑΛΑΙΟ 3 ΤΕΧΝΙΚΕΣ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ	34
3.1 ΕΥΡΕΤΗΡΙΑ	36
3.1.1 ΕΥΡΕΤΗΡΙΟ ΑΝΕΣΤΡΑΜΜΕΝΩΝ ΑΡΧΕΙΩΝ	36
3.1.2 ΣΥΜΠΙΕΣΗ ΤΟΥ ΕΥΡΕΤΗΡΙΟΥ	38
3.2 ΕΠΙΛΥΣΗ ΔΥΑΔΙΚΩΝ ΕΡΩΤΗΜΑΤΩΝ	38
3.3 ΤΟ ΜΟΝΤΕΛΟ ΤΟΥ ΔΙΑΝΥΣΜΑΤΙΚΟΥ ΧΩΡΟΥ	40

ΚΕΦΑΛΑΙΟ 4	ΑΠΟΤΕΛΕΣΜΑΤΙΚΗ ΑΥΤΟΜΑΤΗ ΑΝΑΚΑΛΥΨΗ “ΚΑΥΤΩΝ”	42
4.1	ΑΝΑΚΑΛΥΠΤΟΝΤΑΣ ΔΙΑΦΑΙΝΟΜΕΝΕΣ ΤΑΣΕΙΣ.....	44
4.2	“ΚΑΥΤΑ” ΠΡΟΪΟΝΤΑ	45
4.3	Η ΠΡΟΤΕΙΝΟΜΕΝΗ ΠΡΟΣΕΓΓΙΣΗ	46
4.4	ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ	48
4.4.1	ΣΥΝΘΕΤΙΚΑ ΔΕΔΟΜΕΝΑ	49
4.4.2	ΧΡΟΝΟΙ ΕΚΤΕΛΕΣΗΣ	50
4.5	ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ.....	51
ΚΕΦΑΛΑΙΟ 5	ΔΗΜΙΟΥΡΓΙΑ ΕΝΟΣ ΑΠΟΤΕΛΕΣΜΑΤΙΚΟΥ ΜΗΧΑΝΙΣΜΟΥ	
	ΑΠΟΔΟΣΗΣ ΒΑΡΩΝ ΓΙΑ ΤΗΝ ΕΝΣΩΜΑΤΩΣΗ ΤΟΥ ΜΙΚΡΟΟΙΚΟΝΟΜΙΚΟΥ	
	ΠΑΡΑΓΟΝΤΑ ΣΤΟΥΣ ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΕΩΝ	52
5.1	ΕΙΣΑΓΩΓΗ	54
5.2	ΛΑΜΒΑΝΟΝΤΑΣ ΥΠΟΨΗ ΤΗΝ ΜΙΚΡΟΟΙΚΟΝΟΜΙΚΗ ΔΙΑΣΤΑΣΗ ΤΩΝ ΚΑΝΟΝΩΝ	
	ΣΥΣΧΕΤΙΣΕΩΝ	55
5.2.1	ΔΙΑΤΥΠΩΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ.....	55
5.2.2	Η ΠΡΟΣΕΓΓΙΣΗ ΜΑΣ	57
5.2.2.1	<i>Ένα παράδειγμα απόδοσης βαρών με την προσέγγισή μας.....</i>	<i>60</i>
5.2.3	ΠΕΡΙΘΩΡΙΟ ΕΝΔΙΑΦΕΡΟΝΤΟΣ	64
5.3	ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ.....	65
5.4	ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ.....	66
ΚΕΦΑΛΑΙΟ 6	ΕΝΑΣ ΒΕΛΤΙΩΜΕΝΟΣ ΑΛΓΟΡΙΘΜΟΣ ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ ΚΑΝΟΝΩΝ	
	ΣΥΣΧΕΤΙΣΕΩΝ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΠΟΛΛΑΠΛΕΣ ΤΙΜΕΣ ΚΑΤΩΦΛΙΟΥ	
	ΥΠΟΣΤΗΡΙΞΗΣ 68	
6.1	ΕΙΣΑΓΩΓΗ	70
6.2	Ο ΑΛΓΟΡΙΘΜΟΣ ΚΤΜ.....	71
6.3	ΔΙΑΤΑΞΗ ΤΩΝ ΑΝΤΙΚΕΙΜΕΝΩΝ.....	75
6.4	Η ΔΟΜΗ ΔΕΔΟΜΕΝΩΝ.....	76
6.5	ΣΥΓΚΡΙΣΗ ΑΠΟΔΟΣΗΣ.....	78
6.5.1	ΠΕΙΡΑΜΑΤΑ ΜΕ ΣΥΝΘΕΤΙΚΑ ΔΕΔΟΜΕΝΑ	79
6.5.2	ΑΝΟΜΟΙΟΜΟΡΦΑ ΚΑΤΑΝΕΜΗΜΕΝΑ - SKEWED ΔΕΔΟΜΕΝΑ.....	80
6.6	ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ.....	81
ΚΕΦΑΛΑΙΟ 7	ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΤΗΝ	
	ΥΠΟΣΤΗΡΙΞΗ ΑΝΑΚΑΛΥΨΗΣ ΓΝΩΣΗΣ	82
7.1	ΕΙΣΑΓΩΓΗ.....	84
7.2	ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ	85

7.3	ΤΟ ΠΡΟΤΕΙΝΟΜΕΝΟ ΣΥΣΤΗΜΑ.....	87
7.4	ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΗΝ ΠΡΟΣΕΓΓΙΣΗ ΜΑΣ ΣΑΝ ΕΝΑ ΚΛΑΣΣΙΚΟ ΑΛΓΟΡΙΘΜΟ ΕΞΟΥΥΗΣ ΔΕΔΟΜΕΝΩΝ (BATCH MODULE)	88
7.5	ΔΙΑΔΡΑΣΤΙΚΟ MODULE.....	90
7.5.1	ΥΠΗΡΕΣΙΕΣ ΠΟΥ ΒΑΣΙΖΟΝΤΑΙ ΣΕ ΔΥΑΔΙΚΑ ΕΡΩΤΗΜΑΤΑ	91
7.5.1.1	Εύρεση όλων των αντικειμένων που περιέχουν ένα συγκεκριμένο στοιχείο/α.....	91
7.5.1.2	Προτείνοντας αντικείμενα με βάση παρόμοιες συναλλαγές.....	95
7.5.1.3	Χρήση Δυναδικών Ερωτημάτων συμπληρωματικά προς τους Κλασσικούς Αλγορίθμους.....	98
7.5.2	ΥΠΗΡΕΣΙΕΣ ΠΟΥ ΒΑΣΙΖΟΝΤΑΙ ΣΕ ΕΡΩΤΗΜΑΤΑ ΔΙΑΤΑΞΗΣ.....	99
7.5.2.1	Σημαντικότητα των Αντικειμένων.....	99
7.5.2.2	Η προσέγγιση μας	101
7.5.2.2.1	Απόδοση Βαθών στις Υπάρχουσες συναλλαγές.....	101
7.5.2.2.2	Αξιολόγηση Νέων Συναλλαγών	103
7.5.2.2.3	Φάση Προτάσεων.....	104
7.5.2.3	Δημιουργία Ευρετηρίου ειδικά για Ερωτήματα Διάταξης.....	106
7.6	ΑΞΙΟΛΟΓΗΣΗ.....	107
7.6.1	ΣΥΝΘΕΤΙΚΑ ΔΕΔΟΜΕΝΑ.....	107
7.6.2	ΧΡΟΝΙΚΕΣ ΚΑΙ ΧΩΡΙΚΕΣ ΑΠΑΙΤΗΣΕΙΣ ΓΙΑ ΤΗΝ ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ ΕΥΡΕΤΗΡΙΟΥ	110
7.6.3	ΧΡΟΝΟΙ ΑΠΟΚΡΙΣΗΣ.....	111
7.6.4	ΑΞΙΟΛΟΓΗΣΗ ΠΟΙΟΤΗΤΑΣ.....	112
7.6.5	ΣΥΖΗΤΗΣΗ ΠΑΝΩ ΣΤΗΝ ΑΞΙΟΛΟΓΗΣΗ	114
7.7	ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ.....	116
ΚΕΦΑΛΑΙΟ 8 ΑΥΤΟΜΑΤΗ ΑΝΑΚΑΛΥΨΗ ΤΟΠΙΚΑ ΣΥΧΝΩΝ ΠΡΟΪΟΝΤΩΝ ΣΕ ΣΥΝΔΥΑΣΜΟ ΜΕ ΠΟΛΥ ΣΥΧΝΑ ΠΡΟΪΟΝΤΑ		
119		
8.1	ΕΙΣΑΓΩΓΗ.....	121
8.1.1	ΕΝΔΙΑΦΕΡΟΝΤΑ ΑΝΤΙΚΕΙΜΕΝΑ	121
8.2	ΤΟΠΙΚΑ ΣΥΧΝΑ ΑΝΤΙΚΕΙΜΕΝΑ.....	123
8.2.1	ΕΥΡΕΣΗ ΤΟΥ ΠΛΗΘΟΥΣ ΚΑΙ ΤΟΥ ΜΕΓΕΘΟΥΣ ΤΩΝ ΔΙΑΜΕΡΙΣΕΩΝ	124
8.3	Ο ΑΛΓΟΡΙΘΜΟΣ LFI.....	125
8.4	Ο ΑΛΓΟΡΙΘΜΟΣ P-DIC	129
8.5	ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ	133
8.5.1	LFI ENANTION P-DIC	134
8.5.1.1	Επίδραση του Μεγέθους της Διαμέρισης.....	134
8.5.1.2	Επίδραση του Κατωφλίου Υποστήριξης.....	135
8.5.1.3	Επίδραση του Local Frequency Ratio	136
8.5.2	ΟΙ ΑΛΓΟΡΙΘΜΟΙ LFI ΚΑΙ P-DIC ΣΕ ΣΧΕΣΗ ΜΕ ΤΟΥΣ ΠΑΡΑΔΟΣΙΑΚΟΥΣ ΑΛΓΟΡΙΘΜΟΥΣ.....	136
8.5.3	ΠΟΙΟΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ	141
8.6	ΣΥΜΠΕΡΑΣΜΑΤΑ	145
ΚΕΦΑΛΑΙΟ 9 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ ΤΗΣ ΕΡΕΥΝΑΣ 147		
9.1	ΕΙΣΑΓΩΓΗ.....	148
9.2	Η ΕΡΕΥΝΑ ΣΤΗΝ ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ ΔΙΑΧΡΟΝΙΚΑ	148
9.3	ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ ΤΗΣ ΕΡΕΥΝΑΣ.....	151

9.3.1	ΠΡΟΕΚΤΑΣΕΙΣ ΤΗΣ ΥΠΑΡΧΟΥΣΑΣ ΕΡΓΑΣΙΑΣ ΜΑΣ	151
9.3.2	ΝΕΕΣ ΕΡΓΑΣΙΕΣ	153
9.3.2.1	<i>Ενσωμάτωση της αρνητικής επίδρασης κάποιων αντικειμένων.....</i>	<i>153</i>
9.3.2.2	<i>Χρήση της ακριβούς σειράς εμφάνισης των αντικειμένων στη βάση.....</i>	<i>154</i>
9.3.3	ΆΛΛΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ ΜΕ ΒΑΣΗ ΤΗ ΔΙΕΘΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ	156
9.4	ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ ΚΑΙ ΠΡΟΣΤΑΣΙΑ ΠΡΟΣΩΠΙΚΩΝ ΔΕΔΟΜΕΝΩΝ	158
	ΑΝΑΦΟΡΕΣ.....	163
	ΕΥΡΕΤΗΡΙΑ	173
	ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ	177
	ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ	178
	ΕΥΡΕΤΗΡΙΟ ΌΡΩΝ	179
	ΠΑΡΑΡΤΗΜΑ Α ΑΝΤΙΣΤΟΙΧΙΑ ΑΓΓΛΙΚΩΝ – ΕΛΛΗΝΙΚΩΝ ΌΡΩΝ.....	180
	ΠΑΡΑΡΤΗΜΑ Β ΧΡΗΣΙΜΟΙ ΣΥΝΔΕΣΜΟΙ ΣΧΕΤΙΚΑ ΜΕ ΤΗΝ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ	181

Ευρετήριο Πινάκων

Πίνακας 1: Οι παράμετροι που χρησιμοποιούνται για την δημιουργία των συνθετικών δεδομένων	32
Πίνακας 2: Μια βάση η οποία περιέχει 7 συναλλαγές, όπου κάθε γραμμή είναι και μια συναλλαγή	37
Πίνακας 3: Το ανεστραμμένο αρχείο για την βάση του Πίνακας 2	37
Πίνακας 4: Τεχνητές Βάσεις	49
Πίνακας 5: Μια βάση συναλλαγών	60
Πίνακας 6: Τελικά Αποδιδόμενα Βάρη σε κάθε προϊόν	60
Πίνακας 7: Ένα παράδειγμα βάσης που χρησιμοποιούμε για datamining	89
Πίνακας 8: Μια διατεταγμένη λίστα αντικειμένων	94
Πίνακας 9: Τα αντικείμενα που περιέχονται στις συναλλαγές 4, 7, 10 και 16	95
Πίνακας 10: Οι συναλλαγές που περιέχουν το στοιχείο E	98
Πίνακας 11: Βρίσκοντας τα πιο σημαντικά αντικείμενα σε μια συναλλαγή	104
Πίνακας 12: Ρυθμίσεις παραμέτρων για τις χρησιμοποιηθείσες βάσεις	109
Πίνακας 13: Χαρακτηριστικά των Βάσεων μας	138
Πίνακας 14: Οι παράμετροι που χρησιμοποιούνται για τα πειράματα με τον LFI και τον P-DIC σε σχέση με τους παραδοσιακούς αλγορίθμους	146

Ευρετήριο Εικόνων

Εικόνα 1: Οι κυριότεροι τομείς αλληλεπίδρασης του data mining.....	3
Εικόνα 2: Η διαδικασία ανακάλυψης γνώσης	5
Εικόνα 3: Ο ψευδοκώδικας του αλγορίθμου Apriori	24
Εικόνα 4: Η μείωση των περασμάτων από τη βάση που επιτυγχάνεται με χρήση του αλγορίθμου DIC σε σχέση με τον αλγόριθμο Apriori.....	25
Εικόνα 5: Το κυρίως πρόγραμμα του αλγορίθμου μας.....	47
Εικόνα 6: Σχετική απόδοση του MSApriori σε σχέση με τον αλγόριθμο μας.....	51
Εικόνα 7: Οι εκχωρούμενες τιμές υποστήριξης της προσέγγισης μας και του αλγορίθμου MSApriori σε σύγκριση με την μέγιστη τιμή υποστήριξης	62
Εικόνα 8: Οι τιμές υποστήριξης που αποδίδονται από την προσέγγιση μας και τον αλγόριθμο MSApriori στο αντικείμενο B για κυμαινόμενες τιμές της μεταβλητής β και του LS.....	63
Εικόνα 9: Παράδειγμα εκτέλεσης του αλγορίθμου KTM.....	74
Εικόνα 10: Ένα πλέγμα αντικειμένων	75
Εικόνα 11: Οι μεταβάσεις ανάμεσα στις διαφορετικές καταστάσεις.....	78
Εικόνα 12: Σύγκριση των χρόνων εκτέλεσης σε ποσοστά	80
Εικόνα 13: % Βελτίωση του χρόνου εκτέλεσης	81
Εικόνα 14: Η λειτουργία του συστήματος μας	88
Εικόνα 15: Χρήση του συστήματος μας σαν ένας κλασικός αλγόριθμος που δουλεύει σε βήματα.....	90
Εικόνα 16: Οι συνδυασμοί για την σειρά B – F – D – C.....	92
Εικόνα 17: Βρίσκοντας τα k πιο συχνά αντικείμενα	93
Εικόνα 18: Κατανομή βαρών σε μια βάση	109
Εικόνα 19: Απαιτούμενοι χρόνοι για την δημιουργία και συμπίεση ευρετηρίου	111
Εικόνα 20: Απαιτούμενος χρόνος για την δημιουργία και συμπίεση του ευρετηρίου για αυξανόμενο αριθμό συναλλαγών	111
Εικόνα 21: Ο Αλγόριθμος LFI.....	126
Εικόνα 22: Η διαδικασία candidate-gen	127
Εικόνα 23: Η διαδικασία count_1-itemsets	128
Εικόνα 24: Το “τρέξιμο” του αλγορίθμου P-DIC.....	131
Εικόνα 25: Οι αλλαγές καταστάσεων για το αλγόριθμο P-DIC	133
Εικόνα 26: Η επίδραση της αλλαγής του μεγέθους των διαμερίσεων στην απόδοση	135
Εικόνα 27: Η επίδραση της μεταβολής του κατωφλίου υποστήριξης στην απόδοση.....	135
Εικόνα 28: η επίδραση της μεταβολής του Local Frequency Ratio στην απόδοση.....	136
Εικόνα 29: Οι σχετικοί χρόνοι εκτέλεσης για την βάση T40I10D100K	139
Εικόνα 30: Οι σχετικοί χρόνοι εκτέλεσης για την βάση Korasak.....	139
Εικόνα 31: Οι σχετικοί χρόνοι εκτέλεσης για την βάση BMS-WebView-1_mb.....	140
Εικόνα 32: Οι σχετικοί χρόνοι εκτέλεσης για την βάση BMS-WebView-2_mb.....	140
Εικόνα 33: Η εξέλιξη της ανακάλυψης γνώσης σε σχέση με την μηχανική μάθηση [42].....	149
Εικόνα 34: Ζητήματα κατά την έρευνα ανακάλυψης γνώσης [42].....	150
Εικόνα 35: Τάσεις στην έρευνα ανακάλυψης γνώσης [42]	151

Ευρετήριο Όρων

A

Apriori 21, 22, 23, 24, 25, 26, 28, 54, 64, 65, 68, 75, 76, 80, 84, 86, 94, 107, 108, 116, 127, 130, 131, 132, 135, 140

D

data mining 2, 4, 6, 7, 8, 11, 12, 13, 14, 81, 82, 85, 109, 112, 136, 137, 146, 156, 158, 159, 161, 164

M

minLFR..... 119, 121, 138, 140

MIS 28, 29, 45, 46, 47, 54, 59, 68, 69, 70, 71, 72, 73, 74, 75, 76, 115, 116, 117, 118, 119, 122, 123, 124, 125, 126

MSApriori..... 27, 28, 29, 48, 54, 56, 59, 60, 61, 62, 64, 67, 68, 69, 76, 77, 127, 130, 131, 132, 135, 140

A

Αντικείμενο..... 59, 91, 173

Δ

διαμέριση

Διαμερίσεις 5, 114, 117, 119, 120, 121, 124, 131, 138

δισύνολο 70, 126, 173

K

Κανόνες Συσχετίσεων..... 17, 18, 42, 114

κατηγοριοποίηση 13, 14

Κατώφλι βεβαιότητας..... 173

κατώφλι υποστήριξης 9, 19, 21, 24, 26, 27, 28, 29, 44, 45, 47, 49, 52, 53, 54, 63, 68, 73, 74, 86, 112, 115, 117, 119, 120, 121, 123, 128, 129, 130, 132, 135, 138, 173

Καντά Προϊόντα 42

M

μονοσύνολο 70, 71, 126, 131, 173

Π

Πλήρης Αναζήτηση 173

Σ

σπάνιο αντικείμενο 69, 124

στοιχείο 35, 37, 38, 44, 46, 49, 85, 88, 91, 92, 94, 96, 97, 98, 100, 119, 122, 126, 131

συχνό αντικείμενο..... 70, 97, 124

Συχνό αντικείμενο 173

T

Τοπικά Συχνά 114, 117, 121

τοπικά συχνό..... 117, 119, 120, 122, 123, 124

Υ

υποστήριξη 9, 19, 20, 21, 24, 25, 27, 47, 49, 55, 57, 61, 63, 86, 89, 90, 91, 93, 94, 96, 116, 117, 120, 126

ΠΑΡΑΡΤΗΜΑ Α ΑΝΤΙΣΤΟΙΧΙΑ ΑΓΓΛΙΚΩΝ – ΕΛΛΗΝΙΚΩΝ ΌΡΩΝ

Αγγλικός Όρος	Ελληνική Απόδοση
1. <i>1-item</i>	Στοιχείο - μονοσύνολο
2. <i>2-itemset</i>	Αντικείμενο – δισύνολο
3. <i>k-itemset</i>	κ-σύνολο
4. <i>Classification</i>	Κατηγοριοποίηση
5. <i>Clustering</i>	Ομαδοποίηση – συσταδοποίηση
6. <i>Collaborative filtering</i>	Συνεργατικό φιλτράρισμα
7. <i>Data Mining</i>	Εξόρυξη δεδομένων
8. <i>Explicit voting</i>	Άμεση ψηφοφορία
9. <i>False negatives</i>	Αντικείμενα τα οποία λανθασμένα θεωρήθηκαν σαν πιθανά σπάνια αντικείμενα
10. <i>False positives</i>	Αντικείμενα τα οποία λανθασμένα θεωρήθηκαν σαν πιθανά συχνά αντικείμενα
11. <i>Feedback</i>	Ανάδραση
12. <i>Frequent itemset</i>	Συχνό αντικείμενο
13. <i>Full search</i>	Πλήρης Αναζήτηση
14. <i>Hash table</i>	Πίνακας κατακερματισμού
15. <i>Implicit voting</i>	Έμμεση ψηφοφορία
16. <i>Itemset</i>	Αντικείμενο - Προϊόν
17. <i>Knowledge Discovery in Databases – KDD</i>	Ανακάλυψη Γνώσης από Βάσεις Δεδομένων
18. <i>Minimum confidence</i>	Κατώφλι βεβαιότητας
19. <i>Minimum interest ratio</i>	Κατώτατος λόγος ενδιαφέροντος
20. <i>Minimum local frequency ratio</i>	Κατώφλι τοπικής συχνότητας
21. <i>Minimum support</i>	Κατώφλι υποστήριξης
22. <i>Pattern</i>	Μοτίβο - πρότυπο
23. <i>Partition</i>	Διαμέριση - διαμερίζω
24. <i>Polysemy</i>	Πολυσημία
25. <i>Precision</i>	Ακρίβεια
26. <i>Recall</i>	Ανάκληση
27. <i>Scalable</i>	Κλιμακούμενος
28. <i>Skewed databases</i>	Βάσεις με υπερβολικά ανομοιόμορφες κατανομές εμφανίσεων δεδομένων
29. <i>Synonymy</i>	Συνωνυμία
30. <i>Test audience</i>	Κοινό ελέγχου
31. <i>TID – Transaction Identifier</i>	Αναγνωριστικό συναλλαγής

ΠΑΡΑΡΤΗΜΑ Β ΧΡΗΣΙΜΟΙ ΣΥΝΔΕΣΜΟΙ ΣΧΕΤΙΚΑ ΜΕ ΤΗΝ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

ΕΤΑΙΡΕΙΕΣ – ΣΥΣΤΗΜΑΤΑ	
Ηλεκτρονική Διεύθυνση	Εταιρεία
http://www.heuristics.com/	American Heuristics (Profiler)
http://www.angoss.com/	Angoss software (Knowledge Seeker)
http://www.attar.com/	Attar Software (XpertRule Profiler)
http://www.businessobjects.com/	Business Objects (BusinessMiner)
http://www.datamind.com/	DataMind (DataMind Professional)
http://www.hnccs.com/	HNC Software (DataMarksman, Falcon)
http://www.hyperparallel.com/	HyperParallel (Discovery)
http://www.datamining.com/	Information Discovery Inc. (Information Discovery System)
http://www.isl.co.uk/index.html	Integral Solutions (Clementine)
www.ibm.com/software/data/iminer/	IBM (Intelligent Data Miner)
http://www.lucent.com/	Lucent Technologies (Interactive Data Visualization)
http://www.ncr.com/	NCR (Knowledge Discovery Benchmark)
http://www.neovista.com/	NeoVista Solutions (Decision Series)
http://www.nestor.com/	Nestor (Prism)
http://www.pilotsw.com/	Pilot Software (Pilot Discovery Server)
http://www.holossys.com/	Seagate Software Systems (Holos 5.0)
http://www.spss.com/	SPSS (SPSS)
http://www.think.com/	Thinking Machines (Darwin)

ΔΕΔΟΜΕΝΑ	
Ηλεκτρονική Διεύθυνση	Περιγραφή
http://kdd.ics.uci.edu/	UCI KDD Database Repository – η πληρέστερη και δημοφιλέστερη συλλογή δεδομένων για την μηχανική μάθηση αλλά και για την ανακάλυψη γνώσης γενικότερα.
http://fimi.cs.helsinki.fi/	Μια συλλογή δεδομένων τα οποία προέκυψαν από την πραγματοποίηση των συνεδρίων FIMI (τα οποία είχαν να κάνουν με αποτελεσματικές υλοποιήσεις αλγορίθμων ανακάλυψης γνώσης).
http://www.fedstats.gov/	Μια συλλογή που περιλαμβάνει κύρια δεδομένα στατιστικής φύσεως από διάφορες υπηρεσίες των Η.Π.Α.
http://www.mlnet.org/cgi-bin/mlnetois.pl/?File=datasets.html	MLnet (European Machine Learning Network) λίστα με δεδομένα για διάφορες εργασίες ανακάλυψης γνώσης.
http://www.census.gov/	Δεδομένα διαφόρων απογραφών από τις Η.Π.Α.
http://www.kdnuggets.com/datasets/kddcup.html	Δεδομένα που χρησιμοποιήθηκαν κατά το συνέδριο KDD, για την διεξαγωγή διαγωνισμού απόδοσης αλγορίθμων
http://www.almaden.ibm.com/software/quest/Resources/index.shtml	Η γεννήτρια συνθετικών δεδομένων της IBM από την ομάδα Quest καθώς και κάποια έτοιμα δεδομένα
http://www.miles.cnuce.cnr.it/~palmeri/datam/DCI/	Η γεννήτρια συνθετικών δεδομένων της IBM με μια διαφορετική υλοποίηση
ΣΥΝΕΔΡΙΑ	

Ηλεκτρονική Διεύθυνση	Συνέδριο
www.vldb.org/	VLDB (Very Large Data Bases)
Μεταβάλλεται κάθε χρόνο ανάλογα με το συνέδριο	ICDE (International Conference on Data Engineering)
http://www.acm.org/sigs/sigkdd/conferences.php	KDD (Knowledge Discovery in Databases)
http://www.siam.org	SIAM – SDM (International Conference on Data Mining)
Μεταβάλλεται κάθε χρόνο ανάλογα με το συνέδριο	PAKDD (Pacific-Asia Conference on Knowledge Discovery and Data Mining)
www.flairs.com/	FLAIRS (Florida Artificial Research Society)
http://www.sigmod.org/pods/index.html	ACM SIGMOD/PODS (International Conference on Management of Data/ Principles of Database Systems)
http://ijcai.org/	IJCAI (International Joint Conference on Artificial Intelligence)
Μεταβάλλεται κάθε χρόνο ανάλογα με το συνέδριο	ICML (International Conference on Machine Learning)
http://www.mldm.de/	MLDM (International Conference on Machine Learning and Data Mining)
Μεταβάλλεται κάθε χρόνο ανάλογα με το συνέδριο	IDA (Intelligent Data Analysis)
Μεταβάλλεται κάθε χρόνο ανάλογα με το συνέδριο	ECML/PKDD (European Conference on Machine Learning/ Principles and Practice of Knowledge Discovery in Databases)
www.cikm.org/	CIKM (Conference on Information and Knowledge Management)
http://www.cs.uvm.edu/~icdm/	ICDM (IEEE Conference on Data Mining)

ΠΕΡΙΟΔΙΚΑ	
Ηλεκτρονική Διεύθυνση	Τίτλος Εντύπου
www.elsevier.com/locate/knosys	Knowledge Based Systems (Elsevier)
www.elsevier.com/locate/datak	Data and Knowledge Engineering (Elsevier)
www.elsevier.com/locate/cie/	Computers & Industrial Engineering (Elsevier)
www.elsevier.com/locate/infosys	Information Systems (Elsevier)
www.elsevier.com/locate/ins	Information Sciences (Elsevier)
www.elsevier.com/locate/jss	The Journal of Systems and Software (Elsevier)
www.elsevier.com/locate/ipl	Information Processing Letters (Elsevier)
http://www.kluweronline.com/issn/1384-5810	Data Mining and Knowledge Discovery (Kluwer)
http://www.cs.uvm.edu/~kais	Knowledge and Information Systems (Springer)
http://iospress.metapress.com/app/home/A9	Intelligent Data Analysis (IOS Press)
http://www.computer.org/tkde/index.htm	IEEE Transactions on Data and Knowledge Engineering (IEEE)
http://www3.oup.co.uk/computer_journal	The computer Journal
http://ise.gmu.edu/JIIS/	Journal of Intelligent information Systems (Kluwer)
www.acm.org/tods	ACM Transactions on Database Systems – TODS (ACM)
www.acm.org/tois	ACM Transactions on Information Systems – TOIS (ACM)

ΔΙΑΔΙΚΤΥΑΚΟΙ ΤΟΠΟΙ	
Ηλεκτρονική Διεύθυνση	Περιγραφή
http://www.kdnuggets.com	Η πληρέστερη και δημοφιλέστερη ιστοσελίδα σχετικά με την ανακάλυψη γνώσης.
http://www.db-net.aueb.gr/	Ιστοσελίδα του Οικονομικού Πανεπιστημίου Αθηνών
http://www.mlnet.org/	Συλλογή ιστοσελίδων σχετικές με την έρευνα, τις δημοσιεύσεις, τα λογισμικά, θέσεις εργασίες κ.α. στον τομέα της εξόρυξης γνώσης με έμφαση στη μηχανική μάθηση)
http://www.kdnet.org/control/	