ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ, ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΚΑΙ ΔΙΚΤΥΩΝ

ΑΠΟΚΡΥΨΗ ΣΥΧΝΩΝ ΣΤΟΙΧΕΙΟΣΥΝΟΛΩΝ ΜΕΣΩ ΑΝΑΚΑΤΑΣΚΕΥΗΣ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

ΒΑΡΒΑΡΑ ΘΕΟΚΛΗ
ΒΟΛΟΣ 2011

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ, ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΚΑΙ ΔΙΚΤΥΩΝ

ΑΠΟΚΡΥΨΗ ΣΥΧΝΩΝ ΣΤΟΙΧΕΙΟΣΥΝΟΛΩΝ ΜΕΣΩ ΑΝΑΚΑΤΑΣΚΕΥΗΣ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

ΒΑΡΒΑΡΑ ΘΕΟΚΛΗ
ΒΟΛΟΣ 2011

ΕΠΙΒΛΕΠΟΝΤΕΣ

ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ: ΒΑΣΙΛΕΙΟΣ ΒΕΡΥΚΙΟΣ

ΚΑΘΗΓΗΤΗΣ: ΗΛΙΑΣ ΧΟΥΣΤΗΣ

Contents

List of Algorithms

List of Tables

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον επιβλέπων καθηγητή κ. Βασίλειο Βερύκιο, για τη πολύτιμη καθοδήγηση, τις συμβουλές και την υπομονή πού έδειξε σε όλο αυτό το διάστημα. Επίσης τον καθηγητή κ. Ηλία Χούστη για την πολύτιμη συνδρομή του για την ολοκλήρωση της διπλωματικής εργασίας.

Επίσης ένα μεγάλο ευχαριστώ στο Χρήστο για τις πολύτιμες συμβουλές και την υποστήριξη.

Τέλος ευχαριστώ την οικογένεια μου για τη στήριξη και συμπαράσταση που μου παρείχε σε όλη τη διάρκεια των σπουδών μου.

Πρόλογος

Η διαχείρηση βάσεων δεδομένων έχει μετατραπεί σε ένα κεντρικό συστατικό της καθημερινής ζωής στη σύγχρονη κοινωνία. Όμως προκύπτουν ζητήματα εμπιστευτικότητας των ευαίσθητων προσωπικών δεδομένων που υπάρχουν στις βάσεις δεδομένων και χρησιμοποιούνται από διάφορες επιχειρήσεις και οργανισμούς. Η απόκρυψη της ευαίσθητης γνώσης στις βάσεις δεδομένων είναι ένα πρόβλημα που μπορεί να εξεταστεί μέσω απόκρυψης των συχνών στοιχειοσυνόλων στη βάση δεδομένων. Στην παρούσα εργασία, προτείνεται ένας αναδρομικός αλγόριθμος που στοχεύει στην απόκρυψη των συχνών στοιχειοσυνόλων στη βάση δεδομένων μέσω ανακατασκευής του συνόλου δεδομένων. Ο αλγόριθμος ξεκινά με το καθορισμό των συχνών στοιχειοσυνόλων που θέλουμε να κρύψουμε. Στη συνέχεια ξεκινώντας από τα μεγαλύτερου επιπέδου στοιχειοσύνολα μειώνουμε το μετρητή τους τόσο ώστε να μην είναι συχνά ενώ συγχρόνως μειώνουμε και το μετρητή όλων των προγόνων τους. Για την αξιολόγηση του αλγορίθμου, χρησιμοποίηθηκαν πειραματικά δεδομένα τα οποία επιτυγχάνουν την απόκρυψη των συχνών στοιχειοσυνόλων.

1 Εισαγωγή

Η αλματώδης ανάπτυξη της πληροφορικής και των επικοινωνιών τα τελευταία χρόνια έχει καταστήσει την πληροφορία ως ένα από τα βασικότερα και πολυτιμότερα αγαθά. Για τις επιχειρήσεις αποτελεί πλέον, σημαντικό στοιχείο για την παραγωγικότητα τους. Τα συστήματα βάσεων δεδομένων χρησιμοποιούνται για την αποθήκευση, την επεξεργασία και την παραγωγική εκμετάλλευση, ενός σημαντικού όγκου πληροφοριών και δεδομένων που κινεί και παράγει κάθε επιχείρηση ή οργανισμός καθημερινά. Συγκεκριμένα ο όρος εξόρυξη γνώσεως χρησιμοποιείται για αναφορά στις πραγματικές τεχνικές που χρησιμοποιούνται για την ανάλυση και εξαγωγή συμπερασμάτων που αφορούν τη συμπεριφορά ολόκληρης της βάσης δεδομένων. Για παράδειγμα η αγορά ενός προϊόντος όταν αγοράζεται μαζί με ένα άλλο προϊόν δείχνει, τις προτιμήσεις των πελατών και επομένως μια συσχέτηση μεταξύ των προιόντων. Αυτές οι συσχετήσεις χρησιμοποιούνται από τις υπεραγορές. Εκμεταλλευόμενοι τις προτιμήσεις των πελατών, στέλνονται μηνύματα διαφημιστικού και ενημερωτικού περιεχομένου που αφορούν κυρίως προϊόντα που συμπεριλαμβάνονται στις προτιμήσεις των πελατών. Επίσης είναι εφικτή η καταγραφή της πορείας ενός πελάτη, δηλαδή η παρακολούθηση των συνηθειών του πελάτη, όπως πόσο συχνά επισκέπτεται την υπεραγορά και τι προϊόντα αγοράζει. Αυτές οι πληροφορίες βοηθούν στη διαφήμιση, στο μάρκετιγκ, στην ταξινόμηση των ορόφων, στην καλύτερη τιμολόγηση των προϊόντων καθώς επίσης και στον προσδιορισμό οποιοδήποτε προσφορών.

Παρόλο τα πολλά πλεονεκτήματα που προσφέρει αυτή η εξαγωγή συμπερασμάτων, προκύπτει το ζήτημα της ιδιωτικότητας και εμπιστευτικότητας των δεδομένων. Πολλές φορές τα δεδομένα μπορεί να πουληθούν σε άλλους οργανισμούς και υπάρχει ο κίνδυνος έκθεσης των ευαίσθητων προσωπικών δεδομένων σε τρίτους καθώς και η εκμετάλλευση τους. Το ονοματεπώνυμο, η ηλικία, η διεύθυνση, το email, το τηλέφωνο ενός πελάτη οδηγούν στην πλήρη ταυτοποίηση και σε συνδυασμό με τις προτιμήσεις των πελατών αποκτά μεγάλη αξία για τους πωλητές αγαθών και υπηρεσιών. Επίσης σε βάσεις δεδομένων εταιρειών υπάρχει ανάγκη προστασίας των εταιρικών μυστικών που μπορούν να αποκαλυφθούν κατά την εφαρμογή των αλγορίθμων εξόρυξης γνώσης.

Όλα τα παραπάνω οδήγησαν στην ανάπτυξη ενός νέου πεδίου έρευνας στο τομέα της εξόρυξης γνώσης από βάσεις δεδομένων και στατιστικά στοιχεία. Η έρευνα προσανατολίστηκε προς δυο κατευθύνσεις. Πρώτον, την επεξεργασία της αρχικής βάσης ώστε να εξαλειφθούν τα ευαίσθητα δεδομένα, όπως αναγνωριστικά ονόματα και τα λοιπά. Δεύτερον την προστασία ευαίσθητης γνώσης που θα μπορούσε να εξαχθεί μέσω της χρήσης αλγορίθμων εξόρυξης. Αυτό κυρίως επιτυγχάνεται μέσω αλγορίθμων απόκρυψης, όπως αυτόν που παρουσιάζουμε στην παρούσα εργασία, που μεταβάλλουν

την αρχική βάση με τέτοιο τρόπο ώστε να προστατεύουν τα ευαίσθητα δεδομένα.

Τέλος υπάρχει ανάγκη αυτοί οι αλγόριθμοι να είναι αποδοτικοί ώστε κατά την εφαρμογή τους να μην αλλοιώνουν τη βάση σε τόσο μεγάλο βαθμό, που να αποτρέπει την εξαγωγή σωστών συμπερασμάτων.

2 Σχετική Έρευνα

Το πρόβλημα της απόκρυψης των ευαίσθητων δεδομένων μέσα από την διαδικασία εξόρυξης γνώσης είναι ένας τομέας που αναπτύσσεται συνεχώς και εισήχθη αρχικά από τον Atallah και λοιποί[2]. Ιδιαίτερο ενδιαφέρον παρουσιάζει η εργασία των Orlowska και λοιποί [3]. Είναι πολύ ενδιαφέρουσα η μέθοδος ανακατασκευής της βάσης δεδομένων που χρησιμοποιείται στη συγγεκριμένη εργασία για να δημιουργήσει ένα δείγμα δεδομένων της αρχικής βάσης, με σκοπό να προστατεύσει την ευαίσθητη γνώση.

Χρησιμοποιώντας την απόδειξη της λύσης των Atallah και λοιποί [2] ότι το πρόβλημα με τα ευαίσθητα συχνά στοιχεία μπορεί να έχει μια NP-hard βέλτιστη λύση, αρκετές ακόμα έρευνες έχουν προταθεί για να βελτιώσουν το ευρετικό αλγόριθμο, όπως τους Dasseni και λοιποί [4](και η επέκταση αυτής της εργασίας [15]), των Oliveira και λοιποί [10] και των Kantarcioglu και λοιποί [7]και η πιο πρόσφατη εργασία των Zhu και Du [16]. Οι Saygin και λοιποί [13]πρότειναν μια διαφορετική μέθοδο απόκρυψης, κρύβοντας κάποια στοιχεία, δίνοντας τους άγνωστη τιμή. Με παρόμοιο τρόπο οι Pontikakis και λοιποί [11] [12]πρότειναν μια εναλλακτική μέθοδο απόκρυψης αλλάζοντας μηδενικά σε άσσους και αντίστροφα.

Η εργασία των Sun και Yu [14] εισάγει μια άπληστη προσέγγιση με χρήση των περιθωρίων για την απόκρυψη των συχνών στοιχειοσυνόλων. Η μέθοδος εστιάζει στη συντήρηση της ποιότητας των περιθωρίων που κατασκευάζονται από τα μη συχνά ευαίσθητα στοιχειοσύνολα. Η εργασία των Abul [1]εισάγει την έννοια co-occuring των συχνών στοιχειοσυνόλων (υποσύνολα των συχνών στοιχειοσυνόλων) και πώς η ύπαρξή τους μπορεί να έχει επιπτώσεις στη διαμοίραση των ευαίσθητων δεδομένων.

Οι Moustakidis και Verykios [9]παρουσιάζουν δυο ευρετικούς αλγόριθμους απόκρυψης και χρησιμοποιούν το κριτήριο maxmin και το θεώρημα για το περιθώριο των συχνών στοιχειοσυνόλων. Οι προγραμματιστικές προσεγγίσεις για την απόκρυψη των ευαίσθητων στοιχειοσυνόλων όπως προτείνονται από τον Menou [8], χρησιμοποιούν τη διαδικασία απόκρυψης ως CSP για να προσδιορίσουν τον ελάχιστο αριθμό συναλλαγών. Μια παρόμοια προσέγγιση χρησιμοποιείται από Gkoulalas και Verykios [5] με τη χρησιμοποίηση μιας ακριβούς εγγυημένης μεθοδολογίας για να προσδιοριστεί ο μικρότερος αριθμός στοιχειοσυνόλων για τη διαδικασία επεξεργασίας της βάσης, ώστε να προκαλέσουν τις μικρότερες πιθανές παρενέργειες.

Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει η εργασία των Gkoulalas και Verykios [6]. Η προσέγγισή βασίζεται στα περιθώρια για την απόκρυψη των ευαίσθητων στοιχειοσυνόλων και χρησιμοποιεί μια υβριδική προσέγγιση για τις τεχνικές επεξεργασίας της βάσης και ελαχιστοποίησης της επεκταμένης αρχικής βάσης δεδομένων με την προσθήκη νέων

συναλλαγών, για να παρέχει βέλτιστες λύσεις και λύνοντας προβλήματα των προηγούμενων προσεγγίσεων.

3 Ορισμός του Προβλήματος

Στο παρακάτω κεφάλαιο θα παρουσιάσουμε το πρόβλημα της απόκρυψης ευαίσθητης γνώσης μέσω της απόκρυψης συχνών στοιχειοσυνόλων. Καθώς και τους στόχους που θέλουμε να επιτύχουμε με τη χρήση της μεθόδου μας.

3.1 Βασικές Έννοιες

Έστω μια σχεσιακή βάση δεδομένων D. Ως $I=\{i_1,i_2,....,i_n\}$ ορίζουμε ένα σύνολο γνωρισμάτων που ονομάζουμε στοιχεία. Ορίζουμε ως στοιχειοσύνολο ένα υποσύνολο από στοιχεία επιλεγμένο από το σύνολο της μορφής $i_1i_2..i_n$. Επιπλέον ορίζουμε ως επίπεδο(level) του στοιχειοσυνόλου το από πόσα επιμέρους στοιχεία αποτελείται. Για παράδειγμα το $i_1i_2i_3$ είναι ένα στοιχειοσύνολο τρίτου επιπέδου μιας και αποτελείται από τα στοιχεία i_1,i_2,i_3 και συμβολίζεται ως 3-στοιχειοσύνολο. Από τα προηγούμενα προκύπτει το μέγιστο επίπεδο στοιχειοσυνόλου που μπορούμε να έχουμε, αποτελείται από το μέγιστο άθροισμα όλων των διαθέσιμων στοιχείων. Δηλαδή αν το σύνολο των στοιχείων είναι $I=\{i_1, i_2, i_3, i_4\}$ τότε το μέγιστο στοιχειοσύνολο μας - maxlevel_στοιχειοσύνολο είναι $I=\{i_1i_2i_3i_4\}$.

Ως συχνά στοιχειοσύνολα ορίζουμε τα στοιχειοσύνολα που εμφανίζονται στις συναλλαγές της βάσης με ρυθμό μεγαλύτερο από μια ελάχιστη συχνότητα εμφάνισης s_{min}που ορίζει ο χρήστης και τα συμβολίζουμε με FI (frequent itemsets). Ενώ ως μη συχνά στοιχειοσύνολα ορίζουμε όσα δεν ξεπερνούν το s_{min} και αντιστοίχως τα συμβολίζουμε ως IF (infrequent Itemsets).

Επίσης λέμε πως ένα στοιχειοσύνολο/στοιχείο είναι απόγονος ενός άλλου αν περιέχει όλα τα στοιχεία του και αντίστοιχα ορίζουμε το στοιχειοσύνολο/στοιχείο πρόγονο. Για παράδειγμα το στοιχειοσύνολο $i_1i_2i_3$ είναι απόγονος των i_1, i_2, i_3, i_1i_2, i_1i_3, i_2i_3και αντίστοιχα αυτά τα στοιχειοσύνολα και στοιχεία είναι πρόγονοι του.

Τέλος έστω $D=\{t_1,t_2,...,t_n\}$ να είναι το σύνολο συναλλαγών της βάσης D. Για κάθε στοιχειοσύνολο X λέμε πως περιέχεται στη συναλλαγή t_i αν το X είναι ίσο ή υποσύνολο του t_i.

3.2 Παρουσίαση Προβλήματος

Έστω μια σχεσιακή βάση δεδομένων D και ένα ελάχιστο κατώφλι υποστήριξης s_{min}. Μας δίνεται ένα σύνολο από συχνά στοιχειοσύνολα H $\subset$FI τα οποία πρέπει να κρυφτούν

δηλαδή να χάσουν την ιδιότητα του συχνού στοιχειοσυνόλου για δεδομένο s_{min}. Επιπλέον λόγω του ορισμού των συχνών στοιχειοσυνόλων πρέπει να κρύψουμε και όλους τους απόγονους τους. Ο στόχος μας είναι να ανακατασκευάσουμε την αρχική μας βάση D σε μια νέα βάση D' η οποία θα επιτυγχάνει αυτόν τον στόχο. Επιπλέον για να διασφαλίσουμε ότι τα εξαγώγιμα αποτελέσματα των αλγορίθμων εξόρυξης γνώσης παραμένουν χρήσιμα και αληθή θα πρέπει να διασφαλίσουμε ότι στην τελική μας βάση D' διασφαλίζουμε όσο μπορούμε την ποιότητα του σύνολου των δεδομένων. Δηλαδή μέσω της χρήσης της μεθόδου απόκρυψης που θα χρησιμοποιήσουμε δε πρέπει να δημιουργούμε νέα συχνά στοιχειοσύνολα από μη συχνά ή να κρύβουμε συχνά στοιχειοσύνολα που δε μας έχει ζητηθεί. Επιπλέον πρέπει να φροντίζουμε ώστε οι τελικές υποστηρίξεις όλων των στοιχειοσυνόλων μας, αλλά ιδιαίτερα όσων είναι σημαντικά για τη διαδικασία εξόρυξης γνώσης να μένουν όσο των δυνατόν αναλλοίωτες

4 Παρουσίαση Βασικών Σημείων Μεθοδολογίας

Στο επόμενο κεφάλαιο θα παρουσιάσουμε κάποια βασικά μεγέθη και έννοιες που θα χρησιμοποιήσουμε στην μεθοδολογία μας.

4.1 Βάσεις δεδομένων

Έστω D και D'η αρχική και τελική βάση δεδομένων μας αντίστοιχα. Με |D| θα συμβολίζουμε το συνολικό αριθμό συναλλαγών που περιέχονται στη βάση D.

4.2 Αριθμός συναλλαγών και ελάχιστο κατώφλι υποστήριξης

Στην ανακατασκευή της βάσης δεδομένων με σκοπό την απόκρυψη συχνών στοιχειοσυνόλων προκύπτουν κάποια προβλήματα από τον ορισμό των συχνών στοιχειοσυνόλων συγκεκριμένα:

Αν s_{min} είναι το ελάχιστο κατώφλι υποστήριξης για να είναι συχνό ένα στοιχειοσύνολο, C_i ο αριθμός συναλλαγών της βάσης D που υποστηρίζουν το i και |D| ο αριθμός όλων των συναλλαγών που περιέχονται στη βάση D. Τότε $s_{min} = \frac{C_i}{|D|}$.

Από αυτό προκύπτει ότι για δεδομένο s_{min} και |D| υπάρχει ένα $\min C_i$ για το οποίο είναι ένα στοιχειοσύνολο συχνό.

4.3 Θετικό και αρνητικό περιθώριο

Στη συνέχεια θα ορίσουμε τις έννοιες του αρνητικού και θετικού περιθωρίου. Έστω $FI=\{X \subseteq I : s_D(X) \geq s_{min}\}$

είναι το σετ όλων των συχνών στοιχειοσυνόλων στη D. Ορίζουμε ως αρνητικό περιθώριο του FI, και το συμβολίζουμε $B^-(FI)$, ως το σύνολο όλων των μη-συχνών στοιχειοσυνόλων του D των οποίων όλα τα υποσύνολα ανήκουν στο FI. Δηλαδή $B^-(FI)=\{X \subseteq I : X \notin F \wedge \forall Y \subset X : Y \in F\}$. Ομοίως ορίζουμε το θετικό περιθώριο του FI , $B^+(FI)$ ως το σύνολο όλων των συχνών στοιχειοσυνόλων των οποίων

όλοι οι απόγονοι είναι μη συχνά στοιχειοσύνολα. $B^+(FI)=\{X \subseteq I : X \in F \wedge \forall Y \supset X : Y \notin F\}$. Από αυτούς τους δυο ορισμούς εξάγουμε δυο χρήσιμα συμπεράσματα. Το θετικό περιθώριο περίεχει όλα τα συχνά στοιχειοσύνολα με τα μικρότερα C_i ενώ στους προγόνους τους περιέχονται και όλα τα υπόλοιπα συχνά στοιχειοσύνολα. Ομοίως το $B^-(FI)$ περιέχει τα μη-συχνά στοιχειοσύνολα με τα μεγαλύτερα C_i(μικρότερα βεβαίως από το min C_i). Αν καταφέρναμε να μετακινήσουμε τα στοιχειοσύνολα που περιέχονται στο H στο $B^-(FI)$ και να διατηρήσουμε το $B^+(FI)$ (αφού το αναθεωρήσουμε για τυχόν αλλαγές λόγω H) θα έχουμε επιτύχει το στόχο μας.

4.4 Σύνολο δεδομένων και σύνολο στοιχειοσυνόλων.

Για κάθε βάση D που περιέχει ένα σύνολο από συναλλαγές μπορούμε να πάρουμε το σύνολο των στοιχειοσυνόλων που περιέχονται σε αυτές μαζί με τις φορές που εμφανίζονται σε αυτές (C_i) χρησιμοποιώντας κάποιον αλγόριθμο εξόρυξης σαν τον Apriori. Ονομάζουμε αυτό το σύνολο P(I). Στο "A new framework of privacy preserving data sharing" [3] αποδείχτηκε πως μπορούμε να αντιστοιχήσουμε ακριβώς αυτό το P(I) με το σύνολο δεδομένων μας (βάση D). Στη συνέχεια θα παραθέσουμε κάποια βασικά στοιχεία που είχαν παρουσιαστεί για την ευκολία του αναγνώστη.

Για κάθε στοιχειοσύνολο $X \epsilon P(I)$ ο αριθμός C_x μπορεί να προκύπτει είτε από συναλλαγές που υποστηρίζουν αποκλειστικά το X είτε από συναλλαγές που υποστηρίζουν κάποιο απόγονο του. Ονομάζουμε cardinality f(X) του X τον αριθμό συναλλαγών που υποστηρίζουν αποκλειστικά το X στο σύνολο δεδομένων (βάση D). Η παρακάτω formula μας δίνει τον τρόπο υπολογισμού της.

$f(X) = |T(X) = CX - \sum f(I')\gamma\iota\alpha X \subseteq I'\epsilon P(I)$

Για παράδειγμα έστω $P(I)=\{A_5,B_4,C_3,AB_3,AC_2,BC_1\}$ υπολογίζουμε το cardinality ως εξής

$f(AB)=C_{AB}-f(0)=3$

$f(AC)=C_{Ac}-f(0)=2$

$f(BC)=C_{BC}-f(0)=1$

$f(A)=C_{A}-f(0)=0$

$f(B)=C_{B}-f(0)=0$

$f(C)=C_{C}-f(0)=0$

Από τα παραπάνω προκύπτει ότι το B δεν υποστηρίζεται αποκλειστικά από καμία συναλλαγή στη βάση. Έτσι μπορούμε να συμπεράνουμε ότι για κάθε $f(X)=n$ μπορούμε να έχουμε τρία διακριτά αποτελέσματα.

είτε $n>0$ άρα το X υποστηρίζετε αποκλειστικά από n συναλλαγές στη βάση.

είτε $n=0$ άρα το X δεν υποστηρίζετε αποκλειστικά από καμία συναλλαγή στη βάση.

είτε $n<0$ οπότε το X θα έπρεπε να έχει μεγαλύτερο C_X στο P(I) και πρέπει να κάνουμε αλλαγές.

5 Παρουσίαση μεθοδολογίας

Σε αυτό το κεφάλαιο παρουσιάζουμε τη μεθοδολογία μας για απόκρυψη συχνών στοιχειοσυνόλων. Ως βάση χρησιμοποιήσαμε τη διαδικασία που περιγράφετε στο 'A new framework of privacy preserving data sharing' [3] με ιδιαίτερη έμφαση στην απόκρυψη συχνών στοιχειοσυνόλων μέσω μείωσης της υποστήριξης τους στην τελική βάση.

5.1 Αλγόριθμος Μείωσης

Ξεκινάμε την παρουσίαση της μεθοδολογίας μας με τον αλγόριθμο για τη μείωση των counts (C_i) στο P(I) των συχνών στοιχειοσυνόλων που θέλουμε να κρύψουμε (H).

Στον αλγόριθμο που προτείνουμε πρώτα σχηματίζουμε το τελικό σύνολο H που αποτελείται από όσα συχνά στοιχειοσύνολα μας δόθηκαν για κρύψιμο και τυχόν συχνούς απογόνους τους. Στη συνέχεια ξεκινώντας από τα μεγαλύτερου επιπέδου στοιχειοσύνολα μειώνουμε το C_i στο P(I) τους, τόσο ώστε να μην είναι συχνά ενώ συγχρόνως μειώνουμε και τα C_i όλων των προγόνων τους. Αν λόγω αυτών των μειώσεων κάποια συχνά στοιχειοσύνολα πέσουν κάτω από το όριο του $\min C_i$ τα επαναφέρουμε ώστε να διατηρηθούν.

Algorithm 5.1 Αλγόριθμος Μείωσης

1.input:H,P(I), FI
2.Sort H by highest length
3.for each i in H do{
4.reduceby= C_i- (support-1)
5.reduce the C of i and all of its ancestors by reduceby
6.}
7.for each i ϵFI
8.if(P(I)<$\min C_i$)
9.P(I)=$\min C_i$
10.return P(I)

5.2 Cardinality Check

Στη συνέχεια παρουσιάζουμε τον αλγόριθμο για τον έλεγχο του cardinality. Μετά την εφαρμογή του αλγορίθμου μείωσης δίνουμε το αναθεωρημένο P(I) στον αλγόριθμο

cardinality_check. Ξεκινώντας από τα στοιχειοσύνολα υψηλότερου επιπέδου ο αλγόριθμος υπολογίζει τα cardinality κάθε στοιχειοσυνόλου I σύμφωνα με την εξίσωση $f(X) = |T(X) = CX - \sum f(I')\gamma\iota\alpha X \subseteq I'\epsilon P(I)$ και τα τοποθετεί στον πίνακα C_I(I). Αν προκύψει αρνητική τιμή ο έλεγχος αποτυγχάνει και πρέπει να προσαρμόσουμε αναλόγως το P(I) μας.

Algorithm 5.2 Cardinality_Check Algorithm

```
1.input:C_I(I),pass=1
2.for each i in C_I do{
3.calculate cardinality of i
4.if(C_I(i)<0){pass=0}
5.}
6.return pass.
```

5.3 Construct Database

Αν το cardinality_check είναι επιτυχές τότε χρησιμοποιούμε τον πίνακα C_I(I) που προκύπτει για να κατασκευάσουμε τη νέα βάση D'.

Algorithm 5.3 Αλγόριθμος Κατασκευής Νέας Βάσης

```
1.input:C_I(I)
2.for each i in C_I do
3.add C_I(i) transactions of i to D'
7.return D'.
```

5.4 Βασικός αλγόριθμος μεθοδολογίας

Στη συνέχεια παρουσιάζουμε το βασικό αλγόριθμο της μεθοδολογίας μας. Ιδιαίτερη προσοχή πρέπει να δοθεί σε δυο σημεία τα οποία παρουσιάζουμε στη συνέχεια.

Επαναυπολογισμός minC_i

Μετά τον υπολογισμό της νέας βάσης D'(γραμμές 11-20), είναι αναγκαίο να επαναυπολογίσουμε το minC_i επειδή ο αριθμός συναλλαγών της βάσης μας μπορεί να έχει αλλάξει αλλάζοντας τα δεδομένα που χρησιμοποιήσαμε στους υπολογισμούς μας. Υπάρχουν τρεις διακριτές περιπτώσεις. Το newminC_iθα είναι το ίδιο με το παλιό άρα η βάση μας είναι αποδεκτή. Το newminC_i είναι μικρότερο $\text{newminC}_i < \text{minC}_i$ οπότε πρέπει να αυξήσουμε τον αριθμό συναλλαγών της βάσης μας ώστε να αποφύγουμε να κάνουμε συχνά στοιχειοσύνολα που δεν πρέπει να είναι προσθέτοντας συναλλαγές

συχνών στοιχειοσυνόλων που δε θα δημιουργήσουν side effects. Τέλος το newminC_i είναι μεγαλύτερο από το minC_i οπότε η μεθοδολογία μας αποτυγχάνει αφού πλέον η λύση που έχουμε βρει δεν εξασφαλίζει ότι τα συχνά μας στοιχειοσύνολα έχουν διατηρηθεί συχνά στη νέα βάση κάτι που μπορεί να έχει επιπτώσεις στη διαδικασία εξόρυξης γνώσης.

Αποτυχία Cardinality Check

Το επόμενο σημείο ενδιαφέροντος είναι τι συμβαίνει αν αποτύχει το cardinality check. Σε αυτή τη περίπτωση, ξεκινάμε διορθώνοντας τη τιμή στο P(I) όσων FI έχουν αρνητική τιμή στο cardinality array C_I(I). Στη συνέχεια προσπαθούμε με τη προσθήκη maxlvlitemset συναλλαγών να εξαλείψουμε τις αρνητικές τιμές και στα μη συχνά μας. Για να γίνει καλύτερα κατανοητό το πως επιτυγχάνετε αυτό θα παρουσιάσουμε ένα παράδειγμα. Έστω το

P(I)={A_{10},B_{11},C_{10},D_8,AB_7,AC_6,AD_6,BC_7,$BD_{6,}$,CD_4,ABC_4,ABD_4} και minC_i=4. Μετά το cardinality check to C_I(I) μας θα είναι:

C_I(I)={A_{-3},B_{-3},C_{-3},D_{-4},AB_{-1},AC_2,AD_2,BC_3,$BD_{2,}$,CD_4,ABC_4,ABD_4}

Αν προσθέσουμε 3(ώστε να μην γίνει συχνό αυτό ή οι πρόγονοι του) ABCD (maxlvlitemset) στο P(I) θα έχουμε το ακόλουθο P(I).

P(I)={A_{10},B_{11},C_{10},D_8,AB_7,AC_6,AD_6,BC_7,$BD_{6,}$,CD_4,ABC_4,ABD_4,ACD_3,BCD_3,$ABCD_3$}. το νέο cardinality check θα είναι

C_I(I)={A_{-1},B_0,C_0,D_{-1},AB_2,AC_2,AD_2,BC_3,$BD_{2,}$,CD_1,ABC_1,ABD_1,ACD_0,BCD_0,$ABCD_3$}, οπότε απλά αυξάνουμε και τη τιμή των A και D κατά 1 στο P(I) και περνάμε το cardinality check.

Algorithm 5.4 Βασικός Αλγόριθμος

1.input:D,P(I),minC_i,H,FI,IF,s=minC_i/|D|
2.H=H+descendants
3.FI=FI-H
4.reduction(P(I))
5.For each i in F{
6.if P(i)<minC_i
7.P(i)=minC_i}
8.while TRUE do{
9.C_I(I)=P(I),pass=cardinality_check(C_I(I))
10.if(pass=1){
11. D'=construct_database(C_I(I))
12. newminC_i=s*|D'|
13. if(newminC_i<=minC_i){
14. if(newminC_i<minC_i){
15. add |D|-|D'| FI transactions
16. }
17.output D'
18. }
19. else FAIL
20.}
21.else{#cardinality check failed
22.starting with max rank itemsets
23.for each i in C_I(i){
24. if (C_I(i)<0){
25. if(i∈FI){
26. P(i)=P(i)-(C_I(i))
27. }
28. else{
29.find max_neg=the max negative value of the IF in C_I(I)}
30.if(max_neg!=0){
30.add (-max_neg) or (minC_i-1) of maxlvlitemset into P(I)
31.#and adjust every itemset in P(I) accordingly
32.for each i in P(I){
33.if(P(i)<P(maxlvlitemset)
34.P(i)=P(maxlvlitemset)
35. }
36.}
37.}
38.}

6 Παραδείγματα

Σε αυτό το κεφάλαιο παρουσιάζουμε κάποια αντιπροσωπευτικά παραδείγματα εκτέλεσης της μεθοδολογίας μας.

6.1 Βασικό παράδειγμα εκτέλεσης

Μας δίνεται η παρακάτω βάση

Table 6.1: Example 1 Database

Database				
TR	A	B	C	D
T1	1	0	1	1
T2	0	0	1	0
T3	1	0	1	1
T4	1	1	0	0
T5	0	0	1	1
T6	0	0	1	1
T7	0	1	1	0
T8	1	0	0	1
T9	0	1	1	1
T10	1	0	0	0

Από την οποία λαμβάνουμε το ακόλουθο P(I).

Table 6.2: Example 1 lattice before reduction

lattice
$A^5B^3C^7D^6$
$AB^1AC^2AD^3BC^2BD^1CD^5$
ACD^2BCD^1

Παραδείγματα

Για $\min C_i=2$, $s=\min C_i/|D|=2/10=0.2$ και H={AD}. Ξεκινάμε προσθέτοντας στο H τους συχνούς απογόνους. H={AD,ACD}. Προχωράμε στη μείωση του P(I) για τα H και τους προγόνους τους.

Αρχικά για το ACD 2->1,AC 2->1,AD 3->2,CD 5->4,A 5->4,C 7->6,D 6->5. Ομοίως για το AD 2->1,A 4->3,D 5->4.

Το τελικό μας P(I) μετά την μείωση θα είναι.

Table 6.3: Example 1 lattice after reduction

lattice
$A^3B^3C^6D^4$
$AB^1AC^2AD^1BC^2BD^1CD^4$
ACD^1BCD^1

Ιδιαιτερη προσοχή πρέπει να δοθεί σε όσα από τα συχνά στοιχειοσύνολα θέλουμε να κρατήσουμε και έχουν μειωθεί κάτω από το όριο $\min C_i$, όπως εδώ το AC. Σε αυτή τη περίπτωση θα τα αυξήσουμε ώστε να είναι τουλάχιστον ίσα για να πραμείνουν συχνά.

Στη συνέχεια κάνουμε το cardinality_check μας. Ο πίνακας C_I(I) που παράγεται είναι:

Table 6.4: Cardinality Example 1

lattice
$A^0B^0C^0D^0$
$AB^1AC^1AD^0BC^1BD^0CD^2$
ACD^1BCD^1

Δεν υπάρχουν αρνητικές τιμές άρα είναι αποδεκτό και προχωράμε στην κατασκευή της τελικής βάσης σύμφωνα με το C_I.

Table 6.5: Example 1 Final Database

Database				
TR	A	B	C	D
T1	1	0	1	1
T2	0	1	1	1
T3	1	1	0	0
T4	1	0	1	0
T5	0	0	1	1
T6	0	0	1	1
T7	0	1	1	0

Ελέγχουμε το newminC_i=s*|D'|=0.2*7=1.4=2 (δε μπορούμε να έχουμε μη ακέραιο αριθμό συναλλαγών) άρα αποδεκτή βάση.

6.2 Παράδειγμα failed cardinality

Για το ίδιο P(I).

Table 6.6: Example 2 lattice before reduction

lattice
$A^5B^3C^7D^6$
$AB^1AC^2AD^3BC^2BD^1CD^5$
ACD^2BCD^1

Για minC_i=2,s=minC_i/|D|=2/10=0.2 και H={A}. Έχουμε H={A,AC,AD,ACD} . Μετά τη μείωση το P(I) θα είναι.

Table 6.7: Example 2 lattice after reduction

lattice
$A^1B^3C^6D^4$
$AB^1AC^1AD^1BC^2BD^1CD^4$
ACD^1BCD^1

Ο πίνακας που θα προκύψει από το cardinality check είναι:

Table 6.8: Cardinality array example 2

lattice
$A^{-1}B^0C^1D^0$
$AB^1AC^0AD^0BC^1BD^0CD^2$
ACD^1BCD^1

Υπάρχει αρνητική τιμή άρα πρέπει να διορθώσουμε το P(I) μας. Αυτό το πετυχαίνουμε προσθέτωντας το maxlvlitemset στο P(I) (με τιμή $\min C_i$-1=1) το νέο μας P(I) γίνεται

Table 6.9: Example 2 revised lattice

lattice
$A^1B^3C^6D^4$
$AB^1AC^1AD^1BC^2BD^1CD^4$
$ABC^1ABD^1ACD^1BCD^1$
$ABCD^1$

Το νέο cardinality hash γίνεται

Table 6.10: New cardinality array example 2

lattice
$A^0B^1C^1D^0$
$AB^0AC^0AD^0BC^1BD^0CD^3$
$ABC^0ABD^0ACD^0BCD^0$
$ABCD^1$

Καμιά αρνητική τιμή άρα προχωράμε με τη κατασκευή της νέας βάσης D'

Table 6.11: Example 2 Final Database

Database				
TR	A	B	C	D
T1	1	1	1	1
T2	0	1	0	0
T3	0	1	1	0
T4	0	0	1	0
T5	0	0	1	1
T6	0	0	1	1
T7	0	0	1	1

Ελέγχουμε το newminC_i=s*|D'|=0.2*7=1.4=2 (δε μπορούμε να έχουμε μη ακέραιο αριθμό συνναλαγών) άρα αποδεκτή βάση.

6.3 Παράδειγμα αποτυχίας

Μας δίνεται η παρακάτω βάση

Table 6.12: Example 3 Database

Database					
TR	A	B	C	D	E
T1	1	1	1	0	1
T2	1	1	0	0	1
T3	1	1	0	1	0
T4	1	1	0	1	0
T5	1	1	1	0	0
T6	1	1	1	0	0
T7	0	0	1	1	0
T8	0	0	1	1	0

Από την οποία λαμβάνουμε το ακόλουθο P(I).

Table 6.13: Example 3 lattice

lattice
$A^6B^6C^5D^4E^2$
$AB^6AC^3AD^2AE^2BC^3BD^2BE^2CD^2CE^1$
$ABC^3ABD^2ABE^2ACE^1BCE^1$
$ABCE^1$

Για $\min C_i$=3,s=$\min C_i/|D|$=3/8=0.375 και H={ABC} . Μετά τη μείωση το P(I) θα είναι.

Table 6.14: Example 3 lattice after reduction

lattice
$A^5B^5C^4D^4E^2$
$AB^5AC^3AD^2AE^2BC^3BD^2BE^2CD^2CE^1$
$ABC^2ABD^2ABE^2ACE^1BCE^1$
$ABCE^1$

Ο πίνακας που θα προκύψει από το cardinality check είναι:

Table 6.15: Cardinality array example 3

lattice
$A^{-1}B^{-1}C^{-2}D^0E^0$
$AB^0AC^1AD^0AE^0BC^1BD^0BE^0CD^2CE^0$
$ABC^1ABD^2ABE^1ACE^0BCE^0$
$ABCE^1$

Υπάρχει αρνητική τιμή άρα πρέπει να διορθώσουμε το P(I) μας. Αυτο το πετυχαίνουμε αυξάνοντας στο P(I) τις τιμές των A,B,C (στο 6 επειδή είναι συχνά δεν μας ενοχλεί το να τα αυξήσουμε)το νέο μας P(I) γίνεται

Table 6.16: Example 3 revised lattice

lattice
$A^6B^6C^6D^4E^2$
$AB^5AC^3AD^2AE^2BC^3BD^2BE^2CD^2CE^1$
$ABC^2ABD^2ABE^2ACE^1BCE^1$
$ABCE^1$

Το νέο cardinality hash γίνεται

Table 6.17: new cardinality array example 3

lattice
$A^0B^0C^0D^0E^0$
$AB^0AC^1AD^0AE^0BC^1BD^0BE^0CD^2CE^0$
$ABC^1ABD^2ABE^1ACE^0BCE^0$
$ABCE^1$

Καμιά αρνητική τιμή άρα προχωράμε με τη κατασκευή της νέας βάσης D'

Table 6.18: Example 3 new Database

Database					
TR	A	B	C	D	E
T1	1	1	1	0	1
T2	1	1	0	0	1
T3	1	1	0	1	0
T4	1	1	0	1	0
T5	1	1	1	0	0
T6	1	0	1	0	0
T7	0	1	1	0	0
T8	0	0	1	1	0
T9	0	0	1	1	0

Ελέγχουμε το newminC_i=s*|D'|=0.375*9=3.375=4 (δε μπορούμε να έχουμε μη ακέραιο αριθμό συναλλαγών) άρα μη αποδεκτή βάση αφού για να είναι ένα στιχειοσύνολο συχνό σε αυτή τη βάση πρέπει να έχει newminC_i=4.

7 Σύνοψη-Συμπεράσματα

Σε αυτή την εργασία αναλύσαμε το πρόβλημα απόκρυψης συχνών στοιχειοσυνόλων. Προσπαθήσαμε να δώσουμε τη δική μας εκδοχή στην επίλυση του προβλήματος με τη βοήθεια ενός αλγόριθμου ανακατασκευής του συνόλου δεδομένων αποτελούμενη από τέσσερα μέρη. Δείξαμε ότι η μεθοδολογία μας μπορεί να κρύψει τα συχνά στοιχειοσύνολα χωρίς πρόκληση οποιωνδήποτε παρενεργειών και να συντηρήσει μια υψηλή ποιότητα της συνολικής βάσης δεδομένων με σχετικά μικρές αλλοιώσεις στην υποστήριξη των στοιχειοσυνόλων. Παρόλα αυτά η μεθοδολογία μας δεν επιτυγχάνει να δώσει εγγυημένη λύση, ενώ αρχικές μετρήσεις σε βάσεις δεδομένων με πραγματικά στοιχεία έδειξαν ιδιαίτερα μεγάλους χρόνους εκτέλεσης για χαμηλές υποστηρίξεις, κάτι που οφείλεται στην αναγκαιότητα της γνώσης του P(I). Από τα παραπάνω συμπεραίνουμε ότι η συγγεκριμένη μεθοδολογία θα ήταν καλύτερο να συνδοιαστεί με κάποιον προσθετικό αλγόριθμο απόκρυψης για να δημιουργήσει μία hybrit μεθοδολογία.

Bibliography

[1] Osman Abul. Hiding co-occurring frequent itemsets. In *EDBT/ICDT '09 Proceedings of the 2009 EDBT/ICDT Workshops*, pages 117–125, 2009.

[2] M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios. Disclosure limitation of sensitive rules. In *Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange*, pages 45–52. IEEE Computer Society, 1999.

[3] X. Chen, M. Orlowska, and X. Li. A new framework of privacy preserving data sharing. In *Proceedings of the 4th IEEE International Workshop on Privacy and Security Aspects of Data Mining*, pages 47–56, 2004.

[4] Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid, and Elisa Bertino. Hiding association rules by using confidence and support. In *Proceedings of the 4th International Workshop on Information Hiding*, pages 369–383, 2001.

[5] Aris Gkoulalas-Divanis and Vassilios S. Verykios. An integer programming approach for frequent itemset hiding. In *Proceedings of the 2006 ACM Conference on Information and Knowledge Management (CIKM 2006)*, pages 748 – 757, 2006.

[6] Aris Gkoulalas-Divanis and Vassilios S. Verykios. A hybrid approach to frequent itemset hiding. In *ICTAI '07 Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 01*, pages 297–304, 2007.

[7] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1026–1037, 2004.

[8] S. Menon, S. Sarkar, and S. Mukherjee. Maximizing accuracy of shared databases when concealing sensitive patterns. *Information Systems Research*, 16(3):256–270, 2005.

[9] G. V. Moustakides and Vassilios S.Verykios. A maxmin approach for hiding frequent itemsets. *Data and Knowledge Engineering*, 65(1):75–89, 2008.

[10] S. R. M. Oliveira and O. R. Zaïane. Protecting sensitive knowledge by data sanitization. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003)*, pages 211–218, 2003.

[11] E. D. Pontikakis, A. A. Tsitsonis, and V. S. Verykios. An experimental study of distortion-based techniques for association rule hiding. In *Proceedings of the 18th Conference on Database Security (DBSEC 2004)*, pages 325–339, 2004.

[12] E. D. Pontikakis, V. S. Verykios, and Y. Theodoridis. On the comparison of association rule hiding heuristics. In *Hellenic Database Management Symposium*. 2004.

[13] Y. Saygin, V. S. Verykios, and C. Clifton. Using unknowns to prevent discovery of association rules. *ACM SIGMOD Record*, 30(4):45–54, 2001.

[14] X. Sun and P. S. Yu. A border-based approach for hiding sensitive frequent itemsets. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM 2005)*, pages 426–433, 2005.

[15] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni. Association rule hiding. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):434–447, 2004.

[16] Zutao Zhu and Wenliang Du. K-anonymous association rule hiding. In *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*, pages 305–309, 2010.