ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

Ελευθέριος Κουμάκης

Ηράκλειο, Κρήτη
Ιανουάριος 2004

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

HealthObs: Health Observatory

Ένα Ολοκληρωμένο Σύστημα Εξόρυξης Δεδομένων και Ανακάλυψης Γνώσης από Κατανεμημένες και Ετερογενείς Κλινικές Πηγές Δεδομένων

Εργασία που υποβλήθηκε από τον
Ελευθέριο Κουμάκη
ως μερική εκπλήρωση των απαιτήσεων
για την απόκτηση
ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

Συγγραφέας:

Ελευθέριος Κουμάκης
Τμήμα Επιστήμης Υπολογιστών

Εισηγητική Επιτροπή:

Στέλιος Ορφανουδάκης
Καθηγητής, Επιβλέπων

Δημήτρης Πλεξουσάκης
Αναπληρωτής Καθηγητής, Μέλος

Βασίλης Χρηστοφίδης
Επίκουρος Καθηγητής, Μέλος

Γεώργος Ποταμιάς
Ερευνητής Ι.Π. Ι.Τ.Ε., Επόπτης

Δεκτή:

Δημήτρης Πλεξουσάκης
Αναπληρωτής Καθηγητής,
Πρόεδρος Επιτροπής Μεταπτυχιακών Σπουδών

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Η εξέλιξη του παγκόσμιου ιστού έχει μετασχηματίσει το Διαδίκτυο σε ένα σταθμό ανταλλαγής πληροφοριών. Η πληροφορία αυτή χαρακτηρίζεται από την *ετερογένεια* της και την *κατανεμημένη* φύση της.

Το πρόβλημα εύρεσης τρόπων και μεθοδολογιών διασύνδεσης κατανεμημένων και ετερογενών πηγών πληροφοριών και δεδομένων γίνεται ολοένα και περισσότερο κρίσιμο, ιδιαίτερα σε πεδία όπου ο όγκος των συλλεγομένων πληροφοριών και δεδομένων αυξάνεται με ραγδαίο ρυθμό. Ένα τέτοιο πεδίο είναι ο τομέας της *υγείας*. Οι απαιτήσεις στον τομέα αυτό αυξάνουν την ανάγκη για την ολοκλήρωση στοιχείων από τις κατανεμημένες και ετερογενείς πηγές πληροφοριών με έναν τρόπο που: η άμεση πρόσβαση στα συνοπτικά στοιχεία, και η αξιόπιστη και εμπεριστατωμένη υποστήριξη των διαδικασιών κλινικής λήψης αποφάσεων (evidence-based clinical decision making) μπορεί να βοηθήσει τους ειδικούς να προσανατολιστούν στο χάος των συλλεγομένων πληροφοριών και να εκμεταλλευτούν τις πιο σχετικές από αυτές.

Η πρόσβαση σε πληροφορίες και η συλλογή δεδομένων δεν είναι αυτοσκοπός. Αυτό που είναι επιθυμητό είναι η αξιοποίησή της, άρα η δυνατότητα για εξαγωγή χρήσιμων και κατανοητών συμπερασμάτων. Εστιάζοντας στον τομέα της ιατρικής και με υπόβαθρο το HygeiaNet - The Integrated Health Care Network of Crete (www.hygeianet.gr), προσπαθούμε να δώσουμε ώθηση προς μια νέα κατεύθυνση στη διαχείριση κλινικών δεδομένων από τον ολοκληρωμένο ηλεκτρονικό φάκελο υγείας πολιτών (Integrated Electronic Health Care Record - IEHCR). Από μια ασθενο-κεντρική (patient-oriented) όψη του φακέλου να περάσουμε σε μια *πληθυσμο-κεντρική* (population-oriented) όψη. Κάτι τέτοιο θα προσφέρει τη δυνατότητα *επιδημιολογικών μελετών μέσω του Διαδικτύου* καθώς και την εκμετάλλευση των σχετικών αποτελεσμάτων στη καθημερινή κλινική πρακτική.

Η βασική πρόκληση είναι η λειτουργική ενσωμάτωση μηχανισμών *εξόρυξης γνώσης (data mining for knowledge discovery)* σε ένα κατανεμημένο και ετερογενές περιβάλλον δεδομένων. Για το λόγο αυτό, προτείνεται και υλοποιείται μια πολυσύνθετη διαδικασία ολοκλήρωσης η οποία αντιμετωπίζει θέματα όπως: α) σημασιολογική ομογενοποίηση και ολοκλήρωση ετερογενών δεδομένων, β) ανάπτυξη και προσαρμογή τεχνικών εξόρυξης δεδομένων, γ) παρουσίαση των αποτελεσμάτων μέσω εργαλείων απεικονιστικής (visualization), και δ) υλοποίηση λειτουργικού και φιλικού διάμεσου επικοινωνίας ανθρώπου-μηχανής (human computer interface – HCI).

Το τελικό προϊόν είναι ένα ευέλικτο και λειτουργικό περιβάλλον τυποποίησης και νοήμονης επεξεργασίας κλινικών δεδομένων, το σύστημα **HealthObs**.

ABSTRACT

The evolution of the World Wide Web has transformed Internet into a context capable for information sharing among its users. The information in the various resources over the web is characterized by its *heterogeneity* and its *distributed* nature.

The problem of finding ways and methodologies for interconnection over distributed and heterogeneous sources of information and data have become critical, particularly in fields where the volume of the collected information and data is rapidly increasing. An area that has experienced radical change, and continues to do so, is health-care delivery. The requirements in this field increase the need for integration over the distributed and heterogeneous clinical data sources in a way that: the direct access in data, and the reliable support of clinical decision-making processes (evidence-based clinical decision making) may help experts to orient themselves in the information space and enhance their decision-making capabilities.

The access at information and the collection of data are not end in itself. What is desirable is the exploitation of data, hence the possibility for exporting useful and comprehensible conclusions. Focusing in the field of health, and within the framework of HygeiaNet -The Integrated Health Care Network of Crete (www.hygeianet.gr), we try to give impulse to a new direction of clinical data management, as retrieved from the Integrated Electronic Health Care Record (IEHCR). From patient-oriented view we pass into a *population-oriented* view. This shift will offer the potential of Internet-based epidemiology and the exploitation of the relative results in the daily clinical practice.

The main challenge is the operational incorporation of *data mining* mechanisms for the shake of knowledge discovery from distributed and heterogeneous information sources. Towards this end, a multi-layer data integration and processing approach is proposed that deals with: a) the smooth and seamless semantic homogenisation of heterogeneous data, b) the development and adaptation of knowledge discovery operations, c) the presentation of results via visualization techniques and tools, and d) development of operational and user friendly interface (human computer interface - HCI). The final product is a flexible and functional environment for the standardisation and the intelligent mining of clinical data, the **HealthObs** system.

1. ΕΙΣΑΓΩΓΗ

Η ραγδαία εξάπλωση του παγκοσμίου ιστού ενώνει εκατομμύρια χρήστες σε όλη τη γη διαπερνώντας γεωγραφικά, πολιτισμικά και επιστημονικά όρια. Ένα πεδίο που έχει γίνει δέκτης αυτής της αλλαγής είναι ο τομέας της υγείας. Επιστήμονες από τον τομέα της ιατρικής μπορούν πλέον να δουλέψουν σε ένα *κατανεμημένο* πληροφοριακό περιβάλλον, έχοντας τη δυνατότητα να προσπελαύνουν ταυτόχρονα εκατοντάδες υπηρεσίες και αφθονία πληροφορίας.

Το βασικό μειονέκτημα της ραγδαίας αυτής εξέλιξης είναι ότι οι επαγγελματίες υγείας μπορούν εύκολα να αποπροσανατολιστούν στο τεράστιο όγκο πληροφοριών και δεδομένων. Το πρόβλημα εύρεσης τρόπων και μεθοδολογιών διασύνδεσης κατανεμημένων και *ετερογενών* πηγών πληροφοριών και δεδομένων γίνεται ολοένα και περισσότερο κρίσιμο.

Με βάση πρόσφατη μελέτη στις Ηνωμένες Πολιτείες [59], περίπου 44.000-98.000 θάνατοι οφείλονται σε ιατρικά λάθη τα οποία θα μπορούσαν να έχουν αποφευχθεί - ακόμη και με τις πιο αισιόδοξες εκτιμήσεις περισσότεροι άνθρωποι πεθαίνουν από ιατρικά λάθη παρά από ατυχήματα, καρκίνο ή AIDS!!. Νομίζουμε ότι ένας βασικός λόγος για αυτή τη κατάσταση δεν είναι η ανεπάρκεια του ιατρικού προσωπικού αλλά, η αδυναμία των ιατρών να αντλήσουν και να επεξεργαστούν τη πιο ενημερωμένη, εμπεριστατωμένη (evidential) και χρήσιμη πληροφορία από τις διαθέσιμες πηγές, π.χ., ιατρικές οδηγίες (guidelines) και πρωτόκολλα. Έτσι, πέρα από τη πρόσβαση και άντληση τεραστίων μεγεθών πληροφορίας και δεδομένων, είναι πολύ σημαντικό να μελετηθεί και να αντιμετωπιστεί η γενική ανάγκη *ολοκλήρωσης* των δεδομένων αυτών και η παροχή μηχανισμών και λειτουργιών *ανακάλυψης 'χρήσιμης' γνώσης* από τις αντίστοιχες πηγές. Οι σχετικές υπηρεσίες θα ενισχύσουν και θα υποστηρίξουν τις διαδικασίες *εμπεριστατωμένης κλινικής λήψης αποφάσεων* (evidence-based clinical decision making).

1.1 HYGEIAnet και το Σύστημα HealthObs

Εστιάζοντας στον τομέα της ιατρικής και με αντιπροσωπευτικό μοντέλο το HygeiaNet (The Integrated Health Care Network of Crete; `www.hygeianet.gr`) του Κέντρου Ιατρικής Πληροφορικής και Τηλεματικών Εφαρμογών στην Υγεία του ΙΠ-ΙΤΕ, προσπαθούμε να δώσουμε ώθηση προς μια νέα κατεύθυνση στη διαχείριση κλινικών δεδομένων από το ολοκληρωμένο ηλεκτρονικό φάκελο υγείας πολιτών (Integrated Electronic Health Care Record

- IEHCR). Από μια ασθενο-κεντρική (patient-oriented) όψη του φακέλου να περάσουμε σε μια *πληθυσμο-κεντρική* (population-oriented) όψη.

Επιπλέον, η πρόληψη υγείας και οι επιδημιολογικές μελέτες γίνονται περισσότερο απαιτητικές με τη μεταφορά και ανταλλαγή πληροφοριών. Αυτή η απαίτηση αυξάνει την ανάγκη για την ολοκλήρωση στοιχείων από τις κατανεμημένες και ετερογενείς κλινικές πηγές πληροφοριών με έναν τρόπο που: η άμεση πρόσβαση στα συνοπτικά στοιχεία και η αξιόπιστη υποστήριξη γνώσης (βασισμένη σε στοιχεία) βοηθάει τους ειδικούς να προσανατολιστούν στο χάος της πληροφορίας και να δημιουργήσουν μια γενική άποψη του προβλήματος. Κάτι τέτοιο θα προσφέρει τη δυνατότητα *επιδημιολογικών μελετών μέσω του Διαδικτύου* καθώς και την εκμετάλλευση των σχετικών αποτελεσμάτων στη καθημερινή κλινική πρακτική.

Με το τεχνολογικό υπόβαθρο του IEHCR – όπως προσφέρεται από το HYGEIAnet, το σύστημα **HealthObs** που αναπτύχθηκε προσφέρει:

- ✓ λειτουργίες ***σημασιολογικής ομογενοποίησης*** και ομοιόμορφης αναπαράστασης πληροφοριών με τη χρήση της τεχνολογίας XML,
- ✓ αντικειμενοστρεφή σχήματα δόμησης των δεδομένων και σχετικές λειτουργίες,
- ✓ λειτουργίες ***εξόρυξης δεδομένων*** (data mining) και ***ανακάλυψης γνώσης*** (knowledge discovery) από XML έγγραφα (**XML-MINING**),
- ✓ ένα εύχρηστο και φιλικό ***διάμεσο επικοινωνίας*** χρήστη-υπολογιστή (human-computer interface) για τη διαχείριση των παραμέτρων του προς διερεύνηση προβλήματος και την ***απεικόνιση*** και επίβλεψη των αποτελεσμάτων (visualization).
- ✓ Το σύστημα HealthObs μπορεί πολύ εύκολα να ***προσαρμοστεί*** σε οποιοδήποτε πεδίο εφαρμογής εκτός της ιατρικής, μέσω ευέλικτων δομών και μορφών αναπαράστασης δεδομένων (σχετικά πειραματικά αποτελέσματα εμπεριέχονται στην εργασία) καθώς και σχετικών εργαλείων που έχουν αναπτυχθεί (**DOMAIN EDITOR**)

1.2 Αντικείμενο και Στόχοι της Εργασίας

Η παρούσα μεταπτυχιακή εργασία εστιάζει στο πρόβλημα της ανακάλυψης και εξαγωγής γνώσης από κατανεμημένες και ετερογενείς πηγές πληροφοριών. Η βασική πρόκληση είναι πως λειτουργίες εξόρυξης γνώσης γίνονται λειτουργικές σε ένα κατανεμημένο και ετερογενές περιβάλλον πληροφοριών και δεδομένων. Για το λόγο αυτό, προτείνεται και υλοποιείται μια πολυσύνθετη διαδικασία ολοκλήρωσης η οποία αντιμετωπίζει θέματα όπως:

- αξιόπιστη ομογενοποίηση και ολοκλήρωση των ετερογενών δεδομένων,
- επεξεργασία των δεδομένων - statistical analysis, εξόρυξη δεδομένων, κλπ.,
- παρουσίαση των αποτελεσμάτων,
- υλοποίηση ενός συστήματος φιλικό προς το χρήστη.

Βασική συνεισφορά της εργασίας είναι η χρήση, προσαρμογή, διασύνδεση και συνεργασία μεθόδων εξόρυξης δεδομένων και τεχνολογιών XML. Με την ραγδαία εξέλιξη του παγκόσμιου ιστού και την διαρκή αύξηση των χρηστών του, δημιουργήθηκε μια νέα μορφή *ομοιόμορφης αναπαράστασης δεδομένων*, η XML, η οποία είναι εύκολα προσβάσιμη από οποιαδήποτε πλατφόρμα και μεταφέρσιμη μέσω του δια-δικτύου.

Η βασική παραδοχή της εργασίας είναι ότι:

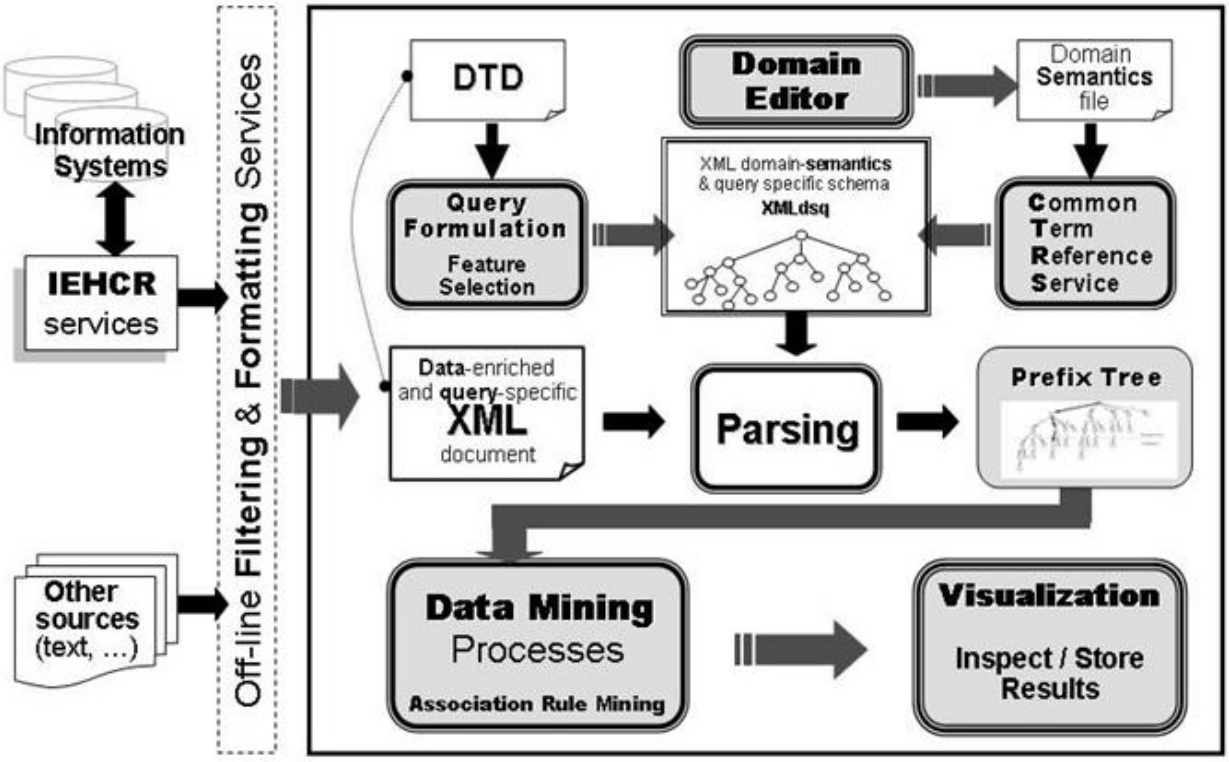
με την χρήση της XML και την υλοποίηση αλγορίθμων εξόρυξης δεδομένων και μηχανικής μάθησης (machine learning) επί XML έγγραφων μπορεί να αντιμετωπιστεί με επιτυχία το πρόβλημα της ανακάλυψης γνώσης από κατανεμημένες και ετερογενείς πηγές δεδομένων.

Η ολοκλήρωση του συστήματος επιτυγχάνεται με την υλοποίηση λειτουργικών διεπαφών συστημάτων καθώς και φιλικών προς το χρήστη διεπαφών οι οποίες κρύβουν την πολυπλοκότητα του όλου συστήματος. Έτσι ο χρήστης μπορεί να ασχοληθεί με τη διερεύνηση μεγάλου όγκου δεδομένων χωρίς να απασχολείται με τη γεωγραφική κατανομή και τη δυνητικά σημασιολογική ετερογένεια τους.

1.3 Αρχιτεκτονική του συστήματος

Στην εργασία αυτή παρουσιάζουμε την μεθοδολογία, και όλες τις διεργασίες που πρέπει να ακολουθηθούν, προκειμένου να πραγματοποιηθούν αυτές οι λειτουργίες. Στο σχήμα 1 παρουσιάζεται η αρχιτεκτονική αναφοράς του συστήματος. Έχοντας ως υπόβαθρο διάφορες πηγές πληροφοριών, εξάγεται ένα XML αρχείο με το αντίστοιχο DTD (αναπαράσταση της δομής ενός XML) του. Το DTD αρχείο επεξεργάζεται και με μια κατάλληλη διεπαφή ο χρήστης μπορεί να επιλέξει τα στοιχεία του XML που θέλει να συμμετέχουν στην εξόρυξη γνώσης. Παράλληλα ένα άλλο αρχείο, το αρχείο «κοινής ορολογίας» (domain semantics) δίνει τη δυνατότητα για σημασιολογική ομογενοποίηση των δεδομένων. Για την υλοποίηση του αρχείου αυτού έχει δημιουργηθεί ένα ειδικό εργαλείο (μέρος της συνολικής εργασίας) το οποίο επιτρέπει σε χρήστες διαφορετικού επιπέδου να διαμορφώσουν σημασιολογικά τα δεδομένα αλλά και να χαρτογραφήσουν αριθμητικές τιμές σε ποιοτικά σταθμά. Έχοντας τα επιλεγμένα από το χρήστη στοιχεία και τη σημασιολογία του πεδίου δημιουργείται το «πρωτότυπο σχήμα» (XMLdsq) των δεδομένων. Το σχήμα αυτό λειτουργεί σαν φίλτρο κατά την επεξεργασία του XML αρχείου. Με αυτό τον τρόπο αγνοείται η περιττή πληροφορία και δίδεται η αναγκαία σημασιολογία. Τα δεδομένα αποθηκεύονται σε μία δενδρική δομή, το prefix-

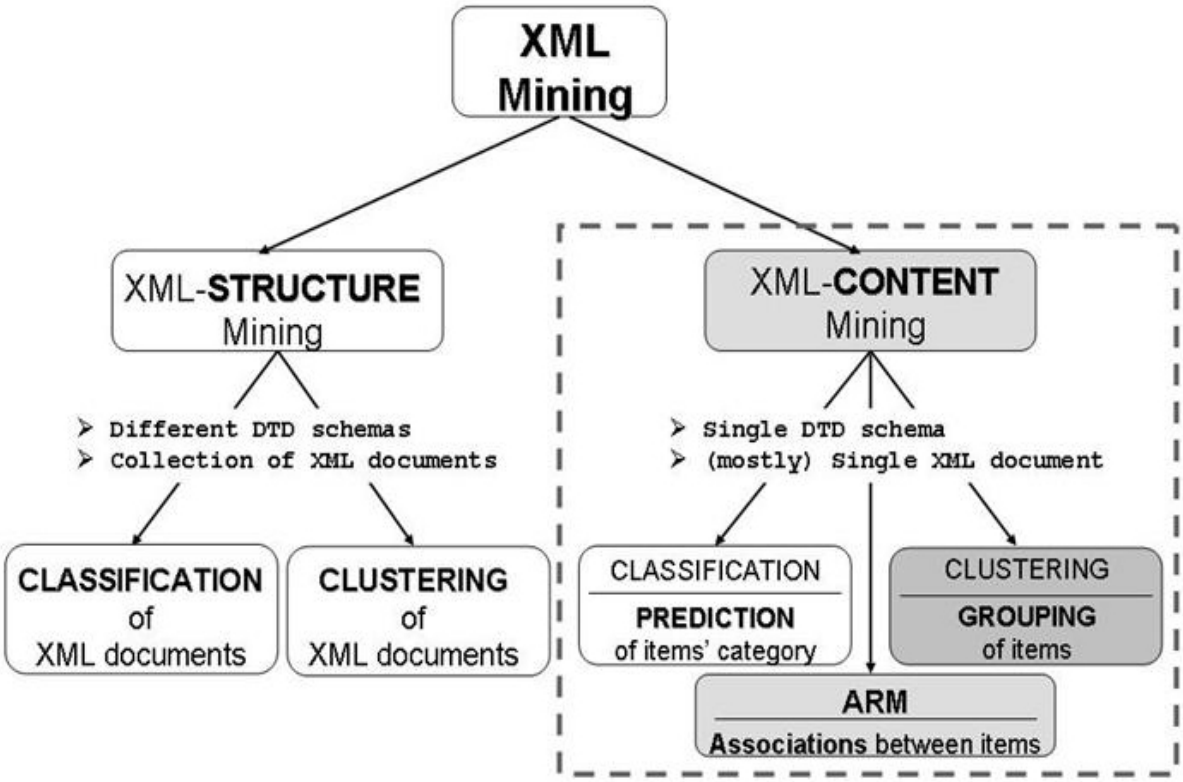
tree. Έπειτα με κατάλληλες διεπαφές ο χρήστης επιλεγεί τον αλγόριθμο που θέλει να εφαρμοστεί στα δεδομένα και ορίζει τις παραμέτρους του. Τέλος τα αποτελέσματα απεικονίζονται γραφικά και ο χρήστης μπορεί να αποθηκεύσει τα αποτελέσματα σε διάφορες μορφές.



Σχήμα 1 : Αρχιτεκτονική αναφοράς HealthOb

1.4 Εξόρυξη γνώσης και ημι-δομημένα δεδομένα

Αν και υπάρχουν πολλές αναφορές και δημοσιεύεις για υλοποίηση εξαγωγής γνώσεων σε βάσεις δεδομένων, λίγη έρευνα έχει γίνει για δεδομένα σε XML μορφή. Οι περισσότερες έρευνες που συνδέουν την εξόρυξη γνώσης με τα ημι-δομημένα δεδομένα εστιάζονται στο πρόβλημα την εξαγωγής αντιπροσωπευτικού σχήματος δομής των δεδομένων. Ελάχιστες, είναι οι έρευνες που συνδέονται με την πληροφορία του περιέχουν τα δεδομένα ημι-δομημένων αρχείων. Η εξόρυξη γνώσης , όπως φαίνεται και στο σχήμα 2, μπορεί να κατηγοριοποιηθεί σε δύο κύριες κατηγορίες. Την εξόρυξη σχήματος από διαφορετικά XML αρχεία και την εξόρυξη γνώσης από τα δεδομένα που περιέχουν τα XML αρχεία. Η παρούσα εργασία εστιάζεται στην δεύτερη κατηγορία υλοποιώντας ένα αλγόριθμο «κανόνων συσχέτισης» (association Rules) και δύο αλγόριθμους «συνάθροισης» (Clustering).



Σχήμα 2: Εξόρυξη γνώσης σε ημι-δομημένα δεδομένα

1.5 Δομή και Περιεχόμενο της Εργασίας

Η παρούσα παρουσίαση της εργασίας είναι δομημένη ως εξής:

- Το Κεφάλαιο 2 (*Τεχνολογίες Υποβάθρου του HealthObs*) επιχειρεί μια περιγραφή του δικτύου τηλεματικών υπηρεσιών υγείας στη περιφέρεια Κρήτης (όπως αναπτύσσεται και συντηρείται από το Κέντρο Ιατρικής Πληροφορικής και Τηλεματικών Εφαρμογών στην Υγεία του ΙΠ-ΙΤΕ) και την ανάδειξη της προστιθέμενης αξίας που το σύστημα HealthObs προσφέρει. Επίσης παρουσιάζονται οι πληροφοριακές τεχνολογίες (information technologies – IT) στη βάση των οποίων έχει σχεδιαστεί και αναπτυχθεί το σύστημα HealthObs. Πιο συγκεκριμένα παρουσιάζονται τα σχετικά πρότυπα διασύνδεσης συστημάτων (όπως παρέχονται από τον οργανισμό του OMG) με μια εκτενή αναφορά στις προδιαγραφές της CORBAmed και των βασικών συστατικών-λογισμικού (software components) για την ολοκλήρωση πληροφοριών. Επιπλέον, επιχειρείται μια σύντομη αναφορά σε ζητήματα ομοιόμορφης και ενιαίας μοντελοποίησης πληροφοριών (uniform information modeling) με έμφαση σε γλώσσες σήμανσης, όπως η XML.

- Στο Κεφάλαιο 3 (Ανακάλυψη Γνώσης από Δεδομένα) παρουσιάζονται οι τεχνολογίες εξόρυξης δεδομένων οι οποίες έχουν κατάλληλα προσαρμοστεί

και ενισχυθεί σε περιβάλλοντα δεδομένων τα οποία δομούνται με βάση το πρότυπο της XML.

- Στο Κεφάλαιο 4 (*Το Σύστημα HEALTHobs*) παρουσιάζονται με λεπτομέρεια τα συστατικά, οι μηχανισμοί, οι τεχνικές, οι αλγόριθμοι και τα εργαλεία τα οποία συνθέτουν το σύστημα HealthObs.

- Στο Κεφάλαιο 5 (*Σχετικές Εργασίες*) παρουσιάζονται σχετικές με τη τεχνολογία του HealthObs εργασίες και επιχειρείται μια συγκριτική μελέτη.

- Στο Κεφάλαιο 6 (Το Περιβάλλον εργασίας του HealthObs) Παρουσιάζεται αναλυτικά το Περιβάλλον εργασίας του HealthObs.

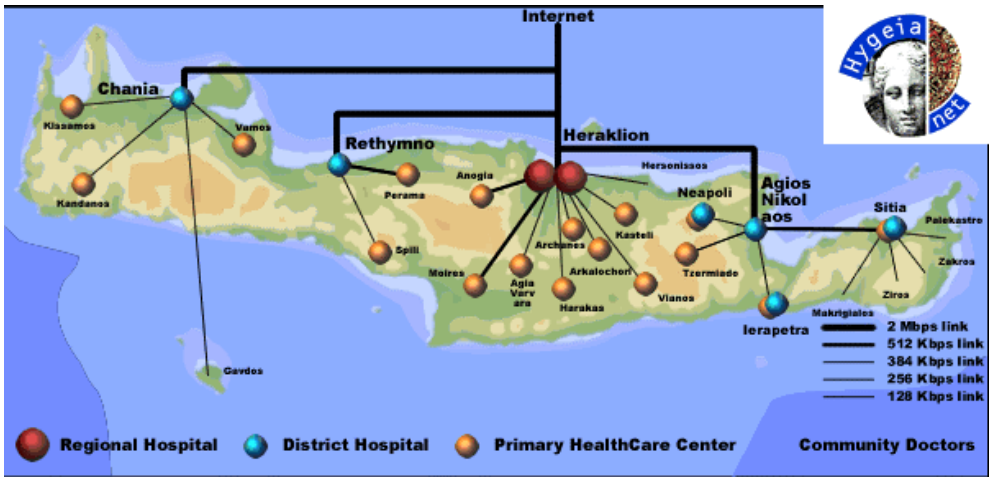
- Στο κεφάλαιο 7 (*Μελέτες Εφαρμογής του HealthObs*), παρουσιάζονται εκτενή παραδείγματα εφαρμογής του HealthObs στο κλινικό πεδίο, όπου διαφορετικές μελέτες-πεδίου (case-studies) και τα σχετικά πειραματικά αποτελέσματα αναλύονται, με ταυτόχρονη παρουσίαση της λειτουργικότητας των σχετικών διαμέσων επικοινωνίας ανθρώπου-μηχανής (human computer interfaces – HCI).

- Το τελευταίο Κεφάλαιο 8 (*Συμπεράσματα και Μελλοντική Ερευνητική & Αναπτυξιακή Εργασία*) αναφέρεται στα συμπεράσματα τα οποία εξήχθησαν από την έρευνα και ανάπτυξη (E&A) κατά τη διάρκεια υλοποίησης και χρήσης του συστήματος, καθώς και στις μελλοντικές E&A εργασίες οι οποίες θα μπορούσαν να αξιοποιήσουν και να ενισχύσουν την συγκεκριμένη υποδομή και τις υπηρεσίες που το HealthObs προσφέρει.

2 ΤΕΧΝΟΛΟΓΙΕΣ ΥΠΟΒΑΘΡΟΥ ΤΟΥ HEALTHOBS

Το HYGEIAnet [9,10] είναι μια μακροπρόθεσμη Ε&Α προσπάθεια του Κέντρου Ιατρικής Πληροφορικής και Τηλεματικών Εφαρμογών στην Υγεία του ΙΠ-ΙΤΕ. Πρόκειται για ένα ολοκληρωμένο περιφερειακό δίκτυο τηλεματικών υπηρεσιών το οποίο αποτελεί πιλότο για παρόμοιες προσπάθειες σε εθνικό ή σε ευρωπαϊκό επίπεδο.

Στα πλαίσια ενός περιβάλλοντος, στο οποίο παρέχεται αυξημένη προστασία της ιδιωτικής ζωής, χτίζεται μια υποδομή με απώτερο στόχο την παροχή και την ποιότητα των υπηρεσιών υγείας που παρέχονται στους πολίτες, όπως επίσης και την ανάπτυξη ολοκληρωμένων ιατρικών υπηρεσιών. Επιπλέον παρέχονται πληροφορίες και υπηρεσίες που θεωρούνται απαραίτητες, ώστε να είναι δυνατή η διαρκής βελτίωση των παρερχομένων υπηρεσιών, και της καλύτερης κατανόησης της κατάστασης της υγείας του πληθυσμού.



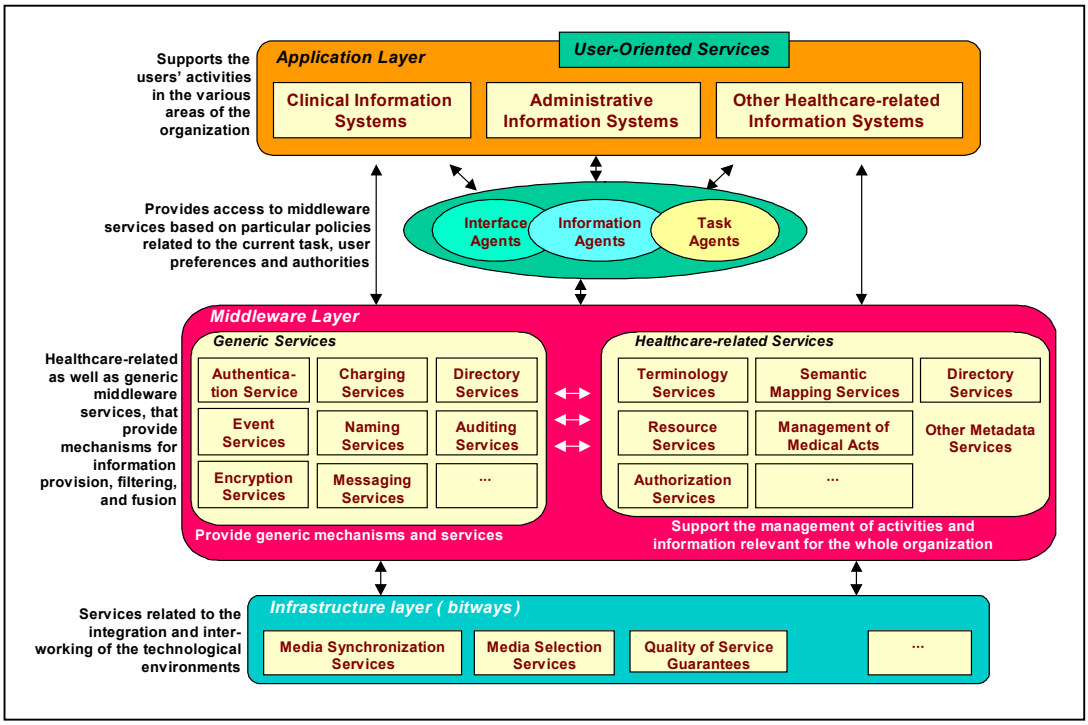
Σχήμα 3. Περιφερειακό Δίκτυο Τηλεματικών Υπηρεσιών της Κρήτης (HYGEIAnet)

Το HYGEIAnet έχει ως στόχο την εξυπηρέτηση των κατοίκων της συγκεκριμένης περιφέρειας, είτε αυτοί είναι ασθενείς, είτε είναι επαγγελματίες στον χώρο της υγείας (π.χ. γιατροί), είτε είναι ερευνητές, είτε διοικητικό προσωπικό. Οι αρχές που ακολουθήθηκαν κατά τον σχεδιασμό του HYGEIAnet, προήλθαν από την πραγματικότητα η οποία ισχύει στο περιφερειακό σύστημα υγείας το οποίο εφαρμόζεται στην Ελλάδα. Επιπλέον μπορεί να παρατηρηθεί ότι οι πολίτες ενδιαφέρονται πλέον για το σύστημα, και θέλουν να πάρουν ένα μέρος της ευθύνης της δικής τους υγείας. Ένας σημαντικός παράγοντας είναι ο

συντονισμός των ποικίλων υπηρεσιών, με απώτερο στόχο την εύκολη διαχείριση και εκμετάλλευση των πληροφοριών από τους επαγγελματίες που δρουν στον χώρο της υγείας. Τα επιμέρους ολοκληρωμένα περιφερειακά δίκτυα τηλεματικών υπηρεσιών αποτελούν τα δομικά τμήματα πάνω στα οποία θα στηριχτεί στο μέλλον η Εθνική Ιατρική Πληροφοριακή υποδομή.

2.1 Αρχιτεκτονική Αναφοράς του HYGEIAnet

Η αρχιτεκτονική αναφοράς του Healthcare Information Infrastructure (HII) [11], και η οποία παρουσιάζεται στο σχήμα 4, καθοδηγεί την ανάπτυξη του δικτύου τηλεματικών εφαρμογών για την παροχή ολοκληρωμένων υπηρεσιών [12,13,14]. Το προσφερόμενο αποτέλεσμα είναι ένας γενικός σκελετός στον οποίο μπορεί να στηριχτεί κάποιος για να ολοκληρώσει τα διάφορα ιατρικά πληροφοριακά συστήματα, έτσι ώστε να παρέχεται μια μεγάλη ποικιλία από υπηρεσίες οι οποίες θα εξυπηρετούν τους διάφορους επαγγελματίες, όπως επίσης και τον ίδιο τον πληθυσμό.



Σχήμα 4. Αρχιτεκτονική Αναφοράς του HYGEIAnet

Όπως παρουσιάζεται στο Σχήμα 4, η HII αποτελείται από τρία βασικά τμήματα, τις εφαρμογές (Application Layer), τις middleware υπηρεσίες (Middleware Layer), και την δικτυακή υποδομή (Infrastructure Layer). Το πρώτο αφορά τις

εφαρμογές οι οποίες υποστηρίζουν τις δραστηριότητες των χρηστών στις διάφορες περιοχές του οργανισμού. Οι εφαρμογές αυτές αλληλεπιδρούν με τις υπηρεσίες που προσφέρονται από το middleware επίπεδο, χωρίς όμως να είναι απαραίτητο για τον χρήστη να γνωρίζει την ύπαρξη τους.

Το δεύτερο επίπεδο (Middleware Layer), παρέχει δύο κατηγορίες υπηρεσιών: (α) τις γενικού σκοπού υπηρεσίες, οι οποίες είναι κοινές σε οποιοδήποτε πληροφοριακό σύστημα και σε οποιαδήποτε περιοχή εφαρμογών, όπως υπηρεσίες καταλόγων, υπηρεσίες ονομάτων κλπ. και (β) τις ιατρικού περιεχομένου υπηρεσίες οι οποίες βρίσκουν εφαρμογή στον δεδομένο χώρο της υγείας - τέτοιου είδους υπηρεσίες είναι οι υπηρεσίες επερώτησης λεξικών, οι υπηρεσίες διαχείρισης ιατρικών πόρων κλπ.

Τέλος το τρίτο επίπεδο (Infrastructure Layer), παρέχει ένα σύνολο υπηρεσιών που σχετίζονται με την ολοκλήρωση και δια-λειτουργικότητα (interoperability) των διάφορων τεχνολογικών περιβαλλόντων.

2.2 Κλινικά Πληροφοριακά Συστήματα στο HYGEIAnet

Η βασικότερη middleware υπηρεσία που διαθέτει το HYGEIAnet είναι η Ομοσπονδία Καταλόγου Κλινικών Δεδομένων Ασθενών (Patient Clinical Data Directory - PCDD) [11]. Ο κύριος στόχος της υπηρεσίας αυτής είναι υποστήριξη του Integrated Electronic HealthCare Record (IEHCR) με στόχο τη συνέπειά του, και τη δυνατότητα επαναχρησιμοποίησής και επεκτασιμότητάς του. Το PCDD ευρετηριάζει (index) ασθενείς, όπως επίσης και πληροφορίες για τα κλινικά αντικείμενα των 'τμημάτων' (segments) του ηλεκτρονικού φακέλου των ασθενών. Στη παρούσα εργασία εκμεταλλευόμαστε το Πληροφοριακό Σύστημα Πρωτοβάθμιας Φροντίδας (PHCCIS) το οποίο έχει εισαχθεί σαν ένα ξεχωριστό `'SystemType'` στο περιβάλλον του IEHCR και της υπηρεσίας PCDD. Τα πληροφοριακά συστήματα τα οποία αποτελούν επιπλέον `'SystemTypes'` του IEHCR του HYGEIAnet φαίνονται στο Πίνακα 1.

Πίνακας 1. Κλινικά Πληροφοριακά Συστήματα στο HYGEIAnet

Ακρώνυμο	**Πλήρης Ονομασία**	**Κλινική Μονάδα**	**Σε λειτουργία**
PSCIS	Πληροφοριακό Σύστημα Προ-νοσοκομειακής Επείγουσας Ιατρικής	9 Κέντρα Ιατρική Περίθαλψης	9 χρόνια
IASO	Πληροφοριακό Σύστημα Παιδοχειρουργικής	1 παιδοχειρουργική κλινική	7 χρόνια
PHCCIS	Πληροφοριακό Σύστημα Πρωτοβάθμιας Φροντίδας	Εθνικό Κέντρο Άμεσης Βοήθειας (ΕΚΑΒ) στο Ηράκλειο Κρήτης	6 χρόνια

Το κλινικό πληροφοριακό σύστημα πρωτοβάθμιας υγείας (PHCCIS) είναι ένα πληροφοριακό σύστημα το οποίο χρησιμοποιείται σε πολλά Κέντρα Πρωτοβάθμιας Ιατρικής Περίθαλψης στην περιφέρεια της Κρήτης και υλοποιεί ένα γενικό ιατρικό αρχείο το οποίο χρησιμοποιείται από τους γιατρούς των κέντρων πρωτοβάθμιας φροντίδας. Η παρούσα εργασία εστιάζεται στο PHCCIS.

2.3 Κλινικό Πληροφοριακό Σύστημα Πρωτοβάθμιας Υγείας: Το Περιβάλλον Εφαρμογής του HealthObs

Το PHCCIS παρέχει περισσότερες από 30 φόρμες για πρόσβαση στα δεδομένα των ασθενών και υποστηρίζει 'Query by Example' λειτουργίες. Αυτό σημαίνει ότι μπορούν να εφαρμοστούν και πολύπλοκες επερωτήσεις. Το PHCCIS παρέχει δύο όψεις για τα κομμάτια του ηλεκτρονικού φακέλου του ασθενή :

- ✓ visit-oriented: Αποτελείται από λίστες από encounters/επισκέψεις και υλοποιείται ως ένας δυσδιάστατος πίνακας. Για κάθε encounter/επίσκεψη, καταχωρούνται οι σχετικές εξετάσεις, διαγνώσεις, και εκτιμήσεις. Πιο συγκεκριμένα ένα encounter μπορεί να περιέχει έναν αριθμό από εξετάσεις, όπως κλινική εξέταση, γυναικολογική εξέταση, βιοχημική ανάλυση, ανάλυση αίματος ή μια ακτινολογική εξέταση, όπως επίσης και μια διάγνωση και την θεραπευτική αγωγή που ακολουθείται.
- ✓ problem-oriented: Το αρχείο του ασθενή διαιρείται σε σειρές από κεφάλαια, στο κάθε ένα από τα οποία αποδίδεται ο τίτλος "πρόβλημα" ή πιο γενικά "προηγούμενη διάγνωση". Το "πρόβλημα"/διάγνωση μπορεί να οριστεί ως ICD9 (International Coding of Diseases) κωδικός ή να περιγραφεί ως απλό κείμενο. Τα δεδομένα του αρχείου του ασθενή ομαδοποιούνται κάτω από ένα ή περισσότερα προβλήματα. Όταν ένα πρόβλημα επιλέγεται, παρουσιάζεται η πληροφορία που είναι σχετική με αυτό. Κάθε πρόβλημα σχετίζεται με ένα ή περισσότερα επεισόδια. Ένα επεισόδιο μπορεί να είναι: (α) ένα *επεισόδιο ενός προβλήματος*, πολλά προβλήματα υγείας δεν είναι συνεχώς ενεργά, αλλά έχουν έναν κύκλο δραστηριότητας - η περίοδος κατά την οποία ένα πρόβλημα είναι ενεργό μπορεί να οριστεί ως ένα επεισόδιο του προβλήματος, και (β) *ένα επεισόδιο υπηρεσίας*: μια συλλογή από γεγονότα κατά την διάρκεια μια καθορισμένης χρονικής περιόδου.

2.4 Η Πληθυσμιακή Όψη των Κλινικών Δεδομένων: Ανάγκες Ολοκλήρωσης και Σημασιολογικής Ομογενοποίησης

Σκοπός της εργασίας αυτής είναι να δώσει μια τρίτη όψη στο σύστημα PHCCIS.

Από ασθενο-κεντρική (visit ή problem oriented) να μεταβούμε σε μια πλυθησμο-κεντρική (population oriented) όψη των κλινικών δεδομένων και πληροφοριών. Επιπλέον, η πρόληψη υγείας και οι *επιδημιολογικές* μελέτες γίνονται περισσότερο απαιτητικές με τη μεταφορά και ανταλλαγή πληροφοριών. Αυτή η απαίτηση αυξάνει την ανάγκη για *την ολοκλήρωση* στοιχείων από *τις κατανεμημένες* και *ετερογενείς* κλινικές πηγές πληροφοριών με έναν τρόπο που: η άμεση πρόσβαση στα συνοπτικά στοιχεία και η αξιόπιστη *υποστήριξη εξαγωγής γνώσης (βασισμένη σε στοιχεία)* να βοηθάει.

Με την βοήθεια του διεθνούς προτύπου COAS (Clinical Object Access Service) το οποίο προσφέρει πρόσβαση στην κλινική πληροφορία και ομοιόμορφη διαμόρφωση των ανακτημένων κλινικών στοιχείων, επιτυγχάνουμε την ανάσυρση δεδομένων από ετερογενείς πηγές κλινικής πληροφορίας. Η CORBA - ως στρώμα επικοινωνίας, και οι διεπαφές COAS - ως σχέδια αντιπροσώπευσης, στοχεύουν να κερδίσουν τον έλεγχο των κατανεμημένων πηγών πληροφοριών σε επίπεδο *μεταδεδομένων (*metadata*)*, επιτρέποντας την αυτονομία των μεμονωμένων συστημάτων στο επίπεδο δεδομένων.

Ωστόσο η CORBA δεν προσφέρει βοήθεια σε επίπεδο παροχής γνώσης καθώς δεν εγγυάται και δεν εξασφαλίζει ότι συγκεκριμένα τμήματα (components) της πληροφορίας μπορούν να λειτουργήσουν μαζί. Επιπρόσθετα μολονότι τα σχετικά IDL-διάμεσα καθορίζουν την αναγκαία σύνταξη για συνεργασία και πρόσβαση σε κατανεμημένες πληροφορίες δεν περιγράφουν την σημασιολογία αυτής της πληροφορίας. Αυτή η λειτουργία έρχεται να πραγματοποιηθεί με την τεχνολογία της XML με την οποία προσεγγίζεται ικανοποιητικά τη λύση στο ζητούμενο πρόβλημα, που δεν είναι άλλο από την *σημασιολογική ομογενοποίηση* της ετερογενούς κλινικής πληροφορίας την οποία επιθυμούμε να επεξεργαστούμε.

2.5 Ολοκλήρωση Κατανεμημένων Πληροφοριακών Συστημάτων: Ανάγκες Για Πρότυπα Διάμεσα Επικοινωνίας

2.5.1 OMG

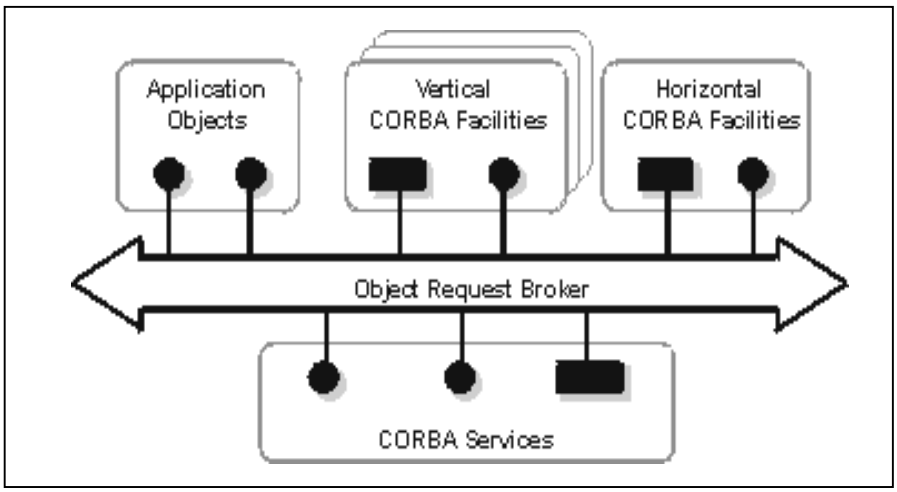
Το Object Management Group (OMG) [1] είναι η μεγαλύτερη κοινοπραξία εταιριών λογισμικού στον κόσμο και αριθμεί περισσότερα από 800 μέλη, υποστηρικτές (developers) και τελικούς χρήστες. Το OMG διαμορφώθηκε για να δημιουργήσει μια αγορά λογισμικού προωθώντας τη τυποποίηση αντικειμένων λογισμικού.

Σκοπός του οργανισμού είναι η καθιέρωση οδηγιών και λεπτομερών

προδιαγραφών για να παρέχει ένα κοινό πλαίσιο για την ανάπτυξη εφαρμογών. Ένα βασικό όφελος ενός αντικειμενοστραφούς (object oriented) συστήματος είναι η δυνατότητά του να λειτουργήσει με υπάρχοντα αντικείμενα και την προσθήκη νέων αντικειμένων στο σύστημα. Η διαχείριση αντικειμένων οδηγεί στη γρηγορότερη ανάπτυξη εφαρμογών, την ευκολότερη συντήρηση, την εξελιξιμότητα και το επαναχρησιμοποιήσιμο λογισμικό. Επίσης μέσα στους στόχους συμπεριλαμβάνονται η μεταφορά, και συνεργασία των οντοκεντρικών συνιστωσών λογισμικού (software components) σε ετερογενή περιβάλλοντα. Προς αυτήν την κατεύθυνση το OMG υιοθετεί προδιαγραφές διεπαφών (interface) και πρωτοκόλλων, βασισμένα κυρίως σε εμπορικά διαθέσιμη τεχνολογία, και τα οποία όλα μαζί ορίζουν το Object Management Architecture (OMA).

2.5.2 Object Management Architecture

Το Object Management Architecture Guide [2] (OMAG) περιγράφει τους τεχνολογικούς στόχους και την ορολογία του OMG, και παρέχει την εννοιολογική υποδομή πάνω στην οποία βασίζονται οι υποστηρικτικές προδιαγραφές. Επίσης περιέχει το OMG Object Model, το οποίο ορίζει την κοινή σημασιολογία για τον καθορισμό των δημόσιων χαρακτηριστικών των αντικειμένων με ένα κοινά αποδεκτό και ανεξάρτητο υλοποίησης τρόπο, και το OMA Reference Model. Το OMG κατηγοριοποιεί τα αντικείμενα σε τέσσερις κατηγορίες : the CORBAservices, CORBAfacilities, CORBAdomain objects, and Application Objects.



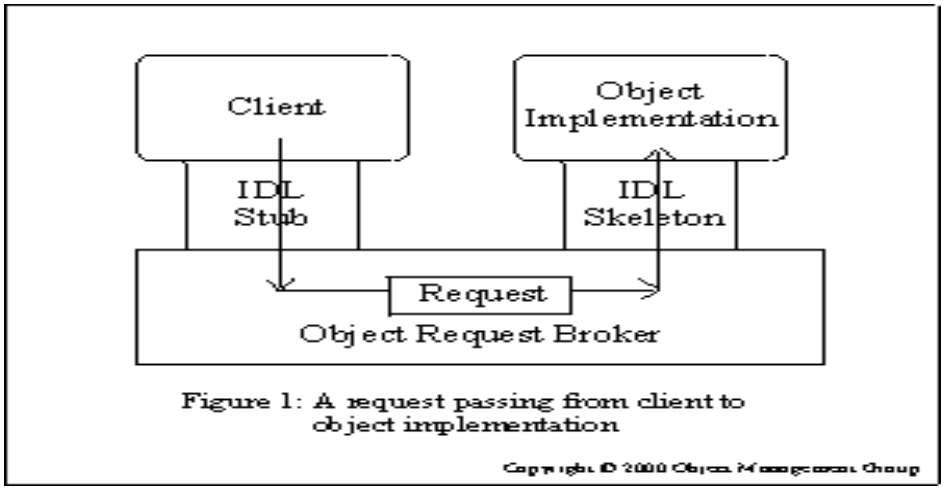
Σχήμα 5: Μοντέλο αναφοράς OMG

Το Μοντέλο Αναφοράς (Reference Model) ,το οποίο απεικονίζεται στο σχήμα 5, καθορίζει και χαρακτηρίζει τις συνιστώσες, τα διάμεσα επικοινωνίας συστημάτων (system interfaces), και τα πρωτόκολλα τα οποία απαρτίζουν το

OMA. Αυτό περιέχει το Object Request Broker (ORB) το οποίο δίνει στους πελάτες και τα αντικείμενα την δυνατότητα να επικοινωνούν σε κατανεμημένα περιβάλλοντα όπως επίσης και στις τέσσερις κατηγορίες από interfaces Η ευρεία αποδοχή του OMA του OMG από την βιομηχανία παρέχει στους υποστηρικτές (developers) και τους τελικούς χρήστες την ικανότητα να αναπτύσσουν συνεργαζόμενα συστήματα λογισμικού τα οποία είναι κατανεμημένα από όλες τις απόψεις, όπως λειτουργικά συστήματα, γλώσσες προγραμματισμού, και επικοινωνία πάνω από δίκτυο. Στις καθορισμένες πλέον προδιαγραφές συμπεριλαμβάνονται η Common Object Request Broker Architecture (CORBA), τα CORBAservices, και τα CORBAfacilities.

2.5.3 CORBA

Μια από τις καθορισμένες πλέον προδιαγραφές του OMA είναι η CORBA (Common Object Request Broker Architecture) [3]. Η CORBA είναι η επικοινωνία υπολογιστών μέσω δικτύου ανεξαρτήτου αρχιτεκτονικής ,υποδομής, λογισμικού και εφαρμογής υπολογιστών. Χρησιμοποιώντας το τυποποιημένο πρωτόκολλο IIOP, ένα πρόγραμμα βασισμένο σε (CORBA) από οποιοδήποτε προμηθευτή, για σχεδόν οποιουσδήποτε υπολογιστή, λειτουργικό σύστημα, γλώσσα προγραμματισμού, και δίκτυο, μπορεί να επικοινωνήσει με ένα άλλο πρόγραμμα βασισμένο σε (CORBA) από το ίδιο ή έναν άλλο προμηθευτή, για σχεδόν οποιοδήποτε υπολογιστή, λειτουργικό σύστημα, γλώσσα προγραμματισμού, και δίκτυο.



Σχήμα 6: Διεπιφάνεια προγραμματισμού μέσω Corba

Η CORBA ορίζει την διεπιφάνεια προγραμματισμού για το Object Request Broker. Το ORB είναι ο βασικός μηχανισμός με τον οποίο τα αντικείμενα

μπορούν να στέλνουν αιτήσεις (requests) και να παίρνουν απαντήσεις (responses) το ένα από το άλλο, είτε βρίσκονται στο ίδιο μηχάνημα, είτε επικοινωνούν πάνω από κάποιο δίκτυο. Οι πελάτες δεν χρειάζεται πλέον να απασχολούνται με τους μηχανισμούς επικοινωνίας και ενεργοποίησης (activation) των αντικειμένων, τον τρόπο με τον οποίο υλοποιήθηκαν αυτά, ή το που βρίσκονται.

Η OMG Interface Definition Language (IDL) παρέχει ένα τυποποιημένο τρόπο για να ορίζουμε CORBA interfaces αντικειμένων. Ο IDL ορισμός είναι το "συμβόλαιο" ανάμεσα σ' αυτόν που υλοποιεί το αντικείμενο και στον πελάτη. Η IDL είναι μια αυστηρά δηλωτική γλώσσα, η οποία είναι ανεξάρτητη από τις γλώσσες προγραμματισμού που μπορεί να χρησιμοποιηθούν. Η CORBA 2.0 είναι μια επεκταμένη και αναδομημένη έκδοση των προδιαγραφών του CORBA 1.2. Η CORBA 2.0 είναι μια οικογένεια από προδιαγραφές που αποτελούνται από τα παρακάτω τμήματα:

- Τον πυρήνα (Core) (περιέχει την σύνταξη και την σημασιολογία της IDL)
- Ένα σύνολο από αντιστοιχίσεις σε διάφορες γλώσσες προγραμματισμού το οποίο μεγαλώνει συνέχεια. Παραδείγματα τέτοιον γλωσσών είναι οι : C, C++, SmallTalk, Ada95, COBOL, Java.

2.5.4 CORBAmed

Το CORBAMed [4, 5] είναι το τμήμα CORBA που αφιερώνεται στην περιοχή της υγειονομικής περίθαλψης. Το CORBAmed ορίζει κοινά αποδεκτά οντοκεντρικές διεπαφές, οι οποίες συμβάλουν στην δια-συνεργασία ανάμεσα σε μια ποικιλία από πλατφόρμες, λειτουργικά συστήματα, γλώσσες, και εφαρμογές. Προορίζεται να δημιουργεί πρότυπα δια-λειτουργικότητας για την υγειονομική περίθαλψη. Η προσέγγιση CORBAMed περιλαμβάνει τα πολλαπλά επίπεδα του MPIs όπως για παράδειγμα το επίπεδο υπηρεσιών, το οργανωτικό επίπεδο, το επιχειρηματικό επίπεδο.

Χρησιμοποιεί τον 'ID Domain Manager' που διαχειρίζεται τον προσδιορισμό και το συσχετισμό του δημογραφικού προφίλ ασθενών για τις διερευνηθείσες πληροφορίες ασθενών. Ο κύριος στόχος του CORBAmed είναι η βελτίωση της ποιότητας των υπηρεσιών, καθώς και η μείωση των εξόδων, με τη χρήση της τεχνολογίας CORBA. Αυτή τη στιγμή το CORBAmed έχει ξεκινήσει τις διαδικασίες για την παραγωγή κοινά αποδεκτών διεπαφών σε διάφορους τομείς της υγείας. Μέχρι στιγμής έχουν ολοκληρωθεί τρεις από αυτές, το Clinical Observation Access Service (COAS), το Person Identification Service (PIDS), και το Lexicon Query Service (LQS).

Το Lexicon Query Service [6] είναι η υπηρεσία που καθορίζει ένα σύνολο κοινών, για ανάγνωση μόνο (read-only), μεθόδων για πρόσβαση στα περιεχόμενα συστημάτων ιατρικής ορολογίας. Ο όρος συστήματα ιατρικής ορολογίας καλύπτει όλο το φάσμα των συστημάτων, από τα απλά που αποτελούνται από λίστες ενός συνόλου από κώδικες και φράσεις, έως και συστήματα δυναμικά, με πολλαπλά σχήματα ιεραρχίας και κατηγοριοποίησης.

Το Person Identification Service [7] ορίζει κατάλληλες διεπαφές, έτσι ώστε, να είναι δυνατός ο μονοσήμαντος προσδιορισμός της ταυτότητας των ασθενών. Το PIDS σχεδιάστηκε, με τέτοιο τρόπο, ώστε να:

- είναι ικανό να αποδίδει ids στα πλαίσια ενός ID Domain, αλλά και να συσχετίζει ids που προέρχονται από διαφορετικά ID Domains.
- είναι ικανό να ψάχνει και να εντοπίζει ασθενείς, ανεξάρτητα από τον αλγόριθμο ταιριάσματος, είτε αυτόματα, είτε με τη βοήθεια κάποιου ειδικού.
- υποστηρίζει ομοσπονδίες από υπηρεσίες απόδοσης ταυτότητας σε ασθενείς (PIDS).
- υποστηρίζει υλοποιήσεις του PIDS οι οποίες θα προστατεύουν το απόρρητο των ασθενών και θα βασίζονται σε μια μεγάλη ποικιλία από πολιτικές και μηχανισμούς ασφάλειας.
- επιτρέπει την εύκολη διασυνεργασία διαφορετικών υπηρεσιών PIDS.
- ορίσει τα διάφορα επίπεδα συμβατότητας, από τα απλά για επερωτήσεις μόνο ID Domains, μέχρι τις σύνθετες ομοσπονδίες από συσχετιζόμενα ID Domains.

2.5.5 Clinical Observation Access Service (COAS): Ένα Διεθνές Πρότυπο Τυποποίησης Κλινικών Δεδομένων

Το Clinical Observation Access Service [8] είναι ένα σύνολο από διεπαφές και δομές δεδομένων με τα οποία οι εξυπηρετητές μπορούν να παρέχουν κλινικές παρατηρήσεις (clinical observations), και είναι πλέον στην τελική του έκδοση από τον Απρίλιο του 1999.

Ο όρος κλινικές παρατηρήσεις ορίστηκε από την CORBAmed ως ένα σημαντικό κομμάτι τις πληροφορίας που καταγράφεται για κάθε ασθενή. Παραδείγματα κλινικών παρατηρήσεων είναι τα εξής: εργαστηριακές εξετάσεις, βιοσήματα, υποκειμενικές και αντικειμενικές παρατηρήσεις και εκτιμήσεις, παρατηρήσεις και μετρήσεις που παρέχει κάποιος ειδικός όπως ένας ακτινολόγος ή ένας παθολόγος ο οποίος αναλύει εικόνες και άλλα δεδομένα (multi-media data).

Μερικά κοινά γνωρίσματα αυτών των παρατηρήσεων είναι τα παρακάτω. Οι κλινικές παρατηρήσεις :

- ✓ ασχολούνται με ένα αντικείμενο στον οποίο παρέχεται ιατρική φροντίδα, όπως ένας ασθενής, ή ένα πληθυσμός.
- ✓ αναπαριστούν την κατάσταση του αντικειμένου στο χρόνο, είτε μια συγκεκριμένη χρονική στιγμή , είτε σε κάποιο συγκεκριμένο χρονικό διάστημα.
- ✓ γίνονται, ή καταγράφονται, από ένα μηχάνημα ή από κάποιον ειδικό.
- ✓ διακρίνονται από κάποιο βαθμό αξιοπιστίας.

Τα παραπάνω γνωρίσματα θα παίξουν σημαντικό ρόλο στην κατανόηση τόσο του συνόλου των λειτουργιών που θα πρέπει να ικανοποιεί μια υπηρεσία πρόσβασης σε κλινικές παρατηρήσεις, όσο και του μοντέλου αναφοράς που θα αναλυθεί στην συνέχεια.

2.5.5.1 Απαιτήσεις Λειτουργικότητας

Το πρώτο και υποχρεωτικό σύνολο λειτουργιών για κάθε υπηρεσία πρόσβασης σε κλινικές παρατηρήσεις αποτελείται από έξι λειτουργίες:

- Οι κλινικές παρατηρήσεις πρέπει να μπορούν να αναζητούνται και να μεταφέρονται.
- Οι κλινικές παρατηρήσεις πρέπει να φιλτράρονται, όπως για παράδειγμα βάση ενός ασθενή, βάση του τύπου της παρατήρησης, βάση της κατάσταση και/ή του χρόνου.
- Πρέπει να υπάρχει μηχανισμός επερώτησης για τις διαθέσιμες παρατηρήσεις.
- Πρέπει να παρέχεται πρόσβαση σε πληροφορίες που αφορούν το "περιβάλλον" (context) μιας παρατήρησης.
- Πρέπει να υπάρχει ένα προκαθορισμένο σύνολο από τύπους παρατηρήσεων.
- Πρέπει να υπάρχει η ικανότητα να χρησιμοποιούνται κοινά αποδεκτά και δημόσια διαθέσιμα λεξικά.

Το δεύτερο και τελευταίο σύνολο λειτουργιών είναι προαιρετικό:

- Οι υπηρεσίες πρόσβασης σε κλινικές παρατηρήσεις μπορούν να παρέχουν ένα μηχανισμό ο οποίος θα παρέχει πρόσβαση σε μελλοντικές παρατηρήσεις.
- Οι υπηρεσίες πρόσβασης σε κλινικές παρατηρήσεις μπορούν να παρέχουν την ικανότητα υποστήριξης δυναμικής ανακάλυψης των υποστηριζόμενων τύπων παρατηρήσεων και δεδομένων.
- Οι υπηρεσίες πρόσβασης σε κλινικές παρατηρήσεις μπορούν να παρέχουν

την ικανότητα γενικού φιλτραρίσματος των επερωτήσεων.

- Οι υπηρεσίες πρόσβασης σε κλινικές παρατηρήσεις μπορούν να παρέχουν την ικανότητα για χρήση των υπηρεσιών Trader (Trader services).
- Οι υπηρεσίες πρόσβασης σε κλινικές παρατηρήσεις μπορούν να παρέχουν την ικανότητα πρόσβασης στο ιστορικό αναπροσαρμογής των δεδομένων.
- Οι υπηρεσίες πρόσβασης σε κλινικές παρατηρήσεις μπορούν να παρέχουν ένα πληροφοριακό μοντέλο αναφοράς και τα αντίστοιχα IDL αρχεία.
- Οι υπηρεσίες πρόσβασης σε κλινικές παρατηρήσεις μπορούν να παρέχουν την ικανότητα χρήσης τόσο τοπικών και περιορισμένων λεξικών, όσο κοινά αποδεκτών και διαθέσιμων.
- Οι υπηρεσίες πρόσβασης σε κλινικές παρατηρήσεις μπορούν να παρέχουν την ικανότητα χρήσης των υπηρεσιών επερώτησης λεξικού (LQS), έτσι ώστε να υποστηρίζονται πολλά λεξικά.

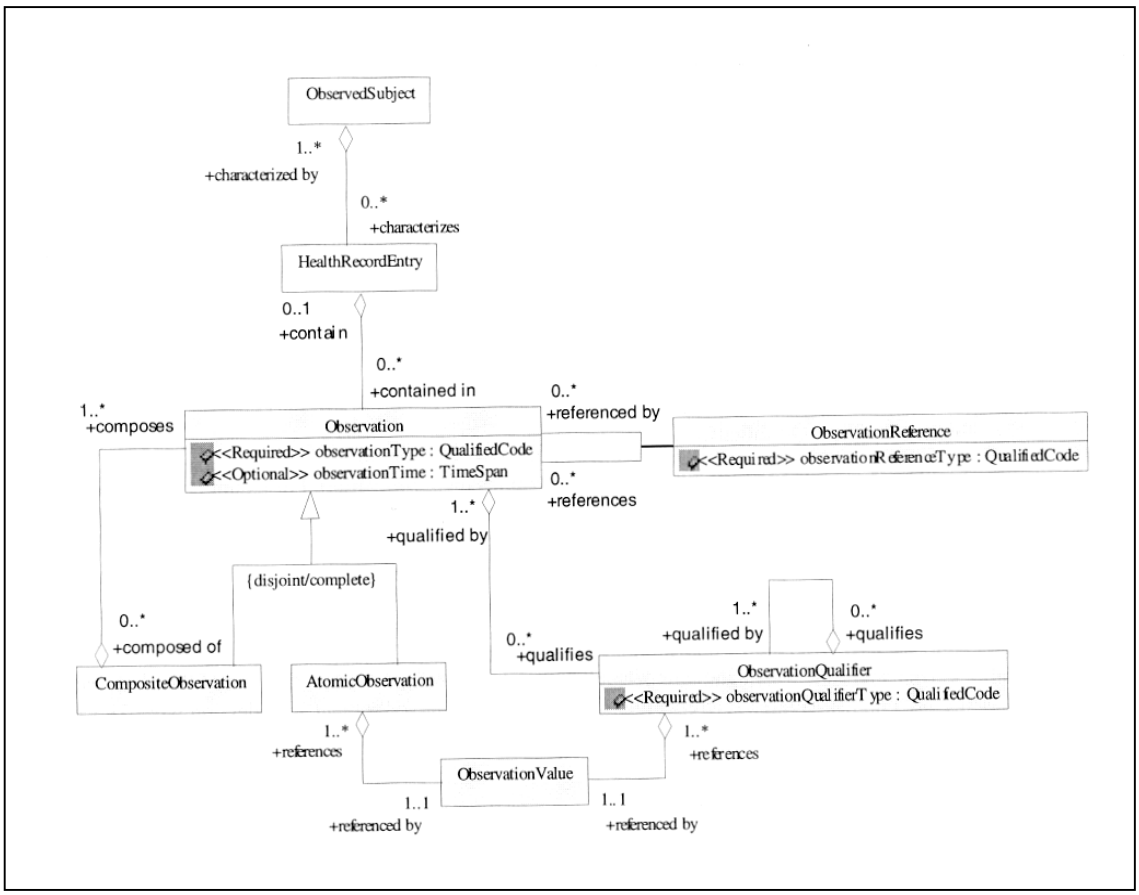
Οι παραπάνω λειτουργίες καλύπτουν σχεδόν όλο το φάσμα των λειτουργιών που μια υπηρεσία πρόσβασης σε κλινικές παρατηρήσεις θα μπορούσε να παρέχει. Αυτό που πρέπει να τονιστεί είναι ότι οι εφαρμογές οι οποίες διατηρούν κλινικές παρατηρήσεις, συνήθως υλοποιούν ένα μόνο μέρος αυτών των λειτουργιών, πράγμα απολύτως φυσιολογικό, διότι κάθε εφαρμογή προσπαθεί να καλύψει τις ανάγκες των χρηστών μέσα στα πλαίσια του περιβάλλοντος που αυτή λειτουργεί.

Το COAS δεν προσδιορίζει τον τρόπο με τον οποίο πρέπει να δομούνται τέτοιού είδους εφαρμογές, αλλά καθορίζει τον τρόπο με τον οποίο αυτές οι εφαρμογές επικοινωνούν με τον υπόλοιπο κόσμο. Τα παραπάνω ισχύουν και για το μοντέλο αναφοράς που αναλύεται στην συνέχεια.

2.5.5.2 Περιγραφή του Μοντέλου Αναφοράς COAS

Σε αυτό το υποκεφάλαιο θα περιγραφεί το πληροφοριακό μοντέλο αναφοράς για τις υπηρεσίες πρόσβασης σε κλινικές παρατηρήσεις. Το μοντέλο όπως φαίνεται και στο Σχήμα 7, είναι αρκετά απλό. Παρόλα αυτά είναι αρκετά ισχυρό, ώστε να προσφέρει την επεκτασιμότητα που είναι απαραίτητο στοιχείο στον χώρο της παροχής ιατρικής φροντίδας.

Οι οντότητες Health Record Entry και Observed Subject αναπαρίστανται στο μοντέλο για να δειχθεί ότι μπορούν να ταιριάξουν σε αυτό. Παρόλα αυτά δεν υποστηρίζονται από το υπάρχον μοντέλο. Συνεπώς μια υπηρεσία πρόσβασης σε κλινικές παρατηρήσεις αποτελείται από τις εξής οντότητες:



Σχήμα 7: Γενική Άποψη του Μοντέλου Αναφοράς COAS

- **Observation.** Το Observation είναι μια αφηρημένη κλάση και τα attributes που περιέχει τα κληρονομούν οι οντότητες Composite Observation και Atomic Observation τα οποία είναι υποκλάσεις του Observation. Επίσης είναι πλήρες (complete) και διακριτό (disjoint). Ο χαρακτηρισμός πλήρες σημαίνει ότι δεν υπάρχουν άλλες υποκλάσεις του εκτός από τις δύο που προαναφέραμε, και ο χαρακτηρισμός διακριτό σημαίνει ότι οι περιπτώσεις αυτού του τύπου μπορούν να έχουν μόνο μία από τις δύο υποκλάσεις ως τύπο. Όσον αφορά τις σχέσεις του με άλλες οντότητες, το Observation σχετίζεται με το Composite Observation (ένα ή περισσότερα Observations συνιστούν μηδέν ή περισσότερα Composite Observations), με το Observation Reference (μηδέν ή περισσότερα Observations σχετίζονται με μηδέν ή περισσότερα Observations), και με το Observation Qualifier (ένα ή περισσότερα Observations προσδιορίζονται από μηδέν ή περισσότερα Observation Qualifiers).
- **Composite Observation.** Το Composite Observation αναπαριστά ένα σύνολο από Observations. Τέτοιου είδους σύνολο είναι μια πλήρης εξέταση αίματος. Στην ουσία πρόκειται για Observations τα οποία αποτελούνται από

άλλα πιο απλά Observations. Αυτό φαίνεται και στο σχήμα, 7 δηλαδή, μηδέν ή περισσότερα Composite Observations αποτελούνται από ένα ή περισσότερα Observations. Το Composite Observation είναι υποκλάση του Observation και συνεπώς κληρονομεί τα attributes του. Τέλος, όπως βλέπουμε το Composite Observation δεν σχετίζεται με κάποια τιμή.

- **Atomic Observation.** Το Atomic Observation αναπαριστά ένα απλό Observation και σχετίζεται με κάποια τιμή (Observation Value). Κάθε Atomic Observation σχετίζεται με μία και μόνο μία τιμή. Τέλος, επειδή το Atomic Observation είναι υποκλάση του Observation κληρονομεί τα χαρακτηριστικά του.
- **Observation Reference.** Το Observation Reference είναι μία κλάση, η οποία προσδιορίζει τον τύπο των σχέσεων ανάμεσα στα Observations.
- **Observation Qualifier.** Το Observation Qualifier έχει ως σκοπό να προσδιορίσει περαιτέρω το Observation καταγράφοντας το "περιβάλλον" του. Το ίδιο ισχύει και για το Observation Value με το οποίο συσχετίζεται. Τέλος, μηδέν ή περισσότερα Observation Qualifiers μπορούν να προσδιορίσουν επιπλέον ένα ή περισσότερα Observation Qualifiers.
- **Observation Value.** Το Observation Value είναι ένας αφηρημένος τύπος. Αυτό είναι απόλυτα λογικό μιας και η τιμή ενός Observation μπορεί να έχει οποιαδήποτε μορφή. Τέτοιες μορφές είναι το απλό κείμενο, ένα νούμερο, κάποια εικόνα κλπ. Το COAS ορίζει ένα σύνολο από τύπους, ως υποκλάσεις του Observation Value, το οποίο ναι μεν είναι διακριτό, αλλά όχι πλήρες ακόμα. Όσον αφορά τις σχέσεις του με άλλες οντότητες του μοντέλου, ένα και μόνο ένα Observation Value σχετίζεται με ένα ή περισσότερα Atomic Observations, και προσδιορίζεται επιπλέον από ένα ή περισσότερα Observation Qualifiers.

Για τις αναλυτικές περιγραφές ο αναγνώστης παραπέμπεται στο [8].

2.6 Ενιαία Μοντελοποίηση Πληροφοριών

Οι γλώσσες γενικής σήμανσης αναπτύχθηκαν για να δώσουν λύση στην έλλειψη συμβατότητας και την αδυναμία της σήμανσης να υποδηλώσει τη δομή του κειμένου. Το πρώτο βήμα προς αυτή την κατεύθυνση έγινε από τον C.F. Goldfarb στη δεκαετία του 1970 [17]. Η πρότασή του έδινε έμφαση σε δύο βασικές αρχές:

- Η σήμανση πρέπει να περιγράφει τη δομή ενός εγγράφου και όχι το στυλ ή τη μορφοποίηση
- Η σήμανση πρέπει να ακολουθεί αυστηρή σύνταξη ούτως ώστε ο κώδικας να είναι κατανοητός από το πρόγραμμα ή το χρήστη.

Το αποτέλεσμα ήταν η document composition facility Generalized Markup Language (GML), η οποία έγινε αποδεκτή ως πρότυπο το 1986 (ISO 8879) ως Τυποποιημένη Γλώσσα Γενικής Σήμανσης (Standard Generalized Markup Language - SGML).

Η SGML θεωρεί ότι τα έγγραφα συντίθενται μέσω της επανάληψης κάποιων βασικών στοιχείων που δίνουν με τρόπο έλλογο αναπαράσταση του περιεχομένου. Αυτό επιτυγχάνεται μέσω της χρήσης ενός βασικού συντακτικού, στο οποίο ανατρέχει το πρόγραμμα για να αντιληφθεί τη δομή του εγγράφου. Το συντακτικό αυτό ονομάζεται 'Ορισμός Τύπου Εγγράφου' (Document Type Definition - DTD) και αποτελεί συνήθως ένα εξωτερικό αρχείο. Περιέχει τον τρόπο που είναι δομημένα τα στοιχεία του εγγράφου και το είδος των δεδομένων που περιέχει. Η αναφορά στον Ορισμό Τύπου δηλώνεται στην αρχή του εγγράφου [18].

Το έγγραφο αποτελείται από στοιχεία (elements) κειμένου. Κάθε στοιχείο περιέχεται μεταξύ δύο όρων (tags) οι οποίοι συνίστανται στο όνομα των στοιχείων εντός γωνιωδών αγκυλών. Οι όροι εισάγονται κατά ζεύγη (υποχρεωτικά μόνο στην XML, σε αντίθεση προς την SGML και την HTML). Ο αρχικός όρος (start-tag) στην αρχή κάθε στοιχείου αντιστοιχεί σε έναν τελικό όρο(end-tag). Η μονή διαφορά στην απόδοσή των δύο αυτών όρων είναι ότι ο τελικός όρος έχει ως πρόθεμα μία κάθετο, στην ουσία σηματοδοτούν την έναρξη και τη λήξη του στοιχείου. Το κείμενο σε ένα έγγραφο σε SGML πρέπει υποχρεωτικά να περικλείεται σε ένα τουλάχιστον ζεύγος όρων.

2.6.1 Χαρακτηριστικά της XML

Η XML [15,16] είναι μία γενικευμένη γλώσσα σήμανσης, η οποία επιτρέπει στους συντάκτες του εγγράφου να ορίσουν οι ίδιοι το σύνολο όρων που θα χρησιμοποιήσουν. Τα έγγραφα σε XML είναι αυτό-περιγραφόμενα (self-describing). Ένα έγκυρο (valid) έγγραφο περιέχει τη σειρά των κανόνων στους οποίους πρέπει να υπακούουν τα δεδομένα που εισάγονται. Ο σκοπός της δεν είναι να αποτελέσει μόνο ένα υποκατάστατο της HTML για τη μετάδοση πληροφοριών στο δια-δίκτυο. Οι προθέσεις των δημιουργών της συνοψίζονται στα παρακάτω σημεία:

- Η XML πρέπει να μπορεί να χρησιμοποιηθεί απ' ευθείας στο δια-δίκτυο. Έπρεπε κατά συνέπεια να απλοποιηθεί η δομή της SGML και να ληφθούν υπόψη οι ανάγκες των εφαρμογών που τρέχουν σε δικτυακό περιβάλλον.
- Η XML θα υποστηρίζει ένα ευρύ φάσμα εφαρμογών. Δεν θα περιορίζεται λοιπόν στις δικτυακές εφαρμογές αλλά θα μπορεί να χρησιμεύσει σε μία ευρεία σειρά προγραμμάτων από επεξεργαστές κειμένου έως Βάσεις

δεδομένων.

- Η XML πρέπει να είναι συμβατή με την SGML. Κάθε έγκυρο έγγραφο σε XML είναι και ένα έγκυρο έγγραφο σε SGML εφόσον είναι υποσύνολό της.
- Πρέπει να είναι εύκολο να γράψει κανείς προγράμματα που να επεξεργάζονται έγγραφα σε XML, ώστε να διαδοθεί ευρύτερα και γρηγορότερα.
- Ο αριθμός των προαιρετικών χαρακτηριστικών της XML πρέπει να περιοριστεί κατά το δυνατόν ή στην ιδεατή περίπτωση να εξαλειφθεί τελείως, ώστε να αποφευχθούν τα προβλήματα ασυμβατότητας που είχαν παρουσιαστεί με την SGML.
- Τα έγγραφα σε XML πρέπει να είναι κατανοητά από τον άνθρωπο και ξεκάθαρα. Στο βαθμό που η XML χρησιμοποιεί απλό κείμενο για να ορίσει τα δεδομένα επιτρέπει στο χρήστη να συντάξει τα έγγραφα σε ένα απλό επεξεργαστή κειμένου.
- Ο σχεδιασμός της XML πρέπει να προχωρήσει γρήγορα, για να αποφευχθεί ο κίνδυνος να παραμεριστεί από κάποια άλλη λύση.
- Ο σχεδιασμός της XML πρέπει να είναι τυπικός (formal) και συνοπτικός με σκοπό να υιοθετηθεί ευρύτερα.
- Τα έγγραφα σε XML πρέπει να δημιουργούνται εύκολα.
- Η λακωνικότητα (terseness) στην XML έχει πολύ μικρή σημασία. Η SGML υποστηρίζει μία σειρά τεχνικών οι οποίες περιορίζουν τον όγκο της δακτυλογράφησης, όπως τη δυνατότητα να παραλείπεται ο τελικός όρος.

Η εκμετάλλευση μεγάλου όγκου πληροφορίας και εξαγωγή γνώσης από αυτή, είναι ένα θέμα το οποίο μέχρι πρότινος αντλούσε πληροφορία από βάσεις δεδομένων. Στις μέρες μας η ικανότητα για παραγωγή και συγκέντρωση δεδομένων έχει αυξηθεί. Με την ραγδαία εξέλιξη του παγκόσμιου ιστού και την διαρκή αύξηση των χρηστών του, δημιουργήθηκε μια νέα μορφή δεδομένων, η XML, η οποία είναι εύκολα προσβάσιμη από οποιαδήποτε πλατφόρμα και μεταφέρσιμη μέσω του δια-δικτύου. Η XML αναπτύχθηκε και καθιερώθηκε πολύ γρήγορα στο χώρο, κυρίως λόγω των πλεονεκτημάτων που έχει έναντι στις βάσεις δεδομένων.

Επιγραμματικά αναφέρουμε τα κυριότερα :

- Ένα XML αρχείο είναι αυτό-περιγραφόμενο.
- Είναι αναγνώσιμο.
- Έχει τις ιδιότητες ενός αρχείου (επεξεργασία, μεταφερσιμότητα).
- Ανεξάρτητο πλατφόρμας (Windows, Macintosh, Unix, Linux)

3 ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ
(KDD- KNOWLEDGE DISCOVERY FROM DATABASES)

Η *εξόρυξη γνώσης* (data mining) η οποία ονομάζεται και *ανακάλυψη γνώσης από βάσεις δεδομένων* (knowledge discovery from databases - KDD), είναι μια μη τετριμμένη διαδικασία για τον προσδιορισμό πραγματικών, πρωτότυπων, χρήσιμων και κατανοητών 'προτύπων' (patterns) μέσα στα δεδομένα [19]. Άλλοι όροι που χρησιμοποιούνται για την ονομασία αυτής της διαδικασίας είναι, η Ανασκαφή Γνώσης από Βάσεις Δεδομένων (Knowledge Mining from Databases), η Εξαγωγή Γνώσης (Knowledge Extraction) και η Ανάλυση Δεδομένων (Data Analysis).

Η σημασία των παραπάνω όρων είναι αν όχι η ίδια, ελάχιστα διαφορετική [20] οι διαφορές συνήθως οφείλονται στον τρόπο αντιμετώπισης της κάθε περίπτωσης. Επίσης πρέπει να αναφερθεί ότι το 'εξόρυξη δεδομένων' είναι ένας τομέας που απασχολεί ερευνητές που προέρχονται από πολλά διαφορετικά πεδία της επιστήμης γενικότερα [19]. Στη παρούσα εργασία χρησιμοποιούμε τον όρο ***'Εξόρυξη Γνώσης'*** με την έννοια ότι τεχνικές ***'Εξόρυξης'*** δεδομένων χρησιμοποιούνται για την ανακάλυψη ***'Γνώσης'***.

Μερικά από αυτά είναι τα εξής:

- Βάσεις Δεδομένων (Database Systems).
- Συστήματα Βάσεων Γνώσης (Knowledge Base Systems).
- Τεχνητή Νοημοσύνη (Artificial Intelligence).
- Μηχανική Μάθηση (Machine Learning).
- Πρόσληψη Γνώσης (Knowledge Acquisition).
- Στατιστική (Statistics).
- Χωρικές Βάσεις Δεδομένων (Spatial Databases).
- Οπτικοποίηση Δεδομένων (Data Visualization).

Η κλασσική προσέγγιση στην επεξεργασία δεδομένων βασίζεται ουσιαστικά σε έναν ή περισσότερους ειδικούς αναλυτές, οι οποίοι αποκτούν στενή σχέση με τα δεδομένα και χρησιμεύουν ως ένα είδος 'διεπαφής' μεταξύ των δεδομένων, των χρηστών και των προϊόντων. Ωστόσο γίνεται εύκολα κατανοητό ότι η χειρονακτική επεξεργασία δεδομένων είναι ιδιαίτερα αργή, ακριβή και υποκειμενική. Με τα μεγέθη μάλιστα των δεδομένων να αυξάνονται με δραματικούς ρυθμούς, η χειρονακτική επεξεργασία καθίσταται αδύνατη. Τα

μεγέθη των βάσεων δεδομένων αυξάνονται για δυο λόγους. Πρώτον, αυξάνεται ο αριθμός των εγγραφών, *n*, ή των αντικειμένων, και δεύτερον αυξάνεται ο αριθμός, *d*, των πεδίων ή των χαρακτηριστικών τους αντίστοιχα. Όταν λοιπόν έχουμε να κάνουμε με επεξεργασία εκατομμυρίων εγγραφών, με δεκάδες ή και εκατοντάδες χιλιάδες πεδίων, η αυτόματη επεξεργασία τους κρίνεται κάτι παραπάνω από αναγκαία.

★ Τα *δεδομένα* μπορούμε να ισχυριστούμε ότι αποτελούν μια ακατέργαστη μορφή πληροφορίας την οποία πρέπει να επεξεργαστούμε περαιτέρω, πολλές φορές με την βοήθεια του υπολογιστή.

★ *Πληροφορία*, με την ορθή και ουσιαστική ερμηνεία του όρου, είναι τα δεδομένα τα οποία έχουν οργανωθεί με τέτοιο τρόπο (από τον άνθρωπο ή τον υπολογιστή), ώστε να είναι σημαντικά, αξιόλογα και χρήσιμα.

★ Η *γνώση* θα μπορούσαμε να πούμε ότι είναι μια μορφή πληροφορίας, η οποία βρίσκεται ένα επίπεδο παραπάνω από αυτήν την "ακατέργαστη" ροή δεδομένων και είναι κάτι το τελείως διαφορετικό από την απλή μετάφραση αυτών των απλών μορφών πληροφορίας. Οι τρέχουσες εφαρμογές απαιτούν πιο σύνθετες δομές, όπως διαδικασίες, λειτουργίες, χρονικές ακολουθίες, στόχους, κίνητρα κ.τ.λ. Ο όρος γνώση λοιπόν περιγράφει την ευρύτερη κατηγορία της πληροφορίας που απορρέει και συνεπάγεται από όλα τα παραπάνω.

Σε ένα αφηρημένο επίπεδο η ανακάλυψη γνώσης από βάσεις δεδομένων(KDD) σχετίζεται με την ανακάλυψη και εύρεση μεθόδων και τεχνικών ώστε να δημιουργείται και να εξάγεται *νόημα* από τα δεδομένα. Το KDD είναι ιδιαίτερα χρήσιμο σε περιπτώσεις όπου τα χαμηλού επιπέδου δεδομένα είναι δύσκολο να κατανοηθούν ή και να μεταφραστούν, είτε λόγω του τεράστιου όγκου τους, είτε λόγω της αυξημένης πολυπλοκότητας τους. Εάν τα δεδομένα εξάγονται από ένα ιδιαίτερα σύνθετο πεδίο, η διαδικασία του KDD συνήθως λαμβάνει χώρα και εκτελείται σε μικρού μεγέθους σύνολα δεδομένων, ανάλογα με την πολυπλοκότητα της λειτουργίας/ διαδικασίας που δημιούργησε τα δεδομένα. Στον πυρήνα και στο επίκεντρο της διαδικασίας του KDD, βρίσκεται η εφαρμογή εξειδικευμένων μεθόδων εξόρυξης δεδομένων για την *ανακάλυψη και αυτόματη εξαγωγή συμπερασμάτων* (automated deduction), *προτύπων* και *συσχετίσεων* (associations).

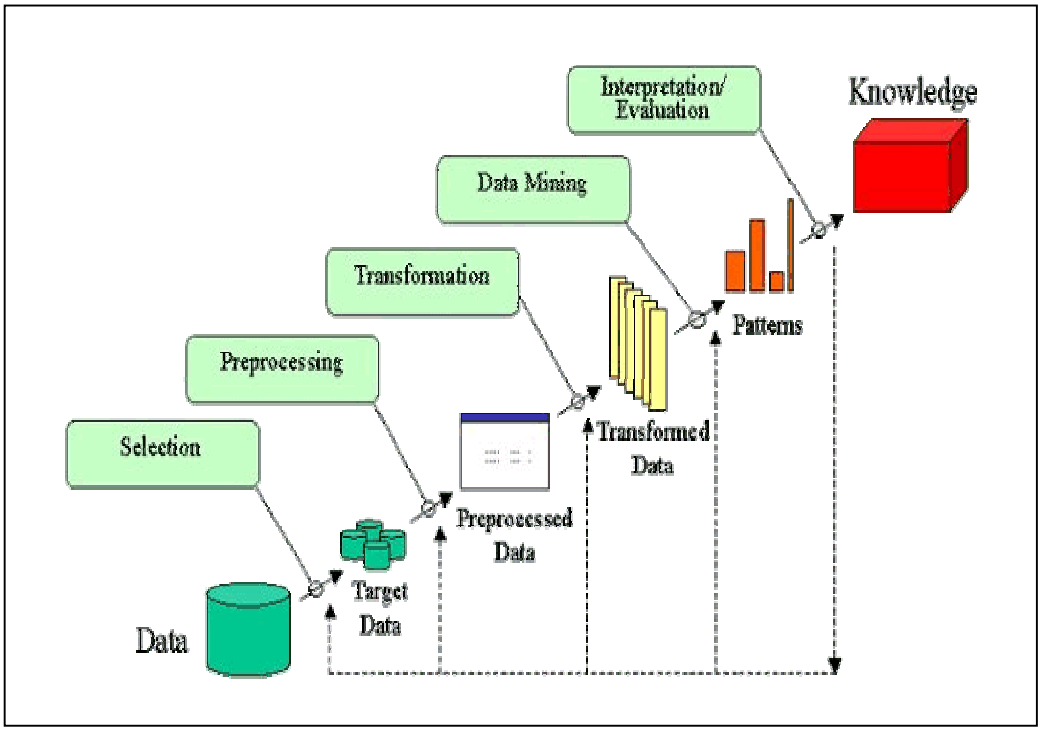
3.1 Η Διαδικασία του KDD

Όπως διακρίνουμε και στο σχήμα 8 η διαδικασία του KDD περιλαμβάνει τα εξής βήματα :

1. *Εφαρμογή* και *κατανόηση* του πεδίου εφαρμογής. Περιλαμβάνει και υποδηλώνει την μελέτη, εκμάθηση και ανάπτυξη σχετικής με το θέμα προηγούμενης γνώσης του χρήστη. Κατανόηση του πεδίου εφαρμογής, τον ξεκάθαρο προσδιορισμό στόχου και αρχικού σκοπού της διεργασίας του KDD, έτσι όπως αυτός διαμορφώνεται και προσδιορίζεται από την πλευρά του χρήστη.
2. Την *δημιουργία* ενός συνόλου *δεδομένων*, το οποίο περιλαμβάνει ένα σύνολο από επιλεγμένα δεδομένα και μεταβλητές πάνω στα οποία θα επικεντρωθεί η διαδικασία της ανακάλυψης και εξόρυξης.
3. *Προ-επεξεργασία* (preprocessing) και "καθαρισμός" των δεδομένων. Στο βήμα αυτό συμπεριλαμβάνονται λειτουργίες όπως απαλλαγή της πληροφορίας από θόρυβο, συγκέντρωση της αναγκαίας και χρήσιμης πληροφορίας και αποφάσεις διαχείρισης και αντιμετώπισης περιπτώσεων όπου λείπουν κάποιες τιμές πεδίων
4. *Ελάττωση* και *προβολή δεδομένων* (data reduction and projection). Η λειτουργία αυτή περιλαμβάνει την εύρεση χρήσιμων χαρακτηριστικών για την αναπαράσταση των δεδομένων. Με την ελάττωση των διαστάσεων ή με μεθόδους μετασχηματισμού, ο αναγκαίος αριθμός μεταβλητών που πρέπει να ληφθούν υπόψη για την αναπαράσταση των δεδομένων μπορεί να ελαττωθεί.
5. Επιλογή της *μεθόδου* που θα χρησιμοποιηθεί για την εκπόνηση του εξόρυξη δεδομένων. Η επιλογή έχει να κάνει με την απόφαση του χρήστη για το εάν θα χρησιμοποιηθεί συνάθροιση, ομαδοποίηση, σύνοψη ή κάποιο άλλο μοντέλο.
6. Επιλογή των *αλγορίθμων* που θα επιτελέσουν το 'εξόρυξη δεδομένων'. Η λειτουργία αυτή είναι υπεύθυνη για την επιλογή των μεθόδων που θα χρησιμοποιηθούν για να αναζητήσουν *"πρότυπες αλληλοσυσχετίσεις"* (association patterns) ανάμεσα στα δεδομένα και θα συνδυάσουν/ ταιριάξουν μια συγκεκριμένη μέθοδο εξόρυξης, με το γενικότερο πνεύμα της διαδικασίας του KDD.
7. *Εξόρυξη δεδομένων* – Περιλαμβάνει την αναζήτηση χαρακτηριστικών του ενδιαφέροντος του χρήστη, είτε σε μια ειδική μορφή αναπαράστασης της πληροφορίας, είτε σε ένα σύνολο από τέτοιες αναπαραστάσεις της, συμπεριλαμβάνοντας την συνάθροιση κανόνων / δένδρων απόφασης, ομαδοποίηση κ.τ.λ.
8. *Επεξήγηση-ερμηνεία* (interpretation). Το στάδιο αυτό περιλαμβάνει την πιθανή επανάληψη των βημάτων 1-7. Επίσης καλύπτει θέματα όπως παρουσίαση των εξαγόμενων χαρακτηριστικών και μοντέλων, αλλά και την

απεικόνιση (visualization) των ίδιων των δεδομένων εισόδου, από τα οποία προκλήθηκαν τα εξαγόμενα αποτελέσματα.

9. *Χρήση* της παραγόμενης *γνώσης*. Το βήμα αυτό περιλαμβάνει απευθείας επαφή και χρήση της προκύπτουσας γνώσης. Αυτό μπορεί να σημαίνει ότι η εξαγόμενη γνώση μπορεί να γίνει είσοδος σε κάποιο άλλο σύστημα για περαιτέρω επεξεργασία (π.χ., σε διαδικασίες multi-strategy learning). Ανάλογα μπορεί να σημαίνει τη μορφοποίηση, δόμηση και παρουσίαση της ληφθείσας γνώσης στο χρήστη με κάποιο έξυπνο και όμορφο τρόπο.
10. Παράλληλα γίνεται και η *αξιολόγηση* του σκοπού που συντελέστηκε το KDD. Η νέα εκμαιευμένη γνώση συχνά χρησιμοποιείται για να τυποποιήσει και να επισημοποιήσει υποθέσεις. Επίσης νέες ερωτήσεις μπορούν να "*γεννηθούν*", χρησιμοποιώντας την επεκτεταμένη και διευρυμένη πλέον γνώση μας πάνω στο αντικείμενο που μελετάμε.



Σχήμα 8 : βασικά βήματα του KDD

3.2 Μορφές Γνώσης

Επαναφέροντας τον προαναφερθέντα ορισμό για το KDD από τον W. Fraweley, η προς ανακάλυψη γνώση είναι *"σιωπηρά εννοούμενη, προηγουμένως άγνωστη και υποθετικά χρήσιμη"*. Με το να είναι 'εννοούμενη' η γνώση, ο ορισμός της επεκτείνεται πέρα από την κλασσική προσέγγιση, σύμφωνα με την οποία η παρουσιαζόμενη γνώση είναι άμεσα κατανοητή και απλά κρατείται

αποθηκευμένη και διαχειρίζεται επιτυχώς από τα DBMS (Data Base Management Systems). Το γεγονός ότι *«ο κ. Χ εργάζεται στην εταιρία Υ και κερδίζει $40.000 ετησίως»*, παρά το πόσο αποκαλυπτικό μπορεί να είναι για το χρήστη, δεν είναι το επιθυμητό αποτέλεσμα για έναν KDD αλγόριθμο.

Το γεγονός αυτό δείχνει ότι αποκαλώντας τη γνώση "προηγουμένως άγνωστη", ο χαρακτηρισμός αναφέρεται τόσο στην προοπτική προσέγγισης από την πλευρά του συστήματος, όσο και στο τρέχων επίπεδο γνώσης του χρήστη. Αυτή η μορφή γνώσης μπορεί να χαρακτηριστεί και ως *μετά-γνώση* (meta-knowledge) και μπορεί να χαρακτηρίζει κρυμμένους νόμους και εμφανίσεις μορφών (structures), οι οποίες δεν χαρακτηρίζονται από ισχυρές συναρτησιακές εξαρτήσεις, αλλά απλά εμφανίζονται με κάποια πιθανότητα. Ο Fraweley στο [21], απαιτεί να λαμβάνει χώρα η διαπίστωση ότι: *η πληροφορία που συνάγεται από τα εκμαιευμένα χαρακτηριστικά είναι απλούστερη από το υποσύνολο της πληροφορίας που εκμαιεύεται από τα ίδια τα αντικείμενα τα οποία την περιγράφουν*.

Ο R. Agrawal στο [22] καθορίζει τρεις τύπους γνώσης προς ανακάλυψη και έρευνα σε βάσεις δεδομένων:

- την **ταξινόμηση** (classification)
- την **συσχέτιση** (associations)
- την **ακολουθία περιοδικών φαινομένων** (sequences of frequent events).

Η *ταξινόμηση* προσπαθεί να χωρίσει τα δεδομένα εισόδου σε ξεχωριστές **κλάσεις** χρησιμοποιώντας τόσο '*supervised*', όσο και '*unsupervised*' (γνωστό ως clustering) μεθόδους μάθησης [23]. Ο στόχος είναι η εύρεση βασικών εννοιών, με όσο το δυνατόν περισσότερο σαφή διαχωριστικά εννοιολογικά σύνορα τα οποία χαρακτηρίζουν μια κλάση αντικειμένων. Έτσι, για μη προκαθορισμένα αντικείμενα σχετικά με την εννοιολογική τους κατηγοριοποίηση, χάρη στην συνάθροιση μπορεί να προβλεφθεί η εννοιολογική κατηγορία και κλάση στην οποία ανήκουν. Ένα super-market για παράδειγμα μπορεί να επιθυμεί την κατηγοριοποίηση των προϊόντων του προκειμένου να αποφασίζει τι προσφορές θα κάνει.

Στο [21] οι W. Frawley και G. Piatetski-Shapiro υποδιαιρούν τη παρούσα λειτουργία σε δυο επιμέρους διαδικασίες:

- Περίληψη-σύνοψη (summarization), όπου αναζητούνται κοινά χαρακτηριστικά για μια κλάση μόνο.
- Διακριτοποίηση (discretization), όπου ο στόχος είναι η εύρεση

χαρακτηριστικών, τα οποία βοηθούν στο διαχωρισμό διαφορετικών κλάσεων ή εναλλακτικά το διαχωρισμό μιας κλάσης από όλες τις υπόλοιπες.

- Όταν ανακαλύπτονται *χρονικές ακολουθίες* (sequences), ο χρόνος, όπως είναι κατανοητό, αποτελεί ένα επιπρόσθετο χαρακτηριστικό. Παραδείγματα τέτοιων εφαρμογών μπορούν να βρεθούν σε αγορές ή σε συμπεριφορές καταναλωτών. Ένα ενδιαφέρον πρόβλημα που αφορά χρονικές ακολουθίες είναι η ανακάλυψη *επεισοδίων* (episodes), δηλαδή συχνά εμφανιζόμενων γεγονότων σε ένα δοθέν διάστημα χρόνου (time-window) [29].

3.3 Εξόρυξη Γνώσεων: Τεχνικές Εξόρυξης Δεδομένων για την Ανακάλυψη Γνώσης

Η 'εξόρυξη δεδομένων' είναι μια μεθοδολογία επίλυσης προβλημάτων, η οποία στόχο έχει να βρει μια τυπική και επίσημη περιγραφή χαρακτηριστικών κάποιων αντικειμένων που προκύπτουν από ένα σύνολο δεδομένων, εν γένει σύνθετης και πολύπλοκης φύσης. Οι Decker και Focardi θεωρούν ότι υπάρχουν διάφορα πεδία αντιπροσωπευτικά για να εφαρμοστούν σε αυτά οι μεθοδολογίες της εξόρυξης δεδομένων, και αναφέρουν ενδεικτικά την ιατρική και τις επιχειρήσεις ως δυο τέτοια πεδία. Η εξόρυξη δεδομένων βασίζεται σε δυο υποθέσεις.

- Πρώτον, ότι οι λειτουργίες που κάποιος θέλει να επιτελέσει και να γενικεύσει, μπορούν να προσεγγιστούν από κάποια απλά υπολογιστικά μοντέλα, σε κάποιο βασικό επίπεδο ακρίβειας.
- Δεύτερον, τα δειγματοληπτικά δεδομένα περιέχουν σε ικανοποιητικό βαθμό την απαιτούμενη πληροφορία για να επιτελεστεί η προσδοκούσα γενίκευση.

Ο Fayyad θεωρεί την εξόρυξη δεδομένων ως την εφαρμογή ειδικών αλγορίθμων για την εξαγωγή ομοιοτήτων και συσχετίσεων από το σύνολο των δεδομένων. Τα πρόσθετα βήματα της διαδικασίας του KDD υπάρχουν για να σιγουρεύουν και να εγγυώνται, όσο το δυνατόν περισσότερο, ότι η πληροφορία που εξάγεται από τα δεδομένα είναι και χρήσιμη. "Τυφλή" εφαρμογή της εξόρυξης δεδομένων γνωστή ως "data dredging", μπορεί εύκολα να οδηγήσει σε παραπλανητική και χωρίς νόημα πληροφορία.

Το τμήμα της εξόρυξης δεδομένων, μέρος της συνολικής διαδικασίας του KDD, οπως είναι φανερό και από το σχήμα 8, συνήθως περιλαμβάνει την επαναλαμβανόμενη εφαρμογή ειδικών και εξειδικευμένων μεθόδων και λειτουργιών. Περιλαμβάνει το αρμονικό ταίριασμα ποικίλων μοντέλων, προκειμένου να παρατηρηθούν αναλυτικά και προσεκτικά τα δεδομένα, ώστε να εκμαιευ\τούν διάφοροι τύποι περιγραφών των χαρακτηριστικών τους. Αυτό

το ταίριασμα των μοντέλων σε τελική ανάλυση περιγράφει και την παραγόμενη και συναγόμενη γνώση. Πολύ συχνά απαιτείται η ανθρώπινη κρίση για να αποφασιστεί εάν τα μοντέλα δείχνουν και μπορούν να παράγουν χρήσιμη και ενδιαφέρουσα γνώση. Δυο μαθηματικοί φορμαλισμοί χρησιμοποιούνται για την αναπαράσταση των μοντέλων, η *στατιστική* και η *λογική*. Ένα μη-ντετερμινιστικό μοντέλο υιοθετείται στην στατιστική προσέγγιση, ενώ η χρήση λογικής συνεπάγεται την χρήση ενός καθαρά ντετερμινιστικού μοντέλου.

Η *στατιστική* προσέγγιση στο χώρο της εξόρυξης δεδομένων είναι περισσότερο διαδεδομένη κυρίως για πρακτικές εφαρμογές, καθώς τα πραγματικά δεδομένα είναι συνήθως συσχετισμένα και ταυτισμένα με έναν σημαντικό βαθμό αβεβαιότητας. Οι περισσότερες λειτουργίες της εξόρυξης δεδομένων είναι βασισμένες σε καλά τεκμηριωμένες και αναπτυγμένες τεχνικές από το χώρο και το πεδίο της Μηχανικής Μάθησης, της Αναγνώρισης Προτύπων και της Στατιστικής (όπως ομαδοποίηση, συνάθροιση κ.τ.λ).

Η εξόρυξη δεδομένων αποτελείται από δυο βασικούς στόχους: *την πρόγνωση/ πρόβλεψη (prediction) και την περιγραφή (description).* Η περιγραφή επικεντρώνεται στην εύρεση ερμηνεύσιμων χαρακτηριστικών τα οποία είτε καθορίζουν ποσοτικά τα υπάρχοντα δεδομένα, είτε ανακαλύπτουν βασικές ιδιότητες ανάμεσα στα δεδομένα. Η πρόγνωση αναφέρεται στην αντιστοίχηση μιας τιμής με μια μεταβολή του ενδιαφέροντος μας, για μια μελλοντικά παρουσιαζόμενη εμφάνισή της. Αν και τα όρια ανάμεσα σε αυτές τις δυο λειτουργίες δεν είναι ξεκάθαρα και ακριβή, ο διαχωρισμός τους είναι αρκετά βοηθητικός στην κατανόηση του συνολικού στόχου και σκοπού της ανακάλυψης γνώσης.

Οι στόχοι τους μπορούν να υλοποιηθούν χρησιμοποιώντας μια ποικιλία από μεθόδους, ειδικές για εξόρυξη δεδομένων, όπως είναι η ομαδοποίηση, η συνάθροιση, η σύνοψη και η αλληλεξάρτηση μοντέλων.

- Η ταξινόμηση (classification) είναι μια τεχνική μηχανικής μάθησης η οποία αντιστοιχίζει ένα αντικείμενο εισόδου σε μια ή περισσότερες προκαθορισμένες ομάδες (classes).
- Η ομαδοποίηση (clustering) χρησιμοποιείται για να καθοριστούν επακριβώς, είτε σαφή και ακριβή, είτε επικαλυπτόμενα υποσύνολα ανάμεσα στα δεδομένα, γεγονός που οδηγεί προφανώς σε βέλτιστη περιγραφή.
- Η παλινδρόμηση (regression) έχει να κάνει με την αντιστοίχηση ενός δεδομένου (data item), σε μια πραγματικά μετρούμενη μεταβλητή.

- Η σύνοψη αποτελείται από διάφορες μεθόδους με σκοπό την ανακάλυψη ενιαίων και "συμπυκνωμένων" χαρακτηριστικών των δεδομένων.

3.4 Τεχνικές Εξόρυξης Γνώσεων

Περιληπτικά και συνοπτικά θα αναφέρουμε μερικές δημοφιλείς τεχνικές mining, όπως είναι (α) τα δένδρα απόφασης (decision trees and rules), (β) οι μέθοδοι ομαδοποίησης (linear regression and classification methods), (γ) οι μέθοδοι βασισμένοι σε ομοιότητες (similarity-based methods), (δ) τα πιθανοκρατικά μοντέλα (probabilistic models) και (ε) τα σχεσιακά μοντέλα εκμάθησης (relational learning models). Η κατανόηση τους σκοπό έχει να βοηθήσει το χρήστη στην βέλτιστη επιλογή μοντέλου ανά πρόβλημα και περίσταση.

Τα δένδρα απόφασης αποτελούνται από κόμβους και ακμές. Κάθε κόμβος περιέχει μια κατάσταση από κάποια χαρακτηριστικά των δεδομένων. Τα δένδρα απόφασης παράγουν ομαδοποιήσεις οι οποίες γίνονται εύκολα κατανοητές και παράγουν σύντομα περιεκτικά μοντέλα. Ωστόσο ο περιορισμός σε ένα συγκεκριμένο δένδρο μπορεί να περιορίσει τη λειτουργικότητα του μοντέλου. Τα δένδρα απόφασης και οι αντίστοιχα παραγόμενοι κανόνες συνήθως χρησιμοποιούνται για λειτουργίες πρόγνωσης, ομαδοποίησης και σύνοψης.

Οι βασισμένες σε μετρικές ομοιότητας μέθοδοι χρησιμοποιούν αντιπροσωπευτικά παραδείγματα για να πιστοποιήσουν και να αποδείξουν την ισχύ ενός μοντέλου. Οι ιδιότητες και τα χαρακτηριστικά νέων παραδειγμάτων προβλέπονται και επιβεβαιώνονται, από τα προηγουμένως γνωστά χαρακτηριστικά γνωστών παραδειγμάτων. Αυτή η μέθοδος έχει αποδειχτεί ιδιαίτερα χρήσιμη στο πεδίο της βιολογίας.

Οι βασισμένες στο νόμο του Bayes μέθοδοι παρέχουν ένα φορμαλισμό για την υποψία ισχύος κάποιων θεωρήσεων υπό συνθήκες. Οι μέθοδοι στηρίζονται βασικά στο θεώρημα και τον τύπο του Bayes, ο οποίος είναι θεμελιώδης στην θεωρία των πιθανοτήτων και χρησιμοποιεί δεσμευμένες πιθανότητες. Καθώς είναι ιδιαίτερα σημαντικό να επεξεργαζόμαστε αβέβαια γεγονότα και ενδεχόμενα, οι μέθοδοι αυτές είναι αρκετά χρήσιμες.

Τα σχεσιακά μοντέλα εκμάθησης (Relational Learning; Συμπεριλαμβανομένων των συστημάτων Επαγωγικής Λογικής- Inductive Logic Programming – ILP) συνδυάζουν την λογική πρώτης τάξης με μεθόδους αυτόματου προγραμματισμού και μηχανικής μάθησης. Τα σχεσιακά μοντέλα μπορεί να έχουν μεγάλη ισχύ στον τομέα της αναπαράστασης της πληροφορίας, ωστόσο το γεγονός αυτό αντισταθμίζεται από το σημαντικά αυξημένο κόστος

αναζήτησης λύσεων.

3.5 Μηχανική Μάθηση και Εξόρυξη Δεδομένων

Θα μπορούσε να υποστηρίξει κάποιος ότι πολλές από τις προσεγγίσεις πού γίνονται σε βασικά θέματα της εξόρυξης δεδομένων δεν διαφέρουν δραματικά από τα πρότυπα προβλήματα της Μηχανικής Μάθησης. Ωστόσο, το γεγονός ότι χρησιμοποιείται μια βάση δεδομένων -ανεξάρτητα από την μορφή της- ως πηγή δεδομένων, δημιουργεί πρόσθετες δυσκολίες. Ενώ στην Μηχανική Μάθηση το μέγεθος των προς επεξεργασία δεδομένων ανά περίπτωση και πρόβλημα, σπάνια ξεπερνά τις μερικές χιλιάδες στοιχεία, οι σύγχρονες πηγές δεδομένων στις περισσότερες περιπτώσεις ξεπερνούν τις εκατοντάδες χιλιάδες ή και τα εκατομμύρια διαφορετικών καταγραφών στοιχείων/ δεδομένων, με συνεχώς αυξανόμενη τάση. Το γεγονός αυτό δημιουργεί τεράστιο πρόβλημα στη διαχείριση τόσο των ίδιων δεδομένων, αλλά πολύ περισσότερο και των ενδιάμεσων αποτελεσμάτων που προκύπτουν κατά την επεξεργασία τους.

Επιπρόσθετα, οι περισσότεροι από τους πρόσφατα προτεινόμενους σχετικούς αλγόριθμους και τεχνικές αντιμετωπίζουν μια βάση δεδομένων σαν ένα καθολικό πίνακα (universal relation). Αυτή η υπόθεση επιβαρύνει ακόμη περισσότερο το δημιουργημένο πρόβλημα λόγο μεγέθους των δεδομένων. Επειδή οι βάσεις οι οποίες έχουν διαχωριστεί σε επιμέρους υπό-πίνακες υφίστανται την εφαρμογή κάποιας κανονικής μορφής με σκοπό την ελάττωση του χώρου αποθήκευσης τους, μέσω την ένωσης τους (join) δημιουργούνται ακόμη μεγαλύτεροι πίνακες, κάνοντας το πρόβλημα ακόμη μεγαλύτερο και μάλλον δισεπίλυτο.

Ένα ακόμη πρόβλημα είναι ότι οι βάσεις δεδομένων δεν δημιουργούνται και δεν συντηρούνται για να εξυπηρετήσουν τις προσδοκίες και τους σκοπούς της εξόρυξης δεδομένων. Τα δεδομένα τους σκοπό έχουν να εξυπηρετήσουν το λόγο ύπαρξης της εκάστοτε εφαρμογής και όχι να διευκολύνουν τις εργασίες που αφορούν τη δυνατότητα μετέπειτα επεξεργασίας τους. Σε περιπτώσεις όπου κάποια ουσιώδη χαρακτηριστικά για τη διαδικασία της ανακάλυψης λείπουν, είναι δυνατόν να προκύψουν άσχημα ή ακόμα και λάθος αποτελέσματα. Μεγάλο είναι το δίλημμα που προκύπτει όταν πρέπει να ληφθούν αποφάσεις για τον χειρισμό 'NULL' (άγνωστων, μη καταχωρημένων) τιμών, είτε ακαθόριστων για κάποιο λόγο τιμών. Οι εναλλακτικές λύσεις είναι, είτε να τις αντικαταστήσουμε με κάποιες εξορισμού τιμές που καθορίζονται από κάποιες πιθανότητες, βασισμένες στις ήδη υπάρχουσες και διαθέσιμες τιμές, είτε να αδιαφορήσουμε τελείως γι' αυτές. Με τη χρήση της XML μπορούν να

αυτόματα να δοθούν εξορισμού τιμές σε αντικείμενα. Τόσο στη μια όσο και στην άλλη περίπτωση υπάρχει ο κίνδυνος να οδηγηθούμε σε λανθασμένα αποτελέσματα.

Τα πραγματικά δεδομένα πολλές φορές είναι εμπλουτισμένα με θόρυβο, ή περιέχουν αντιφατικές πληροφορίες που οφείλονται είτε σε λανθασμένες καταχωρήσεις/ εισαγωγές (data entry), είτε στην ίδια την φύση των δεδομένων. Κάτι τέτοιο όμως δεν είναι επιθυμητό και δεν αποτελεί την καλύτερη δυνατή είσοδο για τους πρότυπους αλγόριθμους μάθησης. Έτσι πολλές φορές χρειάζονται πιθανοκρατικές λύσεις για την αντιμετώπιση τέτοιων δυσκολιών.

4 ΤΟ ΣΥΣΤΗΜΑ HEALTHOBS

Στο κεφάλαιο αυτό παρουσιάζουμε την μορφή των δομών που προκύπτουν από την επεξεργασία των XML αρχείων, εξηγούμε τα οφέλη που προκύπτουν από την χρήση αυτών των δομών και κατ' επέκταση δικαιολογούμε και υποστηρίζουμε την επιλογή χρήσης του προτύπου XML. Έπειτα αναλύουμε τους αλγόριθμους που υλοποιήσαμε και εξηγούμε τις δομές αποθήκευσης των δεδομένων και τον τρόπο υλοποίησης.

4.1 Διακριτοποίηση Αριθμητικών Τιμών

Αν και μπορούμε να έχουμε πολλές μορφές τύπων (και κατά συνέπεια τιμών) στα δεδομένα μας τις περισσότερες φορές επικεντρωνόμαστε σε αριθμητικά στοιχεία. Ιδίως όταν το πεδίο εφαρμογής είναι ιατρικά δεδομένα ενδιαφερόμαστε για την σημασία των τιμών αυτών.

Για παράδειγμα η τιμή της χοληστερίνης για έναν άντρα θεωρείται υψηλή αν είναι πάνω από 45. Κατά συνέπεια στο αντικείμενο χοληστερίνη μεγαλύτερη του 45 (δηλαδή υψηλή) ή μικρότερη του 45 (δηλαδή κανονική).

Οπότε είναι αναγκαίος ένας τρόπος διακριτοποίησης κάποιων αριθμητικών τιμών από τα δεδομένα μας. Ο I. Cengiz [38] πρότεινε έναν αλγόριθμο για να εξαγάγει κανόνες συσχέτισης από στοιχεία που περιέχουν και λογικά και αριθμητικά στοιχεία. Μια αριθμητική τιμή μπορεί να χωριστεί σε διαστήματα και κάθε διάστημα μπορεί να θεωρηθεί ένα λογικό αντικείμενο. Έτσι το ποσοτικό πρόβλημα μπορεί να χαρτογραφηθεί σε πρόβλημα λογικών τύπων. Υιοθετήσαμε αυτήν την τεχνική και δημιουργήσαμε μια φιλική προς το χρήστη εφαρμογή, τον 'DOMAIN EDITOR', η οποία χειρίζεται αριθμητικά στοιχεία των αρχείων XML. Στο κεφάλαιο 4.4 θα παρουσιάσουμε αναλυτικά τον domain-editor και θα δούμε τις λειτουργίες, τον τρόπο χρήσης του καθώς και τις δυνατότητες που προσφέρει για την εύκολη προσαρμογή διαφορετικών πεδίων εφαρμογής στο περιβάλλον του συστήματος HealthObs.

4.2 Ομογενοποίηση Ετερογενών Πληροφοριών

Προφανώς, η συλλογή της πληροφορίας δεν είναι αυτοσκοπός. Αυτό που είναι επιθυμητό είναι η αξιοποίησή της, άρα η δυνατότητα για εξαγωγή χρήσιμων και κατανοητών συμπερασμάτων. Ένας σημαντικός στόχος της πληροφορικής είναι να παρέχει στους τελικούς χρήστες ολοκληρωμένες υπηρεσίες υψηλού

επιπέδου. Για να επιτευχθεί όμως κάτι τέτοιο πρέπει να είναι δυνατή η συγκέντρωση όλης της σχετικής πληροφορίας σε ενιαία μορφή, έτσι ώστε να καταστεί δυνατή η αυτοματοποιημένη επεξεργασία της και ανάλυση της.

Μια πολλά υποσχόμενη προσέγγιση σε αυτό το πρόβλημα ολοκλήρωσης είναι η απόκτηση κεντρικού ελέγχου έναντι των επιμέρους οργανισμών και πληροφοριακών πηγών σε ένα επίπεδο 'μετά-δεδομένων' (meta-data), διατηρώντας ταυτόχρονα την αυτονομία των ατομικών συστημάτων στο επίπεδο των 'ατομικών εγγραφών δεδομένων' (data instance level). Ο αντικειμενικός στόχος του μοντέλου αυτού είναι η επίτευξη ολοκλήρωσης της πληροφορίας/ δεδομένων των κατανεμημένων ετερογενών συστημάτων, ενώ παράλληλα επιτρέπεται σε αυτά τα συστήματα να λειτουργούν ανεξάρτητα και ταυτόχρονα.

Εντούτοις, η επίτευξη ολοκλήρωσης σε σημασιολογικό (semantic) επίπεδο αποτελεί ένα μείζον πρόβλημα, καθότι η λογική, η γνώση, και οι μορφές δεδομένων που χρησιμοποιούνται στα διάφορα συστήματα είναι σύνθετες και συχνά ασύμβατες. Επιπρόσθετα, όσο περισσότερο επιθυμεί και προσπαθεί να κρύψει κάποιος την ετερογενή φύση των δεδομένων, τόσο περισσότερο ασχολείται και εμπλέκεται με θέματα ολοκλήρωσης. Έτσι μια ρεαλιστική λύση είναι: *η απόκρυψη της ετερογένειας στο κορυφαίο επίπεδο (δηλ., διεπαφή με τον χρήστη), κάνοντας ταυτόχρονα τις επιμέρους πηγές της πληροφορίας να εμφανίζονται στους τελικούς χρήστες σαν μια τεράστια συλλογή από αντικείμενα που συμπεριφέρονται ομοιόμορφα*.

Βασικό κριτήριο για την επιτυχία μιας υπηρεσίας πληροφόρησης η οποία προσπελαύνει και ανακλά δεδομένα από κατανεμημένες πηγές πληροφόρησης, είναι η ικανότητα της να χειρίζεται αποδοτικά και έξυπνα την ετερογενή φύση της εκεί αποθηκευμένης πληροφορίας. Ας θεωρήσουμε ως παράδειγμα, στα πλαίσια μιας εφαρμογής, μια ιατρική βάση δεδομένων η οποία βρίσκεται σε διαφορετικές γεωγραφικές περιοχές. Εάν η εφαρμογή δεν διαθέτει ένα μοναδικό και ενοποιημένο σχήμα κωδικοποίησης για διαγνώσεις, τότε υπάρχει η περίπτωση οι ανεξάρτητες και αυτόνομες πηγές αποθήκευσης να έχουν καταγράψει με διαφορετικό όνομα και κωδικό την ίδια διάγνωση. Για παράδειγμα σε μια καρδιολογική μονάδα το όνομα για την 'πίεση' μπορεί να έχει εγγραφεί ως "DIASTOLICPRESSURE", ενώ σε μια άλλη ως "PRESSURE". Σε μια τέτοια περίπτωση, μολονότι έχουμε πρόσβαση και στις δυο πηγές πληροφόρησης, δεν είναι καθόλου προφανές πως μπορούμε να συσχετίσουμε στοιχεία μεταξύ των δύο βάσεων.

Προκειμένου να αντιμετωπιστούν αυτά τα προβλήματα, μια πολυσύνθετη

διαδικασία ολοκλήρωσης (multi-phase data integration) θα πρέπει να ακολουθηθεί, προτού εφαρμοστούν οι διαδικασίες εξόρυξη δεδομένων και μηχανικής μάθησης επάνω στα δεδομένα. Η επιτυχία ενός μέσου ομογενοποίησης της πληροφορίας βασίζεται στην δυνατότητα του να αντιμετωπίσει την ετερογενή φύση της πληροφορίας. Προϋπόθεση είναι η ενσωμάτωση μιας εξαρτώμενης από το πεδίο *οντολογίας*.

4.3 Η Υπηρεσία Κοινής Ορολογίας στο HealthObs

(CCTR – Common Term Reference Service)

Ο μοναδικός τρόπος για να αντιμετωπίσουμε προβλήματα τέτοιας προέλευσης, είναι να συμπεριλάβουμε και να ενσωματώσουμε στο σύστημα μας ένα εξειδικευμένο (ιατρικό στην περίπτωση μας) πεδίο οντολογίας. Για αυτό το σκοπό έχουμε αναπτύξει μια υπηρεσία για την αποθήκευση και την ανάκτηση των κοινώς και παγκοσμίως αποδεκτών ονομάτων και κωδικών ιατρικών όρων - το Common Term Reference Service (CTRS).

Η υπηρεσία εκμεταλλεύεται και χρησιμοποιεί τους όρους και τις σχέσεις από το ICD [55] (διεθνής κωδικοποίηση των ασθενειών), και από το ICPC [56] (διεθνής συνάθροιση συμπτωμάτων για τη πρωτοβάθμιας περίθαλψη). Έχοντας ως γνώμονα τα ICD9 και ICPC, διαφορετικά λεξικά για διαφορετικές γλώσσες και για τα διαφορετικά κλινικά συστήματα πληροφοριών μπορούν να διαμορφωθούν εύκολα και να προσαρμοστούν. Επιπλέον, όταν έχουμε δεδομένα από ιατρικά εργαστήρια (κυρίως μικροβιολογικά εργαστήρια) είναι κρίσιμο να αναθέσουμε τις αριθμητικές εργαστηριακές μετρήσεις σε ποιοτικά σταθμά. Για παράδειγμα αν η τιμή της 'DHL Χοληστερίνης' είναι '50' θεωρείται «Υψηλή» για έναν άνδρα.

Συνοψίζοντας η υπηρεσία αυτή προσφέρει:

- ✓ Αντιστοίχηση διαφόρων *συνωνύμων* όρων σε ένα κοινά αποδεκτό όρο - καθιστώντας ολόκληρο το περιβάλλον εύκολα προσαρμόσιμο στα ετερογενή συστήματα πληροφοριών.
- ✓ Ανάθεση των εργαστηριακών μετρήσεων σε αντίστοιχες κατηγορίες τιμών και αντίστοιχες *διακριτές τιμές* (π.χ., 'χαμηλή', 'κανονική', 'υψηλή').
- ✓ Αντιστοίχηση των ασθενειών/συμπτωμάτων στα τυποποιημένα ονόματα αναφοράς τους μέσω των *ICD9/ICPC*.

Η υπηρεσία CTRS υλοποιεί την αποδοτική χρήση ενός κατάλληλα διαμορφωμένου αρχείου κειμένου με απλή και συγκεκριμένη δομή (Σχήμα 9).

```
#####BIOCHEMICAL_EXAM### Comment

CHOLESTEROL DHLCHOLESTEROL
CALCIUM 9-10.5 CALCIUM_NORMAL
CALCIUM 10.5-10000 CALCIUM_HIGH
CALCIUM 0-9 CALCIUM_LOW
if(PATIENT/GENDERID=1) FERRITIN 39-340 FERRITIN_NORMAL
if(PATIENT/GENDERID=2) FERRITIN 15-140 FERRITIN_NORMAL
```

Σχήμα 9: Παράδειγμα αρχείου για την υλοποίηση της υπηρεσίας CTRS

- Το σύμβολο # δηλώνει ότι ολόκληρη η γραμμή είναι σχόλιο. Αν θέλουμε να αντιστοιχήσουμε συνώνυμα με ένα κοινό όρο, γράφουμε τον κοινό όρο και μετά τα συνώνυμα του (όπως στο παράδειγμα 'CHOLESTEROL DHLCHOLESTEROL'). Όταν θέλουμε να αναθέσουμε ποσοτικές τιμές σε ποιοτικές τιμές γράφουμε <όνομα ποσοτικού αντικειμένου> <πεδίο τιμών> <όνομα αντικειμένου>_<κατηγορία> (όπως στο παράδειγμα 'CALCIUM 0-9 CALCIUM_LOW').
- Πολλές φορές υπάρχει περίπτωση η αντιστοίχηση μιας αριθμητικής τιμής ενός αντικειμένου σε κατηγορία να εξαρτάται από ένα άλλο αντικείμενο. Σε αυτή την περίπτωση προσθέτουμε στην αρχή το αναγνωριστικό if και τη συνθήκη που πρέπει να ισχύει, δηλαδή την τιμή που πρέπει να έχει το αντικείμενο από το οποίο εξαρτάται η τιμή μας. Για παράδειγμα if(PATIENT/GENDERID=1) FERRITIN 39-340 FERRITIN_NORMAL.

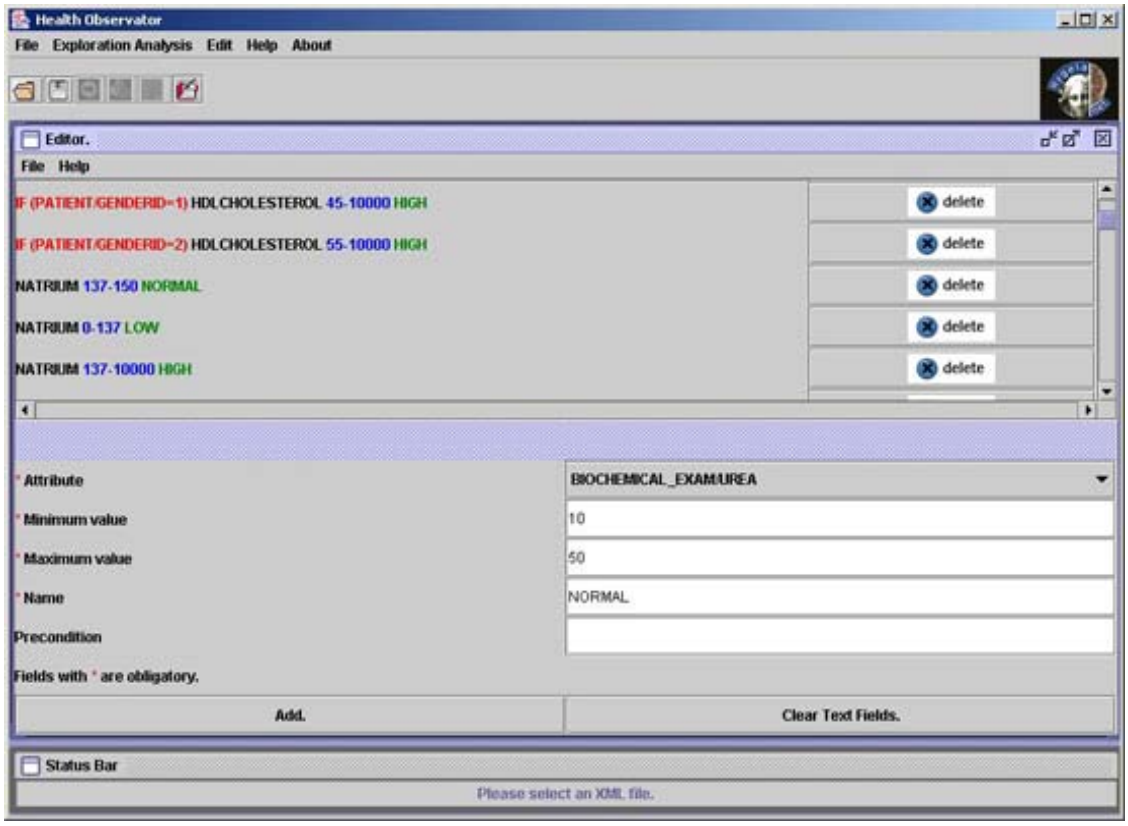
Η μέθοδος της ομογενοποίησης με τη χρήση αρχείου κειμένου βοηθάει στην εύκολη κατανόηση των συσχετίσεων και εξαρτήσεων από απλούς χρήστες. Παρόμοια είναι η προσέγγιση του Synapses [57], ένα ερευνητικό έργο που δίνει την ευκαιρία σε επαγγελματίες του χώρου της υγείας να μοιράζονται δεδομένα και πληροφορίες από σχετικούς ηλεκτρονικούς φακέλους υγείας. Συγκεκριμένα η υπηρεσία 'Manage Patient Indexing' του Synapses έχει υλοποιηθεί με τον ίδιο τρόπο.

4.4 Domain Editor

Η υπηρεσία CTRS μπορεί να χρησιμοποιηθεί όχι μόνο για τον ορισμό της σημασιολογίας του ανά-χείρας πεδίου εφαρμογής, αλλά και για τη *προσαρμογή* διαφορετικών πεδίων εφαρμογής στο περιβάλλον του HealthObs. Προς αυτή τη κατεύθυνση έχει υλοποιηθεί μία ειδική εφαρμογή/εργαλείο, μέρος της συνολικής εργασίας, η οποία τροποποιεί (προσθέτει/διαγράφει) στοιχεία και σχέσεις από το αρχείο κειμένου στο οποίο στηρίζεται η CTRS. Σκοπός του είναι να αποκρύπτει την πολυπλοκότητα της σύνταξης και η δυνατότητα χρήσης από διαφορετικού επιπέδου χρήστες.

Έχοντας ως γνώμονα τα πρότυπα κωδικοποίησης ICD9 και ICPC, διαφορετικά λεξικά για διαφορετικές γλώσσες και για τα διαφορετικά κλινικά συστήματα πληροφοριών μπορούν να διαμορφωθούν εύκολα και να προσαρμοστούν. Επιπλέον, όταν έχουμε δεδομένα από ιατρικά εργαστήρια (π.χ., μικροβιολογικά ή αιματολογικά) είναι κρίσιμο να αναθέσουμε τις αριθμητικές εργαστηριακές μετρήσεις σε *ποιοτικά διαστήματα*.

Η ανάγκη για διαμόρφωση τέτοιου είδους λεξικών και για προσαρμογή λεξικών από ειδικούς στο πεδίο της ιατρικής, μας ώθησε στη δημιουργία ενός εργαλείου εύχρηστου και απλού.



Σχήμα 10. Το HCI του Domain Editor

Αφού έχουμε επιλέξει το αρχείο που θέλουμε να τροποποιήσουμε, τα υπάρχοντα διαστήματα εμφανίζονται στην κορυφή της εφαρμογής, όπως παρουσιάζει το Σχήμα 10.

- Επιλέγοντας από τη γραμμή μενού το **'Edit'** και μετά το υπο-μενού **'interval'** (ή επιλέγοντας το αντίστοιχο κουμπί στην γραμμή εργαλείων) εμφανίζεται ένα παράθυρο διαλόγου που μας ζητάει να ανοίξουμε ένα ήδη υπάρχον λεξικό ή να δημιουργήσουμε ένα νέο με την ονομασία που θα επιλέξουμε.

- Με 'μαύρα' γράμματα έχουμε την ονομασία του στοιχείου, με το 'μπλε' η ελάχιστη και μέγιστη τιμή του αντίστοιχου διαστήματος και με 'πράσινο' το όνομα του διαστήματος. Αν δεν υπάρχουν τιμές τότε έχουμε αντιστοίχηση *συνωνύμων* σε ένα κοινό όρο.

- Εάν θέλουμε να αφαιρέσουμε ένα υπάρχον διάστημα ενός στοιχείου πατάμε απλά στο αντίστοιχο κουμπί δεξιά που γράφει '**Delete**'.

- Στο κάτω σημείο της εφαρμογής εμφανίζεται μια φόρμα που μας επιτρέπει να προσθέσουμε νέα διαστήματα.

- Ο τελευταίος τομέας (Precondition) είναι προαιρετικός. Μερικά διαστήματα εξαρτώνται από άλλες τιμές στοιχείων. Παραδείγματος χάριν το στοιχείο `'HDLCholesterol'` είναι υψηλό για έναν άνδρα όταν είναι πάνω από `'45'` και για μια γυναίκα όταν είναι πάνω από `'55'`(δηλαδή, τα σχετικά διαστήματα για ενήλικες ανάλογα με το φύλο τους, και τα οποία χρησιμοποιούνται από οποιοδήποτε μικροβιολογικό/αιματολογικό εργαστήριο). Μπορούμε να προσδιορίσουμε αυτή την εξάρτηση με τον τομέα προϋπόθεσης (Precondition).

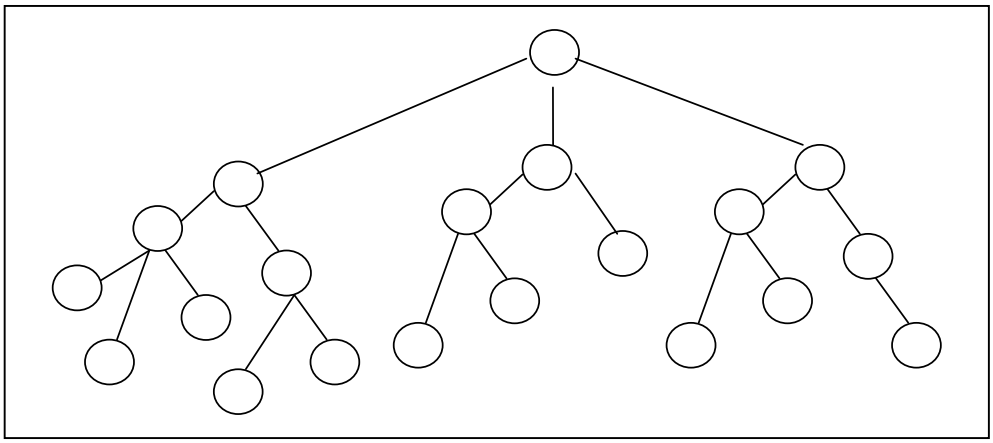
Με βάση τη παραπάνω περιγραφή είναι προφανές ότι μπορούμε πολύ εύκολα να προσαρμόσουμε οποιοδήποτε πεδίο εφαρμογής στο σύστημα HealthObs. Για παράδειγμα, αν έχουμε πρόσβαση σε κλινικά δεδομένα από καρδιολογικές κλινικές τότε το μοντέλο-δεδομένων και σημασιολογίας του ιατρικού πεδίου της καρδιολογίας μπορεί πολύ εύκολα (από έμπειρους ή μη-έμπειρους χρήστες) να ενσωματωθεί.

Γονιδιακά/γενετικά μοντέλα δεδομένων, και σχετικές σημασιολογίες μπορούν επίσης να ενσωματωθούν, διευκολύνοντας τη δημιουργία ενός ενιαίου ***κλινικο-γονιδιωματικού/γενετικού μοντέλου*** δεδομένων – μια απαίτηση

η οποία έχει τεθεί τα τελευταία χρόνια με στόχο την εξατομικευμένη ιατρική (individualized medicine) [60]. Τα σχετικά ερευνητικά και αναπτυξιακά ζητήματα αποτελούν στόχους μελλοντικής εργασίας.

4.5 Αναπαράσταση Δεδομένων στο HealthObs

Η διαδικασία επεξεργασίας της πληροφορίας (filtering, indexing, pre-processing, parsing) με σκοπό την εξαγωγή χρήσιμης πληροφορίας, λαμβάνει χώρα εξολοκλήρου "πάνω" στα παραγόμενα XML αρχεία. Για το λόγο αυτό έχει σχεδιαστεί και υλοποιηθεί ένας parser σε γλώσσα προγραμματισμού java (όπως και όλη η υλοποίηση του συστήματος). Ο κύριος λόγος που επιλέχτηκε αυτή η γλώσσα προγραμματισμού είναι η ιδιότητα της να δημιουργεί εφαρμογές ανεξάρτητες πλατφόρμας.



Σχήμα 11: Η 'δενδροειδής' δομή XML εγγράφων

Ο parser δέχεται σαν είσοδο το XML αρχείο και το αντίστοιχο DTD αρχείο. Ο parser διαβάζει αρχικά το DTD αρχείο και αποθηκεύει τη δομή του XML αρχείου. Αμέσως μετά αρχίζει η «ανάλυση» του XML αρχείου. Τα επιλεγμένα ,από το χρήστη, στοιχεία (στο επόμενο κεφάλαιο θα μιλήσουμε για τον τρόπο με τον οποίο ο χρήστης επιλέγει τα στοιχεία τα οποία θα συμμετέχουν στους αλγόριθμους εξόρυξης δεδομένων) αποθηκεύονται σε μία δενδρική δομή. Η δομή αυτή ,εκτός από την αποθήκευση τιμών των αντικειμένων (attribute values), αποθηκεύει και την *ημι-δομημένη* (semi-structured) μορφή των XML αρχείων. Στη μεταφορά αυτή κάνοντας κατάλληλο 'φιλτράρισμα', διατηρούμε μόνο την αναγκαία προς επεξεργασία πληροφορία, χωρίς να επιβαρύνουμε τη δομή με πληροφορία που δεν πρόκειται να επεξεργαστούμε. Λέγοντας αναγκαία

πληροφορία εννοούμε μόνο τις τιμές των αντικειμένων που έχει επιλέξει ο χρήστης. Στην δομή που αναφέραμε κρατείται όλη η πληροφορία που διαβάζεται από το XML αρχείο. Στο σχήμα 11 φαίνεται η δομή αποθήκευσης του XML αρχείου.

Η «ρίζα» (root) του δένδρου υποδηλώνει το πεδίο αναφοράς του XML αρχείου. Κάθε υπο-δένδρο της ρίζας είναι μία μοναδική πλειάδα τιμών που αντιπροσωπεύει ένα περιστατικό των δεδομένων ή απλά μία «συναλλαγή» (transaction) των δεδομένων. Η δομή κάθε συναλλαγής είναι ένα υποσύνολο της DTD δομής του XML. Το γεγονός ότι κάθε συναλλαγή μπορεί να έχει τιμές σε ένα μέρος μόνο του συνόλου του πεδίου δικαιολογεί την ασυμμετρία που μπορεί να έχει το δένδρο. Οι κόμβοι του δένδρου που είναι και φύλα (δηλαδή καμία ακμή δεν ξεκινάει από αυτά) περιέχουν την τιμή (αλφαριθμητική ή αριθμητική τιμή, υπερ-συνδέσεις (hyperlinks), μορφές εικόνας προσπελάσιμες ως αρχείο κειμένου κλπ) ενός πεδίου. Οι κόμβοι από το φύλο έως τη ρίζα του δένδρου περιέχουν την πληροφορία του πεδίου. Με αυτό τον τρόπο έχουμε αποθηκεύσει όλη την πληροφορία που μας παρέχεται από τα ημι-δομημένα δεδομένα (ιεραρχία αντικειμένων, σύνθετες τιμές).

Η υλοποίηση του parser έγινε σε γλώσσα προγραμματισμού java. Συγκεκριμένα υλοποιήθηκε με την «διεπαφή» (interface) SAX της java. Ο λόγος που επιλέχτηκε αυτή η διεπαφή και όχι η DOM (της java) είναι ο τρόπος προσπέλασης των αρχείων. Προσφέρει ένα σειριακό μηχανισμό προσπέλασης των XML αρχείων σε αντίθεση με τον DOM που αποθηκεύει όλο το αρχείο στη μνήμη (με διάφορες δομές) για την επεξεργασία, πρόσθεση ή διαγραφή δεδομένων. Δεδομένου ότι ο όγκος πληροφορίας πρέπει να είναι πολύ μεγάλος για να έχουμε ενδιαφέροντα και αξιόπιστα αποτελέσματα με αλγόριθμους εξόρυξης δεδομένων, η υλοποίηση έγινε με την διεπαφή SAX που προσφέρει γρήγορη προσπέλαση (μόνο για ανάγνωση) και δεν δεσμεύει μνήμη.

4.6 Οι Διαδικασίες Εξόρυξης Δεδομένων και Ανακάλυψης Γνώσεων στο HealthObs

Στο κεφάλαιο αυτό παρουσιάζονται σε μεγαλύτερη λεπτομέρεια θέματα σχετικά με την εξόρυξη και ανακάλυψη γνώσης. Όπως είπαμε και στο Κεφάλαιο 3, υπάρχουν τρεις τύποι γνώσης προς ανακάλυψη:

- η ταξινόμηση (classification)
- η συσχέτιση (associations)
- η ακολουθία περιοδικά επαναλαμβανόμενων γεγονότων (sequences of

frequent events).

Η ταξινόμηση προσπαθεί να κατηγοριοποιήσει τα δεδομένα εισόδου σε ξεχωριστές κλάσεις χρησιμοποιώντας τόσο 'supervised', όσο και 'unsupervised' (γνωστό ως clustering) μεθόδους μάθησης. Ο στόχος είναι η εύρεση βασικών εννοιών, με όσο το δυνατόν περισσότερο σαφή διαχωριστικά εννοιολογικά σύνορα τα οποία χαρακτηρίζουν μια κλάση αντικειμένων. Έτσι, για μη προκαθορισμένα στοιχεία σχετικά με την εννοιολογική τους κατηγοριοποίηση, χάρη στην ταξινόμηση μπορεί να προβλεφθεί η εννοιολογική κατηγορία και κλάση στην οποία ανήκουν. Όταν ανακαλύπτονται χρονικές ακολουθίες (sequences), ο χρόνος, όπως είναι κατανοητό, αποτελεί ένα επιπρόσθετο χαρακτηριστικό. Παραδείγματα τέτοιων εφαρμογών μπορούν να βρεθούν σε αγορές ή σε συμπεριφορές καταναλωτών. Η τρίτη κατηγορία γνώσης είναι οι κανόνες συσχέτισης. Οι συσχετίσεις μπορεί να είναι αυθαίρετοι κανόνες της μορφής "X =>Y" όπως για παράδειγμα «αν έχει σύννεφα θα βρέξει».

Παρουσιάζουμε τις βασικές ιδιότητες των αλγορίθμων και το βασικό αλγοριθμικό σχήμα που ακολουθείται σε τέτοιες περιπτώσεις και κλείνουμε με μια αναφορά στους βασικότερους προηγούμενους αλγόριθμους στο πεδίο της έρευνας μας.

4.6.1 Ανακάλυψη Κανόνων Αλληλοσυσχέτισης (ARM-Association Rules Mining)

Το πρόβλημα της εξαγωγής κανόνων συσχέτισης (Association Rule Mining-ARM), είναι ένα από τα πιο σοβαρά προβλήματα στη διαδικασία ανακάλυψης γνώσης, και έχει τύχει ιδιαίτερης προσοχής τα τελευταία χρόνια, γεγονός που επιβεβαιώνεται από το πλήθος των σχετικά πρόσφατων δημοσιεύσεων [39, 40, 41, 42, 43]. Ένα σύνηθες πεδίο εφαρμογής τους είναι μια βάση δεδομένων με συναλλαγές καταναλωτών, σε διάφορες αγορές, προκειμένου να εξαχθούν συμπεράσματα που χαρακτηρίζουν από κοινού τη συμπεριφορά τους. Οι εξαγόμενοι κανόνες μπορεί να είναι αποκαλυπτικοί όπως για παράδειγμα ότι *«το 75% των ανθρώπων που αγοράζουν σάλτσα για μακαρόνια αγοράζουν ζυμαρικά και καπνιστό κρέας»*. Και ενώ αυτό το παράδειγμα φαίνεται μάλλον διαισθητικά αναμενόμενο, υπάρχουν αρκετές περιπτώσεις όπου οι εξαγόμενοι κανόνες δεν είναι προφανείς και απαιτείται ένας τρόπος ανακάλυψης τους.

Η συσχέτιση αυτών των κανόνων με εφαρμογές όπως σχεδίαση προϊόντων, marketing, διαφήμιση, προώθηση προϊόντων, μπορεί εύκολα να γίνει κατανοητή. Ωστόσο οι ARM αλγόριθμοι μπορούν να εφαρμοστούν σε ένα ευρύτερο πεδίο εφαρμογών και προβλημάτων και να λάβουν χώρα όπως θα

δείξουμε ακόμα και σε ιατρικά πεδία εφαρμογών και σε ανάλυση και επεξεργασία κλινικών στοιχείων και δεδομένων.

- Ο αριθμός όλων των υποθετικά δυνατών κανόνων αλληλοσυσχέτισης αυξάνει εκθετικά με τον αριθμό των στοιχείων (items) που συμμετέχουν στη διαδικασία. Για 1000 στοιχεία για παράδειγμα, περισσότεροι από 2^{1000} κανόνες πρέπει να θεωρηθούν και να επεξεργαστούν σε μια απλοϊκή προσέγγιση. Έτσι, παρά τις συνεχώς αυξανόμενες και εντεινόμενες προσπάθειες για βελτίωση των εν-χρήση σειριακών αλγορίθμων, στην πράξη οι ARM αλγόριθμοι παραμένουν χρονοβόροι και όχι ιδιαίτερα ικανοποιητικοί για άμεση αλληλεπίδραση με τα αντίστοιχα εργαλεία εξόρυξης κανόνων και εκμαίευσης γνώσης στη γενικότερη της μορφή.

4.6.2 Διαδικασίες ARM

Η εξόρυξη κανόνων αλληλοσυσχέτισης είναι μεταξύ των πιο προηγμένων και ενδιαφερουσών μεθόδων που εισάγονται από τα πεδία της μηχανικής μάθησης και εξόρυξης δεδομένων για την εύρεση των προτύπων και τάσεων των δεδομένων. Οι κανόνες αλληλοσυσχέτισης παρέχουν έναν χρήσιμο μηχανισμό για τον συσχετισμό μεταξύ των στοιχείων. Ο καθορισμός ενός προβλήματος έχει ως εξής:

- Έστω ότι D είναι ένα σύνολο από «περιπτώσεις» (transactions) , όπου κάθε περίπτωση T είναι ένα σύνολο από στοιχεία τέτοια ώστε $T \subset I$. Ένας κανόνας συσχέτισης είναι ένα αποτέλεσμα της μορφής $X \Rightarrow Y$, όπου το X είναι υποσύνολο του I ($X \subset I$), το Y είναι υποσύνολο του I ($Y \subset I$) , και η τομή των X ,Y είναι το κενό σύνολο ($X \cap Y = \varnothing$).
- Ο κανόνας $X \Rightarrow Y$ έχει «εμπιστοσύνη» (confidence) c στο σύνολο των περιπτώσεων D αν c% από τις περιπτώσεις στο D που περιέχουν το σύνολο X περιέχουν επίσης το σύνολο Y.
- Ο κανόνας $X \Rightarrow Y$ έχει «στήριξη» (support) s στο σύνολο των περιπτώσεων D αν s% από τις περιπτώσεις στο D περιέχουν το σύνολο X ένωση Y ($X \cup Y$).
- Ένας κανόνας μπορεί να θεωρηθεί ως η πρόβλεψη ότι: εάν ένα transaction υποστηρίζει το σύνολο X, τότε θα υποστηρίζει και το σύνολο Y με ένα μέτρο confidence του κανόνα, και συμβολίζεται ως conf(R).
- Το confidence ενός κανόνα R, ορίζεται ως η δεσμευμένη πιθανότητα, τέτοια ώστε δοθέντος ότι το transaction T υποστηρίζει το X, τότε θα υποστηρίζει

και το Y. Σε μαθηματική μορφή:

$$conf\ (R) = p\left(Y \subseteq T \mid X \subseteq T\right) = \frac{p\left(Y \subseteq T | X \subseteq T\right)}{p\left(X \subseteq T\right)} = \frac{\sup\ p\left(X \cup Y\right)}{\sup\ p\left(X\right)}$$

- Το support ενός κανόνα R στο D, ορίζεται ως supp(X ∪ Y). Το confidence του κανόνα φανερώνει πως συχνά αναμένεται να εμφανισθεί, ενώ το support του φανερώνει κατά κάποιο τρόπο το πόσο αξιόπιστος είναι αυτός ο κανόνας. Για να είναι σημαντικός και ενδιαφέρων ένας κανόνας θα πρέπει να έχει ικανοποιητικό support και αυξημένο confidence. Θα ισχυριζόμαστε λοιπόν ότι ένας κανόνας R, λαμβάνει χώρα στο D, εάν για τις προκαθορισμένες από το χρήστη ελάχιστες τιμές Cmin και Smin, όπου Cmin το ελάχιστο confidence(minconf) και Smin το ελάχιστο support του κανόνα, ισχύει conf(R) ≥ Cmin και supp(R) ≥ Smin.

Διάφοροι αλγόριθμοι έχουν προταθεί στη διεθνή βιβλιογραφία, όπως ο AIS[41] SETM[44] PARTITION [45] APRIORI [46] , για να κάνουν την αναζήτηση αυτή ταχύτερη και πιο έξυπνη. Όλοι αυτοί οι αλγόριθμοι διαφέρουν κυρίως στην μορφή αναπαράστασης και τον τρόπο αποθήκευσης των δεδομένων, την ενδιάμεση αναπαράσταση των αποτελεσμάτων κατά την διάρκεια της επεξεργασίας και το κόστος Input/Output και CPU (overhead) που προκαλούν. Χαρακτηριστικό τους είναι ότι τις περισσότερες φορές υπάρχει ανάγκη η αρχική βάση δεδομένων να διαβαστεί περισσότερες από μια φορές, γεγονός που αναπόφευκτα προκαλεί πρόσθετες καθυστερήσεις.

➲ ***AIS***. Το πρόβλημα των κανόνων συσχέτισης πρωτοεμφανίστηκε στο [40], με έναν αλγόριθμο ο οποίος στην συνέχεια ονομάστηκε *AIS* από τα ονόματα των συγγραφέων του [41]. Προκειμένου ο AIS να βρει τα frequent sets, δημιουργεί δυναμικά (on-the-fly) υποψηφίους όρους ενώ διαβάζει τη βάση. Αρκετά περάσματα της βάσης κρίνονται αναγκαία, και κατά τη διάρκεια καθενός από αυτά όλα τα transactions της βάσης διαβάζονται το ένα μετά το άλλο. Ένας υποψήφιος δημιουργείται με την προσθήκη νέων όρων σε σύνολα που έχουν ήδη διαπιστωθεί και χαρακτηριστεί ως frequent σε προηγούμενες εξετάσεις της βάσης. Για να αποφευχθούν επανεμφανίσεις υποψηφίων, το item που προστίθεται πρέπει να είναι λεξικογραφικά μεγαλύτερο, από το μεγαλύτερο item του *F*, εφόσον αναφερόμαστε σε 1-extensions. Ο AIS δεν δημιουργεί κάποιον υποψήφιο εάν πρώτα δεν τον εντοπίσει κατά τη διάρκεια ανάγνωσης της βάσης. Η όλη διαδικασία σταματά όταν δεν υπάρχουν πλέον frontier sets, που σημαίνει ότι κανένας

από τους προηγούμενους υποψηφίους δεν ήταν frequent. Αρχικά το μόνο frontier set είναι το $\varnothing$. Δυστυχώς αυτή η στρατηγική δημιουργίας υποψηφίων, προκαλεί την εμφάνιση ενός μεγάλου αριθμού από υποψήφιους και πρόσθετες *τεχνικές κλαδέματος* (pruning) κρίνονται αναγκαίες για την απόφαση της περαιτέρω επέκτασης ή όχι κάποιων υποψήφιων συνόλων. Τέτοιες αποφάσεις όμως προκαλούν πρόσθετο κόστος σε χρόνο και μνήμη, καθώς γίνονται επαναλαμβανόμενα σε πολλά υποσύνολα ενός transaction.

- ***SETM***. Ο *SETM* [44] σχεδιάστηκε για να εκτελεί μόνο βασικές λειτουργίες βάσεων δεδομένων, προκειμένου να βρίσκει τα frequent sets. Για το λόγο αυτό χρησιμοποιεί τη δικιά του αναπαράσταση, με βάση την οποία αποθηκεύει κάθε itemset συνδυασμένο με το αντίστοιχο TID του transaction που το υποστηρίζει. Ο SETM επαναλαμβανόμενα τροποποιεί εξολοκλήρου τη βάση, εκτελώντας τις λειτουργίες δημιουργίας υποψηφίων, μέτρησης των support και διαγραφής των μη συχνά εμφανιζομένων item. Υποθέτουμε ότι όλα τα itemsets που δεν υπερβαίνουν το όριο του *minimum support* έχουν διαγραφεί. Στο επόμενο βήμα διαγράφονται όλοι οι μη-συχνά εμφανιζόμενοι. Αυτό γίνεται αφού πρώτα ταξινομηθούν τα itemsets. Το πρόβλημα με αυτόν τον αλγόριθμο είναι ότι δημιουργεί ίδιους υποψηφίους προερχόμενους από διαφορετικά transactions με αποτέλεσμα να δημιουργούνται τεράστιος αριθμός ενδιάμεσων αποτελεσμάτων. Επιπρόσθετα τα itemsets θα πρέπει να είναι *ταξινομημένα*. Το χειρότερο όμως, όπως αναφέραμε, είναι ότι αυτοί οι τεράστιοι πίνακες που προκύπτουν θα πρέπει να ταξινομούνται δυο φορές προκειμένου να προκύπτουν τα επόμενα κάθε φορά large itemsets.
- ***Apriori***. Ο μεγάλος αριθμός των υποψηφίων που δημιουργούσε ο AIS ώθησε τους δημιουργούς του να αναπτύξουν μια καινούρια στρατηγική δημιουργίας υποψηφίων, η οποία ονομάστηκε Apriori-gen και αποτέλεσε μέρος των αλγορίθμων *Apriori* και *AprioriTid* [46]. Η βασική του αρχή πηγάζει από την ιδιότητα των κανόνων συσχέτισης σύμφωνα με την οποία ο αλγόριθμος δημιουργεί έναν υποψήφιο εάν και μόνο εάν όλα τα υποσύνολα του έχουν προηγουμένως κριθεί ως frequent. Συγκεκριμένα ένας (k+1) υποψήφιος θα γίνει αποδεκτός εάν και μόνο αν όλα τα k-itemsets υποσύνολα του έχουν κριθεί ως frequent.

4.6.3 Ο Apriori σε λεπτομέρεια

Προκειμένου να γίνουν περισσότερο κατανοητά τα παραπάνω ας υποθέσουμε ότι βρισκόμαστε στο στάδιο δημιουργίας των *k+1* υποψηφίων. Ο Apriori-gen παίρνει ως είσοδο όλα τα large/ frequent *k*-itemsets L_k και αναζητά ζεύγη από σύνολα στοιχείων τα οποία έχουν κοινά τα *k-1* μικρότερα στοιχεία τους. Παίρνοντας τα *k-1* κοινά στοιχεία σε συνδυασμό με τα δυο υπόλοιπα, τα δυο αυτά σύνολα ενώνονται για την δημιουργία του υποτιθέμενου υποψηφίου. Επανεμφανίσεις στοιχείων αποκλείονται από την απαίτηση το τελευταίο στοιχείο του δεύτερου συνόλου να είναι μεγαλύτερο. Για την ώρα η παρουσία μόνο δυο υποσυνόλων έχει συντελέσει στην δημιουργία του υποτιθέμενου υποψηφίου. Έτσι οι βελτιώσεις που προκύπτουν από τον Apriori-gen σχετικά με την στρατηγική δημιουργίας υποψηφίων συγκριτικά με τον AIS συνοψίζονται σε δυο σημεία:

- δημιουργούνται λιγότεροι υποψήφιοι
- δεν δημιουργούνται επαναληπτικά για κάθε transaction αλλά μόνο μια φορά.

```
1. L1 = {large 1-itemsets}
2. for (k = 2; Lk-1  ≠ ∅; k++) do begin
3.   Ck = apriori-like-gen(Lk-1);  // Καινούριου Υποψήφιοι
4.   forall transactions t ∈ D do begin
5.       Ct  =  subset(Ck,t);  // Οι  υποψήφιοι  που  περιέχονται
   στο t
6.       forall candidates c  ∈ Ct do
```

Σχήμα 12. Ο αλγόριθμος Apriori

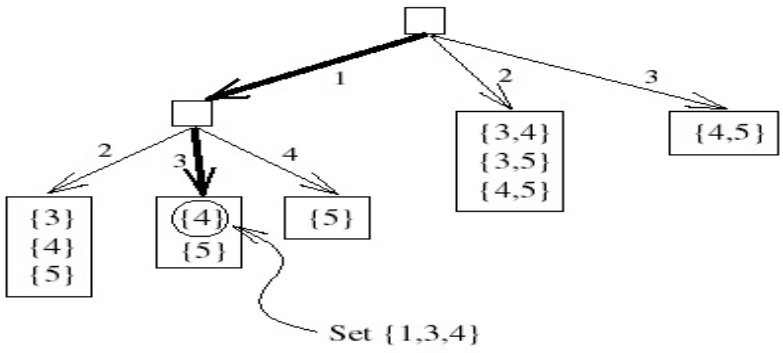
Θεωρώντας τον Apriori τον καλύτερο αλγόριθμο που έχει προταθεί μέχρι σήμερα υλοποιήσαμε τους κανόνες συσχέτισης στο σύστημα μας με αυτόν. Το σχήμα 12 δείχνει τον Apriori αλγόριθμο (σε ψευδοκώδικα) όπως προτάθηκε από το [46] . Στο πρώτο πέρασμα του αλγορίθμου απλά μετριούνται οι εμφανίσεις των στοιχείων για να καθοριστεί ποια από αυτά θα αποτελέσουν τα large *1*-temsets. Κάθε επόμενο πέρασμα, ας πούμε *k*, αποτελείται από δυο φάσεις. Στη πρώτη, τα large L_{k-1} itemsets που βρέθηκαν στο k-*1* πέρασμα χρησιμοποιούνται για να δημιουργήσουν τα υποψήφια itemsets C_k. Στη δεύτερη φάση καθορίζεται το support των υποψηφίων και κατ' επέκταση ευρίσκονται ποια

itemsets είναι και large.

Το βασικό πρόβλημα του Apriori (και του AIS) ήταν ότι έπρεπε να διαβάζει εξολοκλήρου τη βάση σε κάθε "πέρασμα" του. Πολλά από τα items και τα transactions της βάσης από κάποιο σημείο και μετά δεν έχουν καμία ουσιαστική αξία για τα επόμενα "περάσματα" του αλγορίθμου και μόνο επιβάρυνση σε χρόνο και μνήμη προκαλούν. Η δυσκολία πηγάζει από την μορφή αναπαράστασης, η οποία χρησιμοποιείται και από τον αλγόριθμο. Πολλές μελέτες έχουν γίνει σε αυτό το θέμα και έχουν προταθεί αρκετές δομές αναπαράστασης για τον Apriori αλγόριθμο. Παρακάτω θα αναλύσουμε τρεις από αυτές τις δομές, μία εκ των οποίων είναι αυτή που επιλέχτηκε και υλοποιήθηκε για το σύστημα μας.

4.6.4 Δομές Αποθήκευσης για Κανόνες Αλληλοσυσχέτισης

- **Apriori-struct.** Η πρώτη μορφή δομής που προτάθηκε ήταν η Apriori-struct. Σε αυτήν τη μορφή αναπαράστασης τα transactions είναι αποθηκευμένα σε μια ακολουθία ταξινομημένων στοιχείων υπό την μορφή λίστας. Αυτή η μορφή των item-lists κάνει δύσκολη την διαγραφή και απαλλαγή από τα μη αναγκαία τμήματα της πληροφορίας. Η γνώση και η πληροφορία για το ποια items θα μείνουν στη δομή και ποια όχι, είναι διαθέσιμη μόνο αφότου διαβάσουμε μια ακόμα φορά τη βάση και μετρήσουμε τα supports των υποψηφίων. Έτσι μπορούμε να απομακρύνουμε τα περιττά items μόνο σε επόμενο πέρασμα των δεδομένων, αφού αυτά διαβαστούν για μια ακόμα φορά, χωρίς αυτό να είναι αναγκαίο.



Σχήμα 13: Δομή αποθήκευσης Apriori-struct

Το πλεονέκτημα της αναπαράστασης των δεδομένων με τη μορφή των item-lists, είναι ότι το μέγεθος των δεδομένων δεν αυξάνει κατά τη διάρκεια εκτέλεσης του αλγορίθμου και δεν μπορεί να υπερβεί το αρχικό μέγεθος της

βάσης.

- ***AprioriTid***. Ο AprioriTid μπορεί να θεωρηθεί ότι αποτελεί μια βελτιωμένη έκδοση της αρχικής δομής, η οποία δεν στηρίζεται σε βασικές λειτουργίες βάσεων δεδομένων και χρησιμοποιεί την `apriori-gen' για ταχύτερη δημιουργία υποψηφίων. Επιπρόσθετα ο AprioriTid διαβάζει τα δεδομένα μια μόνο φορά και προσπαθεί να τα αποθηκεύσει για όλα τα υπόλοιπα περάσματα. Και σε αυτήν την περίπτωση, κάθε επαναληπτική εκτέλεση του αλγορίθμου αποτελείται από την φάση δημιουργίας των υποψηφίων μέσω της `apriori-gen', ακολουθούμενη από την φάση μέτρησης και τον προσδιορισμό των supports των τρεχόντων (υπό παραγωγή) υποψηφίων.

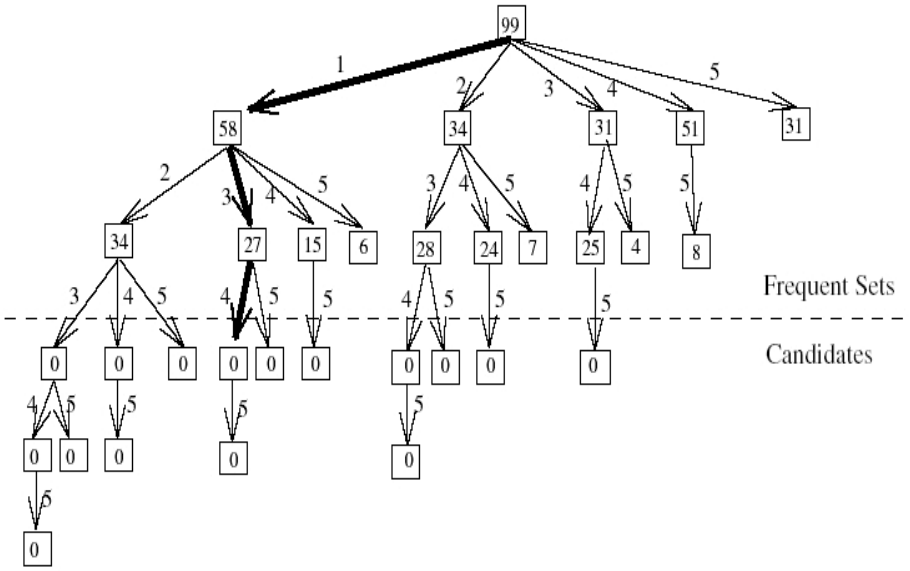
  Το ενδιαφέρον σημείο του αλγορίθμου βρίσκεται στο γεγονός ότι η πηγή δεδομένων δεν χρησιμοποιείται μετά την πρώτη ανάγνωση της για την μέτρηση των υποψηφίων. Σε αντίθεση ένα σύνολο ,το C_k , χρησιμοποιείται γι΄ αυτό το σκοπό. Κάθε μέλος του set $\overline{C}_k$ είναι της μορφής *<TID, {X_k}>*, όπου κάθε X_k είναι ένα υποθετικά large itemset, το οποίο εμφανίζεται στο transaction με αναγνωριστικό TID. Αυτή η μορφή αναπαράστασης είναι γνωστή με το όνομα λίστα υποψηφίων (candidate-lists). Για *k=1*, το $\overline{C}_1$ ανταποκρίνεται στη βάση *D*, με τη διαφορά ότι κάθε item *i*, έχει αντικατασταθεί από το itemset {i}. Ωστόσο για μικρές τιμές του *k*, μια εγγραφή στο $\overline{C}_k$ μπορεί να είναι μεγαλύτερη από το μέγεθος του αντίστοιχου transaction, διότι μια εγγραφή στο $\overline{C}_k$ περιλαμβάνει όλα τα set υποψηφίων μεγέθους *k*, που περιέχονται στο εν λόγω transaction. Λόγω της μορφής αναπαράστασης δεδομένων που υιοθετεί (candidate-list) είναι κυρίως αργός στο δεύτερο πέρασμα, και αν η μνήμη δεν επαρκεί για την αποθήκευση των δεδομένων, η εναλλαγή των αποτελεσμάτων με τη περιφερειακή μνήμη κρίνεται αναγκαία, γεγονός που προκαλεί πρόσθετες καθυστερήσεις

4.6.5 Prefix Tree

Με βάση το δενδρικό μοντέλο που προτάθηκε από τον Apriri-struct και την ιδιότητα του AprioriTid να διαβάζει την πηγή δεδομένων μόνο μία φορά προτάθηκε μια νέα δομή, η prefix-tree. Όπως και στον Apriori-struct, κάθε πέρασμα στο prefix-tree αποτελείται από μια φάση παραγωγής υποψηφίων που ακολουθείται από μια μέτρησης. Η δομή αυτή δεν κάνει είναι καμία διάκριση μεταξύ των εσωτερικών κόμβων και των φύλλων. Οι κόμβοι δεν περιέχουν τα σύνολα, αλλά μόνο τις πληροφορίες για τα σύνολα (π.χ. μετρητές). Κάθε ακμή στο δέντρο αντιπροσωπεύει ένα στοιχείο, και κάθε κόμβος περιέχει τις πληροφορίες για το σύνολο των στοιχείων που βρίσκονται από την πορείας ξεκινώντας από την ακμή ως τη ρίζα. Ένα παράδειγμα

παρουσιάζεται στο σχήμα 14, όπου το σύνολο 1 ..3 ..4 χαρακτηρίζεται από μία πιο πυκνή πορεία. Τα στοιχεία σε ένα καθορισμένο σύνολο X μπορούν να γίνουν κατανοητά όπως μία περιγραφή πορειών από τη ρίζα του δένδρου μέχρι τον κόμβο X. Ταξινόμηση των στοιχείων απαιτείται προκειμένου να αποφευχθεί η επανάληψη συνόλων ; διαφορετικά διάφοροι κόμβοι θα αντιστοιχούσαν στο ίδιο σύνολο.

Το prefix-tree αποθηκεύει τα 'συχνά σύνολα' και τα 'σύνολα υποψηφίων' στο ίδιο δέντρο. Μόλις μετρηθούν οι υποψήφιοι και καθοριστούν να είναι συχνοί, παραμένουν απλά στην κατάλληλη θέση τους στο δέντρο και γίνονται συχνά σύνολα.



Σχήμα 14: Δομή αποθήκευσης prefix-tree

- Στο Σχήμα 14, κάθε κόμβος στο δέντρο περιέχει τις αριθμήσεις συχνότητας για το αντίστοιχο σύνολό του. Η ρίζα αντιπροσωπεύει το κενό σύνολο, και έτσι ο μετρητής του είναι ίσος με τον αριθμό συναλλαγών στη πηγή δεδομένων, δεδομένου ότι όλες οι συναλλαγές υποστηρίζουν το κενό σύνολο. Τα σύνολα υποψηφίων έχουν μια αρίθμηση ίση με μηδέν πριν υποβληθούν σε επεξεργασία. Το σχήμα 14 παρουσιάζει το πλήρες δέντρο, που περιέχει όλα τα υποσύνολα του {1,2,3,4,5} , αλλά στην πραγματικότητα μόνο τα συχνά σύνολα και τα σύνολα υποψηφίων θα αποθηκευτούν στο δέντρο, ενώ τα μη-συχνά σύνολα είτε δεν δημιουργούνται αρχικά είτε διαγράφονται αμέσως. Στο σχήμα 14 όλα τα υποψήφια σύνολα παρουσιάζονται προκειμένου να απεικονιστεί η

ασύμμετρη μορφή του δέντρου, το οποίο είναι ένα άμεσο αποτέλεσμα της ταξινομημένης σειράς των στοιχείων στις ακμές.

Μερικές ιδιότητες μπορούν να δηλωθούν για αυτήν την δομή αποθήκευσης όταν χρησιμοποιείται για να αποθηκεύσει τα συχνά σύνολα:

- **ιδιότητα 1**: ο αριθμός του μετρητή των κόμβων κατά μήκος μιας πορείας δεν αυξάνεται.
- **ιδιότητα 2:** Εάν ένα σύνολο είναι συχνό και επομένως παρουσιάζεται στο δέντρο, κατόπιν όλα τα υποσύνολά του πρέπει να είναι στην κατάλληλη θέση τους στο δέντρο.

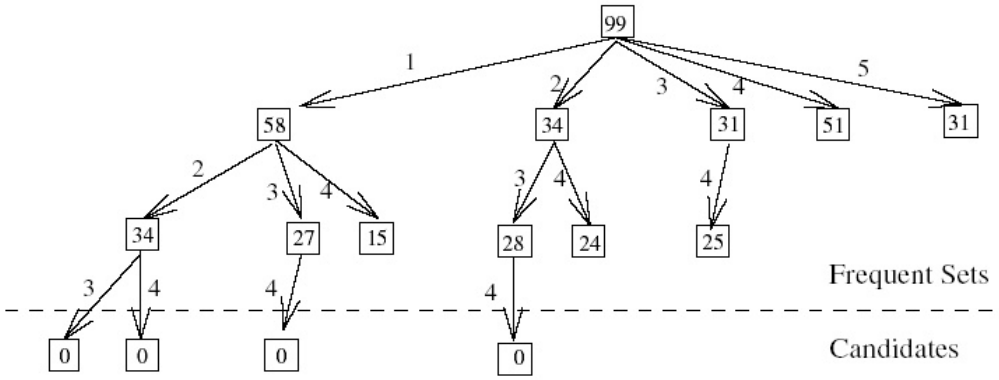
Για την ιδιότητα 1, θεωρήστε ότι ο γονέας ενός κόμβου έχει αριθμό μετρητή (δηλαδή συχνότητα εμφάνισης) μεγαλύτερο από αριθμό μετρητή του κόμβου επειδή η πορεία του γονέα δεν έχει την τελευταία άκρη της πορείας του κόμβο και ένα σύνολο δεν μπορεί να εμφανιστεί συχνότερα στη πηγή δεδομένων από οποιαδήποτε από τα υποσύνολά του. Αυτή η ιδιότητα χρησιμοποιείται στον υπολογισμό της εμπιστοσύνης ενός κανόνα. Σημειώστε ότι η ιδιότητα 2 επεκτείνεται και στα υποσύνολα από την πορεία της ακμής στη ρίζα, που πρέπει φυσικά να υπάρξουν. Παραδείγματος χάριν, εάν το σύνολο {1 ,2 ,3} είναι στο δέντρο, τότε και τα σύνολα {1,2} και {1} είναι σαφώς παρών επειδή είναι κατά μήκος της πορείας.

Για να επεξηγήσουμε τις διαφορές μεταξύ των prefix-tree και της δομής Apriori-struct, χρησιμοποιούμε ένα παράδειγμα που επιδεικνύει πώς η υποστήριξη για τους υποψηφίους μετριέται. Εξετάστε τη συναλλαγή τ = {1 ,4 ,6 ,7 ,9} και υποθέστε ότι οι υποψήφιοι c1 = {1 ,4 ,6 ,7} και c2 = {1 ,4 ,7 ,9} είναι οι μόνοι υποψήφιοι που υποστηρίζονται από το Τ. Υποθέστε περαιτέρω ότι ο Apriori-srtuct έχει αποθηκεύσει τα c1 και c2 σε ένα φύλλο μαζί με διάφορους άλλους υποψηφίους που αποτελούνται από 4 στοιχεία ({1.4.5.6} , {1, 4, 5, 9} , …) που έχουν επίσης το πρόθεμα {1 ,4, .., ..} , αλλά δεν υποστηρίζεται από το τ. Αυτός ο κόμβος μπορεί να προσεγγιστεί στο prefix-tree από τη ρίζα περνώντας πρώτα από την ακμή που ονομάζεται με το στοιχείο 1 και έπειτα την ακμή με το στοιχείο 4. Ο Apriori-struct εξετάζει τα στοιχεία 1 και 4 μία φορά για να φθάσει στο φύλλο, κατόπιν πρέπει να ελέγξει για όλους τους υποψηφίους εκεί εάν υποστηρίζονται από το τ ή όχι. Τα πρώτα δύο στοιχεία (1 και 4) δεν είναι απαραίτητο να εξεταστούν άλλο, αλλά για όλα τα μεγαλύτερα στοιχεία Ι σε ένα υποψήφιο σύνολο πρέπει να ελέγξουμε εάν τα Ι υποστηρίζονται από τα δεδομένα μας Τ. Εάν τα σύνολα αποθηκεύονται ως itemlists, αυτό σημαίνει δύο συγκρίσεις ανά υποψήφιο στο παράδειγμά μας.

Ο αλγόριθμος φθάνει στον κόμβο {1 ,4} όπως Apriori, αλλά έπειτα επιλέγει την άκρη για το στοιχείο 6 άμεσα, και μετά το στοιχείο 7 και αυξάνει την

αρίθμηση για το c1. Η προσπάθεια να βρεθεί η άκρη 9 τον κόμβο {1 ,4 ,6} αποτυγχάνει επειδή το σύνολο {1 ,4 ,6 ,9} δεν είναι υποψήφιος. Ο αλγόριθμος επομένως επιστρέφει στον κόμβο {1,4}, επιλέγει την ακμή 7, κατόπιν την ακμή 9, και αυξάνει την αρίθμηση για το c2. Αυτό ανέρχεται σε συνολικά 5 διαδικασίες επιλογής ακρών που μπορούν να εφαρμοστούν αποτελεσματικά ως hash-table look-ups. Περισσότερες διαδικασίες είναι απαραίτητες για τις μεγαλύτερες συναλλαγές, οπότε ο Apriori θα χρειαστεί επίσης τις πρόσθετες συγκρίσεις για να βρει τα στοιχεία υποψηφίων στη συναλλαγή. Σημειώστε ότι και στους δύο αλγορίθμους, μπορεί να μειωθεί το κόστος υπολογισμών με την αφαίρεση των στοιχείων που δεν είναι μέρος οποιουδήποτε υποψηφίου από μια συναλλαγή.

Συνοψίζοντας, η μέθοδος υποσυνόλων Apriori-struct χρησιμοποιεί ένα δέντρο για να μειώσει τον αριθμό υποψηφίων που πρέπει να εξεταστούν σε μια συναλλαγή, ενώ η μέθοδός του prefix-tree χρησιμοποιεί το δέντρο για να φθάσει ακριβώς στους υποψηφίους που υποστηρίζονται από τη συναλλαγή.

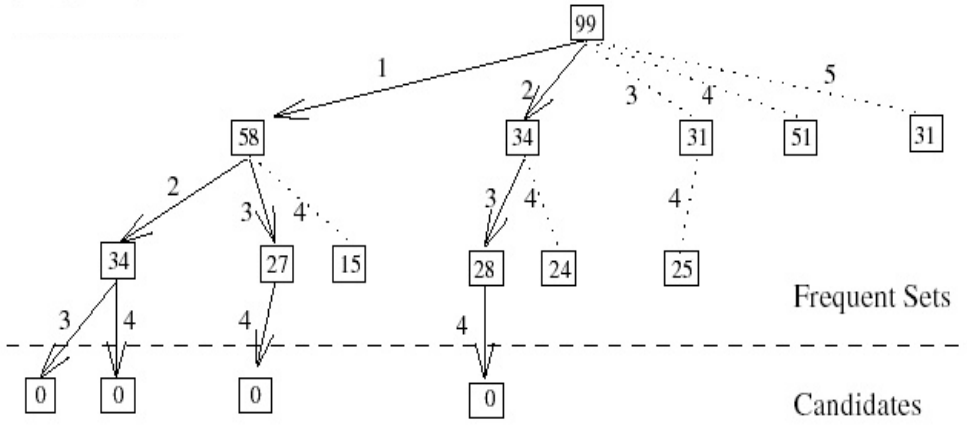


Σχήμα 15 : Δομή prefix-tree χωρίς περιττή πληροφορία

Όπως αναφέρεται παραπάνω, το prefix-tree περιέχει μόνο τα συχνά itemsets και τους τρέχοντες υποψηφίους. Το σχήμα 15 παρουσιάζει το δέντρο από το σχήμα 14 για μια ελάχιστη υποστήριξη 10% (ισοδύναμος με 10 στοιχεία). Πολλά σύνολα έχουν αφαιρεθεί, και λιγότεροι υποψήφιοι δημιουργούνται κατά τη διάρκεια της φάσης παραγωγής υποψηφίων. Εξετάστε το σύνολο {1 ,2 ,5} , παραδείγματος χάριν. Δεδομένου ότι το σύνολο {1 ,5} δεν είναι συχνό σύνολο, αυτό το σύνολο δεν δημιουργείται ποτέ ως υποψήφιος.

4.6.5.1 Νεκροί κλάδοι στο prefix-tree.

Η αποθήκευση και των συνόλων υποψηφίων και των συχνών συνόλων στο ίδιο δέντρο μπορεί να προκαλέσει την αυστηρή υποβάθμιση απόδοσης όταν μετριέται η υποστήριξη(support) για τους υποψηφίους. Στο σχήμα 15, το σύνολο {3,4} δεν έχει κανένα σύνολο υποψηφίων. Η κράτηση του στο δέντρο σημαίνει ότι για κάθε συναλλαγή που περιέχει τα στοιχεία 3 και 4, θα τροποποιούμε τις αριθμήσεις για ανύπαρκτα υποψηφία σύνολα σε εκείνο τον κλάδο. Αυτό είναι σαφώς περιττό. Εντούτοις, το σύνολο {3,4} είναι συχνό, έτσι θέλουμε να το κρατήσουμε στο δέντρο προκειμένου να το χρησιμοποιήσουμε για να παραγάγουμε τους κανόνες αργότερα. Επομένως οι κόμβοι που δεν παράχθηκαν από κανένα υποψήφιο σύνολο, τους οποίους καλούμε νεκρούς κλάδους, κλαδεύονται από το δέντρο και αποθηκεύονται χωριστά κατά τέτοιο τρόπο ώστε μπορούν "να αναβιωθούν" εύκολα όταν απαιτούνται σε επόμενο χρονικό σημείο (όταν παράγουμε τους κανόνες και χρειαζόμαστε την υποστήριξη τους).



Σχήμα 16 : Νεκροί κλάδοι στο prefix_tree

Συσσώρευση περασμάτων. Στο σχήμα 16, εάν μόνο οι υποψήφιοι {1 ,2 ,3} και {1 ,2 ,4} είναι συχνοί μετά από το μέτρημα, ένας νέος υποψήφιος {1 ,2 ,3 ,4} θα πρέπει να ελεγχθεί. Στην πραγματικότητα, αυτό είναι ο μόνος υποψήφιος που έχει μείνει, και ένα ολόκληρο πέρασμα της πηγής δεδομένων θα ήταν απαραίτητο για να μετρήσει ακριβώς αυτό. Για να αποφευχθούν τέτοια φαινόμενα, το prefix-tree επιτρέπει την επέκταση διάφορων επιπέδων του

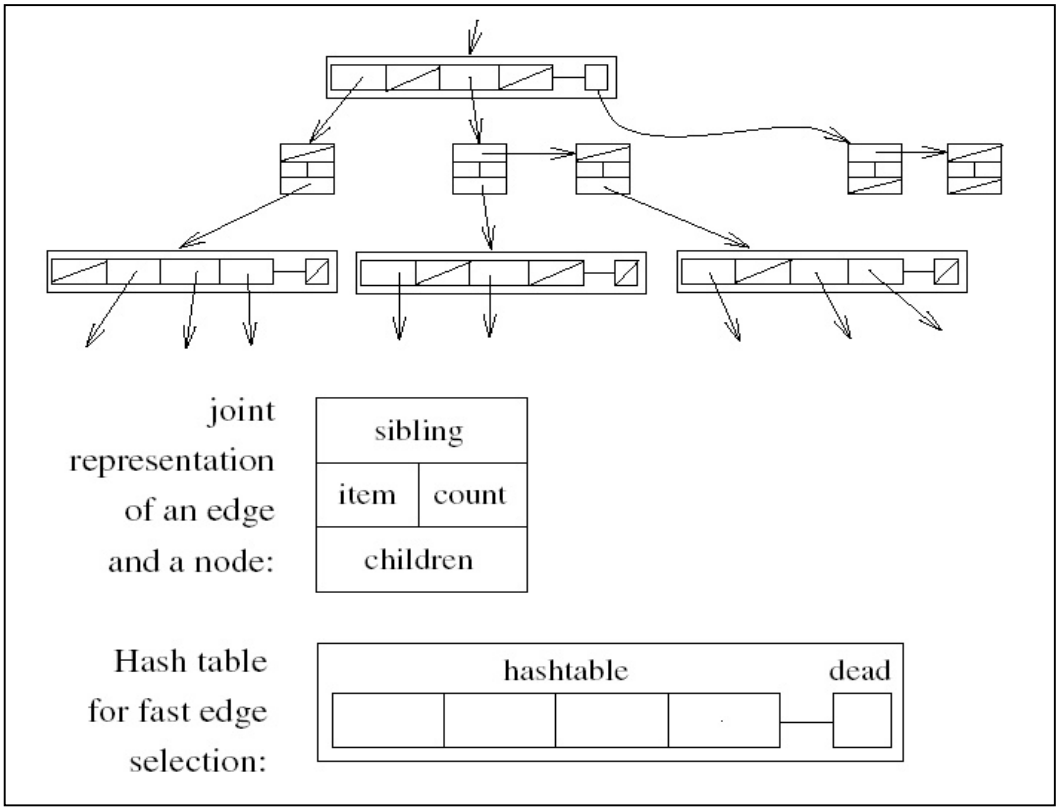
δέντρου πριν το μέτρημα των υποψηφίων. Κατά συνέπεια, και οι τρεις υποψήφιοι στο παράδειγμά μας θα δημιουργηθούν και θα μετρηθούν από κοινού. Αυτή η τεχνική, που καλούμε συσσώρευση περασμάτων, αναφέρθηκε αρχικά στα [47] και [46], αλλά τα αποτελέσματά του δεν εξετάστηκαν εκεί.

Υπάρχει μια ισοζυγία μεταξύ του αριθμού διαδικασιών IO που κερδίζουμε και τον πρόσθετο υπολογισμό που είναι απαραίτητος για να μετρήσει τους πρόσθετους υποψηφίους. Σαφώς, η συσσώρευση περασμάτων διάφορων επιπέδων είναι μόνο επιθυμητή στα τελευταία περάσματα όταν ο αριθμός υποψηφίων είναι μικρός. Αναπτύσσοντας περισσότερα από ένα επίπεδα αρχικά τα αποτελέσματα δεν περικόπτουν σύνολα υποψηφίων. Επομένως η συσσώρευση περασμάτων εφαρμόζεται με τον καθορισμό χαμηλότερου ορίου στον αριθμό υποψηφίων που πρέπει να δημιουργηθούν προτού να επιτραπεί ένα βήμα μέτρησης. Επιλέγοντας τον παράγοντα συσσώρευσης περασμάτων ως κάποιο χαμηλό πολλαπλάσιο του αριθμού στοιχείων, εξασφαλίζουμε ότι τα πρόωρα περάσματα δεν επηρεάζονται (επειδή ο αριθμός υποψηφίων υπερβαίνει το όριο κατά πολύ), ενώ τα τελευταία περάσματα συσσωρεύονται από κοινού.

4.6.5.2 Υλοποίηση του Prefix-tree

Για να παρέχεται γρήγορη επιλογή μιας ακμής διαπερνώντας ένα prefix-tree, ένα hash-table συνδέεται με κάθε κόμβο και περιέχει τα παιδιά του. Όπως διευκρινίζεται στο Σχήμα 17, οι καταχωρήσεις στους hash buckets είναι δομές που περιέχουν αρίθμηση, τον αριθμό στοιχείων και έναν δείκτη hash-table παιδιών. Κάθε hash-table έχει έναν νεκρό-δείκτη(dead pointer) για να αποθηκεύσει έναν συνδεμένο κατάλογο νεκρών στοιχείων που είναι οι ρίζες των νεκρών κλάδων. Κρίσιμο από την άποψη πόρων και η ταχύτητας είναι το μέγεθος των hash-tables. Ενώ ο πίνακας της ρίζας πρέπει να είναι αρκετά μεγάλος, πίνακες βαθύτερα στο δέντρο μπορεί να είναι πολύ μικρότεροι. Δεδομένου ότι όλοι οι υποψήφιοι από ένα φύλλο παράγονται ταυτόχρονα, ο αριθμός τους είναι γνωστός και μπορούμε να το χρησιμοποιήσουμε για να υπολογίσουμε το μέγεθος του hash-table. Το μέγεθος του hash table είναι μια δύναμη του 2 για να επιτραπεί ο γρήγορος υπολογισμός της hash function που χρησιμοποιεί έναν απλό AND (ΚΑΙ) για μια λειτουργία.

Εκτός από το ίδιο το prefix-tree, ένας κατάλογος όλων των στοιχείων που είναι ακόμα σε χρήση (δηλ. όλα τα στοιχεία που είναι ακόμα μέλη των συνόλων υποψηφίων) διατηρείται κατά τη διάρκεια της παραγωγής υποψηφίων. Αυτός ο κατάλογος το καθιστά ικανό να βγάλει έξω όλα τα στοιχεία μιας συναλλαγής που δεν είναι πλέον σχετικά, μειώνοντας κατά συνέπεια τον αριθμό αποτυχημένων hash operations κατά τη διάρκεια του υπολογισμού.



Σχήμα 17. Hash table του prefix-tree

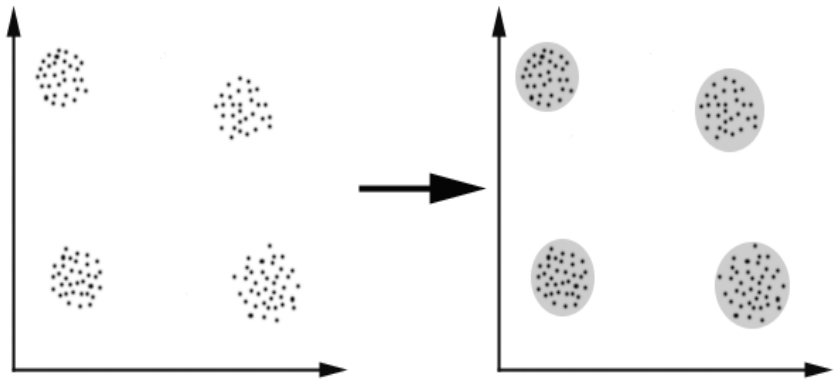
4.7 Αλγόριθμοι Συνάθροισης (clustering)

Όπως είπαμε και προηγουμένως η συνάθροιση χωρίζεται σε δύο μεγάλες κατηγορίες. Την συνάθροιση στοιχείων βασισμένη σε προκαθορισμένη γνώση και την συνάθροιση χωρίς υπόβαθρο γνώσης, γνωστή με την ονομασία *συνάθροιση* (clustering) . Στην συγκεκριμένη εργασία θα ασχοληθούμε μόνο με την δεύτερη κατηγορία λόγο της φύσης των δεδομένων μας. Στις περισσότερες εφαρμογές, ακόμα και στις ποιο δομημένες (βάσεις δεδομένων) δεν υπάρχει αρχική γνώση. Έτσι, για μη προκαθορισμένα στοιχεία σχετικά με την εννοιολογική τους κατηγοριοποίηση, χάρη στην συνάθροιση μπορεί να προβλεφθεί η εννοιολογική κατηγορία και κλάση στην οποία ανήκουν.

Ένας απλός ορισμός της συνάθροισης θα μπορούσε να είναι "η διαδικασία οργάνωσης αντικείμενων σε ομάδες τα μέλη των οποίων είναι παρόμοια με κάποιο τρόπο". Μια συστάδα (cluster) είναι επομένως μια συλλογή των αντικειμένων που είναι "παρόμοια" μεταξύ τους και είναι "ανόμοια" στα στοιχεία

που ανήκουν σε άλλες συστάδες.

Μπορούμε να παρουσιάσουμε αυτό με ένα απλό γραφικό παράδειγμα (Σχήμα 18).



Σχήμα 18: Γραφική απεικόνιση συνάθροισης

Σε αυτήν την περίπτωση προσδιορίζουμε εύκολα τις 4 συστάδες σε τις οποίες τα στοιχεία μπορούν να διαιρεθούν ; το κριτήριο ομοιότητας είναι η απόσταση. Δύο ή περισσότερα στοιχεία ανήκουν στην ίδια συστάδα εάν είναι "στενά" σύμφωνα με μια δεδομένη απόσταση (σε αυτήν την περίπτωση γεωμετρική απόσταση). Αυτό καλείται απόσταση βασισμένη στον συγκέντρωση (distance-based clustering).

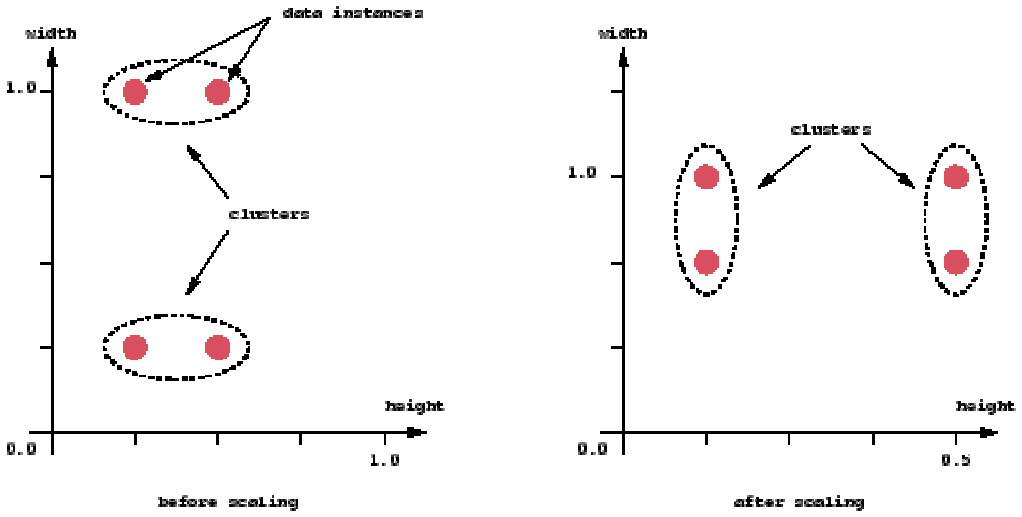
Λόγω της 'ανεπίβλεπτης' (unsupervised) φύσης των δεδομένων, είναι πάρα πολύ εύκολο να εφαρμοστεί η συνάθροιση των αλγορίθμων στα στοιχεία χωρίς ικανοποιητικά αποτελέσματα. Η αντιπροσώπευση των σχεδίων, ο καθορισμός ενός μέτρου εγγύτητας και η ερμηνεία των αποτελεσμάτων αποτελούν την εκτίμηση του αλγορίθμου. Από κοινού σε όλους τους αλγόριθμους εξόρυξης δεδομένων, η ποιότητα των αποτελεσμάτων "μετριέται" από την αντιπροσώπευση των χαρακτηριστικών γνωρισμάτων και των προτύπων. Εάν η αντιπροσώπευση δεν είναι αρκετά εκφραστική τότε με κανένα τρόπο δεν μπορούμε να παράγουμε ικανοποιητικά αποτελέσματα

Αν και τα στατιστικά εργαλεία, όπως η ανάλυση τμημάτων αρχής, μπορούν να βοηθήσουν να βρουν μια βέλτιστη αντιπροσώπευση σε μια έκταση, γενικά δεν υπάρχει καμία θεωρητική οδηγία για την επιλογή μιας κατάλληλης αντιπροσώπευσης για όλες τις περιοχές προβλήματος. Η γνώση περιοχών πρέπει επομένως να χρησιμοποιηθεί για να καθοδηγήσει το σχέδιο μιας αντιπροσώπευσης. Αυτό είναι γνωστό ως επιλογή χαρακτηριστικών γνωρισμάτων (feature selection).

Τα αποτελέσματα μιας διαδικασίας συνάθροισης πρέπει να ερμηνευθούν κατάλληλα στο ευρύτερο πλαίσιο της εφαρμογής. Σε μερικές εφαρμογές, η συνάθροιση είναι μια αρχική εξερεύνηση πριν από το σχέδιο ενός συστήματος στατιστικής που υποθέτει έναν σταθερό αριθμό κατηγοριών. Σε άλλες εφαρμογές, η συνάθροιση χρησιμοποιείται για να υποστηρίξει μια διαδικασία απόφασης, ίσως υπό μορφή συνόλου κανόνα ή δέντρου απόφασης. Σε όλες αυτές τις περιπτώσεις, οι "περιπτώσεις" (των δεδομένων) πρέπει να μετασχηματιστούν κατάλληλα ή να ερμηνευθούν.

4.7.1 Μετρικές Απόστασης

Ένα σημαντικό συστατικό ενός αλγορίθμου συνάθροισης είναι το μέτρο απόστασης μεταξύ των σημείων των στοιχείων. Εάν τα στοιχεία των διανυσμάτων δεδομένων (ανά περίπτωση) είναι όλα στις ίδιες φυσικές μονάδες τότε είναι δυνατό με απλή Ευκλείδεια μετρική απόσταση να ομαδοποιήσουμε επιτυχώς τις παρόμοιες περιπτώσεις στοιχείων. Εντούτοις, ακόμη και σε αυτήν την περίπτωση η Ευκλείδεια απόσταση μπορεί μερικές φορές να είναι παραπλανητική. Το σχήμα 19 επεξηγεί αυτό με ένα παράδειγμα των μετρήσεων *πλάτους* και *ύψους* ενός αντικειμένου. Και παρά τις δύο μετρήσεις που λαμβάνονται στις ίδιες φυσικές μονάδες, μια ενημερωμένη απόφαση πρέπει να ληφθεί ως προς μια σχετική κλιμάκωση. Όπως μπορούμε να διακρίνουμε από το σχήμα 19, η διαφορετική κλιμάκωση μπορεί να οδηγήσει στις διαφορετικές συγκεντρώσεις. Οπότε, η γνώση περιοχών πρέπει να χρησιμοποιηθεί για να καθοδηγήσει τη διατύπωση ενός κατάλληλου μέτρου απόστασης για κάθε ιδιαίτερη εφαρμογή.



Σχήμα 19 : Παράδειγμα χρήσης συνάθροισης

Για δεδομένα με περισσότερες των 2 διαστάσεων έχουν προταθεί διάφορες μετρικές όπως οι Minkowski, Mahalanobis και Hausdorff [51] Οι αλγόριθμοι συνάθροισης μπορούν να κατηγοριοποιηθούν σε δύο ευρείες ομάδες, την ιεραρχική και την τμηματική. Οι ιεραρχικοί αλγόριθμοι παράγουν τις ταξινομήσεις όπου οι ομάδες στοιχείων μπορούν να περιληφθούν από τις μεγαλύτερες ομάδες στοιχείων. Οι τμηματικοί αλγόριθμοι δεν επιτρέπουν τέτοια τοποθετημένη δομή. Οι αλγόριθμοι συνάθροισης μπορεί να διαφέρουν στη λειτουργία τους. Οι συσσωρευτικοί (agglomerative) αλγόριθμοι αυτοί αρχίζουν με ομάδες ενός στοιχείου και τις συγχωνεύουν σε μεγαλύτερες ομάδες. Εναλλακτικά, οι διαχωριστικοί αλγόριθμοι αρχίζουν από μια ομάδα, που αποτελείται από όλα τα στοιχεία, και το χωρίζουν διαδοχικά στις μικρότερες ομάδες [51]. Όλοι οι αλγόριθμοι εξετάζουν όλα τα στοιχεία των διανυσμάτων κατά τον υπολογισμό των μέτρων απόστασης. Οι Jain και Raftery, μεταξύ των άλλων, παρουσιάζουν μια συζήτηση των πολλών περισσότερων παραλλαγών της συγκέντρωσης των αλγορίθμων μαζί με τις δυνάμεις και τις αδυναμίες τους [50, 51, 52, 53, 54].

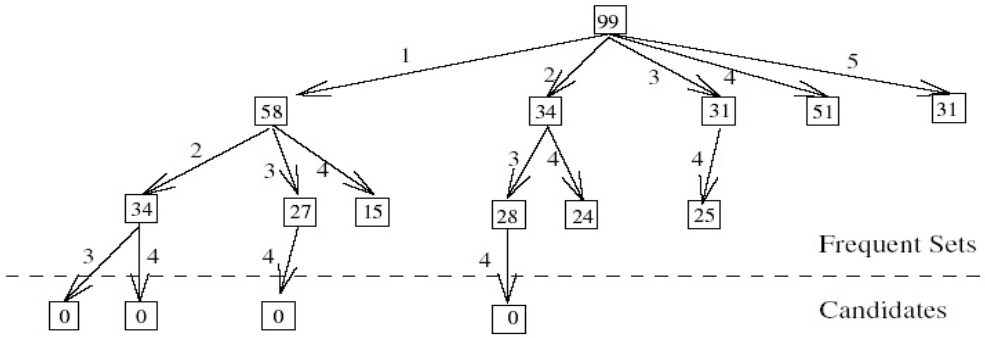
4.7.2 Από Κανόνες Αλληλοσυσχέτισης σε Συνάθροιση

Στην εργασία [48] πρότεινεται η συνάθροιση δεδομένων βασισμένη σε κανόνες αλληλοσυσχέτισης δυαδικών αντικειμένων. Σε αυτή τη μέθοδο ο γράφος συσχέτισης δημιουργήθηκε θεωρώντας τα στοιχεία από τα δεδομένα σαν ένα σύνολο και θέτοντας τις δυαδικές συσχετίσεις σαν κόμβους του γράφου. Το ελάχιστο αντιπροσωπευτικό δέντρο (minimum spanning tree) είναι αυτό που δίνει την συνάθροιση των αντικειμένων. Ελαττώνοντας το γράφο κατάλληλα μπορούμε να έχουμε πολύ καλά αποτελέσματα συνάθροισης , ένα καλό κριτήριο για να συνέχεια του κομματιάσματος είναι πολύ δύσκολο να βρεθεί.

Στην εργασία [49] προτείνεται ένας αλγόριθμος συνάθροισης βασισμένος στα συχνά σύνολα (frequent sets) που βρίσκονται κατά την εξόρυξη κανόνων αλληλοσυσχέτισης. Τα συχνά-σύνολα είναι σύνολα από στοιχεία των δεδομένων τα οποία έχουν «στήριξη» (support) μεγαλύτερη από την ελάχιστη τιμή που έχει επιλέξει ο χρήστης. Με αυτά τα σύνολα δημιουργείται ένας γράφος και ένας κατάλληλος αλγόριθμος τμηματοποίησης χρησιμοποιείται για να βρεθούν σύνολα που ταξινομούν τα δεδομένα.

Βασισμένοι στις δύο προηγούμενες προτάσεις υλοποιήσαμε ένα αλγόριθμο που εκμεταλλευόμενος την γνώση από τους κανόνες συσχέτισης ταξινομεί σύνολα αντικειμένων από τα δεδομένα μας. Χρησιμοποιήσαμε τις ίδιες δομές με τους κανόνες αλληλοσυσχέτισης. Ο αλγόριθμος υλοποιείται όπως παρακάτω.

- Ο parser διαβάζει αρχικά το DTD αρχείο και αποθηκεύει τη δομή του XML αρχείου. Αμέσως μετά αρχίζει η ανάλυση του XMLαρχείου.
- Τα επιλεγμένα, από το χρήστη, αποθηκεύονται σε μία δενδρική δομή την prefix-tree. Η δομή αυτή ,εκτός από την αποθήκευση τιμών των αντικειμένων (attribute values), αποθηκεύει και την ημι-δομημένη δομή των XML αρχείων.
- Στη μεταφορά αυτή κάνοντας κατάλληλο 'φιλτράρισμα', διατηρούμε μόνο την αναγκαία προς επεξεργασία πληροφορία. Οι κόμβοι δεν περιέχουν τα σύνολα, αλλά μόνο τις πληροφορίες για τα σύνολα (π.χ. μετρητές). Κάθε ακμή στο δέντρο αντιπροσωπεύει ένα στοιχείο, και κάθε κόμβος περιέχει τις πληροφορίες για το σύνολο των στοιχείων που βρίσκονται από την πορείας ξεκινώντας από την ακμή ως τη ρίζα.
- Κάθε κόμβος της δενδρικής δομής αντιπροσωπεύει ένα σύνολο από στοιχεία. Επίσης περιέχει τον αριθμό εμφανίσεων του συνόλου των στοιχείων που αντιπροσωπεύει. Οπότε για κάθε σύνολο στοιχείων γνωρίζουμε το ποσοστό εμφάνισης του στο σύνολο των δεδομένων γιατί γνωρίζουμε τον αριθμό εμφανίσεων του αλλά και τον αριθμό συναλλαγών στη πηγή δεδομένων (από τη ρίζα του δένδρου). Συνάθροιση των στοιχείων απαιτείται προκειμένου να αποφευχθεί η επανάληψη συνόλων, διαφορετικά διάφοροι κόμβοι θα αντιστοιχούσαν στο ίδιο σύνολο.



Σχήμα 20. Prefix-tree για διαδικασίες συνάθροισης

Όπως βλέπουμε και από το Σχήμα 20, το ποσοστό εμφάνισης του συνόλου {1,2} στα δεδομένα μας είναι 34/99 ενώ για το {1,4} είναι 15/99. Δίνοντας στο χρήστη τη δυνατότητα να επιλέξει το ελάχιστο ποσοστό εμφάνισης της συστάδας (cluster) στα δεδομένα και τον ελάχιστο αριθμό στοιχείων ανά συστάδα, συναθροίσουμε τα στοιχεία των δεδομένων μας σε σύνολα.

Με τη παραπάνω διαδικασία επιτυγχάνεται:

- Μια πρωτότυπη διαδικασία συνάθροισης δεδομένων, όπου κάθε κανόνας αποτελεί τη περιγραφή κάποιου συνόλου συνάθροισης (cluster), και
- Μια πρωτότυπη διαδικασία 'φιλτραρίσματος' (μείωσης) των επαγομένων κανόνων αλληλοσυσχέτισης – ένα από τα βασικά προβλήματα στη χρήση τεχνολογιών ARM, όπου οι πιο 'σχετικοί' και 'αξιόπιστοι' κανόνες, από ένα δυνητικά μεγάλο σύνολο, πρέπει να επιλεγούν και να επιδειχθούν στο χρήστη.

4.8 Συνάθροιση με τον Αλγόριθμο *k*-Means

Η πιο κοινή αντιπροσώπευση μιας συνάθροισης είναι η εύρεση αντιπροσωπευτικών κέντρων κάθε ομάδας. Αυτό είναι αποτελεσματικό για τις συμπαγείς και ισοτροπικές ομάδες στοιχείων, αλλά όχι για τις επιμηκυμένες ή ανισότροπες ομάδες. Σε ορισμένες εφαρμογές, οι ακρότητες μιας ομάδας χρησιμοποιούνται για να διαμορφώσουν τις συνδετικές εκφράσεις στα σύνολα κανόνα.

Ο *k-Means* είναι ένας από τους απλούστερους και αποδοτικότερους αλγορίθμους εκμάθησης που λύνουν το πρόβλημα της συνάθροισης (χωρίς προκαθορισμένη γνώση). Η πολυπλοκότητα του είναι της τάξεως *O(n)* όπου *n* είναι ο αριθμός της διάστασης των δεδομένων. Η διαδικασία ακολουθεί έναν απλό και εύκολο τρόπο να ομαδοποιηθεί ένα δεδομένο σύνολο στοιχείων μέσω προκαθορισμένου αριθμού ομάδων (υποθέστε τις *k* ομάδες).

- Η κύρια ιδέα είναι να καθοριστούν *k* κέντρα, ένα για κάθε συστάδα. Αυτά τα κέντρα πρέπει να τοποθετηθούν με τέτοιο "έξυπνο" τρόπο γιατί διαφορετικές θέσεις προκαλούν διαφορετικό αποτέλεσμα. Έτσι, η καλύτερη επιλογή είναι να τοποθετηθούν όσο το δυνατόν περισσότερο μακριά ο ένας μακριά από τον άλλον. Το επόμενο βήμα είναι να ληφθεί κάθε σημείο που ανήκει σε ένα δεδομένο σύνολο στοιχείων και να συνδεθεί στο κοντινότερο κέντρο.
- Όταν κανένα σημείο δεν είναι εκκρεμές, το πρώτο βήμα ολοκληρώνεται και γίνεται μια ομαδοποίηση. Σε αυτό το σημείο πρέπει να υπολογίσουμε εκ νέου νέα κέντρα *k* ως βαρύκεντρα των ομάδων ως αποτέλεσμα του προηγούμενου βήματος. Αφότου έχουμε αυτά τα νέα κέντρα *k*, μια νέα σύνδεση πρέπει να γίνει μεταξύ των ίδιων καθορισμένων σημείων στοιχείων και των κοντινότερων νέων κέντρων. Ένας βρόχος έχει παραχθεί. Ως

αποτέλεσμα αυτού του βρόχου μπορούμε να παρατηρήσουμε ότι τα κέντρα *k* αλλάζουν τη θέση τους βαθμιαία έως ότου δεν γίνονται άλλες αλλαγές. Με άλλα λόγια όταν τα κέντρα δεν μεταβάλλονται πλέον. Τέλος, αυτός ο αλγόριθμος στοχεύει στην ελαχιστοποίηση μιας μονάδας μέτρησης όπου σ

λάθους (squared error). Ο τύπος της συνάρτησης αυτής είναι:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

όπου $\left\| x_i^{(j)} - c_j \right\|^2$ είναι μία μετρική απόστασης (Ευκλείδεια , Minkowski, Mahalanobis ή Hausdorff) μεταξύ των $x_i^{(j)}$ και του σημείου του κέντρου c_j. Είναι μία απεικόνιση της απόστασης της «περίπτωσης» δεδομένων από το αντίστοιχο κέντρο της ομάδας.

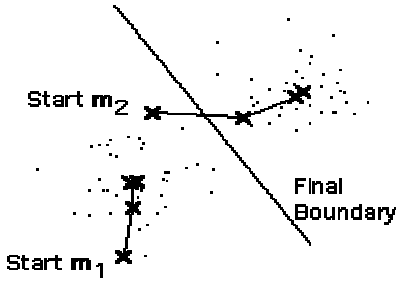
Ο αλγόριθμος αποτελείται από τα παρακάτω βήματα :

1. Θέτουμε K σημεία στο διάστημα που αντιπροσωπεύεται από τα δεδομένα που έχουμε. Αυτά τα σημεία αντιπροσωπεύουν τα αρχικά κέντρα ομάδας.
2. Κατηγοριοποιούμε κάθε «περίπτωση» των δεδομένων μας στην ομάδα με κέντρο το ποιο κοντινό κέντρο προς αυτή.
3. Όταν έχουμε κατηγοριοποιήσει όλες τις «περιπτώσεις» των δεδομένων μας, υπολογίζουμε εν νέου τις θέσεις των κέντρων ως βαρύκεντρο της ομάδας που το αποτελούν.
4. Επαναλαμβάνουμε τα βήματα 2 και 3 έως ότου τα κέντρα δεν κινούνται πλέον. Έτσι δημιουργείται η συνάθροιση των αντικειμένων σε ομάδες.

Παράδειγμα K-Means. Υποθέστε ότι έχουμε τα διανύσματα χαρακτηριστικών γνωρισμάτων *n*, $\boldsymbol{x_1}$, $\boldsymbol{x_2}$…, $\boldsymbol{x_n}$ όλα από την ίδια κατηγορία, και ξέρουμε ότι περιέρχονται στις συστάδες *k, k < n.* Έστω $\boldsymbol{m_i}$ ο μέσος όρος των διανυσμάτων στη συστάδα *i*. Εάν οι συστάδες είναι καλά χωρισμένες, μπορούμε να χρησιμοποιήσουμε έναν μικρότερης-απόστασης ταξινομητή για να τα χωρίσουμε. Δηλαδή μπορούμε να πούμε ότι το αντικείμενο $\boldsymbol{X}$ είναι στη συστάδα *i* εάν // $\boldsymbol{X}$ - $\boldsymbol{m_i}$ // είναι το ελάχιστο όλων των αποστάσεων *k*. Αυτό προϋποθέτει την ακόλουθη διαδικασία για τους *k* μέσους:

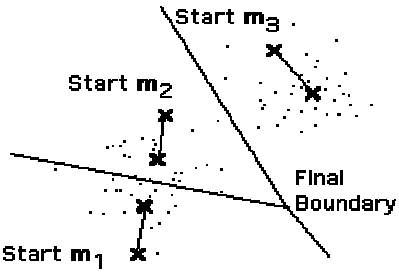
- κάντε τις αρχικές εικασίες για τα κέντρα $m_1, \ldots, m_k$
- έως ότου δεν υπάρχει καμία αλλαγή σε οποιοδήποτε μέσο
  - χρησιμοποιήστε τα κατ' εκτίμηση κέντρα να ταξινομηθούν τα δείγματα στις συστάδες
  - για το i από 1 στο k
    - αντικαταστήστε m_i με το μέσο όρο όλων των δειγμάτων για τη συστάδα

Στο Σχήμα 21, βλέπουμε ένα παράδειγμα πως τα κέντρα δυο συστάδων μετακινούνται με τα βήματα του αλγόριθμου.



Σχήμα 21. Παράδειγμα *k*-Means με δύο κέντρα

Το κυριότερο πρόβλημα του αλγόριθμου *k*-Means είναι ότι πρέπει να προκαθοριστεί ο αριθμός των συστάδων. Τις περισσότερες φορές δεν έχουμε κάποιο στοιχείο που να μας λέει πόσες συστάδες υπάρχουν, οπότε πρέπει να επιλέξουμε τον αριθμό των συστάδων από την αρχή. Στο προηγούμενο παράδειγμα αν είχαμε τρεις συστάδες το αποτέλεσμα θα ήταν διαφορετικό. Στο σχήμα 22 φαίνεται το παράδειγμα με 3 συστάδες.



Σχήμα 22. Παράδειγμα *k*-Means με τρία κέντρα

Δυστυχώς δεν υπάρχει καμία θεωρητική λύση για να βρεθεί ο βέλτιστος αριθμός συστάδων για οποιοδήποτε δεδομένο σύνολο στοιχείων. Μια απλή

προσέγγιση είναι να συγκρίνουν τα αποτελέσματα των πολλαπλών εφαρμογών του αλγόριθμου με τις διαφορετικές κατηγορίες *k* και να επιλέγει ο καλύτερος σύμφωνα με ένα δεδομένο κριτήριο. Πρέπει όμως να είμαστε προσεκτικοί επειδή το αυξανόμενο *k* οδηγεί στις μικρότερες τιμές λάθους εξ ορισμού, αλλά και έναν αυξανόμενο κίνδυνο για πολλές κατηγορίες. Δίνοντας στο χρήστη τη δυνατότητα να επιλέξει το ελάχιστο ποσοστό εμφάνισης της συστάδας (cluster) στα δεδομένα και τον ελάχιστο αριθμό στοιχείων ανά συστάδα, ταξινομούμε τα στοιχεία των δεδομένων μας σε σύνολα.

Αν και μπορεί να αποδειχθεί ότι η διαδικασία τερματίζει, ο αλγόριθμος *k*-Means δεν βρίσκει απαραιτήτως τη βέλτιστη διαμόρφωση που αντιστοιχεί στο ολικό ελάχιστο της συνάρτησης του λάθους. Ο αλγόριθμος είναι επίσης σημαντικά ευαίσθητος στα αρχικά τυχαία επιλεγμένα κέντρα συστάδων. Ο *k*-Means μπορεί να χρειαστεί να τρέχει περισσότερες από μία φορές για να μειωθεί αυτή η επίδραση.

Γενικά, ο *k*-Means είναι ένας απλός αλγόριθμος που έχει προσαρμοστεί σε πολλές περιοχές προβλημάτων. Έτσι, η βασική διαδικασία συνάθροισης που υλοποιείται στο σύστημα HealthObs στηρίζεται στον αλγόριθμο k-Means, με τις κατάλληλες προσαρμογές και επεκτάσεις ώστε να λειτουργεί σε δεδομένα δομημένα σε XML-έγγραφα.

5. ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

Η παρούσα μεταπτυχιακή εργασία παρουσιάζει το πρόβλημα της ανακάλυψης και εξαγωγής γνώσης από κατανεμημένες και ετερογενείς πηγές πληροφοριών. Δομικός λίθος της εργασίας αυτής είναι η μεταπτυχιακή εργασία του Γ. Χριστοφή [24]. Η αναφερθείσα μεταπτυχιακή εργασία αποτέλεσε βασικά μια 'μελέτη βιωσιμότητας' (feasibility-study') σε θέματα ανακάλυψης γνώσεων από κατανεμημένες και ετερογενείς κλινικές πηγές δεδομένων.

Τα μειονεκτήματα της προαναφερθείσας εργασίας αυτής είναι:

- Δημιουργούνται έγγραφα XML από τα ετερογενή δεδομένα τα οποία πρέπει να πληρούν ένα «στατικό» τρόπο αναπαράστασης (συγκεκριμένο DTD).
- Όταν υπάρχει μεγάλος όγκος δεδομένων απαιτείται «κομμάτιασμα» (sampling) των δεδομένων για να λειτουργήσει η εφαρμογή.
- Η υλοποίηση του αλγορίθμου εξόρυξης συσχετίσεων (όπου είναι ο μόνος που υλοποιήθηκε) έχει τα εξής τρία μειονεκτήματα :
- Δεν δίνεται στον χρήστη η δυνατότητα να επιλέξει ποια από τα στοιχεία θέλει να συμμετέχουν στη μέθοδο της συσχέτισης. Έτσι όλα τα στοιχεία που βρίσκονται στα δεδομένα (ακόμα και στοιχεία που δεν έχουν να προσφέρουν πολύτιμη πληροφορία στην εξόρυξη δεδομένων)
- Στα αποτέλεσμα των κανόνων συσχέτισης δεν μπορούμε να έχουμε πάνω από ένα στοιχείο στο «σώμα» του κανόνα.
- Η επιλογή του αλγορίθμου και των δομών δεδομένων απαιτούν την πολλαπλή ανάγνωση των δεδομένων (από το XML αρχείο), πράγμα του καθιστά πάρα πολύ αργό το σύστημα για μεγάλα αρχεία.
- Δεν υπάρχουν οι κατάλληλες διεπαφές χρήστη ούτως ώστε να διευκολύνει τους χρήστες (οι οποίοι θα είναι κυρίως από το πεδίο της ιατρικής).

Με γνώμονα την παραπάνω μεταπτυχιακή εργασία και δίνοντας βάση στα μειονεκτήματα που προαναφέραμε, υλοποιήσαμε ένα σύστημα του οποίου η βασική συνεισφορά είναι η συνεργασία και η τροποποίηση όλων των λειτουργιών εξόρυξης δεδομένων , ώστε να είναι απόλυτα εφαρμόσιμες στα XML έγγραφα. Συγκεκριμένα, επικεντρωνόμαστε και αντιμετωπίζουμε το πρόβλημα εξόρυξης γνώσης μεταξύ των δεδομένων που είναι αποθηκευμένα σε κατανεμημένα και ετερογενή συστήματα. Το περιβάλλον εφαρμογής της προσέγγισης μας είναι ,και εδώ ,το HYGEIAnet: The Integrated Health Care Network of Crete. Η βασική πρόκληση είναι 'πως οι λειτουργίες οι οποίες προέρχονται από τον χώρο του εξόρυξη δεδομένων και της Μηχανικής

Μάθησης, υιοθετούνται και γίνονται λειτουργικές σε ένα περιβάλλον ημι-δομημένων δεδομένων'.

Για το λόγο αυτό, προτείνεται και υλοποιείται μια πολυσύνθετη διαδικασία ολοκλήρωσης, η οποία αντιμετωπίζει θέματα όπως:

- αξιόπιστη ομογενοποίηση και ολοκλήρωση των ετερογενών δεδομένων
- επεξεργασία (statistical analysis, εξόρυξη δεδομένων, κ.τ.λ) των δεδομένων
- παρουσίαση των αποτελεσμάτων
- υλοποίηση ενός συστήματος φιλικό προς το χρήστη.

Αυτή η προσέγγιση ολοκλήρωσης υπαγορεύεται από τον συνδυασμό και την παρουσία πολλών τεχνολογιών και λειτουργιών. Οι λειτουργίες αυτές συσχετισμένες με σύγχρονες, προχωρημένες και αποτελεσματικές αναπαραστάσεις μοντέλων, διαμορφώνουν και προσδιορίζουν ένα σκελετό και ένα περιβάλλον, στο οποίο μπορούν να εκπονηθούν πλέον έξυπνα και αποτελεσματικά όλες οι απαιτούμενες και αναγκαίες KDD διεργασίες.

Γενικά, η παρούσα εργασία και το σύστημα HealthObs επεκτείνει την προαναφερθείσα εργασία σε ένα ολοκληρωμένο περιβάλλον το οποίο είναι:

- περισσότερο ***ευέλικτο*** - με την έννοια της εύκολης προσαρμογής διαφορετικών πεδίων εφαρμογής, όπως παρέχεται από τις σχετικές λειτουργίες σημασιολογικής μοντελοποίησης και ομογενοποίησης ετερογενών πληροφοριών);
- περισσότερο ***πλούσιο*** – με την έννοια της ενσωμάτωσης διαφορετικών, περισσότερο αποδοτικών, και περισσότερο παραμετρικοποιήσεων μεθόδων ανακάλυψης γνώσεων;
- περισσότερο ***φιλικό*** στη χρήση – με την έννοια της σχεδίασης και υλοποίησης ενός *πρωτότυπου και ευέλικτου διάμεσου επικοινωνίας ανθρώπου-συστήματος*, και στο επίπεδο *διαμόρφωσης των ερωτημάτων* από πλευράς του χρήστη αλλά και στο επίπεδο της ***απεικόνισης*** των παραγωμένων αποτελεσμάτων (βλέπε τα δύο επόμενα κεφάλαια όπου το HCI του HealthObs παρουσιάζεται με λεπτομέρεια και ως προς τη λειτουργικότητά του και ως προς τη χρήση του σε συγκεκριμένα παραδείγματα εφαρμογής).

5.1. Σχετικές Εργασίες ARM

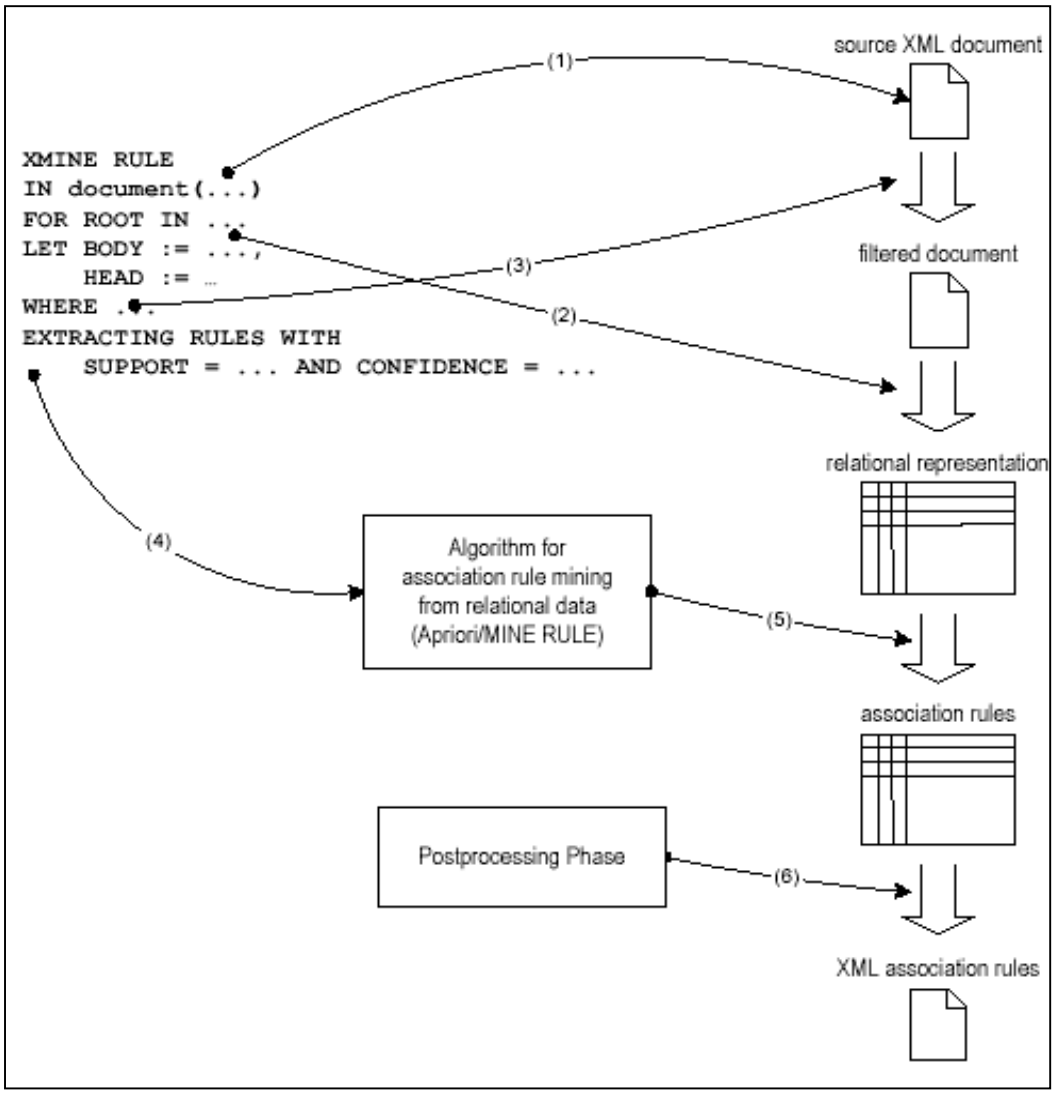
Αν και υπάρχουν πολλές αναφορές και δημοσιεύεις για υλοποίηση εξαγωγής γνώσεων σε βάσεις δεδομένων, λίγη έρευνα έχει γίνει για δεδομένα σε XML μορφή. Οι [25] προσέγγισαν το πρόβλημα της εξόρυξης σχήματος από ημι-δομημένα δεδομένα (XML αρχεία) και πρότειναν ένα αλγόριθμο εξαγωγής κανόνων αλληλοσυσχέτισης όταν ένα μονοπάτι θεωρείται αντικείμενο για εξόρυξη σχήματος. Οι [26], [27] και [28] προσέγγισαν το πρόβλημα εξόρυξης σχήματος ημι-δομημένων δεδομένων βασιζόμενοι σε δενδροειδές «πρότυπα».

Οι περισσότερες έρευνες που συνδέουν την εξόρυξη δεδομένων με τα ημι-δομημένα δεδομένα εστιάζονται στο πρόβλημα της εξαγωγής αντιπροσωπευτικού σχήματος δομής των δεδομένων. Ελάχιστες είναι οι έρευνες που συνδέονται με την πληροφορία του περιέχουν τα δεδομένα ημι-δομημένων αρχείων. Οι [29] υλοποίησαν μια μέθοδο με την οποία εξάγουν κανόνες αλληλοσυσχέτισης ενοποιώντας δομημένες τιμές στοιχείων με έννοιες που εξάγονται από ημι-δομημένα στοιχεία (ιστοσελίδες). Η υλοποίηση στηρίζεται στην προεπεξεργασία των ιστοσελίδων και την δημιουργία προτύπων για τις ιστοσελίδες αυτές. Μια ακόμη έρευνα που υλοποιεί αλγόριθμο μηχανικής μάθησης σε ημι-δομημένα δεδομένα είναι η [30]. Βασίζεται σε κανόνες ταξινόμησης των XML δεδομένων χρησιμοποιώντας τη συχνότητα εμφάνισης υποσυνόλων των δεδομένων.

Παράλληλα με αυτή την μεταπτυχιακή εργασία υλοποιήθηκε μια άλλη μεταπτυχιακή εργασία [31] στο Dipartimento di Elettronica e Information Politecnico di Milano. Σκοπός της είναι η εξόρυξη κανόνων αλληλοσυσχέτισης από XML αρχεία. Βασίζεται στο [32] ,εμπνευσμένο από το [33] και την εργασία του [34] από το πεδίο των βάσεων δεδομένων. Έχοντας ως υπόβαθρο την XPath υλοποίηση του [35] μπορούν να γίνουν επερωτήσεις στα δεδομένα του XML με γραμμή εντολών (command line). Το αποτέλεσμα της επερώτησης αποθηκεύεται σε ένα σχεσιακό πίνακα (όπως και στις βάσεις δεδομένων) και πάνω σε αυτό το σχεσιακό πίνακα υλοποιείται ένας αλγόριθμος αλληλοσυσχετίσεων. Ο αλγόριθμος που επιλέχτηκε είναι ο Apriori. Όταν έχουν εξαχθεί οι κανόνες αποθηκεύονται σε ένα αρχείο σε μορφή XML. Υπάρχουν δύο ακόμη δημοσιεύσεις σε συνέδρια, οι οποίες παρουσιάζουν την εργασία αυτή, οι [36] και [37].

Στο σχήμα 23 απεικονίζονται τα στάδια της επεξεργασίας των δεδομένων (το σχήμα είναι από την δημοσίευση [37]). Είναι φανερό ότι οι κανόνες αλληλοσυσχέτισης υλοποιούνται σε ένα σχεσιακό πίνακα. Με αυτό τον τρόπο το πρόβλημα αναγάγετε σε εξόρυξη γνώσεων από σχεσιακό πίνακα, οπότε οι

ιδιότητες της XML χάνονται. Το σύστημα δεν είναι ανεξάρτητο πλατφόρμας , αφού χρειάζεται σχεσιακός πίνακας (και συνεπώς βάση δεδομένων) για την αποθήκευση των δεδομένων. Επίσης τα δεδομένα σε αυτή τη μορφή δεν είναι μεταφέρσιμα, επεξεργάσιμα και αναγνώσιμα.



Σχήμα 23 XRuleMine – τα στάδια της επεξεργασίας δεδομένων.

Μία ακόμη σχετική εργασία είναι αυτή των Wan και Dobbie [61]. Στόχος της εργασίας αυτής είναι να εξάγει κανόνες συσχέτισης από XML αρχεία χρησιμοποιώντας μόνο την XQuery. Η υλοποίηση των κανόνων συσχέτισης έγινε εξολοκλήρου σε XQuery γλώσσα. Το κύριο πρόβλημα της υλοποίησης αυτής είναι ότι δεν λειτουργεί για οποιαδήποτε «έγκυρη» (valid) XML μορφή. Συγκεκριμένα αν η δομή του XML αρχείου περιέχει σε κάποιο κόμβο

υποκόμβους και «φύλλα» το αρχείο χρειάζεται αναδόμηση σε μία πιο απλή μορφή. Επίσης για την υλοποίηση των κανόνων συσχέτισης απαιτούνται πολλαπλά περάσματα στα δεδομένα. Το σύστημα δεν έχει δοκιμαστεί σε μεγάλο όγκο δεδομένων.

Μια ακόμη σχετική εργασία είναι αυτή του Andrew Edmonds [62]. Σε αυτή την εργασία υλοποιήθηκε ο αλγόριθμος ID3 για την εξόρυξη γνώσης από XML αρχεία. Ο αλγόριθμος αυτός είναι ένας «εποπτικής μάθησης» (supervised learning) αλγόριθμος. Χρειάζονται κάποια κατηγοριοποιημένα παραδείγματα των δεδομένων τα οποία λαμβάνονται σαν γνώση με σκοπό την κατηγοριοποίηση των δεδομένων. Επίσης η υλοποίηση έγινε σε java με την «διεπαφή» (interface) DOM της java. Η διεπαφή DOM αποθηκεύει όλο το αρχείο στη μνήμη (με διάφορες δομές) για την επεξεργασία, πρόσθεση ή διαγραφή δεδομένων. Δεδομένου ότι ο όγκος πληροφορίας πρέπει να είναι πολύ μεγάλος για να έχουμε ενδιαφέροντα και αξιόπιστα αποτελέσματα με αλγόριθμους εξόρυξης δεδομένων, η υλοποίηση αυτή δεσμεύει μεγάλο ποσοστό μνήμης και καθιστά το σύστημα ασταθές για πολύ μεγάλο όγκο δεδομένων. Ένα ακόμη αρνητικό της υλοποίησης αυτής είναι ο χειρισμός των άγνωστων τιμών. Συγκεκριμένα όταν κάποιο «στοιχείο» (item) δεν έχει τιμή αγνοείται ολόκληρη η «περίπτωση» (instance) του περιέχει το στοιχείο αυτό.

Συνοψίζοντας το σχήμα 24 απεικονίζει τις διαφορές των έξι σχετικών υλοποιήσεων στον τομέα της εξόρυξης γνώσης από XML αρχεία.

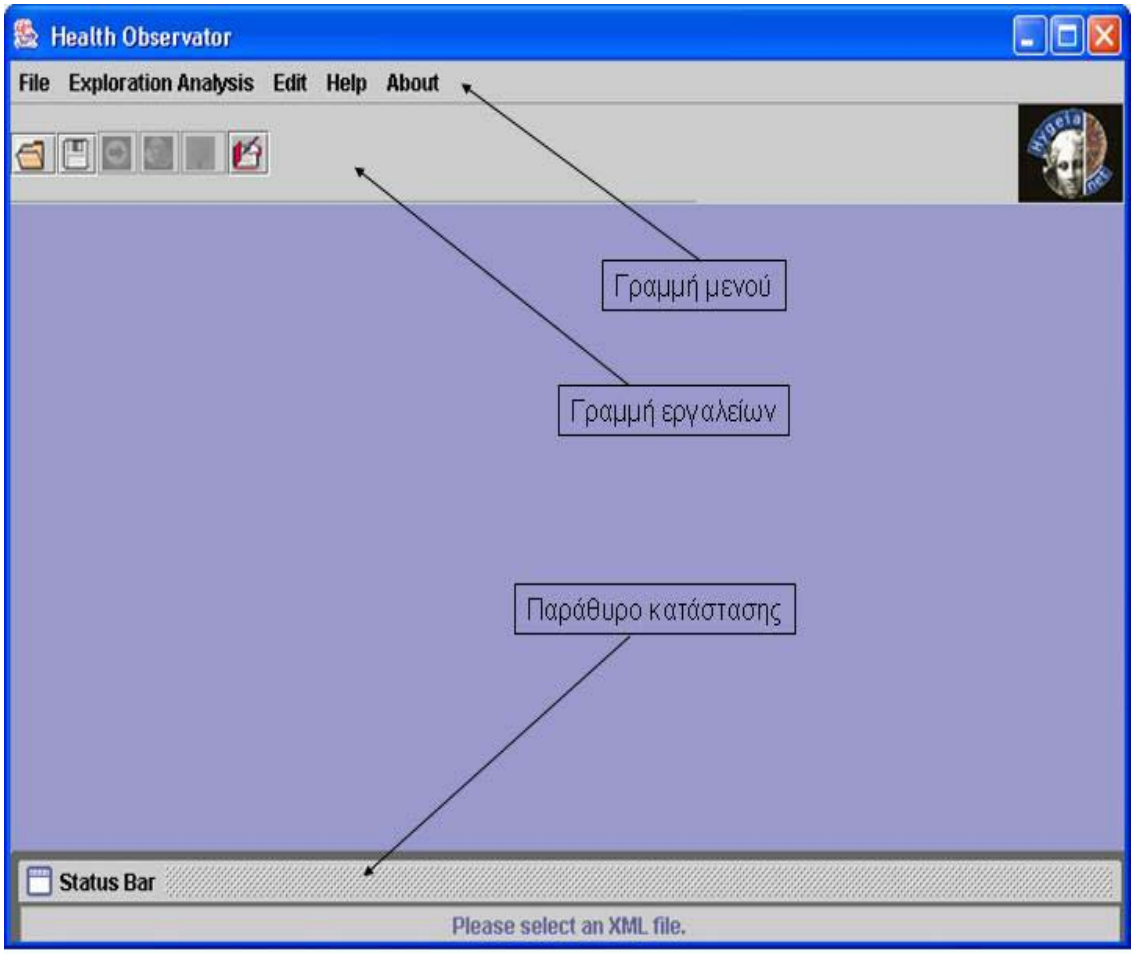
	K. Χριστοφής	R. Carnevale	J.Wan, G.Dobbie	A. Edmonds	HealthOb
1 Προσπέλαση δεδομένων μονο μια φορά	✗	✗	✗	✓	✓
2 If-then Επιλογή	✗	✗	✗	✗	✓
3 Λειτουργία με μεγάλο ογκο δεδομένων	✗	✓	✗	✗	✓
4 Προσαρμογή δεδομένων στις ανάγκες του χρήστη	✗	✗	✓	✓	✓
5 Γραφική απεικόνιση αποτελεσμάτων	✗	✗	✗	✗	✓
6 Διεπαφές Χρήσης	✗	✗	✗	✗	✓

Σχήμα 24 XRuleMine – τα στάδια της επεξεργασίας

6. Το Περιβαλλον Εργασιας του HealthObs: Εγχειριδιο Χρησης

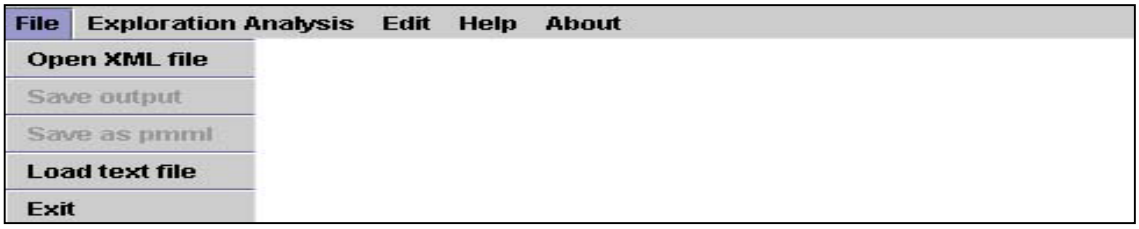
Σε αυτό το κεφάλαιο θα περιγράψουμε τις λειτουργίες του συστήματος HealthObs παρουσιάζοντας τις διεπαφές και τις ευκολίες χρήσης του μας προσφέρει. Το σύστημα απευθύνεται σε απλούς χρήστες ,στη συγκεκριμένη περίπτωση σε ανθρώπους από το χώρο της ιατρικής. Για το λόγο αυτό κρίνεται αναγκαία η δημιουργία απλών και εύχρηστων διεπαφών που αποκρύπτουν την πολυπλοκότητα του συστήματος και ταυτόχρονα παρέχουν όλες τις λειτουργίες του συστήματος.

Η πρώτη επαφή που έχουμε με το σύστημα είναι το περιβάλλον εργασίας του HealthObs.



Σχήμα 25: Το περιβάλλον εργασίας του HealthObs

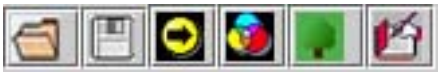
Όπως βλέπουμε και στο σχήμα 25 το περιβάλλον εργασίας αποτελείται από μια γραμμή μενού, μια γραμμή εργαλείων και το παράθυρο κατάστασης.



Σχήμα 26 Γραμμή μενού

Η γραμμή μενού (σχήμα 26) περιέχει τις βασικές επιλογές από τις οποίες εμφανίζονται πτυσσόμενα μενού με πρόσθετες σχετικές επιλογές. Οι επιλογές που περιέχει η γραμμή μενού είναι:

- File: Περιέχει τις υπό-επιλογές για τη διαχείριση των αρχείων δεδομένων. Από αυτές τις επιλογές είναι δυνατή η εκκίνηση της διαδικασίας εξόρυξης γνώσης για κάποιο XML αρχείο (**Open XML file**), η αποθήκευση των αποτελεσμάτων ,εφόσον έχει τελειώσει η διαδικασία εξόρυξης, σε απλό αρχείο κειμένου (**Save output**) ή σε μορφή σύμφωνα με τα διεθνή πρότυπα της pmml (**Save as pmml**). Επίσης με την επιλογή του υπό-μενού (**Load text file**) μπορούμε να δούμε την γραφική απεικόνιση κάποιον αποτελεσμάτων που έχουμε αποθηκεύσει ως απλό κείμενο ή με μορφή pmml.
- Exploration Analysis: Περιέχει τις υπό-επιλογές για την επιλογή του αλγορίθμου που θέλουμε να εφαρμόσουμε στα δεδομένα μας. Οι επιλογές είναι (**Association Rules**) για εφαρμογή κανόνων συσχέτισης ,(**Clustering**) για εφαρμογή κατηγοριοποίησης μέσω κανόνων συσχέτισης και (**Classification**) για την εφαρμογή κατηγοριοποίησης του K-Means αλγόριθμου.
- Edit: Περιέχει μόνο την επιλογή (**intervals**) η οποία μας δίνει τη δυνατότητα να επεξεργαστούμε κάποιο αρχείο ομογενοποίησης δεδομένων με την υπηρεσία Common clinical term Reference (CCTR) όπως την παρουσιάσαμε στο προηγούμενο κεφάλαιο.
- Help: Περιέχει την υπό-επιλογή (help) από την οποία γίνεται κλήση της άμεσης βοήθειας του συστημάτων HealthObs.
- About: Περιέχει την υπό-επιλογή (about) η οποία παρέχει πληροφορίες για το σύστημα.



Σχήμα 27 Γραμμή Εργαλείων

Αμέσως μετά τη γραμμή μενού υπάρχει η γραμμή εργαλείων (σχήμα 27). Η γραμμή εργαλείων είναι συντομεύσεις για κάποιες λειτουργίες του συστήματος. Υπάρχουν έξι συντομεύσεις. Ξεκινώντας από αριστερά προς τα δεξιά έχουμε συντόμευση για το (**Open XML file) ,** (**Save output**) , (**Association Rules**) ,(**Clustering**) , (**Classification**) και (**intervals**).

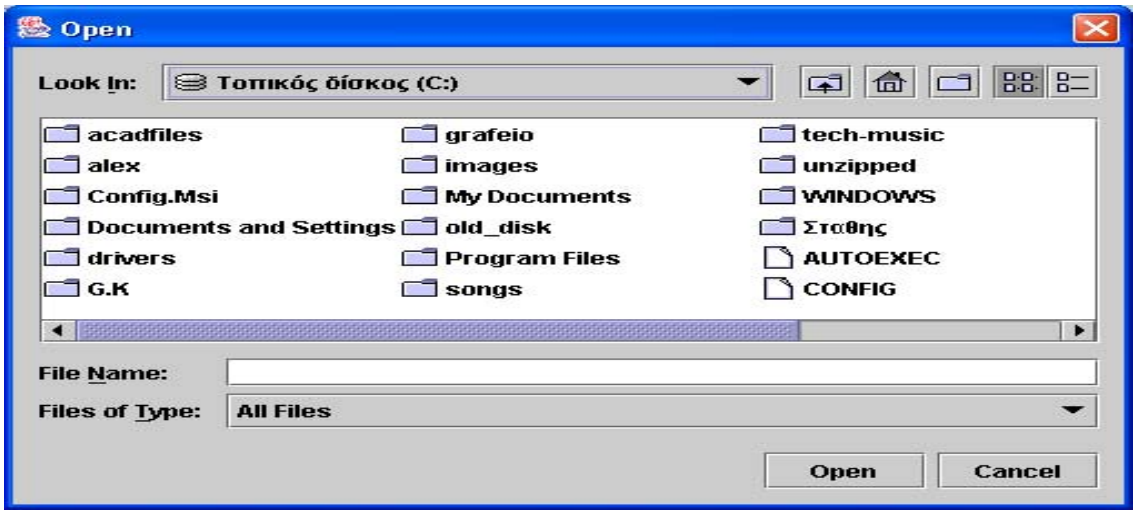
Το τελευταίο στοιχείο που βλέπουμε στο περιβάλλον εργασίας είναι το παράθυρο κατάστασης (σχήμα 28) . Το παράθυρο κατάστασης εκτελεί δύο λειτουργίες.

Σχήμα 28 Παράθυρο κατάστασης

Μας δίνει πληροφορίες για το επόμενο βήμα που πρέπει να κάνουμε εμφανίζοντας κάθε φορά τα κατάλληλα μηνύματα στο κέντρο του παραθύρου, και λειτουργεί ως μπάρα προόδου (progress bar) κατά την διαδικασία της ανάγνωσης των XML αρχείων και της επεξεργασίας των δεδομένων. Ένας προσδιορισμός της προόδου σε αλγόριθμους εξόρυξης δεδομένων είναι υποχρεωτικός γιατί ο χρόνος που είναι ανάλογος των δεδομένων και πολύ μεγάλος (τα δεδομένα για αξιόπιστη εξόρυξη γνώσης πρέπει να είναι πολλά).

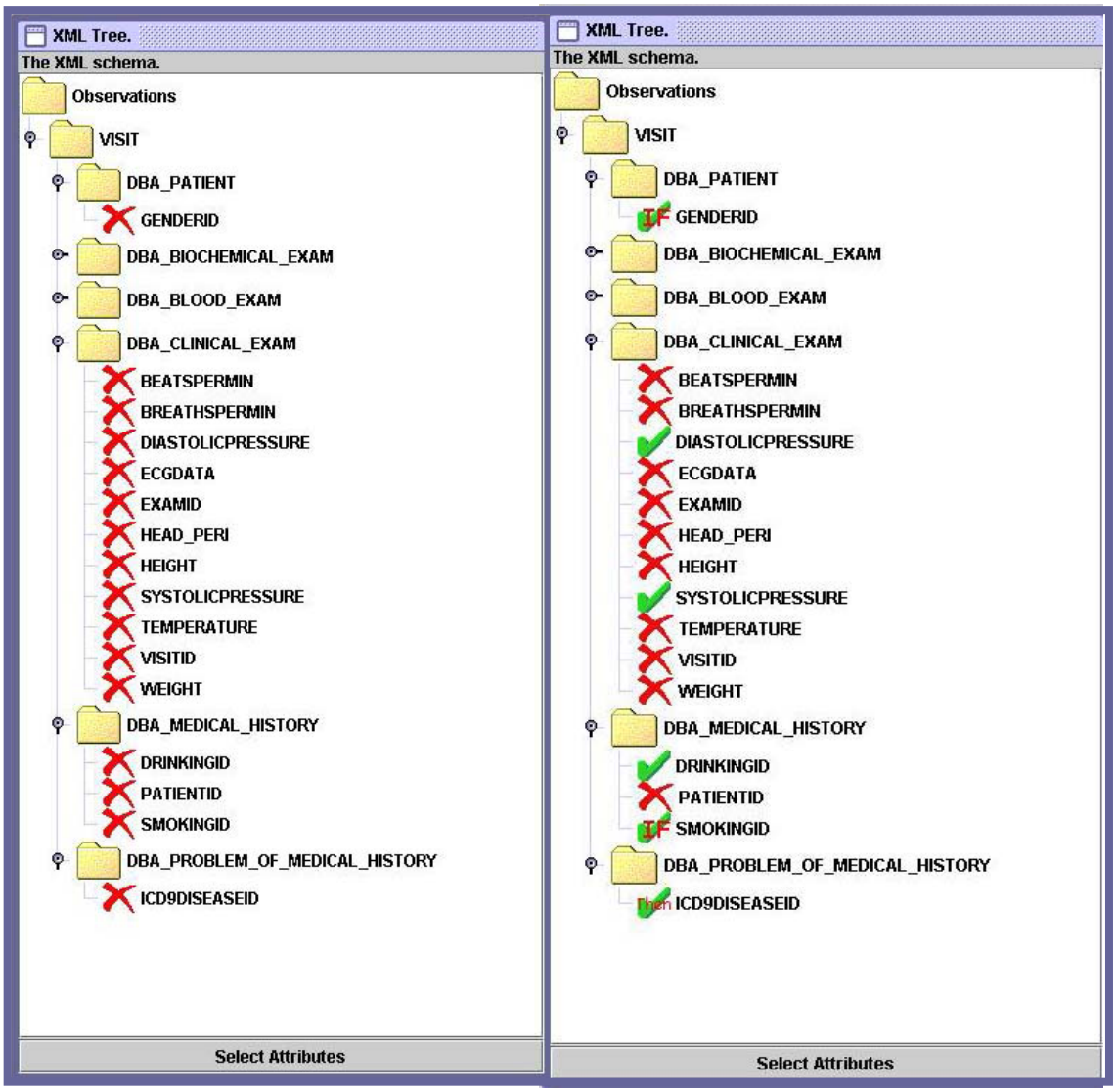
6.1. **Εισαγωγή Δεδομένων και Επιλογή Στοιχείων:** Υποστήριξη στη Διαμόρφωση Ερωτημάτων (query formulation)

Για να ξεκινήσουμε μια διαδικασία εξόρυξης γνώσης από τα δεδομένα μας πρέπει να δώσουμε στο σύστημα το XML αρχείο με τα δεδομένα. Επιλέγοντας το (Open XML file) από την γραμμή μενού ή το αντίστοιχο κουμπί από την γραμμή εργαλείων εμφανίζεται ένα παράθυρο διαλόγου όπως φαίνεται στο σχήμα 29. Βρίσκουμε το XML αρχείο προς επεξεργασία και πατάμε το κουμπί "Open" κάτω δεξιά παράθυρο διαλόγου (βασική προϋπόθεση είναι να βρίσκεται στον ίδιο κατάλογο με το XML αρχείο και το αντίστοιχο DTD αρχείο).



Σχήμα 29 Παράθυρο διαλόγου

Αμέσως μετά στο σύστημα εμφανίζεται το παράθυρο στοιχείων.



Σχήμα 30. Attribute-Selection στο HealthObs

Το παράθυρο των στοιχείων είναι μια γραφική απεικόνιση του σχήματος του XML (από τα δεδομένα μας). Η ιεραρχική δομή των δεδομένων απεικονίζεται στο σχήμα 30, όπως και η δομή αρχείων σε ένα υπολογιστή, με φακέλους που μπορούν να περιέχουν υπό-φακέλους ή και τελικά στοιχεία. Οι γραφικές αναπαραστάσεις που μπορεί να έχει κάθε στοιχείο είναι 4.

 Υποδηλώνει ότι το στοιχείο αυτό περιέχει ένα η περισσότερα υπό-στοιχεία.

 Υποδηλώνει ότι είναι τελικό στοιχείο ,δηλαδή έχει τιμή (αριθμητική ή αλφαριθμητική) , και δεν το έχουμε επιλέξει οπότε το στοιχείο αυτό δεν θα συμμετέχει στον αλγόριθμο εξόρυξης γνώσης που θα επιλέξουμε αργότερα.

 Υποδηλώνει ότι είναι τελικό στοιχείο και το έχουμε επιλέξει οπότε το στοιχείο αυτό θα συμμετέχει στον αλγόριθμο εξόρυξης γνώσης που θα επιλέξουμε αργότερα.

 Υποδηλώνει ότι είναι τελικό στοιχείο και το έχουμε επιλέξει, οπότε το στοιχείο αυτό θα συμμετέχει στον αλγόριθμο εξόρυξης γνώσης. Αν ο αλγόριθμος που θα επιλέξουμε είναι κανόνες συσχέτισης τότε το στοιχείο αυτό θα υπάρχει υποχρεωτικά σε όλους τους κανόνες, στο «κεφάλι» κάθε κανόνα.

 Υποδηλώνει ότι είναι τελικό στοιχείο και το έχουμε επιλέξει, οπότε το στοιχείο αυτό θα συμμετέχει στον αλγόριθμο εξόρυξης γνώσης. Αν ο αλγόριθμος που θα επιλέξουμε είναι κανόνες συσχέτισης τότε το στοιχείο αυτό θα υπάρχει υποχρεωτικά σε όλους τους κανόνες, στο «σώμα» κάθε κανόνα.

Πατώντας πάνω σε κάποιο στοιχείο το οποίο περιέχει άλλα στοιχεία εμφανίζεται ή αποκρύπτετε η λίστα με τα υπό-στοιχεία του. Αν είναι τελικό στοιχείο τότε πατώντας πάνω του το στοιχείο παίρνει μία από τις τέσσερις καταστάσεις τελικού στοιχείου σειριακά. Στο σχήμα 30, έχουμε δύο αναπαραστάσεις του ίδιου XML σχήματος. Αριστερά βλέπουμε το XML σχήμα όπως εμφανίζεται αρχικά (χωρίς να έχουμε επιλέξει κανένα τελικό στοιχείο). Δεξιά βλέπουμε το ίδιο XML σχήμα με επιλεγμένα τα τελικά στοιχεία `DIASTOLICPRESSURE`, `SYSTOLICPRESSURE`, `DRINKINGID`, τα στοιχεία `GENGER`, `SMOKINGID` που θα εμφανίζονται υποχρεωτικά στο κεφάλι των κανόνων (αν έχουμε επιλέξει κανόνες συσχέτισης) και το στοιχείο `ICD9DISEASEID` που θα εμφανίζεται υποχρεωτικά στο σώμα των κανόνων. Όλα τα υπόλοιπα στοιχεία ,τα οποία δεν έχουν επιλεχθεί, δεν θα συμμετέχουν στους αλγόριθμους εξόρυξης γνώσης. Όταν πλέον έχουμε αποφασίσει ποια στοιχεία θέλουμε να συμμετέχουν ,και με ποιο τρόπο, στην εξαγωγή γνώσης τα επιλέγουμε και πατάμε το κουμπί που βρίσκεται στην κάτω πλευρά του παραθύρου στοιχείων και γράφει **Select Attributes**.

6.2. Αλγόριθμοι και Παραμετροποίηση

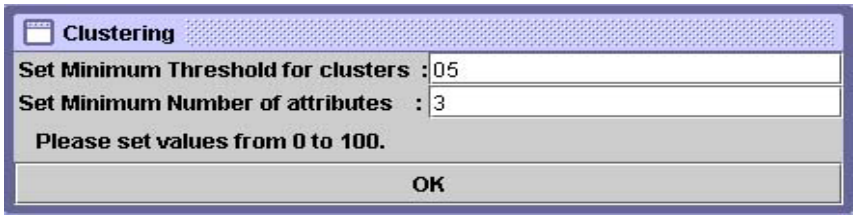
Αφού έχουμε επιλέξει τα στοιχεία μας πρέπει να επιλέξουμε και τον αλγόριθμο που θέλουμε να εφαρμόσουμε στα δεδομένα μας. Μόλις πατήσουμε το **Select Attributes** από το παράθυρο στοιχείων εμφανίζεται από την γραμμή μενού το μενού Exploration Analysis με τα τρία υπομένουν των τριών αλγορίθμων. Αλγόριθμο μπορούμε να επιλέξουμε και από τη γραμμή εργαλείων πατώντας το κατάλληλο εικονίδιο για κάθε αλγόριθμο. Μετά από την επιλογή και του αλγορίθμου το σύστημα ξεκινάει την ανάλυση των δεδομένων. Μόνο τα επιλεγμένα ,από το χρήστη, στοιχεία θα αποθηκευτούν στις δομές του συστήματος. Κατά τη διάρκεια της ανάλυσης των δεδομένων το παράθυρο κατάστασης μας ενημερώνει για το ποσοστό του αρχείου που έχει επεξεργαστεί (τα αρχεία είναι αρκετά μεγάλα με μέγεθος πολλών MB).

Κάθε αλγόριθμος έχει κάποιες παραμέτρους που το σύστημα ζητάει από το χρήστη να τις προσδιορίσει. Μόλις τελειώσει η ανάλυση των δεδομένων εμφανίζεται ένα παράθυρο διαλόγου το οποίο ζητάει από τον χρήστη να δώσει τιμές στις παραμέτρους του αλγορίθμου που επέλεξε.

Για τους κανόνες συσχέτισης ο χρήστης πρέπει να δώσει τιμές για την ελάχιστη "εμπιστοσύνη" (*confidence*) και "στήριξη" (*support*). Όπως φαίνεται και από το σχήμα 31 υπάρχουν κάποιες αρχικές τιμές 5% και 15% τις οποίες ο χρήστης αν θέλει μπορεί να τις αλλάξει. Τελειώνοντας πρέπει να πατήσει το κουμπί **OK** για να ξεκινήσει η διαδικασία εξαγωγής κανόνων συσχέτισης.

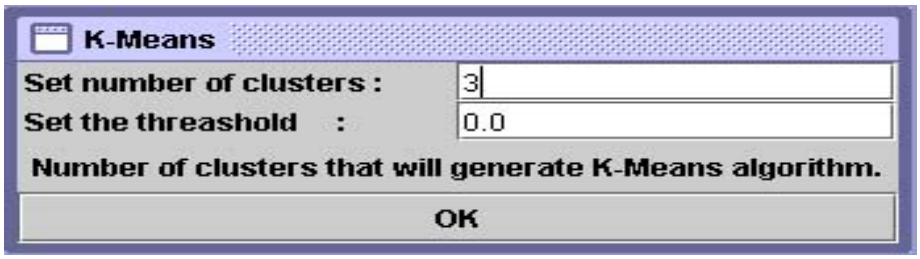
Σχήμα 31. Παράθυρο παραμέτρων κανόνων συσχέτισης

Για την συνάθροιση μέσω κανόνων συσχέτισης ο χρήστης πρέπει να δώσει τιμές για την ελάχιστη "στήριξη" (*support*) της συστάδας των ταξινομημένων στοιχείων και τον ελάχιστο αριθμό στοιχείων ανά συστάδα. Όπως φαίνεται και από το σχήμα 32 υπάρχουν κάποιες αρχικές τιμές 5% και 3 τις οποίες ο χρήστης αν θέλει μπορεί να τις αλλάξει. Τελειώνοντας πρέπει να πατήσει το κουμπί **OK** για να ξεκινήσει η διαδικασία εξαγωγής συνάθροιση μέσω κανόνων συσχέτισης.



Σχήμα 32. Παράθυρο παραμέτρων συνάθροισης

Για την συνάθροιση με τον αλγόριθμο *k*-Means ο χρήστης πρέπει να δώσει τιμές για τον αριθμό των συστάδων που θέλει και το ελάχιστο ποσοστό εμφάνισης για τα στοιχεία που αποτελούν τις συστάδες. Όπως φαίνεται και από το σχήμα 33 υπάρχουν κάποιες αρχικές τιμές 3 και 0 τις οποίες ο χρήστης αν θέλει μπορεί να τις αλλάξει. Τελειώνοντας πρέπει να πατήσει το κουμπί **OK** για να ξεκινήσει η διαδικασία εξαγωγής συνάθροιση μέσω κανόνων συσχέτισης.



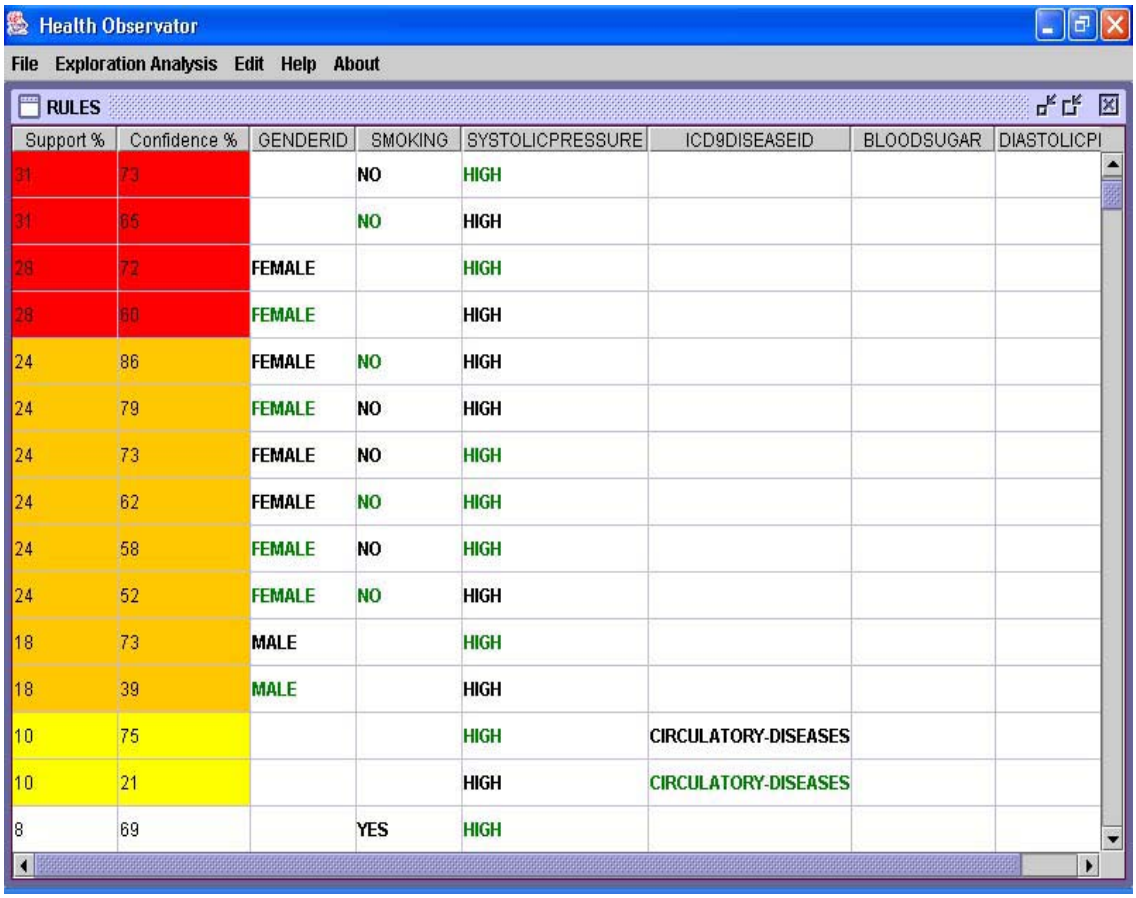
Σχήμα 33. Παράθυρο παραμέτρων K-Means

Έχοντας προσδιορίσει τις παραμέτρους για τον εκάστοτε αλγόριθμο, ξεκινάει η εφαρμογή του αλγορίθμου στα δεδομένα με τις παραμέτρους που έχουν επιλεχθεί. Κάθε αλγόριθμος έχει διαφορετική πολυπλοκότητα και επηρεάζετε διαφορετικά με την αύξηση των δεδομένων. Μεγάλος όγκος δεδομένων (για να εφαρμόσουμε τεχνικές εξόρυξης γνώσης πρέπει να έχουμε μεγάλο όγκο δεδομένων) έχει σαν αποτέλεσμα την καθυστέρηση του αλγορίθμου. Γι αυτό το λόγο υπάρχει το παράθυρο κατάστασης που μας προσδιορίζει την πρόοδο του αλγορίθμου.

6.3. **Εμφάνιση Αποτελεσμάτων** (visualization)

Το σύστημα προσφέρει μια νέα προσέγγιση για την απεικόνιση των εξαγόμενων κανόνων συσχέτισης. Καταρχήν, οι κανόνες συσχέτισης ταξινομούνται (με φθίνουσα σειρά) σύμφωνα με τη δύναμη υποστήριξής (support) τους. Επίσης οι κανόνες ταξινομούνται σε τέσσερις κατηγορίες ανάλογα με το ποσοστό «υποστήριξης» τους. Οι κανόνες που έχουν ταξινομηθεί στην υψηλότερη (μπορούν να θεωρηθούν πολύ ισχυροί) κατηγορία σύμφωνα με το ποσοστό υποστήριξης τους έχουν κόκκινο χρώμα στα πεδία support και confidence. Οι κανόνες που έχουν ταξινομηθεί στην επόμενη κατηγορία (μπορούν να θεωρηθούν αρκετά ισχυροί) έχουν πορτοκαλί χρώμα στα πεδία support και confidence. Οι κανόνες που έχουν ταξινομηθεί στην τρίτη κατηγορία (μπορούν να θεωρηθούν μέτρια ισχυροί) έχουν κίτρινο χρώμα στα πεδία support και confidence. Οι κανόνες που έχουν ταξινομηθεί στην τελευταία κατηγορία (μπορούν να θεωρηθούν μη ισχυροί) έχουν λευκό χρώμα στα πεδία support και confidence. Τα πεδία που εμφανίζονται στην πάνω μεριά του συστήματος είναι το support το confidence και όλα τα στοιχεία που είχε επιλέξει ο χρήστης, τα οποία θα

συμμετείχαν στην εξόρυξη των κανόνων. Κάθε γραμμή είναι ένας κανόνας με την τιμή της «υποστήριξης» του (support) την τιμή της «εμπιστοσύνης» (confidence) και την ποσοτική τιμή του στα αντίστοιχα πεδία στοιχείων. Το «κεφάλι» του κανόνα παρουσιάζεται με μαύρα γράμματα, ενώ το «σώμα» του κανόνα παρουσιάζεται με πράσινη γράμματα. Ένα παράδειγμα της γραφικής αυτής απεικόνισης φαίνεται στο σχήμα 34.

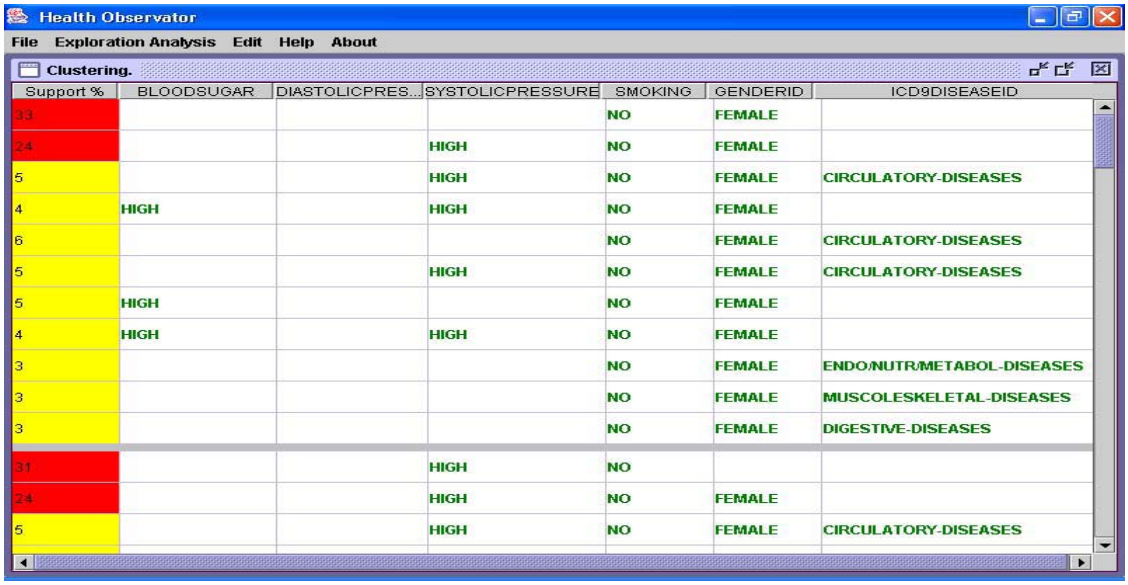


Support %	Confidence %	GENDERID	SMOKING	SYSTOLICPRESSURE	ICD9DISEASEID	BLOODSUGAR	DIASTOLICPI
31	73		NO	HIGH			
31	65		NO	HIGH			
28	72	FEMALE		HIGH			
28	60	FEMALE		HIGH			
24	86	FEMALE	NO	HIGH			
24	79	FEMALE	NO	HIGH			
24	73	FEMALE	NO	HIGH			
24	62	FEMALE	NO	HIGH			
24	58	FEMALE	NO	HIGH			
24	52	FEMALE	NO	HIGH			
18	73	MALE		HIGH			
18	39	MALE		HIGH			
10	75			HIGH	CIRCULATORY-DISEASES		
10	21			HIGH	CIRCULATORY-DISEASES		
8	69		YES	HIGH			

Σχήμα 34. Απεικόνιση αποτελεσμάτων κανόνων συσχέτισης

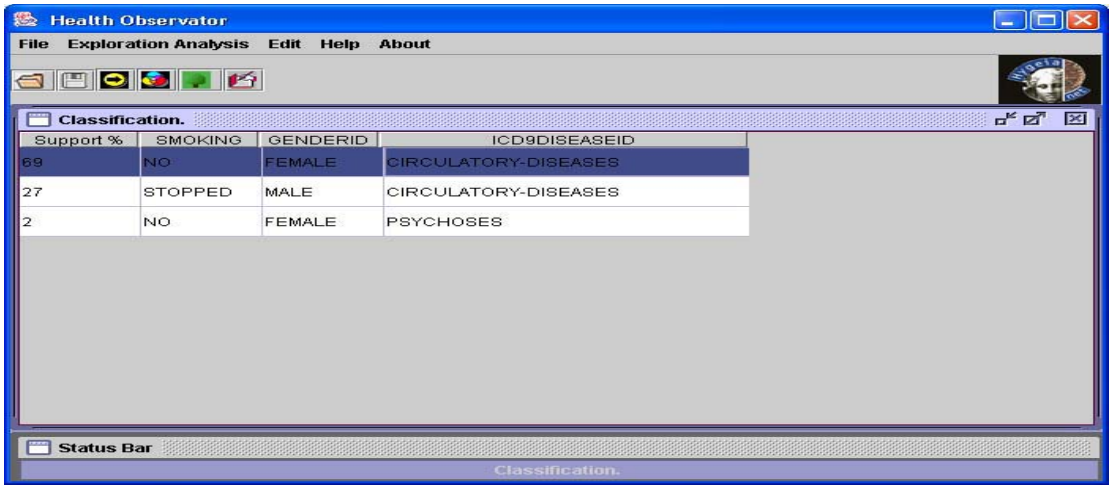
Ένα παρόμοιο τρόπο απεικόνισης έχουμε και για τα αποτελέσματα της συνάθροισης μέσω κανόνων αλληλοσυσχέτισης. Η διαφορά είναι ότι δεν υπάρχει το πεδίο confidence. Έχουμε φθίνουσα συνάθροιση των στοιχείων και κατηγοριοποίηση όπως και στους κανόνες συσχέτισης. Εκτός από την φθίνουσα συνάθροιση έχουμε και ιεραρχική συνάθροιση των στοιχείων. Δηλαδή ξεκινώντας από μια κατηγορία με Ν στοιχεία παρουσιάζουμε όλες τις Ν+1 υπερ-κατηγορίες , με ένα νέο στοιχείο και Ν τα Ν της προηγούμενης. Όταν εξαντληθούν όλες οι ιεραρχικές κατηγορίες συνάθροισης της αρχικής κατηγορίας εμφανίζεται μια γκρι γραμμή και επαναλαμβάνεται το ίδιο με

την επόμενη κατηγορία (σε φθίνουσα σειρά σύμφωνα με την «υποστήριξη»). Η γραφική απεικόνιση παρουσιάζεται στο σχήμα 35.



Σχήμα 35. Απεικόνιση αποτελεσμάτων συνάθροισης

Στην συνάθροιση με τον αλγόριθμο *k*-Means έχουμε ακριβώς τότες κατηγορίες όσες είχε δηλώσει ο χρήστης στις παραμέτρους . Και εδώ υπάρχει φθίνουσα συνάθροιση των κατηγοριών. Στο σχήμα 36 φαίνεται η γραφική απεικόνιση των αποτελεσμάτων.



Σχήμα 36. Απεικόνιση αποτελεσμάτων K-Means

Και στις τρεις απεικονίσεις των διαφορετικών αλγορίθμων μπορούμε να επιλέξουμε μία ή περισσότερες γραμμές και με την εντολή Ctrl-C να αντιγράψουμε και να τα επικολλήσουμε σε οποιαδήποτε κειμενογράφο για περαιτέρω επεξεργασία.
Σημείωση Από κάθε σημείο του συστήματος μας ο χρήστης έχει την δυνατότητα να εκτελέσει τον ίδιο ή άλλο αλγόριθμο με νέες παραμέτρους αλλά με τα στοιχεία που

είχε επιλέξει αρχικά. Για επιλογή νέων στοιχείων πρέπει να ξεκινήσει η διαδικασία από την αρχή, δηλαδή να ανοίξουμε το XML αρχείο που θέλουμε και να διαλέξουμε τα στοιχεία από το παράθυρο στοιχείων.

6.4. Αποθήκευση στοιχείων

Η αποθήκευση των αποτελεσμάτων, από την επεξεργασία των δεδομένων, μπορεί να γίνει με δύο τρόπους. Είτε σαν απλό αρχείο κείμενου με συγκεκριμένη μορφοποίηση για κάθε αλγόριθμο είτε σαν αρχείο της μορφής pmml.

Η pmml (Predictive Model Markup Language) περιγράφει μοντέλα εξόρυξης γνώσης σε XML (Extensible Markup Language) γλώσσα. Είναι ένα κοινά αποδεκτό μοντέλο αναπαράστασης γνώσης το οποίο δημιουργήθηκε από την Data Mining Group DMG [58]. Η PMML παρέχει την προδιαγραφή XML για διάφορα είδη προτύπων εξόρυξης δεδομένων. Με αυτό το μοντέλο τα αποτελέσματα της εξόρυξης γνώσης μπορούν να επαναχρησιμοποιηθούν, να εκδοθούν (ως έγγραφο XML) και να αναλυθούν από διαφορετικούς ερευνητές ή ακόμα και διαφορετικά συστήματα που υποστηρίζουν την pmml περιγραφή. Η pmml υποστηρίζει σχεδόν όλα τα μοντέλα εξόρυξης γνώσης και μηχανικής μάθησης, όπως για παράδειγμα κανόνες συσχέτισης, συνάθροιση, ακολουθίες επαναλαμβανόμενων γεγονότων.

Η αποθήκευση των δεδομένων στο σύστημα μας γίνεται αφού επιλέξουμε από τη γραμμή μενού το μενού **File** και μετά το υπό-μενού **Save output** (για αποθήκευση σαν απλό κείμενο) ή το υπό-μενού **Save as pmml** (για αποθήκευση με το μοντέλο αναπαράστασης pmml). Εμφανίζεται ένα παράθυρο διαλόγου το οποίο μας ζητάει το όνομα και τη «διαδρομή» (path) του αρχείου προς αποθήκευση.

Έχοντας εξάγει κανόνες συσχέτισης από τα δεδομένα μας μπορούμε να τους αποθηκεύσουμε σαν απλό κείμενο με συγκεκριμένη μορφή. Όπως φαίνεται και από το παρακάτω σχήμα αρχικά γράφεται η πληροφορία για τα δεδομένα εισόδου (το όνομα του αρχείου εισόδου).

Μετά έχουμε την πληροφορία για την συνάθροιση των κανόνων στις τέσσερις κατηγορίες και τέλος τους κανόνες συσχέτισης με συγκεκριμένη μορφοποίηση. Κάθε κανόνας συσχέτισης εμφανίζεται στο αρχείο με την εξής μορφή :

<Αριθμός Στήριξης> <Αριθμός εμπιστοσύνης> <στοιχεία στο κεφάλι του κανόνα> ➔ <στοιχεία στο σώμα του κανόνα>

Rules generated from D:\koum\data\archanes.xml

Classification Levels.

6 12 14 0

Sup.	Conf.				
33	85	GENDERID_FEMALE	==>	SMOKING_NO	
33	79	SMOKING_NO	==>	GENDERID_FEMALE	
31	73	SMOKING_NO	==>	SYSTOLICPRESSURE_HIGH	
31	65	SYSTOLICPRESSURE_HIGH	==>	SMOKING_NO	
28	72	GENDERID_FEMALE	==>	SYSTOLICPRESSURE_HIGH	
28	60	SYSTOLICPRESSURE_HIGH	==>	GENDERID_FEMALE	
24	79	SYSTOLICPRESSURE_HIGH	SMOKING_NO	==>	GENDERID_FEMALE
24	73	SMOKING_NO	GENDERID_FEMALE	==>	SYSTOLICPRESSURE_HIGH
24	62	GENDERID_FEMALE	==>	SYSTOLICPRESSURE_HIGH	SMOKING_NO
24	58	SMOKING_NO	==>	SYSTOLICPRESSURE_HIGH	GENDERID_FEMALE
24	52	SYSTOLICPRESSURE_HIGH	==>	SMOKING_NO	GENDERID_FEMALE
18	73	GENDERID_MALE	==>	SYSTOLICPRESSURE_HIGH	
18	39	SYSTOLICPRESSURE_HIGH	==>	GENDERID_MALE	
10	75	ICD9DISEASEID_CIRCULATORY-DISEASES	==>	SYSTOLICPRESSURE_HIGH	
10	21	SYSTOLICPRESSURE_HIGH	==>	ICD9DISEASEID_CIRCULATORY-DISEASES	

8	69	SMOKING_YES	==>	SYSTOLICPRESSURE_HIGH	
8	68	SMOKING_YES	==>	GENDERID_MALE	
8	33	GENDERID_MALE	==>	SMOKING_NO	
8	32	GENDERID_MALE	==>	SMOKING_YES	
8	20	SMOKING_NO	==>	GENDERID_MALE	
8	17	SYSTOLICPRESSURE_HIGH	==>	SMOKING_YES	

Σχήμα 37. Αποθήκευση αποτελεσμάτων κανόνων συσχέτισης σε μορφή απλού κειμένου

Ένα παράδειγμα της αποθήκευσης κανόνων συσχέτισης με το μοντέλο αναπαράστασης pmml φαίνεται στο σχήμα 38.

```
<?xml version="1.0" ?>
<PMML version="2.0" >
      <Header copyright="Koum"
       description="Rules generated from D:\koum\data\archanes.xml"/>
      <DataDictionary numberOfFields="2" >
      <DataField name="transaction" optype="categorical" />
      <DataField name="item" optype="categorical" />
</DataDictionary>
<AssociationModel
      functionName="associationRules"
      numberOfTransactions="28739" numberOfItems="28"
      minimumSupport="0.03" minimumConfidence="0.11"
      numberOfItemsets="314" numberOfRules="157"><MiningSchema>
      <MiningField name="transaction"/>
      <MiningField name="item"/>
```

Σχήμα 38. Αποθήκευση αποτελεσμάτων κανόνων συσχέτισης σε μορφή pmml

Με παρόμοιο τρόπο αποθηκεύονται και τα αποτελέσματα συνάθροισης. Στη συνάθροιση μέσω κανόνων συσχέτισης αρχικά γράφεται η πληροφορία για τα δεδομένα εισόδου. Μετά έχουμε την πληροφορία για την συνάθροιση των κανόνων στις τέσσερις κατηγορίες και τέλος την συνάθροιση με συγκεκριμένη μορφοποίηση. Κάθε αποθήκευση ενός αποτελέσματος συνάθροισης είναι της μορφής: <Αριθμός Στήριξης> <στοιχείο 1> <στοιχείο 2> ….. <στοιχείο N>.

Παρόμοια μορφοποίηση έχουμε και στην συνάθροιση με τον αλγόριθμο K-Means. Αρχικά γράφεται η πληροφορία για τα δεδομένα εισόδου. Μετά έχουμε την συνάθροιση με συγκεκριμένη μορφοποίηση. Κάθε αποθήκευση μιας συνάθροισης είναι της μορφής: <Αριθμός Στήριξης> <στοιχείο 1> <στοιχείο 2> ….. <στοιχείο N>

```
Clustering generated from D:\koum\data\archanes.xml

Classification Levels.
0.06437245554820975 0.0 0.0 0.0
```

Sup.				
24	SYSTOLICPRESSURE_HIGH	SMOKING_NO	GENDERID_FEMALE	
5	SYSTOLICPRESSURE_HIGH	SMOKING_NO	GENDERID_FEMALE	ICD9DISEASEID_CIRCULATORY-DISEASES
4	BLOODSUGAR_HIGH	SYSTOLICPRESSURE_HIGH	SMOKING_NO	GENDERID_FEMALE
6	SMOKING_NO	GENDERID_FEMALE	ICD9DISEASEID_CIRCULATORY-DISEASES	
5	SYSTOLICPRESSURE_HIGH	SMOKING_NO	GENDERID_FEMALE	ICD9DISEASEID_CIRCULATORY-DISEASES
6	SYSTOLICPRESSURE_HIGH	SMOKING_NO	GENDERID_MALE	
6	SYSTOLICPRESSURE_HIGH	SMOKING_NO	ICD9DISEASEID_CIRCULATORY-DISEASES	
5	SYSTOLICPRESSURE_HIGH	SMOKING_NO	GENDERID_FEMALE	ICD9DISEASEID_CIRCULATORY-DISEASES
...	...	...	...	...

Σχήμα 39. Αποθήκευση αποτελεσμάτων συνάθροισης σε μορφή απλού κειμένου

7. ΠΑΡΑΔΕΙΓΜΑΤΑ: ΤΟ HEALTHOBS ΣΤΗ ΠΡΑΞΗ

Στο κεφάλαιο αυτό περιγράφουμε κάποια παραδείγματα με αποτελέσματα από κάποιες εκτελέσεις των αλγορίθμων σε πραγματικά ιατρικά δεδομένα. Παρουσιάζουμε τρία συνολικά παραδείγματα. Το πρώτο σκοπό έχει να δείξει αναλυτικά τον τρόπο εκτέλεσης των αλγορίθμων, γι' αυτό και παρουσιάζεται σε μεγάλη λεπτομέρεια. Στα δυο επόμενα παραδείγματα και έχοντας κατανοήσει πλήρως τον τρόπο εκτέλεσης, παρουσιάζουμε τα αποτελέσματα από τους τρεις αλγόριθμους εξόρυξης δεδομένων, τα οποία εκμαιεύουμε από παραδείγματα

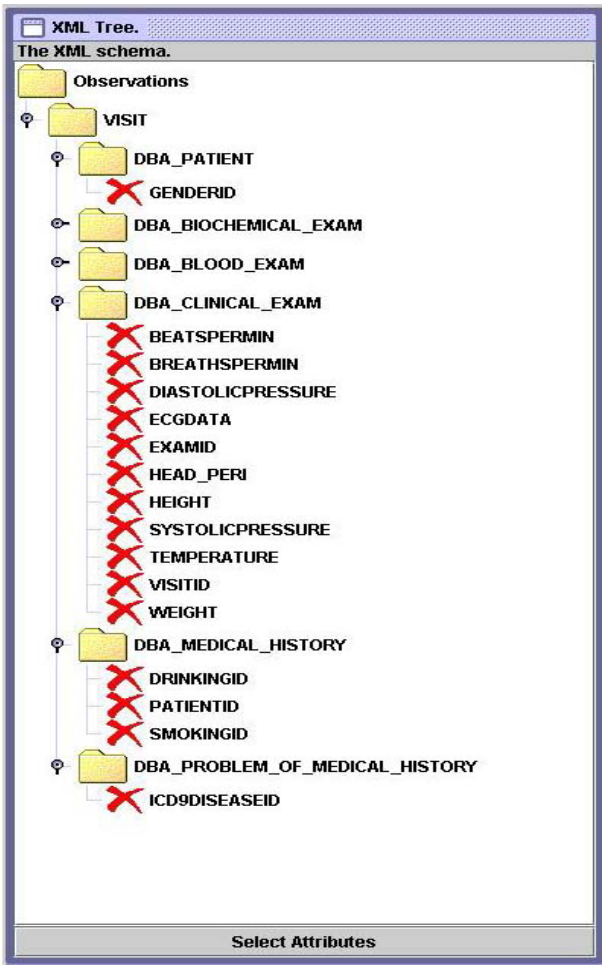
7.1. Παράδειγμα Πρώτο

Στο παράδειγμα αυτό προκειμένου να γίνουν κατανοητές τόσο η ευστάθεια όσο και η εκτέλεση των αλγορίθμων, παρουσιάζουμε αναλυτικά τα δεδομένα μας, αλλά και κάποια βήματα που περιγράφηκαν στα προηγούμενα κεφάλαια. Το πρώτο παράδειγμα αντιπροσωπεύει 560 επισκέψεις σε δύο ιατρικά κέντρα της Κρήτης. Σύμφωνα με αυτή την προσέγγιση κάθε επίσκεψη θεωρείται μια ξεχωριστή «περίπτωση» των δεδομένων μας. Οπότε ο ίδιος ασθενής μπορεί να έχει πολλές καταχωρήσεις στα δεδομένα. Αν θέλαμε να καταχωρούμαι μόνο μια φορά κάθε ασθενή, μπορούσαμε μέσω του COAS server να δημιουργούμε τις περιπτώσεις με «κλειδί» τα πεδία VisitId και PatientId. Σκοπός του παραδείγματος αυτού είναι να κατανοήσουμε την ροή του προγράμματος και να αναδείξουμε την ομογενοποίηση ετερογενών πηγών δεδομένων. Το XML αρχείο που έχουν αποθηκευτεί τα δεδομένα μας (με την βοήθεια του COAS server από το ιατρικό σύστημα PHCCS) είναι 1 MB.

```
<!ELEMENT AIMOPETALIA ( #PCDATA ) >
<!ELEMENT ALBOUMINI ( #PCDATA ) >
<!ELEMENT ALEMFO ( #PCDATA ) >
<!ELEMENT ANISOKYTAROSI ( #PCDATA ) >
<!ELEMENT ANOXIGLIKOZIS120 ( #PCDATA ) >
<!ELEMENT ASBESTIO ( #PCDATA ) >
<!ELEMENT ATYPA ( #PCDATA ) >
<!ELEMENT A_AMILASI ( #PCDATA ) >
<!ELEMENT B12 ( #PCDATA ) >
<!ELEMENT BASEOF ( #PCDATA ) >
<!ELEMENT BASEOFILY_STIKSI ( #PCDATA ) >
<!ELEMENT BEATSPERMIN ( #PCDATA ) >
```

Σχήμα 40. Η δομή του DTD στο HealthObs

Αρχικά το σύστημα διαβάζει το, αντίστοιχο του XML, DTD αρχείο και εμφανίζει το αντίστοιχο παράθυρο των στοιχείων. Στο Σχήμα 40, βλέπουμε ένα μεγάλο μέρος του DTD αρχείου και στο Σχήμα 41, βλέπουμε την γραφική απεικόνιση του ίδιου DTD σχήματος από το σύστημα μας.



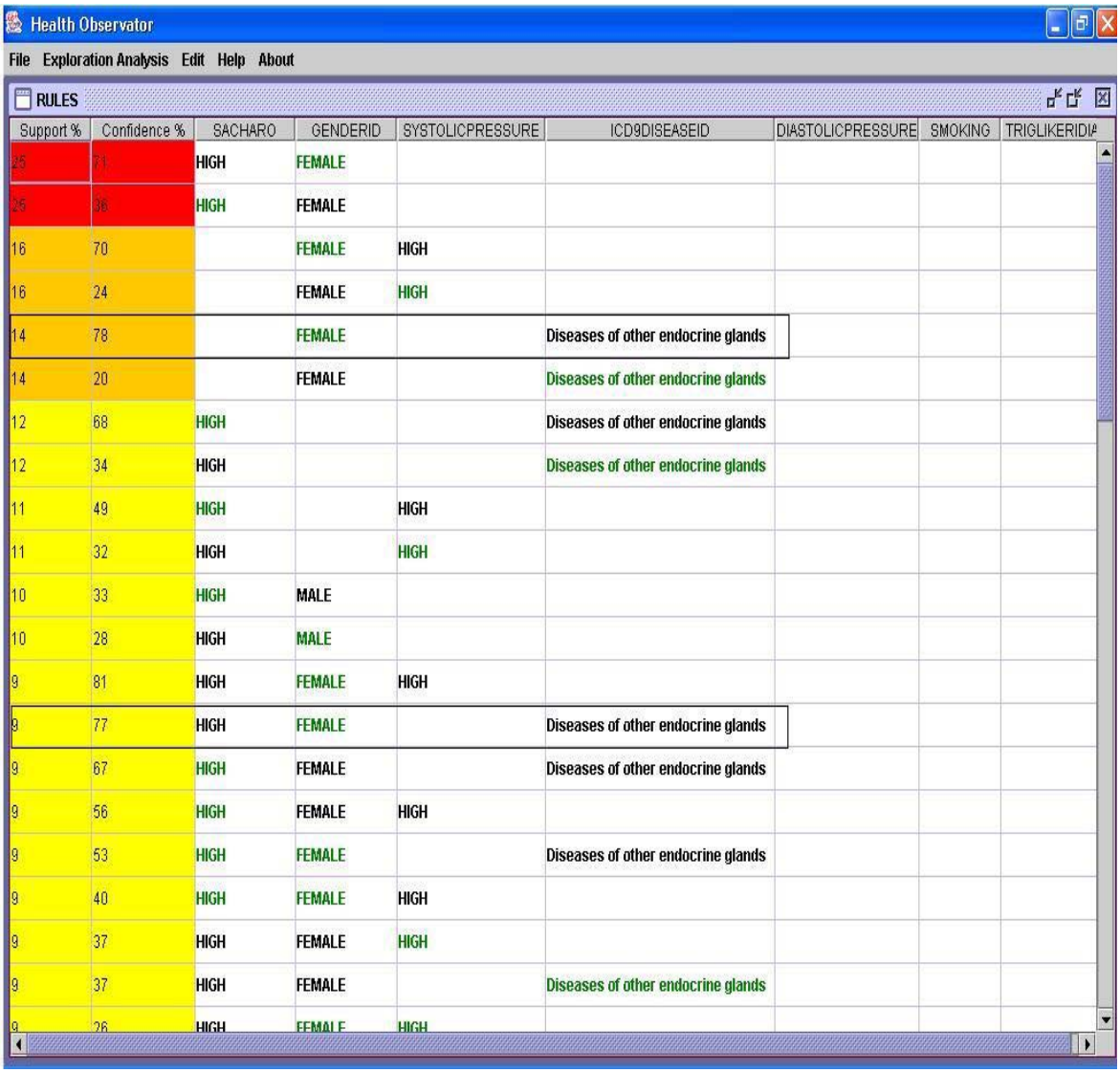
Σχήμα 41. Γραφική απεικόνιση του DTD σχήματος

Επόμενο βήμα είναι η επιλογή των στοιχείων που επιθυμούμε να συμμετέχουν στην εξόρυξη γνώσης. Στο συγκεκριμένο παράδειγμα επιλέξαμε 7 στοιχεία. Από το πεδίο `DBA_PATIENT` επιλέξαμε το στοιχείο `GENDERID`, από το πεδίο `DBA_CLINICAL_EXAM` επιλέξαμε τα στοιχεία `SYSTOLICPRESSURE` και `DIASTOLICPRESSURE` , από το πεδίο `DBA_BIOCHEMICAL_EXAM` τα στοιχεία `SACHARO` και `TRIGLIKERIDIA`, από το πεδίο `DBA_MEDICAL_HISTORY` επιλέξαμε το στοιχείο `SMOKING` και τέλος από το πεδίο `DBA_PROBLEM_OF_MEDICAL_HISTORY` επιλέξαμε `ICD9DISEASEID`. Στο σημείο αυτό πρέπει να επισημάνουμε ότι η πολυπλοκότητα του εκάστοτε αλγορίθμου αυξάνεται με το μέγεθος των δεδομένων (πόσες καταχωρήσεις έχουμε και συνεπώς

πόσο μεγάλο είναι το XML αρχείο) , και τον αριθμό των στοιχείων που συμμετέχουν.

Αν κάποιο στοιχείο έχει υποστεί τμηματοποίηση (με την υπηρεσία Common clinical term reference) σε πολλές επιμέρους κατηγορίες τότε η πολυπλοκότητα του αλγορίθμου αυξάνει ανάλογα με τον αριθμό των κατηγοριών. Για παράδειγμα το στοιχείο `GENDERID` αποτελείται από τις κατηγορίες `MALE, FEMALE`. Το στοιχείο `ICD9DISEASEID` αποτελείται από τις 17 γενικές κατηγορίες του ICD9. Θα μπορούσαμε να έχουμε πιο ειδικές κατηγορίες στο ICD9 οι οποίες είναι 110. Σε αυτή την περίπτωση η πολυπλοκότητα θα αυξανόταν δραματικά.

Έχοντας επιλέξει τα 7 αυτά στοιχεία πρέπει να επιλέξουμε και τον αλγόριθμο εξόρυξης γνώσης. Ο πρώτος αλγόριθμος που θα δούμε είναι οι κανόνες συσχέτισης. Οι παράμετροι που δώσαμε για το συγκεκριμένο παράδειγμα είναι : «στήριξη» (support) 5% και «εμπιστοσύνη» (confidence) 10%. Η γραφική απεικόνιση των αποτελεσμάτων φαίνεται στο σχήμα 42.



Support %	Confidence %	SACHARO	GENDERID	SYSTOLICPRESSURE	ICD9DISEASEID	DIASTOLICPRESSURE	SMOKING	TRIGLIKERIDIA
25	71	HIGH	FEMALE					
25	36	HIGH	FEMALE					
16	70		FEMALE	HIGH				
16	24		FEMALE	HIGH				
14	78		FEMALE		Diseases of other endocrine glands			
14	20		FEMALE		Diseases of other endocrine glands			
12	68	HIGH			Diseases of other endocrine glands			
12	34	HIGH			Diseases of other endocrine glands			
11	49	HIGH		HIGH				
11	32	HIGH		HIGH				
10	33	HIGH	MALE					
10	28	HIGH	MALE					
9	81	HIGH	FEMALE	HIGH				
9	77	HIGH	FEMALE		Diseases of other endocrine glands			
9	67	HIGH	FEMALE		Diseases of other endocrine glands			
9	56	HIGH	FEMALE	HIGH				
9	53	HIGH	FEMALE		Diseases of other endocrine glands			
9	40	HIGH	FEMALE	HIGH				
9	37	HIGH	FEMALE	HIGH				
9	37	HIGH	FEMALE		Diseases of other endocrine glands			
9	26	HIGH	FEMALE	HIGH				

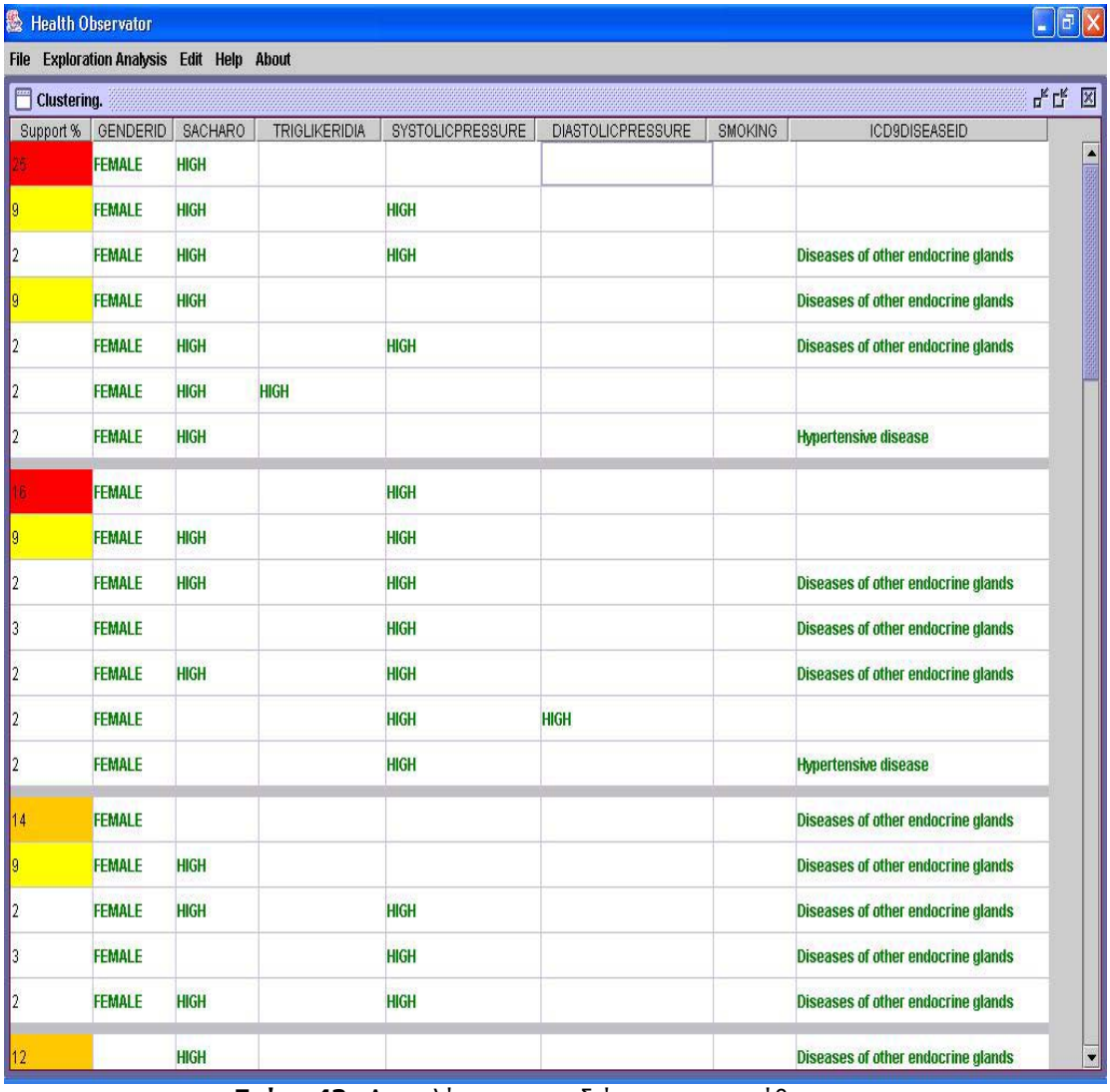
Σχήμα 42. Αποτελέσματα παραδείγματος από κανόνες συσχέτισης

Έχοντας την γραφική απεικόνιση των αποτελεσμάτων και την γνώση ειδικών , στην περίπτωση μας ανθρώπων από τον τομέα της ιατρικής, μπορούν να εξαχθούν ενδιαφέροντα συμπεράσματα. Η αξιολόγηση των αποτελεσμάτων είναι καθαρά θέμα ειδικών στον εκάστοτε τομέα. Ορισμένοι κανόνες μπορεί να είναι προφανείς για τον άνθρωπο, όπως για παράδειγμα «Αν ένας άνθρωπος είναι έγκυος τότε είναι γυναίκα». Υπάρχουν και κανόνες που δεν είναι προφανείς και μπορεί να θεωρηθούν πολύ ενδιαφέροντες από τους ειδικούς.

Μπορεί για παράδειγμα κάποιος να θεωρήσει αρκετά ενδιαφέρον τους κανόνες που εμφανίζονται στο παραπάνω σχήμα με περίγραμμα. Ο πρώτος κανόνας λέει ότι *«αν το γένος είναι θηλυκό τότε παρουσιάζει πρόβλημα ενδοκρινικών αδένων»* με στήριξη 14% και εμπιστοσύνη 78. Ο δεύτερος κανόνας λέει ότι *«αν το γένος είναι θηλυκό τότε παρουσιάζει πρόβλημα ενδοκρινικών αδένων και υψηλά ποσοστά ζαχάρου»* με στήριξη 9% και εμπιστοσύνη 77. Αν ένας ειδικός θεωρούσε τους κανόνες αυτούς ενδιαφέροντες και ήθελε να εστιάσει περισσότερο το πρόβλημα της ασθένειας θα μπορούσε να επαναλάβει το πείραμα προσθέτοντας και τις υποκατηγορίες του προβλήματος των ενδοκρινικών αδένων με την υπηρεσία Common clinical term reference.

Ο χρόνος που χρειάστηκε για να εξαχθούν αυτοί οι κανόνες συσχέτισης με τα συγκεκριμένα 7 στοιχεία και ένα υπολογιστή ταχύτητας 550 MHz με μνήμη 384MB ήταν λιγότερο από 2 λεπτά. Ο χρόνος αυτός μπορεί να κριθεί ως πάρα πολύ καλός για αλγόριθμο κανόνων συσχέτισης με 560 παραδείγματα και πάνω από 20 στοιχεία (θεωρώντας κάθε κατηγοριοποίηση του αρχικού στοιχείου σαν ένα στοιχείο).

Ο επόμενος αλγόριθμος που θα εξετάσουμε είναι η συνάθροιση μέσο κανόνων συσχέτισης. Στην συνάθροιση αυτή ο αλγόριθμος είναι μια απλοποιημένη μορφή του αλγόριθμου των κανόνων συσχέτισης οπότε και έχουμε μικρότερο χρόνο εκτέλεσης του. Τα αποτελέσματα του αλγόριθμου αυτού για τα ίδια στοιχεία με αυτά του προηγούμενου πειράματος και με παράμετρο στήριξης 2% φαίνονται στο σχήμα 43. Όπως είπαμε και στο προηγούμενο κεφάλαιο έχουμε φθίνουσα συνάθροιση των συστάδων με κατηγοριοποίηση (χρώματα στο πεδίο support) και ιεραρχική συνάθροιση των συστάδων. Και σε αυτό το παράδειγμα βλέπουμε ότι η συστάδα «γένος θηλυκό και πρόβλημα ενδοκρινικών αδένων» με στήριξη 14% (βρίσκεται πρώτη στην τρίτη ιεραρχική ομάδα συστάδων) υπάρχει στα αποτελέσματα μας.



Support %	GENDERID	SACHARO	TRIGLIKERIDIA	SYSTOLICPRESSURE	DIASTOLICPRESSURE	SMOKING	ICD9DISEASEID
25	FEMALE	HIGH					
9	FEMALE	HIGH		HIGH			
2	FEMALE	HIGH		HIGH			Diseases of other endocrine glands
9	FEMALE	HIGH					Diseases of other endocrine glands
2	FEMALE	HIGH		HIGH			Diseases of other endocrine glands
2	FEMALE	HIGH	HIGH				
2	FEMALE	HIGH					Hypertensive disease
16	FEMALE			HIGH			
9	FEMALE	HIGH		HIGH			
2	FEMALE	HIGH		HIGH			Diseases of other endocrine glands
3	FEMALE			HIGH			Diseases of other endocrine glands
2	FEMALE	HIGH		HIGH			Diseases of other endocrine glands
2	FEMALE			HIGH	HIGH		
2	FEMALE			HIGH			Hypertensive disease
14	FEMALE						Diseases of other endocrine glands
9	FEMALE	HIGH					Diseases of other endocrine glands
2	FEMALE	HIGH		HIGH			Diseases of other endocrine glands
3	FEMALE			HIGH			Diseases of other endocrine glands
2	FEMALE	HIGH		HIGH			Diseases of other endocrine glands
12		HIGH					Diseases of other endocrine glands

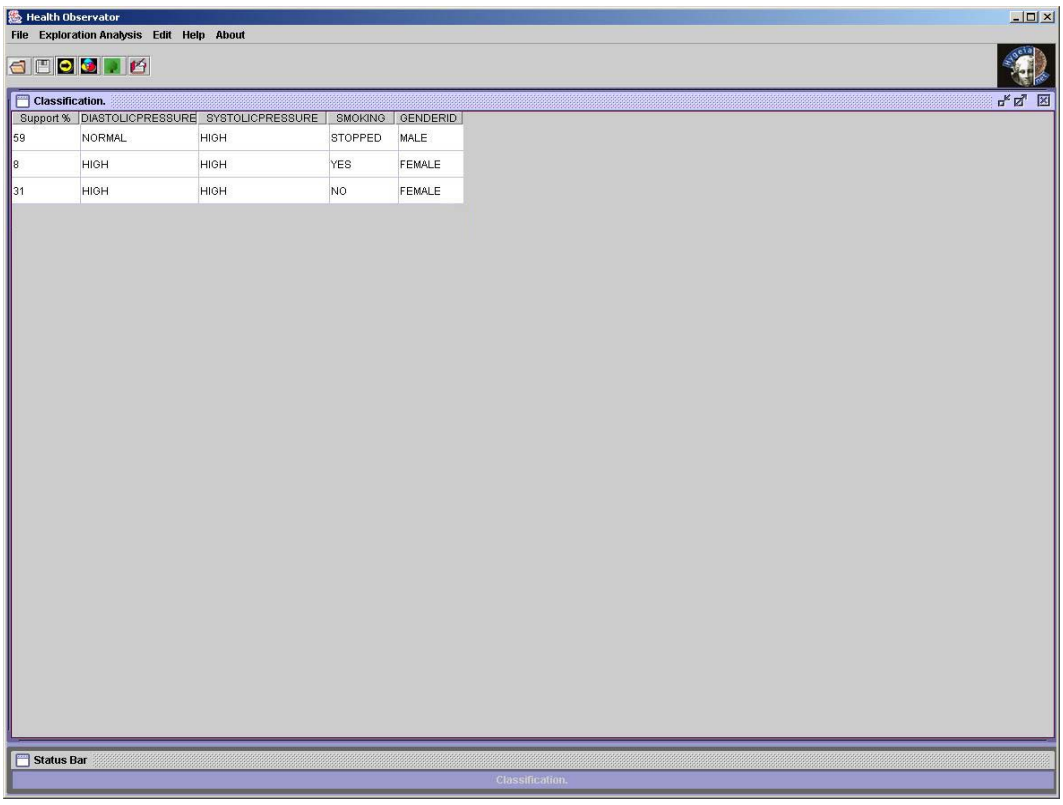
Σχήμα 43. Αποτελέσματα παραδείγματος με συνάθροιση

Ο τρίτος αλγόριθμος συνάθροισης είναι η συνάθροιση με την μέθοδο K-Means. Η βασική διαφορά των δύο αλγορίθμων συνάθροισης είναι ότι ο K-Means δημιουργεί συστάδες στοιχείων με σκοπό τον απόλυτο διαχωρισμό των τιμών κάθε στοιχείου σε μία συστάδα ενώ η συνάθροιση μέσο κανόνων συσχέτισης μπορεί να δημιουργεί και διαφορετικές συστάδες με κείνο πεδίο τιμών ενός η και παραπάνω στοιχείων. Μπορούμε να πούμε ότι ο ένας αλγόριθμος κάνει διαχωριστική συνάθροιση ενώ ο άλλος κάνει επικαλυπτόμενη συνάθροιση.

Η μέθοδος K-Means με τα συγκεκριμένα δεδομένα μπορεί να εξάγει ένα μικρό αριθμό συστάδων. Αυτό οφείλεται στην τμηματοποίηση που έχουν υποστεί τα δεδομένα μας. Λόγο της φύσης του αλγορίθμου, στοιχεία τα οποία χωρίζονται σε ένα μικρό αριθμό κατηγοριών δημιουργούν δυσκολία στον αλγόριθμο. Αυτό οφείλεται στη μετρική απόστασης που χρησιμοποιεί ο αλγόριθμος. Όταν τα στοιχεία έχουν εύρος τιμών 2-3 (για παράδειγμα το στοιχείο GENDERID μπορεί να πάρει τις τιμές MALE ή FEMALE) δεν μπορούμε να έχουμε και σύγκληση στα κέντρα συνάθροισης του

αλγορίθμου. Οπότε τα κέντρα περιορίζονται σε ένα μικρό αριθμό διακριτών σημείων. Λόγο της φύσης των ιατρικών δεδομένων οι απόλυτες τιμές κάποιων στοιχείων δεν έχουν ιδιαίτερο νόημα. Η ανάγκη γνώσης για ένα στοιχείο, για παράδειγμα για την Χοληστερίνη, εστιάζεται στο αν είναι «Χαμηλή» «Φυσιολογική» ή «Υψηλή» και όχι στην τιμή που μετρήθηκε από κάποιο εργαστήριο το οποίο μπορεί να έχει διαφορετικές φυσιολογικές τιμές από οποιοδήποτε άλλο εργαστήριο. Οπότε στην προκειμένη περίπτωση ο αλγόριθμος με την φύση των δεδομένων μας έρχονται σε ρήξη. Αποτέλεσμα αυτού είναι ο αλγόριθμος να εξαρτάται από την αρχική επιλογή των κέντρων και κάποιες φορές να τερματίζει όχι λόγο σύγκλησης αλλά λόγο επαναληπτικής μεταφοράς κάποιων κέντρων σε δύο συγκεκριμένες θέσεις.

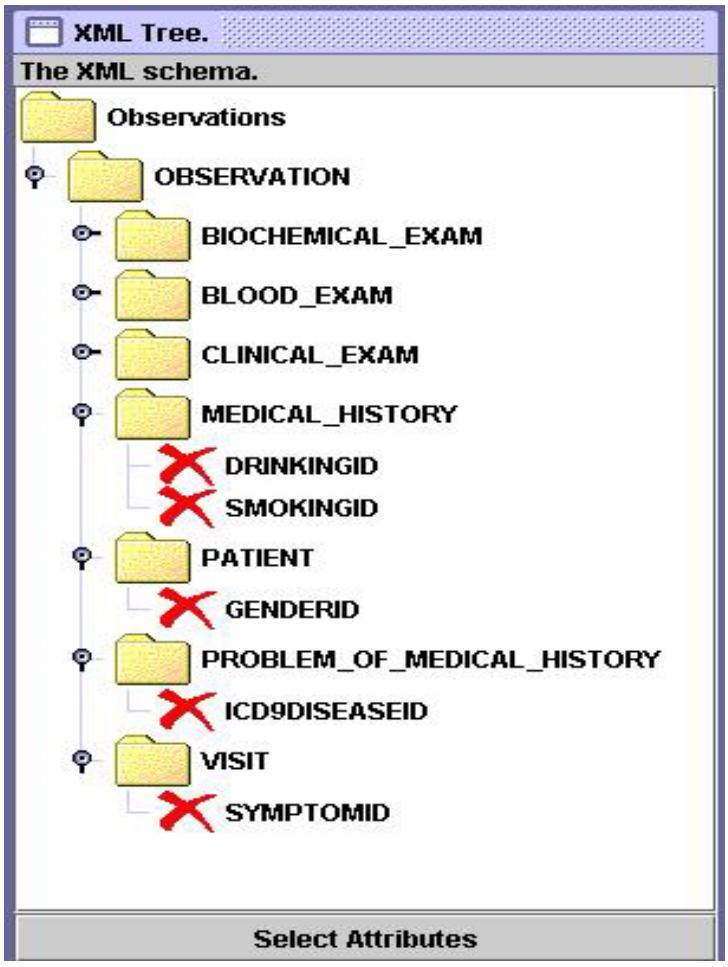
Μία ακόμη ιδιαιτερότητα του αλγορίθμου ,αντιθέτως με τους δύο προηγούμενους, είναι ότι πρέπει όλες οι καταχωρήσεις να έχουν τιμές για τα στοιχεία που έχουμε επιλέξει προς επεξεργασία. Αποτέλεσμα αυτού είναι στο συγκεκριμένο παράδειγμα να έχουμε μόνο 260 καταχωρίσεις που έχουν τιμές και στα 7 στοιχεία που επιλέξαμε. Στο συγκεκριμένο παράδειγμα ο αλγόριθμος χρησιμοποιεί 260 από τις 560 καταχωρίσεις. Τα αποτελέσματα του εμφανίζονται στο σχήμα 44. Το πρώτο πεδίο (support) μας δίνει το ποσοστό των καταχωρίσεων (από τις 260) που ταξινομήθηκαν στην συγκεκριμένη κατηγορία. Δεξιά από το ποσοστό αυτό εμφανίζονται οι τιμές για τα κέντρα που επέλεξε ο αλγόριθμος.



Σχήμα 44. Αποτελέσματα παραδείγματος με K-Means

7.2. Παράδειγμα Δεύτερο

Το παράδειγμα αυτό αντιπροσωπεύει 28.720 επισκέψεις σε ένα ιατρικό κέντρο της Κρήτης. Σκοπός του παραδείγματος αυτού είναι να δοκιμάσουμε το σύστημα σε δεδομένα με μεγάλο όγκο, να εξάγουμε ακόμα ποιο αξιόπιστα αποτελέσματα και να δούμε την συμπεριφορά του συστήματος. Το XML αρχείο που έχουν αποθηκευτεί τα δεδομένα μας (με την βοήθεια του COAS server από το ιατρικό σύστημα PHCCS) είναι 25,8 MB. Αρχικά το σύστημα διαβάζει το, αντίστοιχο του XML, DTD αρχείο και εμφανίζει το αντίστοιχο παράθυρο των στοιχείων. Στο σχήμα 45 βλέπουμε την γραφική απεικόνιση του DTD σχήματος από το σύστημα μας.

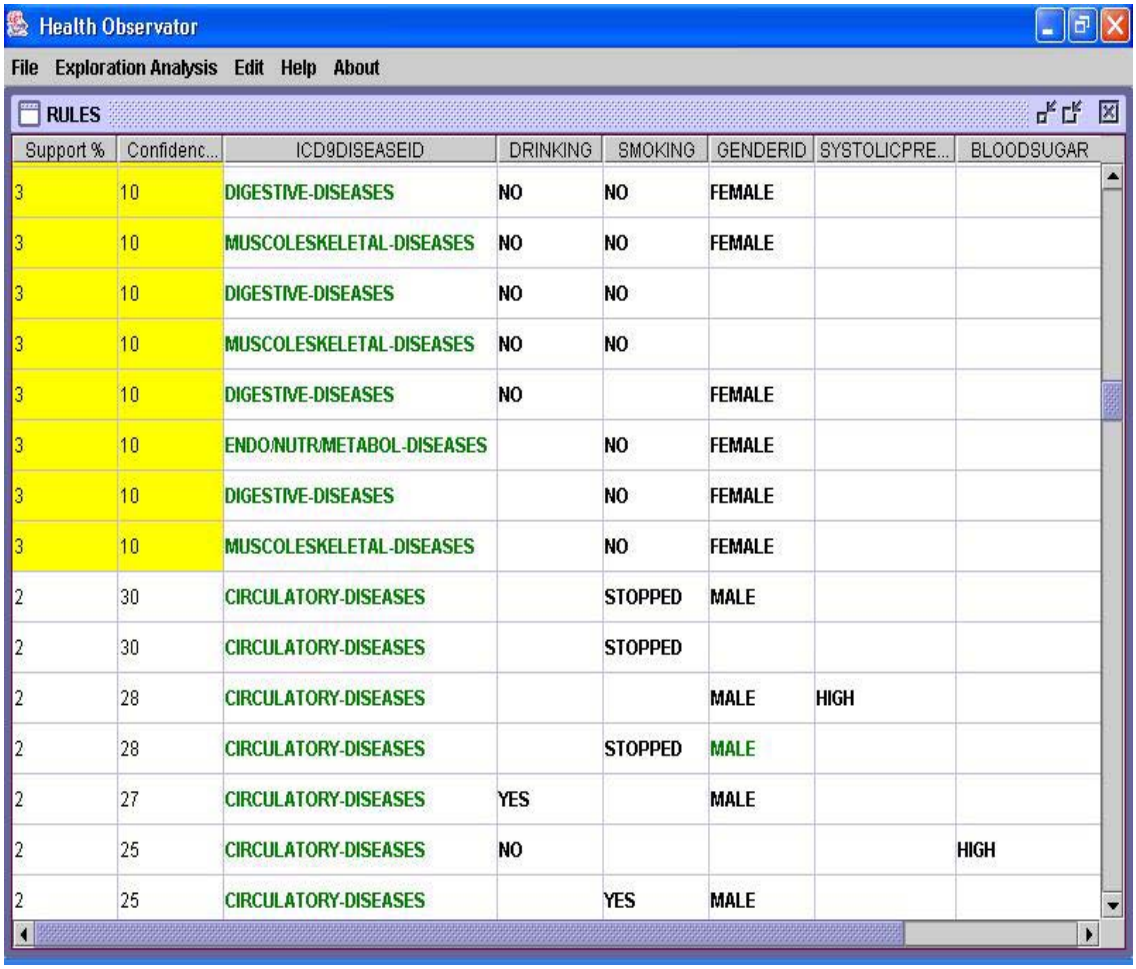


Σχήμα 45. Γραφική αναπαράσταση του DTD

Στο συγκεκριμένο παράδειγμα επιλέξαμε 7 στοιχεία. Από το πεδίο `PATIENT` επιλέξαμε το στοιχείο `GENDERID`, από το πεδίο `CLINICAL_EXAM` επιλέξαμε το στοιχείο `DIASTOLICPRESSURE`, από το πεδίο `BIOCHEMICAL_EXAM` τα στοιχεία `BLOOD_SUGAR`, `DHLCHOLISTEROL`, `CHOLISTEROL`, `UREA`, `URIC_ACID`, `TRIGLIKERIDIA`, από το πεδίο `MEDICAL_HISTORY` επιλέξαμε το στοιχείο

SMOKING από το πεδίο PROBLEM_OF_MEDICAL_HISTORY επιλέξαμε ICD9DISEASEID και τέλος από το πεδίο VISIT το στοιχείο SYMPTOMID. Το στοιχείο IC9DISEASEID το επιλέξαμε με τέτοιο τρόπο (όπως είπαμε και στο προηγούμενο κεφάλαιο) ώστε αν ο αλγόριθμος που θα επιλέξουμε αργότερα είναι κανόνες συσχέτισης αυτό θα υπάρχει υποχρεωτικά σε όλους τους κανόνες, στο «σώμα» κάθε κανόνα. Για το διάβασμα του αρχείου και την αποθήκευση των δεδομένων στις κατάλληλες δομές χρειάσθηκαν λιγότερο από 5 λεπτά.

Επόμενο βήμα είναι η επιλογή του αλγορίθμου. Για τους κανόνες συσχέτισης οι παράμετροι που θέσαμε ήταν 2% «στήριξη» (support) και 10% «εμπιστοσύνη» (confidence). Η επεξεργασία των δεδομένων για τα 13 αυτά στοιχεία (τα οποία έχουν τμηματοποιηθεί σε 40 ποιοτικές τιμές) χρειάσθηκε λιγότερο από 10 λεπτά. Ο χρόνος αυτός μπορεί να κριθεί σαν πολύ καλός για αλγόριθμο κανόνων συσχέτισης με 28.720 καταχωρίσεις και 40 στοιχεία (οι ποιοτικές κατηγοριοποιήσεις).



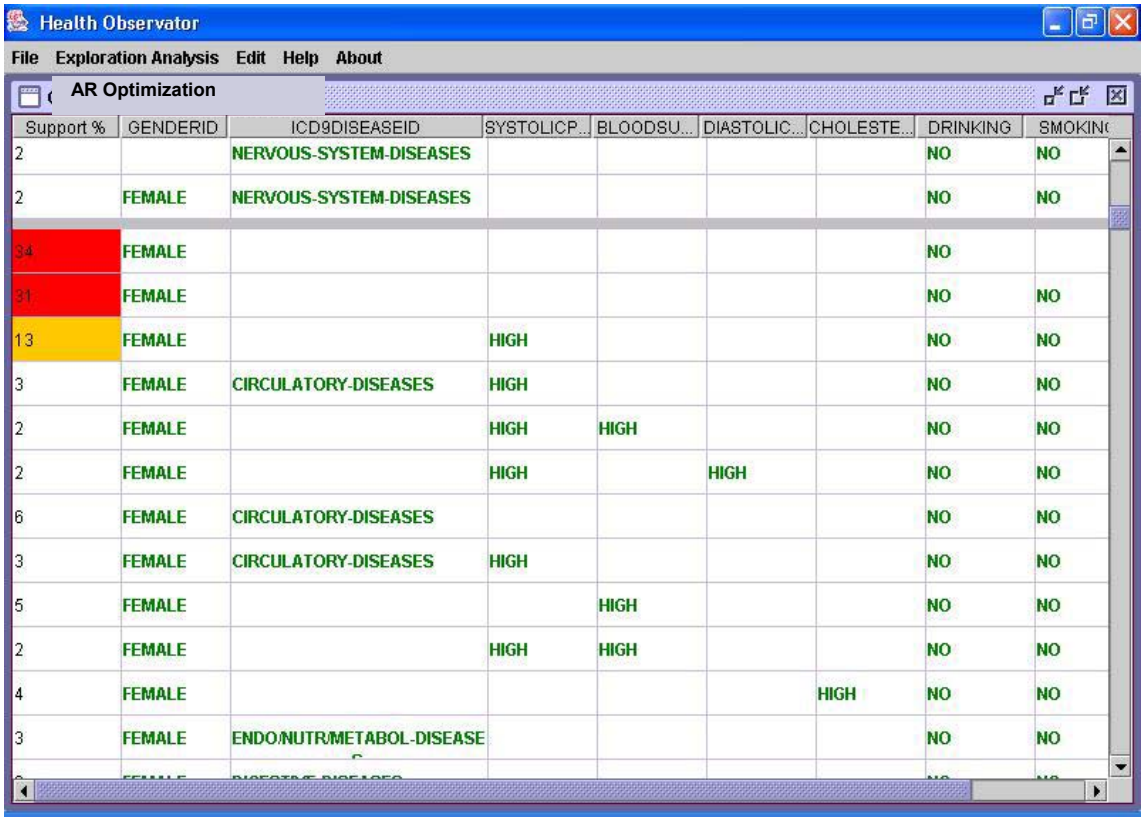
Support %	Confidenc...	ICD9DISEASEID	DRINKING	SMOKING	GENDERID	SYSTOLICPRE...	BLOODSUGAR
3	10	DIGESTIVE-DISEASES	NO	NO	FEMALE		
3	10	MUSCOLESKELETAL-DISEASES	NO	NO	FEMALE		
3	10	DIGESTIVE-DISEASES	NO	NO			
3	10	MUSCOLESKELETAL-DISEASES	NO	NO			
3	10	DIGESTIVE-DISEASES	NO		FEMALE		
3	10	ENDO/NUTR/METABOL-DISEASES		NO	FEMALE		
3	10	DIGESTIVE-DISEASES		NO	FEMALE		
3	10	MUSCOLESKELETAL-DISEASES		NO	FEMALE		
2	30	CIRCULATORY-DISEASES		STOPPED	MALE		
2	30	CIRCULATORY-DISEASES		STOPPED			
2	28	CIRCULATORY-DISEASES			MALE	HIGH	
2	28	CIRCULATORY-DISEASES		STOPPED	MALE		
2	27	CIRCULATORY-DISEASES	YES		MALE		
2	25	CIRCULATORY-DISEASES	NO				HIGH
2	25	CIRCULATORY-DISEASES		YES	MALE		

Σχήμα 46. Αποτελέσματα παραδείγματος με κανόνες συσχέτισης

Τα αποτελέσματα των κανόνων συσχέτισης με τις παραμέτρους που προαναφέραμε μπορούν να φανούν από το σχήμα 46. Όπως βλέπουμε το πεδίο ICD9DISEASEID

έχει σε όλους τους κανόνες ποιοτική τιμή στο «σώμα» κάθε κανόνα (πράσινος χρωματισμός γραμματοσειράς του πεδίου). Τα αποτελέσματα, και σε αυτή την περίπτωση ,πρέπει να ελεγχθούν από ειδικούς και να αξιοποιηθούν κατάλληλα.

Ο επόμενος αλγόριθμος που θα εξετάσουμε είναι η συνάθροιση μέσο κανόνων συσχέτισης. Τα αποτελέσματα του αλγόριθμου αυτού για τα ίδια στοιχεία με αυτά του προηγούμενου πειράματος. Οι παράμετροι που θέσαμε για το συγκεκριμένο αλγόριθμο είναι στήριξη 2%. Ο χρόνος που χρειάσθηκε ο αλγόριθμος ,κατά την αρχική καταχώρηση των δεδομένων στις κατάλληλες δομές, ήταν λιγότερη από 5 λεπτά. Και σε αυτό το παράδειγμα βλέπουμε ότι αρκετές συστάδες τις οποίες εντοπίσαμε και σαν κανόνες συσχέτισης υπάρχουν στα αποτελέσματα μας. Η γραφική απεικόνιση των αποτελεσμάτων φαίνεται στο σχήμα 47.



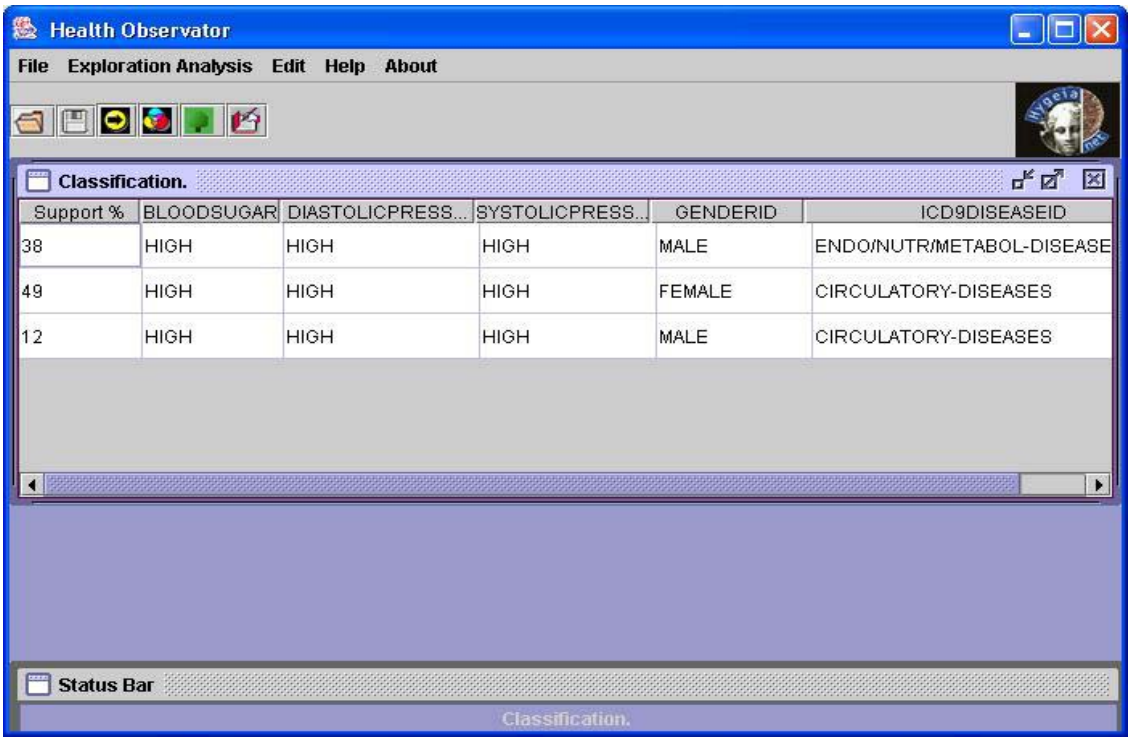
Support %	GENDERID	ICD9DISEASEID	SYSTOLICP...	BLOODSU...	DIASTOLIC...	CHOLESTE...	DRINKING	SMOKING
2		NERVOUS-SYSTEM-DISEASES					NO	NO
2	FEMALE	NERVOUS-SYSTEM-DISEASES					NO	NO
34	FEMALE						NO	
31	FEMALE						NO	NO
13	FEMALE		HIGH				NO	NO
3	FEMALE	CIRCULATORY-DISEASES	HIGH				NO	NO
2	FEMALE		HIGH	HIGH			NO	NO
2	FEMALE		HIGH		HIGH		NO	NO
6	FEMALE	CIRCULATORY-DISEASES					NO	NO
3	FEMALE	CIRCULATORY-DISEASES	HIGH				NO	NO
5	FEMALE			HIGH			NO	NO
2	FEMALE		HIGH	HIGH			NO	NO
4	FEMALE					HIGH	NO	NO
3	FEMALE	ENDO/NUTR/METABOL-DISEASE					NO	NO

Σχήμα 47. Αποτελέσματα παραδείγματος συνάθροισης

Ο τελευταίος αλγόριθμος του παραδείγματος μας είναι η συνάθροιση με την μέθοδο *k*-Means. Λόγο των προβλημάτων που αναφέραμε στο προηγούμενο παράδειγμα, και συγκεκριμένα λόγο του μεγάλου μεγέθους των καταχωρήσεων που δεν λάμβανε υπόψιν του , αναγκαστήκαμε να επιλέξουμε ένα νέο σύνολο στοιχείων. Τα στοιχεία που επιλέχθηκαν ήταν, από το πεδίο `PATIENT` το στοιχείο `GENDERID`, από το πεδίο `CLINICAL_EXAM` επιλέξαμε το στοιχείο `DIASTOLICPRESSURE`, από το πεδίο `BIOCHEMICAL_EXAM` τα στοιχεία `BLOOD_SUGAR`, και τελος από το πεδίο

PROBLEM_OF_MEDICAL_HISTORY επιλέξαμε ICD9DISEASEID.

Ο αριθμός των καταχωρήσεων που έμειναν για εκτέλεση από τον *k*-Means ήταν 680. Παράμετρος για τον αλγόριθμο ήταν ο αριθμός των συστάδων που επιλέξαμε τον αριθμό τρία. Τα αποτελέσματα εμφανίζονται στο σχήμα 48.



Σχήμα 48. Αποτελέσματα παραδείγματος με K-Means

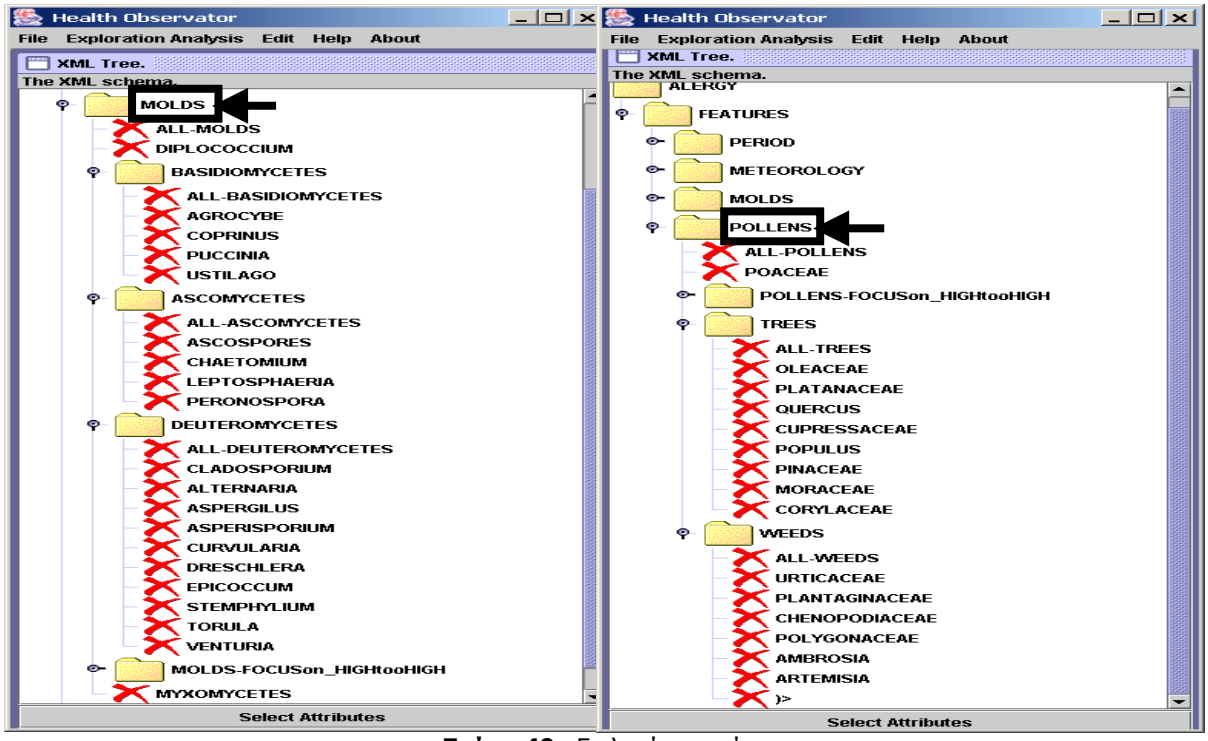
7.3. Παράδειγμα Τρίτο: Συσχέτιση Συγκεντρώσεων Αλλεργιογόνων και Μετεωρολογικών Συνθηκών

7.3.1. Το Υπόβαθρο και οι Στόχοι της Εφαρμογής

Η συγκεκριμένη μελέτη εστιάζεται στην ανάλυση δεδομένων συγκέντρωσης αλλεργιογόνων στην ατμόσφαιρα και η συσχέτισή τους με τις επικρατούσες μετεωρολογικές συνθήκες. Τα δεδομένα έχουν συλλεχθεί (και συνεχίζουν να συλλέγονται) από τον ιατρό κ. Μιχάλη Γωνιανάκη μέσω σχετικών σταθμών 'σύλληψης' και μέτρησης *κόκκων αλλεργιογόνων* στην ατμόσφαιρα της πόλεως του Ηρακλείου Κρήτης.

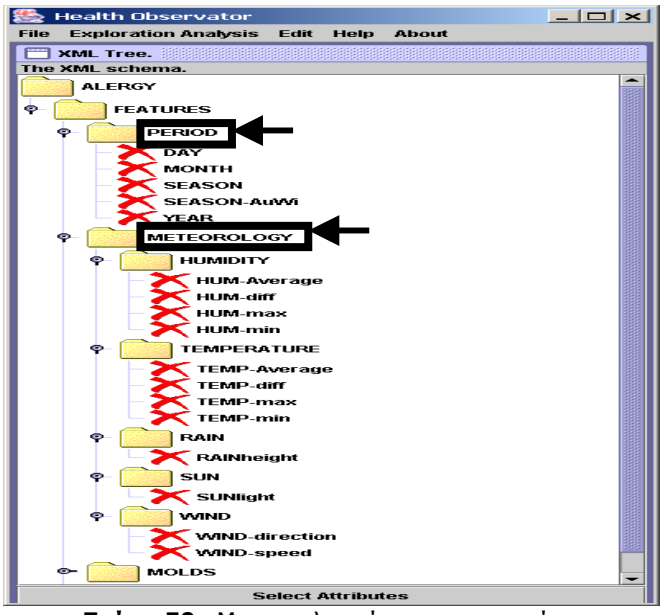
Τα αλλεργιογόνα, αφού συλληφθούν, ταυτοποιούνται (μέσω μικροσκοπικής επίβλεψης και ανάλυσής τους) ως προς τη κατηγορία/υπο-κατηγορία στην οποία ανήκουν. Η μελέτη εστιάζεται σε δύο-(2) τέτοιες μεγάλες κατηγορίες: (α) **μύκητες** – **molds** ('μούχλα' από χλωρίδα) και (β) **γύρεις** – **pollens** (από την ανθοφορία ανθέων, δένδρων και σποράς αγρωστωδών [weeds]). Κάθε μία από αυτές τις κατηγορίες χωρίζεται σε μια σειρά από αντίστοιχες υπο-κατηγορίες. Στο σχήμα 49 (το

οποίο αποτελεί και το σχετικό HCI επιλογής χαρακτηριστικών του συστήματος HealthObs) φαίνονται οι σχετικές κατηγορίες/υπο-κατηγορίες.



Σχήμα 49. Επιλογή στοιχείων

Στο σχήμα 50 φαίνονται τα χαρακτηριστικά *μετεωρολογικών συνθηκών* και *χρονικής περιόδου* με τα οποία αναζητούνται ενδιαφέροντες συσχετίσεις με τις συγκεντρώσεις αλλεργιογόνων.



Σχήμα 50. Μετεωρολογικά χαρακτηριστικά

Οι εισαγωγές των παραπάνω σχημάτων συνθέτουν το συνολικό μοντέλο-δεδομένων της εφαρμογής και του συγκεκριμένου προβλήματος.

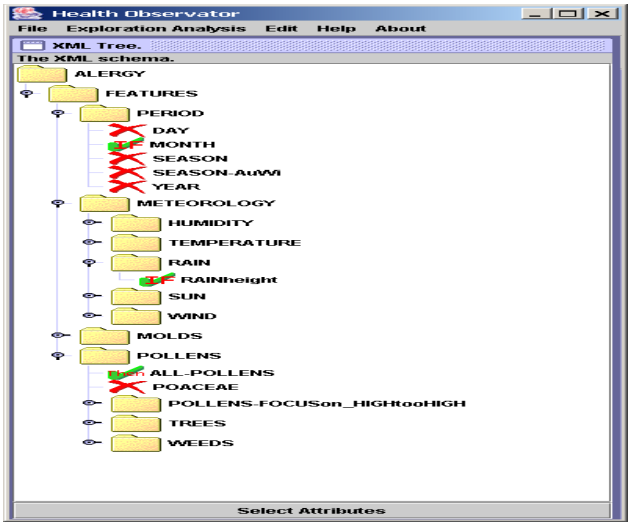
Ο μακροπρόθεσμος στόχος της συγκεκριμένης έρευνας αποβλέπει:

- Στην ενημέρωση πολιτών και **τουριστών** για την καθημερινή διακύμανση των αλλεργιογόνων (δελτίο πρόβλεψης κυκλοφορούντων αλλεργιογόνων), σχεδόν όλο το χρόνο, για λόγους πρόληψης ‘η/και πιο ορθολογικής αντιμετώπισης του αλλεργικού ασθενούς. Δημιουργία βάσεων δεδομένων και μηχανισμών διαχείρισης τους για τα συλλεγόμενα δεδομένα και τα μοντέλα πρόβλεψης.
- Δημιουργία του Web-portal “**eAllergy**”, για πληροφόρηση γύρω από τα αλλεργιογόνα σε διαφορετικές περιοχές και πόλεις, διασύνδεσής του με τις βάσεις δεδομένων για τη δημιουργία μηχανισμών και υπηρεσιών αυτόματης ενημέρωσης και συντήρησής τους. Μελέτη ενσωμάτωσης και ένταξη της προβλεπόμενης διαδικτυακής υπηρεσίας σε δήμους, νομαρχίες και περιφέρειες της χώρας, καθώς και σε τουριστικούς πράκτορες.

7.3.2. Ενδιαφέροντες Συσχετίσεις: Ερωτήματα και Διερεύνηση τους

Μέσω ενός συγκεκριμένου σεναρίου αναζήτησης ενδιαφερόντων αλληλοσυσχετίσεων διαμορφώσαμε το ακόλουθο ερώτημα: *«Επηρεάζει; πόσο και ποια χρονική περίοδο η βροχή τη συγκέντρωση γύρεων (pollens)».*

Έτσι, εστιάζουμε τη διερεύνηση και επιλέγουμε τα χαρακτηριστικά: `ALL-POLLENS` (την άθροιση των συγκεντρώσεων όλων των υπο-κατηγοριών γύρεων- εδώ εστιάζουμε στις συγκεντρώσεις όλων των δένδρων, `ALL-TREES`, και στις συγκεντρώσεις όλων των αγροστοδών, `ALLWEEDS`), `PERIOD/MONTH` (χρονική περίοδος) και `RAINFALL` (ύψος σε χιλιοστά της βροχής); Όπως βλέπουμε και στο σχήμα 51, όπου στο HCI του HealthObs τα χαρακτηριστικά RAIN και PERIOD/RAIN απαιτείται να βρίσκονται στο ‘IF’ σκέλος των κανόνων-αλληλοσυσχέτισης και το χαρακτηριστικό ALL-POLLENS στο ‘THEN’ σκέλος.



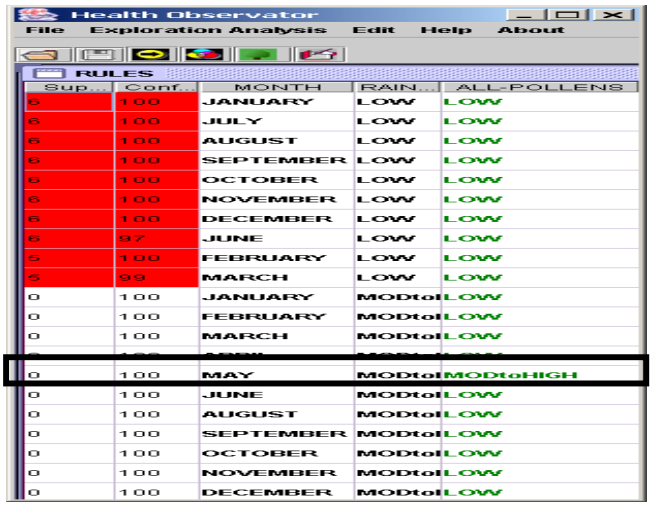
Σχήμα 51. Επιλογή στοιχείων

Εδώ να σημειώσουμε ότι οι τιμές των παραπάνω χαρακτηριστικών έχουν ***διακριτοποιηθεί*** (discretised) σε συγκεκριμένα διαστήματα τιμών στα οποία έχει αποδοθεί φυσική σημασία. Για παράδειγμα, και για τους σκοπούς της συγκεκριμένης μελέτης, η σχετική διακριτοποίηση και απόδοση φυσικής σημασίας στα αντίστοιχα διαστήματα τιμών φαίνεται στο παρακάτω πίνακα.

Πίνακας 2. Διακριτοποίηση τιμών και απόδοση φυσικής σημασίας στα σχετικά διαστήματα τιμών

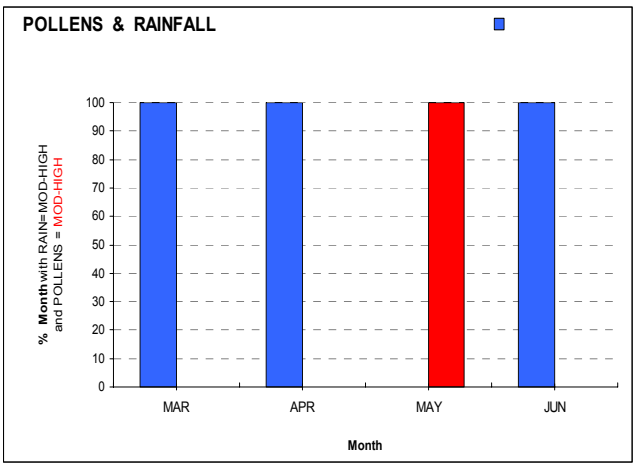
	LOW	**MODERATEtoHIGH**
ALL-POLLENS	(0 – 159.4]	(159.4 – 1894.0+)
ALL-TREES (δένδρα)	(0 – 163.5]	(163.5 – 1347.3+)
ALL-WEEDS (αγρωστωδή)	(0 – 37.1]	(37.1 – 309.9+)
RAIN	(0 – 12.3]	(12.3 – 69.6+)
PERIOD/**MONTH**	12 τιμές; JAN, FEB, MAR, … DEC	

Ενεργοποιώντας τη λειτουργία 'Association Rules' του HealthObs, προέκυψαν οι κανόνες-αλληλοσυσχέτισης που φαίνονται στο σχήμα 52. Γνωρίζοντας ότι οι συγκεντρώσεις γύρεων μετρούνται κατά τη περίοδο της άνοιξης και αρχές καλοκαιριού.



Σχήμα 52. Αποτελέσματα κανόνων συσχέτισης

Γνωρίζοντας ότι οι συγκεντρώσεις γύρεων μετρούνται κατά τη περίοδο της άνοιξης και αρχές καλοκαιριού, το σχήμα 53 είναι το αντίστοιχο μπαρόγραμμα (bar-chart). Από τον σχετικό κανόνα και το μπαρόγραμμα είναι σαφές ότι η βροχή επηρεάζει τη συγκέντρωση γύρεων κατά το μήνα Μάιο – *«IF* `MONTH=MAY` *&* `RAIN=MODtoHIGH` *THEN* `POLLENS-MODtoHIGH` *– 100%"*.



Σχήμα 53. Σύνοψη αποτελεσμάτων 4 μετρήσεων

Με βάση το παραπάνω αποτέλεσμα θα θέλαμε να αναζητήσουμε ποιες υποκατηγορίες γύρεων επηρεάζονται από τη βροχή και πόσο. Έτσι, τέσσερις διαφορετικές αναζητήσεις κανόνων αλληλοσυσχέτισης μορφοποιούνται με στόχο την εύρεση του 'προφίλ' συγκέντρωσης των υποκατηγοριών `TREES` και `WEEDS` όταν δεν λαμβάνεται υπόψη η 'βροχή' (`TREES_unconditioned`, `WEEDS_uncoditioned`) και όταν λαμβάνεται υπόψη (`TREES_with_RAIN=MODtoHIGH`, `WEEDS_with_RAIN=MODtoHIGH`). Όπως φαίνεται και από τα σχετικά μπαρογράμματα των κανόνων αλληλοσυσχέτισης που προέκυψαν στο σχήμα 53 παρατηρούμε:

- Απόκλιση του 'προφίλ' υψηλών συγκεντρώσεων (MODtoHIGH) των TREES για τις βροχερές ημέρες Απριλίου και Μαΐου.
  - Τις βροχερές ημέρες του Απριλίου <u>δεν</u> παρατηρούμε υψηλές συγκεντρώσεις σε σχέση με τις 'uncoditioned' συγκεντρώσεις:

IF `MONTH=APRIL` *THEN* `TREES=MODtoHIGH` ***– 9%***

vs.

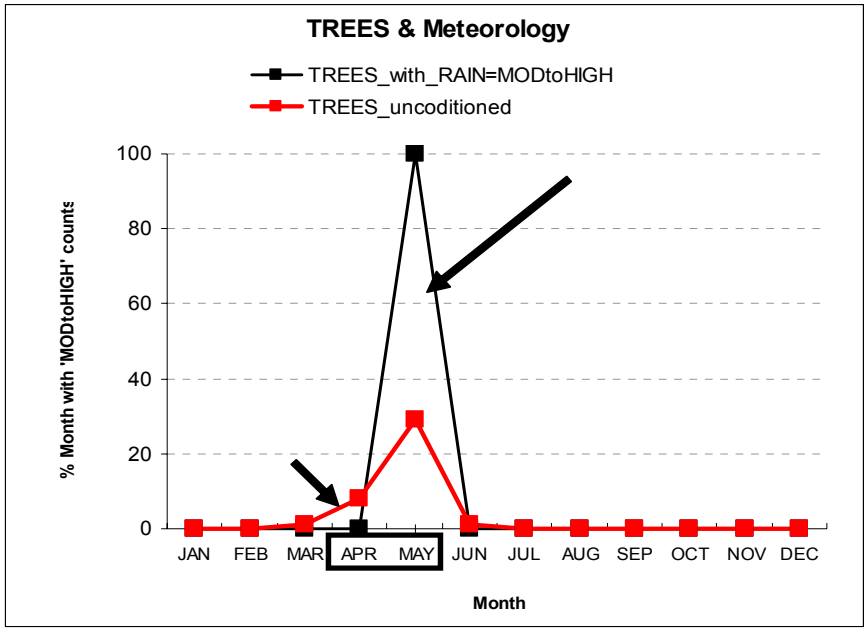
IF `MONTH=APRIL` *&* *`RAIN=MODtoHIGH`* *THEN* `TREES=MODtoHIGH` ***– 0%***

  - Τις βροχερές ημέρες του Μαίου παρατηρούμε αύξηση των υψηλών συγκεντρώσεων σε σχέση με τις 'uncoditioned' συγκεντρώσεις

IF `MONTH=MAY` *THEN* `TREES=MODtoHIGH` ***– 33%***

vs.

IF `MONTH=MAY` *&* `RAIN=MODtoHIGH` *THEN* `TREES=MODtoHIGH` ***– 100%***



Σχήμα 54. Προφιλ για βροχερές μερες με TREES

- Απόκλιση του 'προφίλ' υψηλών συγκεντρώσεων (MODtoHIGH) των WEEDS για τις βροχερές ημέρες της άνοιξης (Μάρτιο, Απρίλιο και Μάιο), τον Ιούνιο και τον Σεπτέμβριο.
  - Τις βροχερές ημέρες του Μαρτίου, Απριλίου, Ιουνίου και Σεπτεμβρίου παρατηρούμε μείωση των υψηλών συγκεντρώσεων σε σχέση με τις 'uncoditioned' συγκεντρώσεις:

IF MONTH=APRIL/MAY/JUNE/SEPT *THEN* WEEDS=MODtoHIGH – ***20/57/38/10%***

vs.

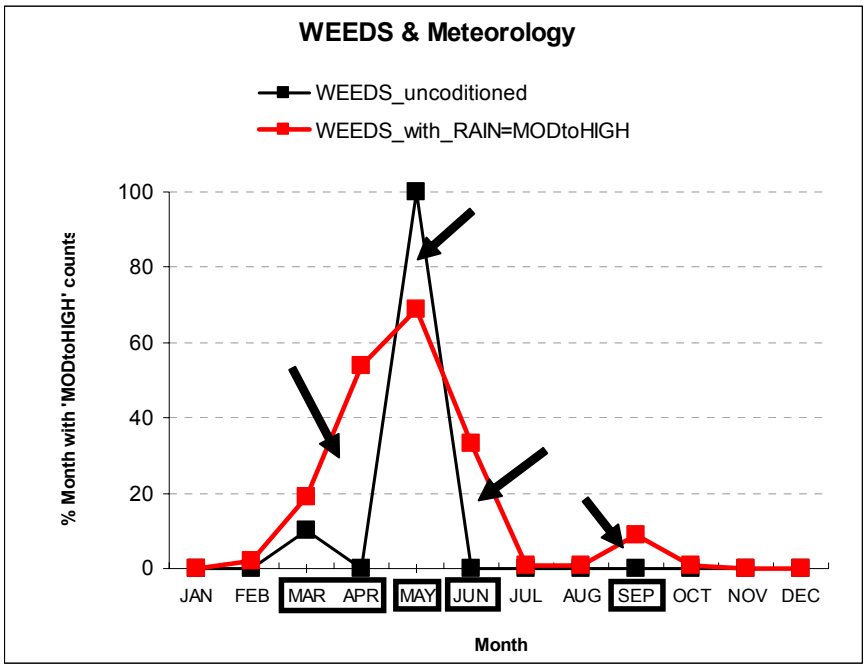
IF MONTH=APRIL *&* *RAIN=MODtoHIGH* *THEN* WEEDS=MODtoHIGH – ***10/0/0/0%***

  - Τις βροχερές ημέρες του Μαίου παρατηρούμε αύξηση των υψηλών συγκεντρώσεων σε σχέση με τις 'uncoditioned' συγκεντρώσεις

IF MONTH=MAY *THEN* WEEDS=MODtoHIGH – ***70%***

vs.

IF MONTH=MAY *&* RAIN=MODtoHIGH *THEN* WEEDS=MODtoHIGH – ***100%***



Σχήμα 55. Προφίλ για βροχερές μέρες με WEEDS

Από τη παραπάνω μελέτη, εκτός των χρήσιμων συμπερασμάτων (για το συγκεκριμένο πεδίο εφαρμογής), μπορούν να διαφανούν και οι πιθανές μελλοντικές επεκτάσεις του συστήματος HealthObs, όπως για παράδειγμα η ενσωμάτωση κλασσικών στατιστικών συγκρίσεων και παραγωγής διαγραμμάτων.

8. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ Ε&Α ΔΟΥΛΕΙΑ

8.1. Συμπεράσματα και Παρατηρήσεις

Κάνοντας κάποιος έναν απολογισμό, καταλήγει σε μία σειρά από πολύτιμα συμπεράσματα, που αφορούν τις τεχνολογίες που χρησιμοποιήθηκαν, τα πρότυπα που μελετήθηκαν, και τη χρησιμότητα του συστήματος ειδικότερα στα πλαίσια του HYGEIAnet. Στο υποκεφάλαιο που ακολουθεί θα σχολιαστούν αυτά τα συμπεράσματα, και κλείνοντας θα αναφερθεί η δουλειά που θα μπορούσε να γίνει μελλοντικά, έτσι ώστε να βελτιωθεί η συγκεκριμένη υποδομή.

Οι βασικές τεχνολογίες στις οποίες στηρίχτηκε αυτή η μελέτη ήταν δύο. Οι γλώσσες γενικής σήμανσης και συγκεκριμένα τα αυτό-περιγραφόμενα XML και οι τεχνικές εξόρυξης γνώσης. Η XML με την ραγδαία ανάπτυξη καθιερώθηκε πολύ γρήγορα στο χώρο, κυρίως λόγο των πλεονεκτημάτων που έχει έναντι στις βάσεις δεδομένων. Επιγραμματικά αναφέρουμε ότι ένα XML αρχείο είναι αυτό-περιγραφόμενο, αναγνώσιμο, έχει τις ιδιότητες ενός αρχείου (επεξεργασία, μεταφερσιμότητα), και είναι ανεξάρτητο πλατφόρμας. Εκμεταλλευόμενοι τις ιδιότητες της XML δημιουργήσαμε ομοιόμορφη πρόσβαση σε κατανεμημένα δεδομένα και με λειτουργίες σημασιολογικής ομογενοποίησης αποκρύψαμε την ετερογένια των πηγών.

Η δεύτερη τεχνολογία (εξόρυξη γνώσης) είναι η διαδικασία για τον προσδιορισμό πραγματικών, πρωτότυπων, χρήσιμων και κατανοητών προτύπων μέσα στα δεδομένα. Ξεφεύγοντας λίγο από τα τετριμμένα, που θέλουν τις πηγές γνώσης να είναι κατά κύριο λόγο βάσεις δεδομένων, προσαρμόσαμε αλγόριθμους μηχανικής μάθησης έτσι ώστε να λειτουργούν σε XML αρχεία.

Με την χρήση XML αρχείων και την υλοποίηση αλγορίθμων εξόρυξης γνώσης, και μηχανικής μάθησης, στα παραγόμενα XML έγγραφα δημιουργείται ένα σύστημα εξόρυξης γνώσης από ετερογενείς και κατανεμημένες πηγές πληροφορίας. Η ολοκλήρωση του συστήματος έρχεται με την υλοποίηση λειτουργικών και φιλικών προς το χρήστη διεπαφών, οι οποίες κρύβουν την πολυπλοκότητα του συστήματος. Η γνώση από τις διάφορες πηγές εξάγεται με τη μορφή XML αρχείου. Με την βοήθεια της υπηρεσίας Common Clinical Term Reference (που αναφέραμε στο κεφάλαιο 4) δημιουργούνται τα κατάλληλα λεξικά τα οποία ομογενοποιούν την πληροφορία. Αμέσως μετά οι αλγόριθμοι εξόρυξης γνώσης εκτελούνται απευθείας στο XML αρχείο. Συγκεκριμένα υλοποιήθηκαν τρεις αλγόριθμοι.

- Κανόνες συσχέτισης με τον αλγόριθμο Apriory και την δενδρική δομή prefix-tree. Ο συνδυασμός του αλγόριθμου αυτού με την συγκεκριμένη

δενδρική δομή δίνει την δυνατότητα προσπέλασης των δεδομένων μόνο μια φορά και ευέλικτης αποθήκευσης της πληροφορίας.

- Αλγόριθμος συνάθροισης βασισμένος στην δενδρική δομή prefix-tree και υλοποιεί συνάθροιση μέσω κανόνων συσχέτισης
- Ο γνωστός αλγόριθμος συνάθροισης K-Means.

8.2. Μελλοντική δουλειά

Τα μελλοντικά σχέδια επέκτασης της εργασίας αυτής, συνοπτικά θα μπορούσαμε να αναφέρουμε ότι κινούνται σε τέσσερις κατευθύνσεις:

- Το σχεδιασμό και την ανάπτυξη κατάλληλων *διεπαφών ανθρώπου-υπολογιστή* (human computer interface), αποθήκευση και την περαιτέρω επεξεργασία αποτελεσμάτων, λαμβάνοντας υπόψη θέματα που σχετίζονται με την *προσωπικότητα του εκάστοτε χρήστη* (***user profile***, ***personalization***).
- Προσθήκες νέων αλγορίθμων μηχανικής μάθησης και εξόρυξης γνώσης, όπως δένδρα αποφάσεων και ακολουθία περιοδικά επαναλαμβανόμενων γεγονότων.
- Πρόσθεση νέων πηγών δεδομένων, όπως απλά αρχεία κειμένων τα οποία με κατάλληλη επεξεργασία (text retrieval) να προσφέρουν γνώση.
- Δημιουργία εργαλείου το οποίο συγκρίνει και παρουσιάζει ομοιότητες / διαφορές σε δύο πειράματα τα οποία έχουν ακριβώς τις ίδιες παραμέτρους αλλά διαφορετικό πλήθος καταχωρήσεων (για παράδειγμα πειράματα με ίδιες συνθήκες από ίδιες πηγές αλλά με μεγάλη χρονική διαφορά). Με αυτό τον τρόπο θα μπορούσε να γίνει στατιστική ανάλυση πάνω σε εξαγόμενα δεδομένα γνώσης.

ΒΙΒΛΙΟΓΡΑΦΙΑ / ΑΝΑΦΟΡΕΣ

1. The Object Management Group (http://www.omg.org).
2. The Object Management Architecture (http://www.omg.org/oma/)
3. Common Object Request Broker Architecture (http://www.corba.org)
4. CORBAmed Health Care Domain Task Force of the Object Management Group (http://www.omg/org/corbamed).
5. Object Management Group, "The CORBAmed Roadmap", Revised Submission, OMG TC Document CORBAmed/98-02-03, February, 1998.
6. 3M Health Information Systems, and Protocol Systems, Inc., "Lexicon Query Service RFP Response", Revised Submission, OMG TC Document CORBAmed/98-03-22, March, 1998.
7. 2AB, Care Data Systems, Inc., CareFlow|Net, Inc., HBO & Company, HealthMagic, Inc., HUBlink, Inc., IDX Systems Corporation, IONA Technologies PLC, Oacis Healthcare Systems, Protocol Systems, Inc., Sholink Corporation, "Person Identification Service (PIDS)", OMG CORBAmed DTF Adopted Submission, OMG TC Document corbamed/98-02-29, February 1998.
8. 3M, Care Data Systems, Inc., CareFlowNet, Inc., HBO & Company, Philips Medical Systems, Protocol Systems, Inc., "Clinical Observation Access Service (COAS)", OMG Document corbamed/99-03-25, April 1999.
9. Integrated Health Telematics Network Of Crete, http://www.hygeianet.gr
10. Dimitrios G. Katehakis, Manolis Tsiknakis, Stelios C. Orphanoudakis Enabling Components of HYGEIAnet TERP 2001 Boston, MA, May 8-13, 2001, pp. 146-153
11. Katehakis D.G., Chronaki C.E., et al., "Towards a Virtual Electronic Healthcare Record: The Patient Clinical Data Directory", Version 3.09, 1999.
12. Tsiknakis M., Chronaki C.E., Kapidakis S., Nikolaou C., and Orphanoudakis S.C., **"**An Integrated Architecture for the Provision of Health Telematic Services based on Digital Library Technologies"**,** International Journal on Digital Libraries, Special Issue on "Digital Libraries in Medicine", vol. 1(3), 257-277, 1997. (http://www.ics.forth.gr/ICS/acti/cmi_hta/publications/dglib97/dglib97.html)
13. Orphanoudakis S.C., Chronaki C.E., Tsiknakis M., and Kostomanolakis S., "Telematics in Healthcare", Chapter 10, In "Biomedical Image Databases," Biomedical Image Databases, S. Wong (editor), Sharon Fletcher, Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Ma 02061
14. Orphanoudakis S.C., "Integrated Telemedicine Networks and Added-Value Services", Proc. VIII Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON'98), Lemesos, Cyprus, June 14-17, 1998(http://www.ics.forth.gr/ICS/acti/cmi_hta/publications/medicon98/medicon98.html).
15. XML Extensible Markup Language (http://www.xml.com)
16. World Wide Web Consortium (W3C) (http://www.w3.org)
17. PARDI, W., *XML in Action*, Microsoft Press, 1999.
18. OLSTAD, V., *Ibid*
19. Fayyad M. Usama, (1996), "Data mining and Knowledge Discovery: Making Sense Out of Data", IEEE EXPERT, Microsoft Research.
20. Chen Ming-Syan, Han Jiawei, and Yu S. Philip, (1996), " Data mining: An Overview from a Database Perspective", Ieee Trans. On Knowledge And Data Engineering.
21. William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. Knowledge discovery in databases: An overview. In Gregory Piatetsky-Shapiro and William J. Frawley, editors, Knowledge Discovery in Databases, pages 1-30, AAAI/MIT, 1991

22. Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Database mining: a performance perspective. IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, December 1993:914-925, 1993

23. Karsten M. Decker and Sergio Focardi. Technology Overview: A Report on Data mining. CSCS TR-95-02, May 29, 1995

24. ΚΩΝΣΤΑΝΤΙΝΟΣ Α. ΧΡΙΣΤΟΦΗΣ ΣΥΣΤΗΜΑ ΕΞΑΓΩΓΗΣ ΓΝΩΣΗΣ ΑΠΟ ΚΑΤΑΝΕΜΗΜΕΝΕΣ ΚΑΙ ΕΤΕΡΟΓΕΝΕΙΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ: ΕΦΑΡΜΟΓΗ ΣΕ ΙΑΤΡΙΚΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ. Master thesis July 2000. University of Crete Computer sience Department.

25. K. Wang ,H. Lui, Schema Dischovery from semi-stuctured Data. In Procc KDD 97 , pp 271-274 , 1997

26. Tetsuhiro Miyahara, Yusuke Suzuki, Takayoshi Shoudai, Tomoyuki Uchida, Kenichi Takahashi, Hiroaki Ueda: Discovery of Frequent Tree Structured Patterns in Semistructured Web Documents. In Proc PAKDD 2001, pp47-52, 2001

27. A. Tatsuya, A Kenji ,K. Shinji ,A. Hiroki ,S. Hiroshi, A. Setsuo Efficient Substructure Discovery from Large Semi-Structured Data. 2nd Annual SLAM Symposium on Data mining ,SDM2002

28. Tetsuhiro Miyahara, Takayoshi Shoudai, Tomoyuki Uchida, Kenichi Takahashi, and Hiroaki Ueda Discovery of Frequent Tree Structured Patterns in Semistructured Web Documents In Proc. PAKDD-2002, 341--355, 2002

29. Lisa Singh, Peter Scheuermann, Bin Chen Generating Association Rules from Semi-Structured Documents Using an Extended Concept Hierarchy. Proceedings of the International Conference on Information and Knowledge Management, November 1997

30. Mohammed J. Zaki _,Charu C. Aggarwal XRules: An Effective Structural Classifier for XML Data. 9th International Conference on Knowledge Discovery and Data Mining, Washington, DC, August 2003

31. R. Carnevale. Algoritmi per stemming ed estrazione di regole di associazione su documenti xml. Master's thesis, Apr. 2002. Master thesis supervised by Marco Colombetti and,Pier Luca Lanzi. (available in Italian).

32. R. Meo, G. Psaila, and S. Ceri. An extension to SQL for mining association rules. *Εξόρυξη δεδομένων and Knowledge Discovery*, 2(2):195 ñ 224, 1998

33. World Wide Web Consortium. XQuery 1.0: An XML Query Language (W3C Working Draft). http://www.w3.org/TR/2001/WD-xquery-20011220, Dec. 2001.

34. World Wide Web Consortium. XML Path Language (XPath) Version 1.0 (W3C Recommendation). http://www.w3c.org/tr/xpath/, Nov. 1999.

35. The Apache Software Foundation. The Apache XML Project. http://xml.apache.org/xalan-j/.

36. D. Braga, A. Campi, M. Klemettinen, and P. L. Lanzi. Mining association rules from xml data. In *Proceedings ofthe 4*ηº *International Conference on Data Warehousing andKnowledge Discovery (DaWaK 2002). September 4-6, Aixen-Provence, France*, 2002. accepted.

37. Daniele Braga , Alessandro Campi , Stefano Ceri , Mika Klemettinen , Pier Luca Lanzi. A Tool for Extracting XML Association Rules from XML Documents. In Proceedings of IEEE-ICTAI 2002, Washington DC, USA, November 2002

38. Ilker Cengiz. Mining Association Rules. Department of Computer Engineering & Information Sciences 06533 Bilkent ,Ankara ,Turkey.

39. Rakesh Agrawal, Sakti Ghost, Tomasz Imielinski, and Arun Swami. An interval classifier for database mining applications. In 18th Int'l Conf. On Very Large Databases(VLDB), Vancouver, Canada pages 560-573, 1992.

40. Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Database mining: a performance perspective. IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, December 1993:914-925, 1993.

41. Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining Association Rules between sets of items in large databases. In SIGMOD, Washington D.C, pages 207-216, May 1993.

42. Marko Bohanec and Ivan Bratko. Trading accuracy for simplicity in decision trees. Machine

Learning, 15:223-250, 1994.

43. Ramakrishnan Srikant, Quoc Vu and Rakesh Agrawal. Mining Association Rules with Item Constraints. 1997.
44. Houtsma and Arun Swami. Set-oriented mining of association rules. Technical Report RJ 9567, IBM Research Report, Oct. 1993.
45. Ashok Sarasere, Edward Omiecinsky, and Shamkant Navathe. An efficient algorithm for mining association rules in large databases. In 21th Int'l Conf. On Very Large Databases(VLDB), Zurich, Switzerland, Sept. 1995. Also Gatech Technical Report No. GIT-CC-95-04.
46. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In 20th Int'l Conf. On Very Large Databases(VLDB), Santiago. Chile, Sept. 1994. Expanded version available as IBM Research Report RJ9839, June 1994.
47. Heikki Mannila, Hannu Toivinen, and A. Inkeri Verkamo. Improved methods for finding association rules. In AAAI Workshop on Knowledge Discovery, Seattle, Washington, pages 181-192, July 1994
48. A. J. Knobbe and P. W. Adriaans. Analyzing binary associations. In Proc. of the 2nd Intl. Conf. on Knowledge Discovery and Data Mining (KDD), 1996, pp. 311314 A.J Knobbe (from Clustering Based on Association Rule Hypergraphs)
49. Eui-Hong (Sam) Han, George Karypis, and Vipin Kumar, " Clustering Based on Association Rule Hypergraphs," in Proceedings of the SIGMOD'97 Workshop on Research Issues in Εξόρυξη δεδομένων and Knowledge Discovery. 1997, ACM.
50. Frank Hoppner, Frank Klawonn, Rudolf Kruse, and Thomas Runkler. *Fuzzy Cluster Analysis, Methods for Classification, Data Analysis and Image Recognition*. ISBN 0-471-98864-2 John Wiley & Sons Ltd, 1999.
51. A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264-323, 1999.
52. Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image Processing, Analysis, and Machine Vision*. PWS Publishing, 2nd edition, 1998.
53. H RTizhoosh .Fuzzy Image Processing .http://watfor.uwaterloo.ca/tizhoosh/fip.htm June 1997.
54. Oren Zamir and Oren Etzioni. Web document clustering: A feasibility demonstration. In *Research and Development in Information Retrieval*, pages 46-54, 1998.
55. http://www.who.int/whosis/icd10/
56. http://www.fmrc.org.au/icpc-i.htm
57. http://www.cs.tcd.ie/synapses/public/
58. www.dmg.org
59. IOM issued *To Err Is Human: Building a Safer Health System*, Washington, DC: National Academy Press; 2000
60. Potamias, G. (2003). Utilizing Genes Functional Classification in Microarray Data Analysis: a Hybrid Clustering Approach. In K. Margaritis and I. Pitas (eds) *Procs 9th Panhellenic Conference in Informatics*, Thessaloniki, Greece, pp. 417-430.
61. Jacky W. W. Wan, Gillian Dobbie: Extracting association rules from XML documents using XQuery. WIDM 2003: 94-97
62. On data mining Tree structured data in XML, A.N. Edmonds, Recent Advances in Soft Computing, Nottingham, UK, Dec. 12th 2002