In Procs Machine Learning in the New Information Age, MLnet workshop,
European Conference of Machine Learning - ECML2000, Barcelona, Spain, pp. 27-38.

Mining Distributed and Heterogeneous Data Sources: A Project in the Medical Domain

K. Hristofis[1,2], G. Potamias[1,2], M. Tsiknakis[1], V. Moustakis[1,3], and S. Orphanoudakis[1,2]

[1]Institute of Computer Science, FORTH, Vassilika Vouton, P.O. Box 1385, GR-711 10, Heraklion, Crete, Greece, {hristof, potamias, tsiknaki, moustaki, orphanou}@ics.forth.gr
[2]Dept. of Computer Science, University of Crete, P.O. Box 1470, GR-714 09, Heraklion, Crete, Greece.
[3]Dept. of Production Engineering and Management, Technical University of Crete, GR-731 00, Chania, Crete, Greece, moustaki@logistics.tuc.gr

1. Introduction

With the current explosion of data, the problem of how to combine *distributed* and *heterogeneous*- D&H information sources becomes more and more critical. Besides collecting enormous amount of data it is very important to consider the general need of *semantic integration* and *knowledge discovery* from these sources, an important and necessary challenge for machine learning- ML, and data mining/knowledge discovery- DM/KDD researchers. The main differences here, and consequently the grand challenges with respect to single, static and homogeneous information sources, are: (a) the scale of the problem is much larger than anything attempted before in ML and DM/KDD, and (b) the raising need for integrating multiple knowledge representations (e.g., domain ontologies and data-models) are more important and vital (Wah *et.al*., 1993).

If the distributed nature of data has a more-or-less clear definition (even hard, and most of the times tedious to achieve), *heterogeneity* is a more complex concept. Consider for example the situation where, the same or different database applications are installed and run at different remote locations. In such a set-up users may enter and record data in a non pre-specified and non-homogeneous format. This is a common situation in an *Integrated Electronic Health Care Record* (I-EHCR) environment (Forslund, and Kilman, 1996; InterCare, 1999, pp. 7-13; Grimson *et.al*., 1997). A physician that accesses a patient's healthcare record needs an overview of the patient's EHCR segments, since in most cases only a small fraction of the complete record will be selected and presented in detail. That also means that when accessing a particular clinical information system there is a need for extracting only a subset of the information stored in it. The real issue here is not only how to access specific information systems that maintain EHCR segments, but also how to identify and *index* the essential information in them. A promising approach to this integration problem is to gain control of the organization's information resources at a *meta-data* level, while allowing autonomy of individual systems at the data instance level. The objective of the meta-database model is to achieve enterprise information integration over distributed and potentially heterogeneous systems, while allowing these systems to operate independently and concurrently (Hsu, 1992). However, achieving integration at the *semantic level* is a challenging problem mainly because the logic, knowledge, and data structures used in various systems are complex and often incompatible (Sciore, 1994). In addition, the further someone wishes to hide heterogeneity, the more he/ she has to deal with semantic integration issues. Thus, a

realistic solution should hide heterogeneity at the top level, while making the individual sources of information appear to end users as a large collection of objects that behave uniformly (Baldonado, 1996).

This paper presents the problem of discovering and acquiring knowledge form D&H *clinical* data sources. In particular, we tackle the problem of inducing interesting *associations* between data items stored in remote clinical information systems. The test-bed environment of our approach is the *HYGEIAnet:* The *Integrated Health Care Network of Crete* (Tsiknakis, 1997; HYGEIAnet Web site). One of the basic healthcare services offered within the HYGEIAnet network is the access to patients' clinical information stored in autonomous (legacy) clinical information systems. Even if the focus is on the medical domain, the proposed methodology and solutions could be smoothly extended to cover the general case of other application domains.

In the next section, we present the architecture of an integrated environment for mining D&H data sources. Section 3, presents the basic technology for accessing distributed and structured data sources, as well as the processes for the semantic homogenization and integration of heterogeneous data sources and items. In section 4, we present the information and data representation framework; based on the XML framework and technology. Section 5, presents the machine learning and data mining processes, which are being adapted on flexible data representation structures. In section 6, some preliminary experimental results are presented. In the last section we conclude, and discuss on the future research and development agenda.

2. General Architecture

To tackle the problem of mining H&D data sources, a *multi-phase* data integration procedure should be followed. The main challenge is how: data mining and machine-learning operations are *adapted* and made operational. A rational approach to this procedure should efficiently confront and cope with: (1) efficient *access* to structured and distributed data sources; (2) reliable *homogenization* and *integration* of heterogeneous data; with a dedicated domain *ontology* and respective ontological operations playing an important role; (3) effective and reliable data *processing* operations (e.g., traditional statistical analysis, *data mining*, etc); and (4) *presentation* of results (e.g., *visualization* operations).

Figure 1 below, shows our view on the general architecture of an integrated environment for mining distributed and heterogeneous data sources. The involved operations and their relations (i.e., workflow) are also indicated. The overall schema resembles an information/ data brokering architecture, to be realized by respective *mediation services*. The operations could be assigned to autonomous *agents*, resulting into a system of *integrated and co-operating agents* (Knoblock and Yigal 1994; Haverkamp and Gauch, 1998; Weiderhold, 1992). The access operations lying between, (i) the autonomous clinical information systems and the *mediator*, and (ii) the data warehouse operations (offered via a dedicated server; the Patient Clinical Data Directory), are already in place and operational within the integrated health Telematic network of the Crete region.

Our work expands the architecture by adding (a) the semantic indexing operations, (b) the DTD/XML generation, and parsing operations, (c) the object-oriented data representation schemas and operations (elaborated around a *forest-like structure*), and (d) the adaptation of KDD operations- realized by a special devised Associations Rule Mining – ARM algorithm.

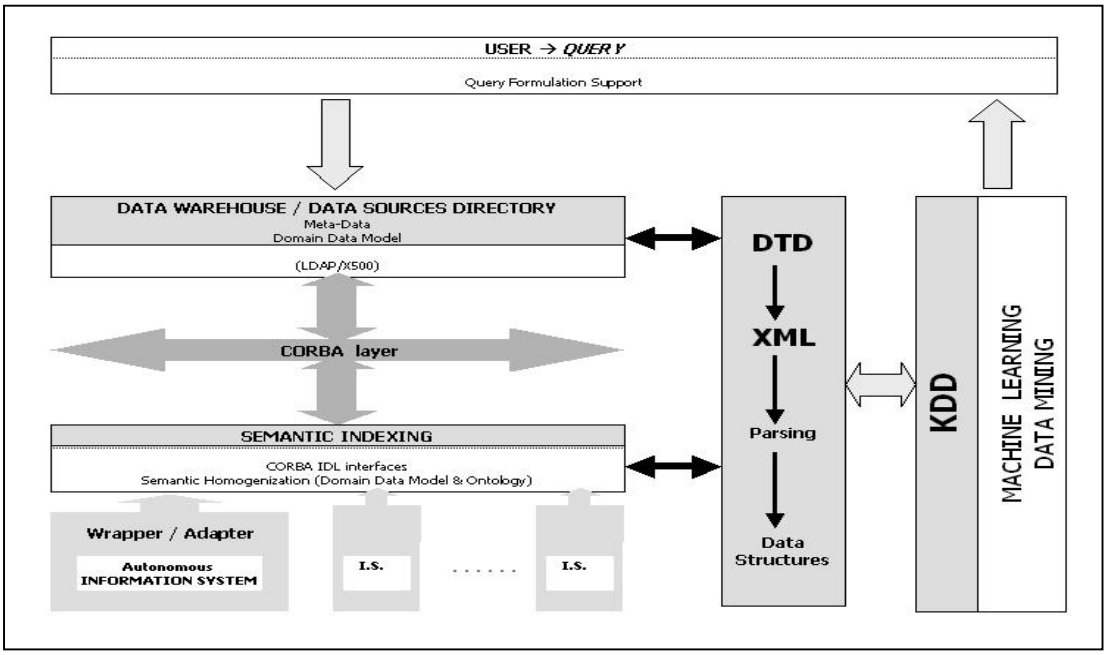

Figure 1. *The General Architecture and Workflow of an Integrated Environment for Mining Distributed and Heterogeneous Data Sources*

3. Data Access & Integration

The ultimate goal of an information brokering enterprise is the integration and fusion of D&H information and data sources on the Internet, and the *delivery* of the integrated and processed data to the user seekers. To this end, the uniform access to D&H data sources is a fundamental and inevitable condition. It has to be noted that the gathering of data is not an intention *per se*. What it matters is the exploitation of the data towards the extraction of useful and meaningful conclusions.

3.1. Access to Distributed Data

The first requirement to information and data access via the Internet is the development of the appropriate communication and inter-networking infrastructure. The Object Management Group's- OMG's CORBA- *Common Object Request Broker Architecture* (OMG group, CORBA Web site) environment offers the specifications of the standards towards uniform and seamless communication between distributed information systems (independently of the hardware or operating system platform). Moreover, the LDAP/X500 (ITU, 1993; Shuh, 1997) directory servers offer *fast* and *effective* identification mechanisms of references within the data.

The *Patient Clinical Data Directory*- PCDD (InterCare, 1999) is an implementation of the I-EHCR in the Crete region. It is used in order to access and retrieve patients' clinical information stored into distributed and remote areas in the region. Via PCDD authorized users (physicians, nurses and other healthcare personnel) have access to the patients' distributed healthcare *'segments'*. Communication and access is based on the X500 protocol utilizing an LDAP directory server. The segments of the patients' integrated healthcare records are appropriately *indexed* following a standard *encounter*-based model for representing and storing patients' visits to particular healthcare units. The adopted schema is based on the universal and unified SOAP- *Subjective - Objective - Assessment - Planning*

model, a unified framework and standard for representing medical encounters. Our specific implementation of the SOAP model is shown in Figure 2 below.

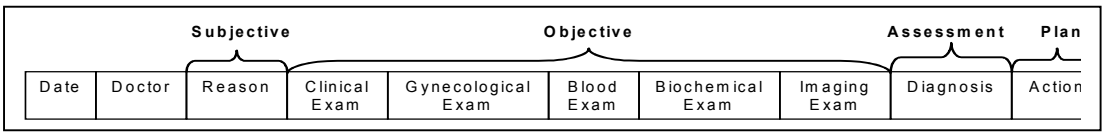

Figure 2. *The implemented SOAP model for representing patients' encounter information*

PCDD stores just *pointers* to specific patient's encounters (i.e., URLs), without reproducing and replicating the respective clinical data, being stored and kept in the autonomous and distributed clinical information systems. What PCDD, and especially the SOAP model offers is a *first-level meta-data abstraction of the distributed data sources and of the respective data items*. In other words, it provides the necessary information about '*where to search for*'. The *semantic mapping* of these pointers to patients' clinical data is realized via specially devised *wrappers* of the respective clinical information systems that belong to the *federation* (e.g., if a patient have a `'blood test'` examination performed in some of the distributed clinical units then, the wrapper of the respective clinical information system records a respective `'Blood Exam'` link/ pointer in the PCDD store.

3.2. Semantic Homogenization

Having identified a meta-data layer that points to particular data sources, i.e., to the respective segments of patients' integrated electronic health care records, the problem is how to extract the respective data. This is achieved via customized CORBA IDL interfaces.

While CORBA makes it possible for developers to independently contribute to a library of components across platforms and languages, it offers little or no help with the *knowledge-level* task of ensuring that particular components actually can work together. Furthermore, although IDL specifies the syntax necessary for interoperation and access to distributed information, it does not describe the *semantics* for what this information stands for. The inclusion of semantics would provide what is currently missing from IDLs: information about the meaning of a component; information about what the component will accomplish; and information about the relationship between a method's inputs and outputs (Gennari *et.al*., 1996). This is a task to be accomplished by: (a) the introduction and utilization of a domain *data-model*, and (b) the incorporation of respective *domain ontology*.

Clinical data-model. If some way of specifying semantics- realized and linked with a *domain-ontology*, could be layered on top of IDL and a CORBA component implementation, this architecture could provide a rich and flexible environment for developing systems that semantically integrate heterogeneous information sources and data items. To this end we have adopted and elaborated on a specific IDL interface that cope with access to clinical information and data, i.e., the COAS- *Clinical Object Access Service* interface (COAS, 1999). COAS is a standard interface provided by CORBAmed- OMG's medical sub-group, used to access *clinical observations* from distributed clinical information systems. Observations can be quantitative, qualitative, and recordings, e.g. vital signs and clinical laboratory results, trends in measured values, impressions from a clinical exam, correlation of several qualitative impressions, and images. For the purposes of our information model and the derived IDL, a clinical observation includes any clinically related item that has the necessary

context information to enable it to be queried from a COAS server. The general schema of the COAS interface is shown in Figure 3 below.

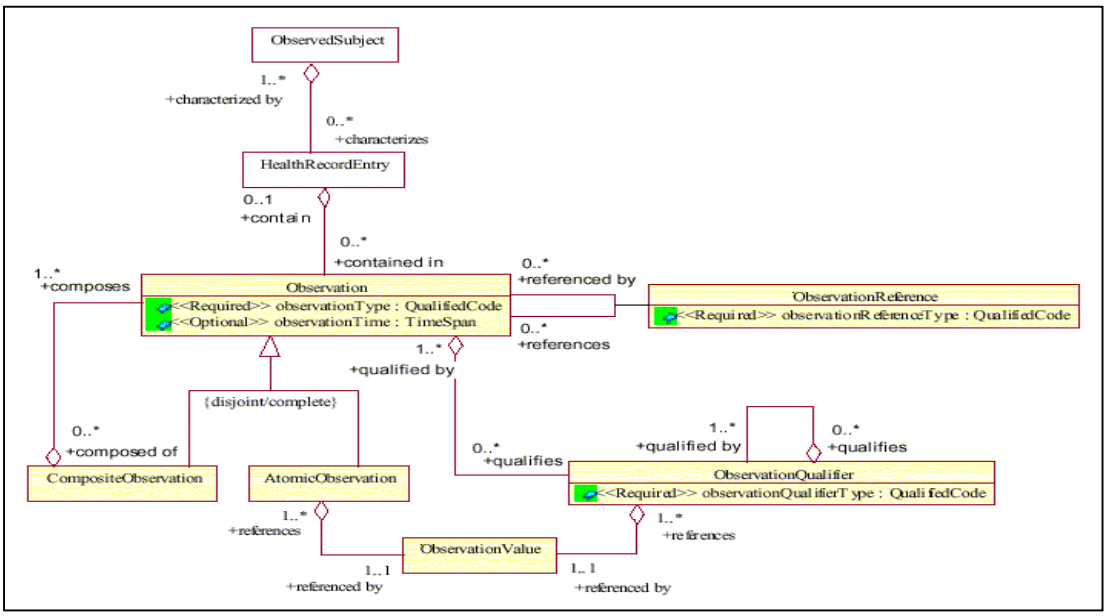

Figure 3. *The COAS (Clinical Object Access Service) Classes/Objects Diagram*

As it may be observed in the figure above, COAS makes no commitment for the exact (clinical) meaning of the information represented by respective clinical observation objects. The basic service offered by COAS is a *hierarchical* organization of clinical information (i.e., `'composite observation'` → `'atomic observation'`). It also provide some *qualifiers* for each observation, for example: patient's id, place and time that the observation was recorded etc. COAS presents a simple, standard and operational interface that accesses, and at the same time *structures, on a second-level of abstraction the information and the data* stored into various distributed clinical information systems.

Medical domain ontology. The success of an information brokering service that access and retrieve distributed data stores depends heavily on its ability to cope with the heterogeneous nature of the stored data. Take for example a medical database application stored into different locations. If the application does not provide a unique and unified *coding* schema for diagnoses then, there may be the case where, the remote data stores record the same data items (e.g., diagnoses and symptoms) with different names. So, even if we have access to both data stores, it is not profound how an associations' rule mining process could be activated in order to discover associations between diagnoses and symptoms. On the other hand, we may face the situation of integrating two (or, more) databases. Take for example the situation where a clinical information system records the clinical findings of a group of patients, with another one recording laboratory examination findings for the same group of patients (e.g., a laboratory or, a radiological information system). How, could we couple the data from both databases in order to discover interesting associations between diagnoses/diseases and laboratory findings?

The only way to cope with this problem is the incorporation of a *domain specific ontology*. Towards this goal, we have designed the format of a file that stores *common* and *universally accepted* names and codes of medical terms, the *Common Clinical Term Representation* file - CCTR. The specifications and the construction of the CCTR file were made according to the UMLS: *Unified Medical Language System*

(UMLS, 2000), and to the ICD: *International Coding for Diseases* standards. With reference to the ICD code different *lexicons*- for different languages and for different clinical information systems could be easily elaborated and adapted. In particular the UMLS meta-thesaurus system offers such mechanisms and functions; having devised a lexicon within UMLS then, we may easily transform it into the CCTR format. For the purpose of uniformity, we have devised a DTD grammar that automatically generates an XML document for the respective CCTR file.

4. Uniform Data Representation

Having accessed and uniformly organized the distributed clinical data the problem is how we represent it. The XML standardized infrastructure (XML Web site) could serve this need. Towards this end, we have developed and implemented a *COAS-compliant DTD grammar* in order to automate the generation of XML documents, the content of which corresponds to the remotely accessed and retrieved data.

Information processing takes place exclusively on top of the XML documents. For this purpose a special XML *parser* is devised (in Visual C++). The parser, (i) reads (scans) the XML document, (ii) identifies composite/atomic observations and their corresponding values, and (iii) constructs a *forest-like* structure for storing and retrieving the XML content. Each *tree* of the forest corresponds to a specific *composite observation*, starting from the root of the tree (e.g., ObsNAME: `"SYSTOLICPRESSURE"`, ValueObsNAME: `"20"`, for clinical examinations and symptoms). Each observation corresponds to an *attribute* (or, *feature*) in ML/DM terminology. In figure 4 below, an example of the forest-like data structure is shown.

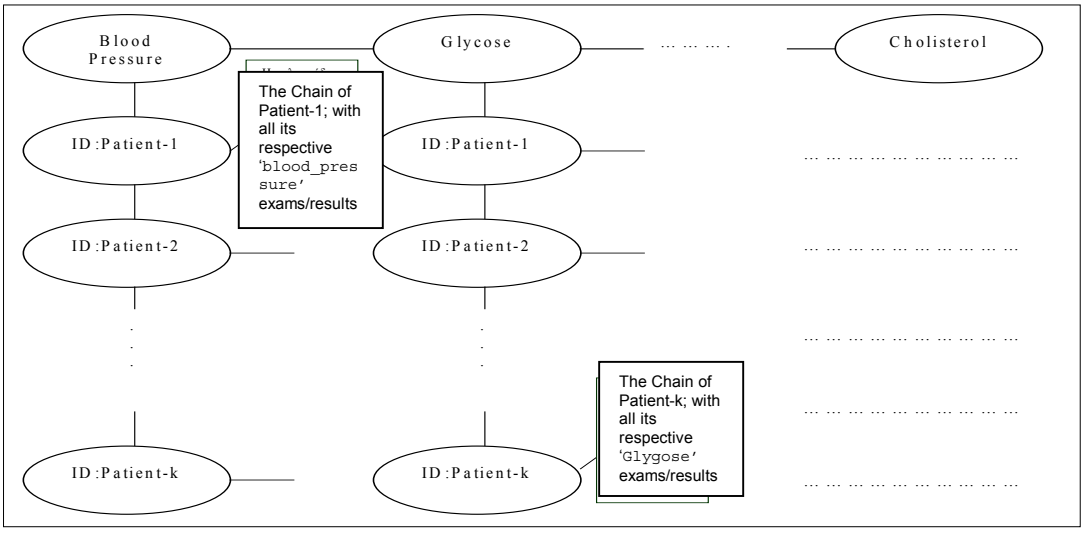

Figure 4. *The 'forest-like' data structure for encoding clinical observation values, stored in respective XML documents*

During XML parsing, the atomic observations (e.g., a symptom name like `'SYSTOLICPRESSURE'` or, a diagnosis name like `'HYPERTENTION'`) are uniformly encoded and recorded in the forest-like structures according to: (1) their respective *common reference name* in the CCTR file, and (2) their respective ICD code. In this way, the accessed and retrieved data are semantically homogenized, represented and stored.

We argue that *future databases will use XML-like structures in order to store and retrieve data*. Some of the most popular and standard ones in the market (e.g., ORACLE) exploit XML-like schemas for exporting a database. Using these data

structures it is very easy to construct and apply global or, partial *querying operation* over the whole database, and prepare customized data formats appropriate for applying machine learning and data mining operations. In this respect, the forest-like structure renders a powerful and much promising approach for hosting/ adapting data mining operations on top of databases (as it will become clear in the sequel).

5. Mining Distributed & Heterogeneous Data

With the multi-phase data integration process described so far, we have on our disposition a set of (patients' clinical) data, homogenized and kept in appropriate forest-like data structures. We may now proceed on the specifics of data mining processes to be adapted on top of these structures. In particular, we focus on the problem of discovering *interesting associations* (i.e., *association rules*) between the recorded patients' clinical data items.

Association Rule Mining- ARM, is among the most advanced and interesting methods introduced by machine learning and data mining research (Agrawal, *et.al*., 1993; Agrawal and Srikant, 1994; Mueller, 1995). The definition of an ARM problem has as follows:

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of literals, called *items*. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. We say that a transaction T *contains* X, a set of some items in I, if $X \subseteq T$ .An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in the transaction set D with *confidence c* if *c*% of transactions in D that contain X also contains Y. The rule $X \Rightarrow Y$ has *support s* in the transaction set D if *s*% of transactions in D contains $X \cup Y$.

Given a set of transactions D, the ARM problem is to discover (identify and form) the associations that exhibit support and confidence values higher that the user specified minimum support- *minsup*, and minimum confidence- *minconf* levels, respectively. Note that, the exploration for association rules does not confined by the representation of D; being a plane ascii file or, a relational database. In our case, we rely of the XML-generated forest-like data structures described in the previous section. The forest-like structure allows for the *dynamic creation of chains of trees' nodes (see figure 4 above)*- with each chain representing an item, adding a high degree of *flexibility* in the involved (and demanding!) search operations.

5.1. Items, Transactions and Distributed Medical Data: Assumptions and Conventions

At this point it has to be explained what is the interpretation that we assign to the concepts of item and transaction. In other words, we have to make some assumptions for the ARM problem(s) to be defined and formed, as related to the application domain (i.e., medicine), and to the distributed nature of the data as well.

- Each transaction corresponds to a specific patient encounter (i.e., identifiable visit of a patient in a healthcare unit of the federation). Each encounter is *uniquely identified* by reference to four attributes, i.e., `'Patient_Id'`, `'Information_System'`, `'Visit_Id'`, and `'Date'`. Information about these attributes is recorded in the distributed clinical information systems of the federation, and it is part of their respective CORBA IDL wrapping interfaces.
- An item is represented by the triplet `<Atomic_Observation, Value, Interval>` where,
  i. `'Atomic_Observation'` as described in section 4 (e.g., *`'cholesterol'`*);

ii. `'Value'`, the value recorded for a specific atomic observation (e.g., *`'251'`*, as a value for *`'cholesterol'`*); and
iii. `'Interval'`, the interval in which `'Value'` belongs (e.g., *`'Interval-2'`*: [120-200] = *`'Normal'`*).

Numeric attributes and items. In the medical domain most of the items are numeric. In order to discover associations between items we have to transform these numeric values into nominal ones, at least for the association rule discovery procedures that cope only with nominal attributes. In medicine, there are always intervals that specify the *low*, *normal*, and *high* status of each measurement. These intervals may come from expert medical advice and/ or form established and universally accepted clinical protocols and guidelines. In the case that for some attributes this information is lacking, we proceed into an automatic *discretisation* (or, *nominalization*) of numeric attributes. Note here that, in the deployed version of the system the user may interact with it and revise the computed intervals and/ or form his/ her own.

5.2. The AprioriXML Algorithm: An Apriori-like ARM Algorithm

The ARM techniques that we have implemented rely on the principles of the *Apriori* algorithm, and its offspring enhancements, *AprioriTid* and *AprioriHybrid* (Agrawal and Srikant, 1994). Taking advantage of the employed rich and dynamic forest-like data structures we have added some extra features to these algorithms. With these revisions our ARM algorithm, called *AprioriXML*, is enhanced with object-oriented search operations able to work on top of XML-structured data and respective representation formalisms. Figure 5 below, shows an outline of our AprioriXML algorithm.

One of the fundamental operations of Apriori-like algorithms is the generation of *large itemsets*. This is achieved by making multiple scans on the input data. The AprioriXML algorithm follows a similar- but more economic operation, by reducing the size of the database (more precisely, of the forest-like structure) after each scan. This composes the major revision to the standard Apriori-like algorithms.

```
1.  L1 = {large 1-itemsets};
2.  for (k=2; Lk-1≠∅; k++)
    {
3.        Ck = apriori-like-gen(Lk-1);                  // phase 1
4.        New_D = fix_update_old_D(Lk-1);               // phase 2
5.        ∀ c ∈ Ck                                      // phase 3
       {
6.          ∀ remain&&update transaction t ∈ New_D
              {
7.            find_itemset = look_for(c, t)
8.            if (find_itemset = = 1)
9.                  c.count ++;
              }
       }
10.       Lk = {c ∈ Ck | c.count >= minsup}            // phase 4
     }
11. Answer = ∪k Lk;
```

Figure 5. *Outline of the AprioriXML ARM algorithm*

Step-1. Scan of the input data (i.e., scan of the forest); the algorithm counts the occurrences of each item; the result is the generation of the *large 1*-itemesets.

Steps 2-k. In each of the subsequent *k* steps the following operations are performed.

i. ***Phase-1.*** *Generation of candidate itemsets.* The large itemsets L_{k-1} found in the *k-1* step of the algorithm are used in order to compute and form the corresponding C_k *candidate* itemsets. The basic function is, `apriori-like-gen`, which takes as argument the set of large itemsets L_{k-1} and returns a superset of the *k*-itemsets.

ii. ***Phase-2.*** *Elimination of Redundant Information.* Making use of the knowledge about the L_{k-1} large itemsets, proceed into the *dynamic revision* of the database, and the restructuring of the instance forest-like structure. Revision is taken place by cutting the nodes that will not affect the generation of itemsets in the next steps. The basic function is `fix_update_old_D`. For k=2, the new database is the initial one by *cutting* all the nodes that contain no large *1*-itemsets. So, in the very early steps of the algorithm our database (i.e., instance of the forest-like structure) is freed from information that will play no role in the future steps. For k>2, if all candidate itemsets does not occur in a transaction then, the *whole transaction is deleted*. The major advantage of this phase is that, in the next *counting* phase there will be *no redundant comparisons*. With this operation the size of the database is gradually decreasing, and we gain in both space and time.

iii. ***Phase-3.*** *Counting.* The supports of all C_k candidate itemsets are computed. The basic function is `look_for`, which takes as input the *k*-itemsets, and taking advantage of the forest-like structure, it efficiently search the transactions' space to find if they occur in the current transaction *t*.

iv. *Phase-4.* The large candidate itemsets are kept (i.e., the ones that passes the *minsup* criterion).

After the generation of large k-itemsets the *'if-then*' rules are generated just as in the Apriori-like algorithms.

6. Experimentation and Preliminary Results

A full scenario of activating and using the presented data mining operations include the following steps (as a reference, see Figure 1 presented in section 2):

1. The user, via the PCDD, posts a specific *query*. For example he/she may be interested for all patients' encounters (present in the federation) with pre-specified values for: clinical findings; for laboratory results; and for recorded diagnoses values. Furthermore, the user specifies the desired 'minsup' and 'minconf' levels.
2. The PCDD server identifies the *links to the encounters* that relate with the specified clinical findings, laboratory tests, and diagnoses.
3. The respective autonomous clinical information systems are *accessed* (via the CORBA IDL wrappers and interfaces), and the details of patients' encounters that match the pre-specified values are retrieved and recalled.
4. The *DTD/XML generation* operations are activated, and the respective XML query-specific document is generated.
5. The generated XML is *parsed* and, (a) it is *semantically homogenized*, and (b) the respective *forest-like structures* are generated.
6. The *KDD/ARM operations* (as realized by the AprioriXML algorithm) are activated in order to discover and form association rules.

Transactions. In the presented experiment, and for making transparent, and easy to follow the involved operations we rely of ten- (10) selected patients' visits in two remote healthcare centers- HCC of Crete (areas of Spili and Anogia; the respective medical encounters are stores and accessed from the local information systems via

PCDD). The specific exploratory query posted was (in natural language): "*access and retrieve all bio-chemical examinations of patients' visits*".

Items. Table 1 below, shows all the bio-chemical exams manipulated by the respective HCCs' information systems. The numeric-valued observations are passed through a discretisation routine, and all values are grouped into three intervals reflecting the *low*, *normal*, and *high* status of the specific medical lab-examinations' measurements (a natural intervals' arrangement for medicine).

Table 1. *Recorded & Retrieved Biochemical examinations & their discretised code values*

Atomic Observation Type	Min Value	Max Value	Code For Low Values	Code For Normal Values	Code For High Values
GLYCOSE	82.00	282.00	1	2	3
OURIA	17.00	66.00	4	5	6
KREATINI	0.45	1.30	7	8	9
CHOLESTEROL	136.00	355.00	10	11	12
HDL_CHOLESTEROL	35.90	90.50	13	14	15
TRIGLIKERIDIA	33.00	565.00	16	17	18
OURIKO	2.10	6.50	19	20	21
NA	137.00	148.00	22	23	24
ALKALIKI_FOSFATASI	94.00	162.00	25	26	27
SIDIROS	0.00	84.00	28	29	30

Data-Base/ Transactions. Table 2a, below, shows the encoded database (with the 10 selected transactions). In Table 2b, the same database is shown after the 2nd scan, when two (2 and 5) transactions were eliminated, and all others were significantly reduced. The database of transactions (measured with the total number of retained items) is reduced about **55%**. The utility and effectiveness of the '*Elimination of Redundant Information'* operation of the AprioriXML algorithm (phase-2; see above) it may be easily verified.

Table 2a. *Initial Database*

T	Original
1	< 1 , 4, 8, 12, 15, 16>
2	< 3 , 5>
3	< 1 , 5, 8, 10, 13, 18, 19>
4	< 1 , 4, 7, 10, 14, 16, 19, 23>
5	< 3 , 6, 9, 11, 18, 21, 22, 25>
6	< 1 , 4, 8, 10, 13, 16, 21, 24, 27>
7	< 1 , 4, 8, 11, 14, 17,19, 23, 26, 30>
8	< 1 , 4, 8, 10, 15, 16, 19, 23, 25>
9	< 1 , 4, 8, 11, 15, 16, 23, 25>
10	< 2 , 4, 8, 10, 14, 16, 20, 23>

Table 2b. *Database After the 2nd Scan*

After 2nd Scan	Status
< 1 , 4, 8, 16>	*Reduced*
----	*Eliminated*
< 1 , 8, 10, 19>	*Reduced*
< 1 , 4, 10, 16, 19, 23>	*Reduced*
----	*Eliminated*
< 1 , 4, 8, 10, 16>	*Reduced*
< 1 , 4, 8, 19, 23>	*Reduced*
< 1 , 4, 8, 10, 16, 19, 23>	*Reduced*
< 1 , 4, 8, 16, 23>	*Reduced*
< 4 , 8, 10, 16, 23>	*Reduced*

Table 3 below, shows the discovered association rules that meet the following threshold requirements: *minsup* >= ***60%*** and *minconf* >= ***80%;*** four- (4) rules were discovered (with lower thresholds more rules could be discovered).

Table 3. *The discovered Association Rules for the Transactions of Table 2a. (minsup=60%, minconf=80%)*

	Supp. %	**Conf.** %
OURIA[Low] => GLYCOSE[Low]	60	85.7
KREATINI[Normal] => GLYCOSE[Low]	60	85.7
KREATINI[Normal] => OURIA[Low]	60	85.7
TRIGLIKERIDIA[Low] => OURIA[Low]	60	100

7. Conclusion and Future Work

We have presented a methodology, its respective architectural setting and operational framework, for mining structured distributed and heterogeneous data sources.

The realization of the proposed architecture is not an easy work, and a multi-phase data integration/processing approach should be followed. This approach is realized by the coupling of multi-disciplinary technologies ranging from, CORBA based seamless access to distributed data to, semantic data homogenization operations- based on the appropriate utilization of a domain specific ontology, and to advanced DTD/XML operations. These operations, coupled with advanced and effective data representation models forms a framework in which efficient and effective ML/KDD operations could be performed.

The fundamental contribution of our work is the incorporation and customization of KDD/ARM operations on top of appropriately generated DTD/XML documents. In a recent presentation from Rakesh Agrawal (Agrawal, 1999), the need for DTD/XML modeling of databases, and the appropriate customization, and adaptation of ML/KDD operations are also pointed. Based on the argument that, *future databases will use XML-like structures in order to store and retrieve data* then, our work presents a promising architecture and framework towards this direction.

Our future work plans are moving towards four directions: (a) large scale experiments (within the clinical information systems' federation of the integrated electronic healthcare record service of Crete), in order to test the effectiveness of our approach, (b) the design and development of appropriate human computer interfaces, accompanied with user-profiling capabilities for personalized delivery ML/KDD results, (c) customization of other ML/KDD data analysis methods in our framework (e.g., clustering, decision rules etc), and (c) porting of the architecture in other domains, e.g., mining financial and economic information/data sources.

Acknowledgements. The work presented in this paper was (and is) carried out in the context, and with the support of the following EU projects: InterCare (Health Telematics, HC 4011), and IRAIA (IST-1999-10602).

References

1. Agrawal R., Imielinski T., and Swami A. Mining association rules between sets of items in large databases. In SIGMOD, Washington D.C., pp. 207-216, May 1993.
2. Agrawal R., and Srikant R. Fast Algorithms for Mining Association Rules. *Proc. of the 20th Int'l Conference on Very Large Databases*, Santiago, Chile, Sept. 1994. Expanded version available as IBM Research Report RJ9839, June 1994.
3. Agrawal R. Data Mining: Crossing the Chasm. Invited talk at the *5th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD-99)*, San Diego, California, August 1999.

4. Baldonado Wang M.Q., and Cousins S.B. Addressing Heterogeneity in the Networked Information Environment. Technical Report, *Computer Science Department, Stanford University*, December 1996.
5. COAS. *Clinical Observations Access Service (COAS)*. Final Submission, OMG Document: corbamed/99-03-25, 1999.
6. CORBA. Web site, http://www.corba.org.
7. Forslund D., and Kilman D. The Virtual Patient Record: A Key to Distributed Healthcare and Telemedicine. *Los Alamos National Laboratory*. February 29, 1996 http://www.acl.lanl.gov/TeleMed/Papers/virtual.html
8. Gennari J.H., Stein A.R., and Musen M.A. Reuse For Knowledge-Based Systems and CORBA Components. Proceedings of *10th Knowledge Acquisition Workshop*, Banff, Alberta, Canada, 1996.
9. Grimson W., Berry D., Grimson J., Stephens G., Felton E., Given P., and O'Moore R. Federated Healthcare Record Server - the Synapses paradigm. *Web document*, http:// www.cs.tcd.ie/synapses/public/html/ technicaldescription.html, 1997.
10. Haverkamp D., Gauch S. Intelligent Information Agents: Review and Challenges for Distributed Information Sources. *Journal of the American Society for Information Science*, 49:4, pp. 304-311, April 1998
11. Hsu C., Bouziane M., Cheung W., Rattner L., and Yee L. Metadatabase Modeling for Enterprise Information Integration. *Journal of Systems Integration*, 2:1, pp. 5-37, 1992.
12. HYGEIAnet Web site. Integrated Health Care Network of Crete, http://www.hygeianet.gr.
13. InterCare. InterCare End-user Applications, *Deliverable D4.1*, *Health Telematics program, Europe, HC 4011 project*. November 1999.
14. ITU. Recommendation X.500 (11/93) - Information technology - Open Systems Interconnection - The directory: Overview of concepts, models, and services, 1993.
15. Knoblock C.A., Yigal A. An architecture for information retrieval agents. *AAAI Spring Symposium on Software Agents*, Stanford, 1994.
16. Mueller A. Fast Sequential and Parallel Algorithms for Association Rule Mining: A Comparision. Technical report CS-TR-3515, dept. of Computer Science, University of Maryland, Vollege Park, MD, August 1995.
17. OMG group. Web site, http://www.omg.org.
18. Sciore E., Siegel M., and Rosenthal A. Using Semantic Values to Facilitate Interoperability Among Heterogeneous Information Systems. *ACM Transactions on Database Systems*, Vol. 19, No. 2, pp. 254-290, June 1994.
19. Shuh B. Directories and X.500: An Introduction. *Network Notes #45*, ISSN 1201-4338, Information Technology Services, National Library of Canada, http://www.nlc-bnc.ca/pubs/netnotes/notes45.htm, March 1997.
20. 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), pp. 257-277, 1997.
21. UMLS. *UMLS 2000 Documentation*. Web document, http://www.nlm.nih.gov/ research/umls/UMLSDOC.HTML
22. Wah B.W., Huang T.S., Joshi A.K., Moldovan D., Aloimonos J., Bajcy R.K., Ballard D., DeGroot D., DeJong K., Dyer C.R., Fahlman E., Grishman R., Hirschman L., Korf R.E., Levinson S.E., Miranker D.P., Morgan N.H., Nirenburg S., Poggio T., Riseman E.M., Stanfill C., Stolfo S.J., Tanimoto S.L., and Weems C. Report on Workshop on High Performance Computing and Communication for Grand Challenge Applications: Computer Vision, Speech and Natural language Processing, and Artificial Intelligence. *IEEE Transactions on Knowledge and Data Engineering*, 5:1, pp. 138-154, February 1993.
23. Wiederhold G., Mediators in the architecture of future information systems. *IEEE Computer*, 25:3, 1992.
24. XML Web site. W3C main XML document. http://www.w3.org/XML/