CHAPTER 6

ONTOLOGY BASED ASSOCIATION RULE MINING
ALGORITHMS FOR INTELLIGENT DECISION MAKING

6.1 INTRODUCTION

Data mining has become an international leading-edge research field. Web technology has changed the way of people’s publish, access and use of information dramatically since its emergence in 1990s. Especially in recent years, the emergence of a new generation of Web environment based on XML [55], compatible with the existing Web applications well, and can better to achieve information sharing and exchange in Web. The convenience and semi-structured characteristics of its text-based make XML widely applied in many fields like, in information management, electronic commerce, personalized publishing, mobile communication, online education, exchange of electronic documents and is still expanding its range of applications. XML has become the defacto standard of Data representation and exchange on the Internet [108]. For these adequate data using XML document format to store, exchange and performance, there is a growing need for further in depth knowledge except the existing information extraction, Web search and information processing methods. This means that data mining is necessary to be done.

The data available on Web is so much unstructured, heterogeneous and time variant that it is only interpretable by human mind and the machine support for this purpose is very limited.

The semantic web addresses the World Wide Web issues by trying to make the machine understandable. Search engines are quite powerful but still often return excessively large or inadequate lists. Machine understandable information can summit the search engine to the relevant pages and can improve both precision and recall.
The semantic Web is a step towards Web Information. It is based on languages that make more of the semantic content of the webpage available in machine-readable formats for agent based computing. One of the components of semantic Web techniques is to use ontology for marking up Web resources and assisting the generation and processing of semantic markup.

The healthcare environment is generally categorized into two types i.e. information rich but knowledge poor. Wealth data are available within the health care system. Effective analysis tool is used to discover the hidden relationships and tends in data. Health care organization, medical centers and hospital are facing major challenges in the provision of quality services at reasonable costs. Diagnosing the patients’ disease is the quality service that implies the effective administering treatments. With respect to healthcare organizations, there is a wealth of hidden information available that is largely untapped. The maximum percentages of data on the web are so much unstructured, heterogeneous, distributed and time options which are understood by humans. However the amount of data is gigantic that it can only be processed efficiently by machines. Therefore the great success of the current form of WWW leads to new challenges.

**Data Interpretation using Machines**

A huge amount of data available online is interpretable by humans only. Machine support in terms of interpretation of data available online is very much limited.

**Intelligent Data Processing**

Since the size of the data in web is enormous it can only be processed by machine efficiently. However efficiency does not guarantee intelligence. The machine
support in terms of processing the data intelligently for determining useful patterns and knowledge is limited.

**Personalization of Information**

It is very much likely that people differ in the contents and presentations they prefer while interacting with the Web. More precisely, Web organizers must know what the patients do and want. Furthermore, it can be categorized in keeping track of individual preferences and facilitating recommendations as well.

This chapter suggests a novel ontology and association rule based algorithm. This algorithm discusses the domain ontology and hash technology procedure of emerging frequent item set and generation of the association rules. Finally the algorithm transforms the operation of the database.

**6.2 ONTOLOGY BASED WEB MINING**

Internet has become an indispensable tool for everyone, Web usage mining correspondingly becomes a hotspot, which uses huge amounts of data in the web server and other relevant datasets for mining analysis and gains valuable knowledge model about usage of relevant Web site. Nagi et al describe a work utilizes association rule mining integrated with fuzziness factor in order to analyze weblog data.

The target is to find pages that are accessed together by majority of the users and hence should be linked in a proper way in order to maximize user satisfaction by providing to the users access flow. This way the number of visitors to the analyzed website will be maximized and hence the target will be achieved. Zhu et. al. [184] propose a new vector space retrieval algorithm based on association diagram extension of key words. By using key words and the related words appearing
simultaneously in a large scale, the algorithm allows generating the association diagram which indicates the simultaneous relationship between key words. The degree of association between any two key words is represented by mutual information.

In addition, the algorithm derives the weight of key words in retrieval vector via association diagram, thus the vector space retrieval based on association diagram extension of key words is realized. Martinez-de-Pison et. al. [115] propose an experience based on the use of association rules from multiple time series captured from industrial processes. The main goal is to seek useful knowledge for explaining failures in these processes. An overall method is developed to obtain association rules that represent the repeated relationships between predefined episodes in multiple time series, using a time window and a time lag.

First, the process involves working in an iterative and interactive manner with several preprocessing and segmentation algorithms for each kind of time series in order to obtain significant events. In the next step, a search is made for sequences of events called episodes that are repeated among the various time series according to a pre-set consequent, a pre-established time window and a time lag. Extraction is then made of the association rules for those episodes that appear many times and have a high rate of hits. Ho et. al. [81] propose artificial intelligence methodology which provides investors with the ability to learn the association among different parameters. After the associations are extracted, investors can apply the rules in their decision support system.

The movement of Hang Seng Index, which is associated with other economic indices including the gross domestic product (GDP) index, the consumer price index (CPI), the interest rate and the export value of goods from Hong Kong, is learnt by the proposed method. With the rapid growth of computer and Internet technologies, e-learning has become a major trend in the computer assisted teaching and learning fields. Most past researches web-based learning focused on the issues of adaptive presentation, adaptive navigation support, curriculum sequencing and intelligent
analysis of student's solutions. These systems commonly neglect to consider whether a learner can understand the learning courseware and generate misconception or not.

To neglect learner's learning misconception will lead to obviously reducing learning performance, thus generating learning difficult. In order to discover common learning misconceptions of learners, Chen et al. [39] study employs the association rule to mine the learner profile for diagnosing learners' common learning misconceptions during learning processes. The association rules that occurring misconception ‘A’ implies occurring misconception ‘B’ can be discovered utilizing the proposed association rule learning diagnosis approach. Meanwhile, this study applies the discovered association rules of the common learning misconceptions to tune courseware structure through modifying the difficulty parameters of courseware in the courseware database so that learning pathway is appropriately tuned. Besides, they also present a remedy learning approach based on the discovered common learning misconceptions to promote learning performance.

At present, a lot of work has to be done in the positive association rules in web mining but negative association rules is more important, Yang et al. [179] have applied negative association rules technology to Web usage mining in the course of the experiment they have proved that negative association rules have a more important role on access pattern to web visitors, give the mining algorithms, to solve the deficiencies in related to positive association rules.

A variety of the web-mining techniques are now being extensively utilized to extract useful knowledge about customer behaviors on the Internet. However, the naive interpretation of the web-mining results would lead to poor decision on customer behaviors, which is likely to cause waste of time and efforts on managing electronic commerce strategy. To overcome this pitfall, Lee et. al. [102] propose using the cognitive map-based interpretation of the web mining results. Conventional approach to obtain the web mining results is based on the association rule approach
(ARA), while the cognitive map approach (CMA) is believed to provide more robust support in interpreting the web mining results.

Therefore, to compare the interpretation capability of the two approaches, the four constructs such as perceived usefulness, causality, information richness, users' attitude. Intention to use the approaches are adopted in the research model and tested against the questionnaire data. To avoid returning irrelevant web pages for search engine results, technologies that match user queries to web pages have been widely developed. Du et al. [49] propose a new study, in which web pages for search engine results are classified as low-adjacence or high-adjacence sets. To match user queries with web pages using formal concept analysis, a concept lattice of the low-adjacency set is defined and the non-redundancy association rules defined by Zaki for the concept lattice are extended. OR and AND Rules between non-query and query keywords are proposed. An algorithm and mining method for these rules are proposed for the concept lattice. The time complexity of the algorithm is polynomial. The amount of ontologies and semantic annotations available on the Web is constantly increasing the type of complex and heterogeneous graph-structured data raises lot of new challenges in the data mining community.

Nebot et al. [130] present a novel method for mining association rules from semantic instance data repositories expressed in RDFs and OWL. They take advantage of the schema-level knowledge encoded in the ontology to derive just the appropriate transactions which will later feed traditional association rules algorithms. This process is guided by the analyst requirements, expressed in the form of a query Pattern experiments performed on real world semantic data that results are much useful.
6.3 CLASSIFICATION OF WEB MINING

Web Mining extracts knowledge from Web data, where at least one of the structure hyperlink or usage (Web log) data is used in the mining process with or without other types of Web data.

Web Mining is a triple WM= (D, P, $). Where D is the set of Web documents and activities; P is the set of patterns and $ is the mapping between D and P.
Web mining classification is illustrated in the following Figure 6.1 with the respective techniques in each category.

Figure 6.1: Classification of Web Mining
6.3.1 ONTOLOGY BASED MINING ASSOCIATION RULES

Ontology is a representation of formal knowledge. It provides a clear and consistent representation of terminology and methods that help people to observe the problems and dealing with affairs, provide public vocabulary of areas and define different levels of formal meanings of terms and relationships between terms. It is organized by taxonomy and includes the typical model of the original language of the ontology and can provide a public and consistent understanding of the field. It overcomes the semantic content of the communication mismatch problem. Ontology structure is divided into the following six stages [178]:

- Identify the purpose and scope of the ontology application: establish the field of study: establish the corresponding domain ontology or process ontology.
- Ontology analysis: define the relationship between all terms and ontology meaning.
- Represent ontology: to select a proper method of ontology according to the system need.
- Ontology test: mainly test the clarity, consistency, integrity, scalability of ontology.
- Ontology building: test the ontology according to the above criteria, to meet the requirements, store the file form, otherwise switch to next.
- Ontology representation: In order to describe and represent ontology, in recent years appeared a variety of ontology languages.

TOWL [122] is designed by the World Wide Web Consortium Web Ontology Working Group. Its syntax is very similar to DAML and OIL, and can easily be converted to the latter. OWL can be used to clearly express the meaning of the vocabulary entries as well as the relationship between these entries. This express of vocabulary entries and relationship between them is called Ontology. OWL has more mechanism to express semantics than XML, RDF and RDF Schema, so it exceed capability of XML, RDF and RDF Schema that are only expressed machine readable document content online.
6.3.2 OPTIMIZATION AND TRANSPLANTATION OF APRIORI ALGORITHM

The proposed mining algorithm is obtained based on the improved Apriori algorithm. The proposed migration and optimization program introduce to the field ontology and hash technology and improve the operation of the frequent item sets and association rules of operation generated, and use the hash table to store the relevant domain ontology. This will make the classic Apriori algorithm database operations into operations on the XML tree memory. The advantage is better to play the XML's strengths, can be operated from the relational database. The search ontology generate candidate item sets and frequent item sets that can direct access to memory. It reduces disk I/O operation, greatly improving the efficiency of the algorithm. Of course, this program has some limitations. For example, large memory space, high space complexity, etc.

First of all, use the Apriori algorithm to find the transaction item sets from the records, then find frequent item sets, search for all to meet the minimum support and minimum confidence of the strong association rules and calculate its confidence. Apriori algorithm in this research was taken to improve the efficiency of the rule mining, which introduces the domain ontology. Program implementation is to use Java to apportionment XML data source into the XML tree and then calculate. Store the candidate item sets and frequent item sets are in the local disk or memory. When the candidate set is too large, to save memory, it is stored as a local temporary file, removed frequent item sets when it is end, and no longer occupy the storage space. In-memory XML tree XML_TREE_SUPPORT used to store frequent item sets. As when calculating the confidence level need the corresponding services confidence, build index of XML_TREE_SUPPORT, applicative hash technology on each transaction storage.
6.3.3 APRIORI ONTOLOGY ALGORITHM

Apriori is one of the first mining association rules algorithms. It always has four stages: For a given database $D$, minimum support (min-sup) and the minimum confidence (min-confidence).

The following steps illustrate the concept of Apriori Ontology Algorithm

```
Step 1: Scan the database $D$, find the items collection is greater than or equal to minimum support, set the project $L_1$.

Step 2: Use large set length k-1 of projects to generate the candidate item sets.

Step 3: Calculation of all candidate item set support, to determine whether it is greater than or equal to the minimum support, the candidate will meet the conditions set excavation project, form a frequent set of k length.

Step 4: Repeat the above steps until no new candidate item sets generated then stop.
```

Mining object of traditional Apriori algorithm is transaction-oriented database. The Ontology-based XML mining association rule is an object-oriented XML data source. The XML data source stored in to the hash table during the data pre-processing time to speed up the speed of traversing the entire XML tree. Meanwhile, the algorithm does three improvements on the traditional Apriori algorithm:
First, the XML data source shorthed in to the hash table and to each transaction and each data item is also added to all the parent of this transaction. Then add the parent to candidate item sets as normal data, when traversing XML trees, also count on the parent class.

Second, according to the pre-existence table hierarchy of domain ontology, can get parent of each item sets and remove the parent class which not appear in any concentration of the candidate.

Third, pruning also includes key items of the item x and its ancestor set.

6.3.4 STEPS OF THE ALGORITHM

Resolve document XML to Tree XML and after resolving ontology to Hash table, execute the following steps to find the frequent item sets.
Step 1: Based on the thought of Apriori, creating candidate item set C. Count each item point with the simple periodicity of XML Tree. Setting the min support and usually it is assigned by the user. Assure the frequent the assemble L in item set 1, and any of the frequent count of the program in L should not be less than the minsupport is used to produce candidate item set C.

Step 2: Execute candidate item set 2. L. Each item set in C is produced by a connection of two frequent item sets which belong to L. Search Hashtable and delete the item set which has the term x and its ancestor item and also the parent classes which do not appear in any candidate item set. Scan the transaction data XML Tree and compute each support of item set in C and use a temporary file XML for C.

Step 3: Select the transaction, the frequent count of whose is not less than the minsupport, so as to make sure the Assemble L in frequent item set 2. And at the same time, compute the Hash function corresponding to the transaction and add to the frequent item set XML with the XML_TREE_NODE_ID combined with item 2 using the characteristics of Apriori algorithm and pruning technique to remove all candidate set such that its subsets is not frequent item sets.
By converging the semantic web and web mining techniques, e-Healthcare evolved a consumer-centered model of health care where various stakeholders collaborate utilizing information and communication technologies (ICTs) including Internet technologies to manage and care for individual’s health. Following are the important factors that equip patients with the essential knowledge for managing and maintaining better health.

The concept of internet based Health facilitates changes the attitude of the patient with masses of information available on the Internet. The patients expose themselves to new information and possibly new value systems from individuals who put this information online.

Access to the knowledge base of medicine and medical history records with clinical test information has empowered patient and healthcare provider relationship.
Patient has access to the same knowledge base that is accessed by healthcare provider. This makes as a compulsion for the physician to regularly update the relevant information to offer better health services that gives a consumer a new sense of empowerment.

Internet Health supports the patients to obtain better care, prevent medical mistreatment and save their own lives at critical times. These aspects have basically extended the scope of health care beyond its conventional boundaries making it a tool in the hands of the patient which they can utilize as knowledge for managing health through association rule mining algorithms.
The following Figure 6.2 shows the internet-based healthcare system.

Figure 6.2: Internet-based Healthcare Knowledge Management
The knowledge management empowers health care sector in predicting future health risks and educating patients online.

6.4.1 INTERNET HEALTHCARE FRAMEWORK

Following are some examples of health care cases that can provide vital information using this proposed framework.

• High number of count for White Blood Cell: Outlier analysis.

• Elderly women suffering from diabetes leading to Miscarriage: Association technique.

• Fever Categorization: Classification technique.
The following Figure 6.3 illustrates the idea behind the internet Based Healthcare System.

Figure 6.3: Internet based Healthcare Framework
There are various tools used for building the suitable ontology structure for the proposed Internet based Healthcare system.

It is an open source ontology management and learning infrastructure targeted for business applications. It includes a comprehensive tool suite allowing easy ontology creation supported by machine learning algorithms. Frame Logic provides a visual metaphor for representing the conceptual structure.

Description logic (DL) is one of the most popular frameworks. DLs are subsets of first order logic which aim at being as expressive as possible while still being decidable.

Protege-2000 is a platform-independent environment for creating and editing ontologies and knowledge bases. Like KAON, it has an extensible plug-in structure.

Sesame is architecture for efficient storage and expressive querying of large quantities of RDFs data.
6.4.2 SEMANTIC WEB LAYER ARCHITECTURE

The semantic web layer architecture suggested by Mr. Tim Berner is as shown in the following Figure 6.4.

Figure 6.4: Semantic Web Layers
Uniform Resource Identifiers (URI) provides a standard way to refer to entities

Unicode is a standard for exchange of symbols

Extensible Markup Language (XML) provides syntax for structured document output

XML document can refer to the different namespaces to make explicit the context of different tags.

Resource Descriptor Framework is the first layer where information becomes machine understandable. W3C recommended that Resource Descriptor Framework is a foundation for processing metadata. It provides interoperability between applications that exchange machine understandable information on the Web. Resource Descriptor Framework documents consist of three types of entities: Resources, properties and statements. Resources may be Web pages, parts or collections of Web pages, or any real-world objects which are not directly part of the World Wide Web. In Resource Descriptor Framework, resources are always addressed by URIs. Properties are specific attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource forms an Resource Descriptor Framework statement. A value is a literal, a resource or another statement. Statements can thus be considered as object-attribute value triples. In short, Resource Descriptor Framework represents semantics of the web page contents.

Ontology is an explicit formalization of a shared understanding of a conceptualization. It consists of the following four terms.

**Conceptualization**

Abstract model extracted from real world to represent some application
Explicit

It has explicit definition for concepts, relations, properties, functions and constraints

Formal

It can be understood and executed by machine.

Share

Reflects commonly recognized domain knowledge

It provides a shared and common understanding of a domain that can be communicated between people and heterogeneous and widely spread application systems. It has a similar functionality as database schema with following differences.

- Ontology language is syntactically and semantically richer than common approaches for databases.
- It consists of semi-structured natural language texts and not tabular information.
- Ontology must be shared because it is used for information sharing and exchange.
- Ontology provides a domain theory and not the structure of a data container.
Figure 6.5: Creation of Ontology Process
Following are the guideline for building the Ontology.

• Identification of the aim and the scope of the ontology
• Reuse existing vocabularies
• Enumerating most important term in the ontology
• Defining the classes and their hierarchies
• Defining the properties of the classes
• Defining the features of the properties
• Creating Instances

6.5 A Priori and Semantic Search Experimental Results

The experiments were designed to compare the results obtained by three different search approaches:

**Keyword search**

A conventional keyword-based retrieval model, using the Jakarta Lucene library.

**Ontology-only search**

The ontology-based retrieval model explained without including the final step of ranking combinations.

**Semantic search**

The completed semantic retrieval model explained including the combination of keyword-based and ontology-based retrieval results.
6.5.1 RESULTS

This section reports and discusses the observed results based on the examples selected through the generated queries. It shows different levels of performance for different characteristic cases; where it has intentionally chosen examples where the ontology-based method does not always return the best results. The overall performance over the queries is also shown on average over the whole test set. The metrics are based on a manual ranking of all documents for each query, on a scale from 0 to 5. In the experiments, all the query variables were given a weight of one. The measurements are subjective and limited, yet indicative of the degree of improvement that can be expected, and in what cases, with respect to a keyword-based engine. The results are shown in Figure 6.6 and Figure 6.7.

Query a. “News about banks that trade on NASDAQ, with fiscal net income greater than two billion dollars.” In this example the semantic retrieval algorithm outperforms keyword-based search because the limited expressive power of the latter fails to express all the conditions in the query. Furthermore, the KB contains many instances of banks, some of which match the query, and news about these banks are recognized as relevant by the semantic retrieval algorithm as soon as their name is mentioned in the document, even if the text does not mention trade markets or fiscal incomes. With keyword-based search, only the documents that explicitly contain words like “bank” and “NASDAQ” are ranked highly. These are typical results when a search query involves a region of the ontology with a high degree of completeness in terms of instances and annotations. These cases yield a high precision up to almost maximum recall. However, the KB does not contain all banks, which explains the decrease of precision at 100% recall. If more instances were added, precision would stand at high levels for all the recall values.
Figure 6.6: Ontology-based retrieval Results
Figure 6.7: Ontology-Based Search Vs Keyword-Based Search
Query b. “News about telecom companies.” In this example, the ontology KB has only a few instances of telecom companies, so not all documents relevant to the query are annotated. This causes low precision values for the ontology-based approach, which drops to zero for higher recall. The example shows how the combination of semantic and keyword-based result retain the efficiency of the latter when the former fails. Furthermore, in the areas where semantic retrieval does work (here, at low recall), the combined approach takes advantage of these few good results to perform better than the keyword-based techniques.

The examples described in this section are representative of the typical behavior of the proposed techniques in characteristic cases. Overall, a significant improvement achieved by the new approach that can be observed in the global comparison provided by the histogram and the average precision curve. Although a systematic efficiency testing has not yet been conducted, the average informally observed response time on a standard professional desktop computer is below 30 sec. A main bottleneck in our first implementation was the traversal of annotations to retrieve the document vectors, the cost of which grows linearly with the size of the result sets ($|T_q|$ and $|R_q|$, where $R_q = \{d : \cdot \mid \text{sim}(d, q) > 0\}$). was drastically reduced by storing the annotations in a separate database.

6.5.2 DISCUSSIONS

The added value of semantic search with Apriori based information retrieval with respect to traditional keyword-based retrieval, as envisioned in the proposed approach, relies on the additional explicit information: type, structure, relations, classification and rules, about the concepts referenced in the documents, represented in an ontology-based KB, as opposed to classic flat keyword-based indices. Semantic retrieval introduces an additional step with respect to classic information retrieval models: instead of a simple keyword index lookup, the semantic retrieval system
processes a semantic query against the KB, which returns a set of instances. This can be seen as a form of query expansion, where the set of instances represent a new set of query terms, leading to higher recall values. Further implicit query expansion is achieved by inference rules and exploiting class hierarchies. The rich concept descriptions in the KB provide useful information for disambiguating the meaning of documents. In summary, proposal achieves good improvements with respect to keyword-based search using Apriori concept.

6.6 EXPERIMENTAL RESULTS OF HEALTH CARE RULES

To verify the method validity of the excavation proposed in this chapter using Java implemented to test the algorithm. XML documents using Microsoft’s MSXML parser to quickly resolve in order to generate a document tree. The experimental use of domain ontology purposes that the primary standard for research to use. As currently there are no dedicated mining XML documents, experimental test data used is from in the simulation library data. Data set contains 5000 multiple services, and after transformation, the XML document size is about 11.5 MB. As the proposed algorithm uses hash table to store the XML document tree, to traverse the entire data set in a short time, so there is no need to repeat the scan XML documents, for the mining process saves a lot of overhead.

Experiments show that the method by introducing domain ontology of the XML document frequent generalized data can effectively reduce the size of XML documents, mining the association rules and make the algorithm easier to be understood. Algorithm memory usage is to depend on actual data storage in addition to the introduction of a parent class of each, algorithm to traverse the XML tree, the item count when a corresponding increase in performance.
The following Figure 6.8 shows the diversification between the numbers of the frequent sets which will be mined by the algorithm and the threshold of minimum support.

Figure 6.8: Diversification between the number of the Frequent item sets and minimum support
From the experimental results of the above diagram, the following are the major observations for the searching process. If the minimum support value is less than the number of items for search, the association rules is high. When the minimum support value is high, the rules and frequent items are less. The proportion for this searching process is reverse implication of items and minimum support threshold.

For any health care dataset, the data items of the database played vital role to find the decision making process for the disease prediction and to fix the complications of the patients according to their history data.

6.7 SUMMARY

XML documents of past mining algorithms are specific items of data mining, allowing users a higher level in the excavation, mining results generated layers, help users better decision making. From these results and discussions, improve the operation of emerging frequent item sets associations to store the domain ontology. At last the algorithms transform the operation of the healthcare database into useful results using association rule mining operations and semantic search in order to make intelligent decision making process for web related data.