Secured Ontological Association Mining

GEETHA MARY A., J. MANEESH, NIRANJAN Y, N.CH.S. N. IYENGAR

School of Computing Sciences and Engineering,
VIT University, Vellore, T.N, India
erpgeetha@gmail.com,,maneesh_jnm@yahoo.com,
{niru1988, nchsniyengar48}@gmail.com

&
ABHINAY TIWARI, AKSHAY AGARWAL, GAURAV MODI

School of Information Technology and Engineering,
VIT University, Vellore, T.N, India
{abhinaytiwari87 ,  akshay.agrawal1101,modi.gauravmodi}@gmail.com

ABSTRACT

With the increase of database applications, mining interesting information from huge databases has become a huge concern. Ontology is a scientific tool used to explain about the techniques used in data mining. Suitable Privacy preserving algorithm needs to be used to protect the data that is to be mined. All these technologies when integrated together yield an efficient and user friendly system.

The system thus developed has Ontology, to explain the users about the domain and technical details, Security to preserve the data, and finally association rule mining is done on the datasets. List of Heart patients is taken as dataset and association rule mining is performed on it. The generated association rule reveals major causes for heart diseases and the patterns of the risk factors which majorly led to heart disease. The datasets are converted to binary format, so memory usage is less and obviously leads to high performance.

Key words: Data Mining, Data Dredging, Knowledge Discovery in data bases, Aprori algorithm, Ontology, Privacy Preserving Data Mining, XML data base, Association rule mining, frequent item sets, support and confidence.
1. INTRODUCTION

Kamber refers data mining as extracting or “mining” knowledge from large amounts of data\(^3\). Other terms which refers to the process of data mining are, knowledge mining from data, knowledge extraction, data/pattern analysis, data archeology and data dredging. There are different types of data mining techniques which lead to different kind of knowledge identification. Types of data mining techniques are:

- Classification groups items based on a predefined attribute.
- Association making correlations between items and individuals, deducing rules that define relationships.
- Clustering group’s items based on a previously undefined attribute.
- Prediction forecasts trends and
- Estimation examines trends for clues to deduce another characteristic\(^1\).

Security is the main concern in today’s life, so security need to provided and data need to be preserved.

Wang specifies two types of privacy,

- Output Privacy
- Input Privacy\(^7\)

**Input Privacy**: Data is manipulated so that the mining result is not affected or minimally affected.

**Output Privacy**: Data is altered so that the mining result will preserve certain privacy.

Zhang classified output privacy techniques into,

- Perturbation based approach: adding noise directly to the original data values.
- Aggregation based approach: Data are generalized according to the domain hierarchy.
- Blocking based approach: Sensitive attributes are truncated.
- Projection based approach: Reduces the dimension, but retains the minimum information for creating data mining model\(^8\).
- Swapping based approach: transferring the data between two or more fields.

2. Literature Survey

As per Kamber steps in data mining process:

Step 1: Data cleaning – To remove noise and inconsistent data.

Step 2: Data Integration – Multiple data sources are combined.

Step 3: Data Selection – Data relevant to the analysis task are retrieved from the database.

Step 4: Data Transformation – Data are transformed into forms appropriate for mining for performing summary or aggregation operations.

Step 5: Data Mining – Essential process where intelligent methods are applied to extract data patterns.

Step 6: Pattern Evaluation – To identify truly interesting patterns representing knowledge based on some interestingness measures.

Step 7: Knowledge Presentation – Visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Zheng describes the phases for building the typical application for Data Mining and they are:

- Identifying the need of Data Mining in some specific areas.
- Communication with domain experts and do request analysis.
- System design, which includes data preparation.
- System implementation, which includes model training and building.
- System evaluation.

Issues addressed by Zheng are:

- The common divergence between users’ understanding of Data Mining techniques.
- Choice of the core Data Mining algorithm.
- Design guidelines for Data Mining system architecture.
- Methods to incorporate domain experts’ knowledge.

Khaing proposed a framework in which the data are got from XML source since it is interoperable and also XML allows a variety of data to be represented. Thus XML specifically defines the data formats of each and every data, so it behaves as an efficient database. Association rule is applied and frequent item sets are generated and knowledge is gained.

Sridharan has developed ontology for web based learning. The flow followed contains the steps:

- Knowledge creation.
- Knowledge extraction.
- Knowledge classification.
- Knowledge retrieval.
- Knowledge sharing and use.
According to Luigi, in the domain of wastewater treatment, many different experts work together. Their disciplines of expertise are: artificial intelligence, chemical engineering, chemistry, computer science, environmental engineering, microbiology.

- Each expert uses a particular vocabulary (a precise common terminology does not exist).
- There are no rules helping in the use of each term.
- Synonyms exist.
- Some term can be used in different disciplines with similar, but not identical, meanings (semantic differences appear using the same term in different disciplines)\(^5\).

All these reasons suggest the need for the creation of a unified, complete and consistent terminology, which can be used in different formal contexts and applications related to the wastewater treatment domain. Ontologies are a practical way to achieve this goal. So these domain concepts could be used along with data mining to get a clear picture of the application.

Pawel\(^1^3\) has done an Clustering analysis based on the ontology. Ontology is used to describe all the compared objects. Three dimensions were taken into consideration to calculate the similarity between individual objects and they are taxonomy similarity, relationship similarity and attribute similarity. Similarity was calculated through a formula using these dimensions. Using protégé tool, owl file is generated to describe the objects. The similarity measures were calculated and a matrix is produced. This matrix is further processed using external tools.

In\(^1^2\) Ling liu discusses about the geometric data perturbation. She has discussed about the different methods of implementing the perturbation protocols. Three protocols, simple, negotiation and space adaptation are evaluated. The Risk breaches for these 3 protocols were done.

Wang explores the data generalization concept from data mining as a way to hide detailed information, rather than discover trends and patterns. Once the data is masked, standard data mining techniques can be applied without modification\(^1^0\). This method used in this paper not only hides the sensitive data but also provides a appropriate mining result.

J afari reveals that certain specific sensitive association rules are hidden by decreasing its support or confidence than the pre-defined minimum support and minimum confidence. To decrease the confidence of a rule \(X \Rightarrow Y\), either increases the support of \(X\), i.e., the left hand side of the rule, but not support of \(X \cup Y\), or decrease the support of the item set \(X \cup Y\). For the second case, if we only decrease the support of \(Y\), the right hand side of the rule, it would reduce the confidence.
faster than simply reducing the support of $X \cup Y$.

3. Problem Description

Different Data Mining techniques results in different form of knowledge. Depending upon the type of knowledge to acquire, specific data mining algorithm need to be applied. Many terms could point out to same meaning, so ontology could be used in such cases. Not all users are aware of the technical and domain details. Ontology is used to explain about the technical and domain details. Specific privacy techniques need to be applied for a particular data mining algorithm.

![Diagram](image)

Figure 1. Secured Ontological Association Mining: High Level

4. Architecture of the Proposed Work

XML FILE

Nowadays XML files are used as a database due to their interoperability nature. And also XML allows a variety of data to be represented. It also specifically defines the type of data and also format of data. Thus XML specifically defines the data formats of each and every data, so it behaves as an efficient database. Mining is usually applied on database since it acts as basic storage of data, but nowadays XML is also used as basic medium of storage. Mining of XML datasets also became necessary to find some useful pattern and to discover knowledge from these patterns. The fast growing nature of XML based documents on web made to discover new techniques to get patterns from the XML documents.

PARSER

The xml file acts as the database. Nowadays XML is also used as basic medium of storage. Mining of XML datasets also became necessary to find some useful pattern and to discover knowledge from these patterns. The fast growing nature of XML based documents on web made to discover new techniques to get patterns from the XML documents. The xml file is browsed and parsed and needed details are obtained as datasets.

ONTOLEGY

Ontology is provided on all the details, i.e. both technical and domain details used in the project. As per the ontology rules, the concepts are divided into super class and sub class composition. The Domain class has types and causes as subclasses. Types explain the different types of heart pain present and causes explains the
main icons which results or indirectly leads to heart diseases. The properties present in each class explain the terms used in detail. The technical details include two classes, one is on Association rule and other on Security features provided. The sub-classes of association rule are frequent itemset generation and generation of the association rule. The subclasses of security are authentication, role based access and privacy of data. Properties, called as slots can be inserted to classes, which explains or adds information for the class.

SECURITY TECHNIQUE

The records fetched from the XML file are passed as data sets. Three types of security is provided,

- Authentication: Only authorized users are allowed to access the software
- Role Based Access: certain users are restricted to access certain parts of the software.
- Privacy of Data: Swapping technique is used among the hiding technique.
The data of two or more fields are swapped and stored in XML file, so when the XML file is taken by any other competitor then the mining results leads to wrong judgments.

**DATA MINING TECHNIQUE**

The datasets obtained from xml file is converted to binary format. i.e. if the necessary item is present then the item is represented as ‘1’ and if the item is absent then it is represented as ‘0’

Association rule is implemented by two main steps:

- **Generating the frequent itemsets** - The item sets which are frequent is found out and minimum support is used as a threshold in finding it. 1 frequent, 2 frequent and until n frequent item sets are generated until no item qualifies for the minimum support. Minimum support is specified by user at each operation of mining.

- **Generating the association rule** - the frequent item sets are used and confidence is used as threshold in generating the association rule.

Minimum Confidence is mentioned by user at each operation of the mining task.

Confidence is calculated for each frequent itemsets.

Confidence : The confidence (\( \alpha \)) for an association rule \( X \rightarrow Y \) is the ratio of the number of transactions that contain \( X \cup Y \) to the number of transactions that contain \( X \).

Expression:1 Confidence

\[
\text{Confidence (} X \rightarrow Y, t) = \frac{|\forall (x \cup y), t|}{|\forall (x)|}
\]

The confidences which passes the minimum confidence level, generates
association rule. The confidences are interpreted and displayed as Association rule.

**KNOWLEDGE**

The pattern i.e. the Association rules generated are the knowledge inferred from the XML data file. These patterns are the knowledge. This knowledge is displayed in GUI format for the user.

5. Implementation

Medical field is taken as the domain of the study and Heart diseases are taken in particular. Data are interpreted and analyzed with help of a doctor and the requirements are also gathered.

The chest pain type such as typical angina, atypical angina, non anginal chest pain and asymptotic chest pain are taken as cases. The major risk factors which lead to these cases are diabetes, habit to smoke and high cholesterol. These risk factors are analyzed and association is formed so as to gain knowledge that how these risk factors together and also by inter linkage lead to chest pain.

5.1 SELECTION OF THE RISK FACTORS

**Smoke**

Toxins in Cigarettes, gets along with blood and becomes a factor for occurrence of atherosclerosis. Atherosclerosis is constant hardening of arteries which is caused by deposition of fatty plaques. Artery wall inflammation and increase in blood clots, hinders blood flow and initiates stokes or heart attacks.

**Diabetes**

Even though a person has a full control on glucose levels, there is a threat of heart disease; threats are even higher for uncontrolled blood sugar. Diabetes for a long time leads to rise in sediment of fatty substance on walls of blood vessels. These sediments affect the blood flow and blood vessels will get hardened.

**Cholesterol**

Cholesterol is a lipid and lipids are fats. The low density lipoproteins (LDL) are responsible for clogging up and blocking arteries, resulting in hardening of the arteries (atherosclerosis).

5.2 RESULTS AND DISCUSSIONS

Ontology is used to describe about the domain, mining technique and different kind of security provided. Datasets from XML are taken and required data are parsed and converted to binary format. Apriori algorithm is applied and frequent item sets are generated. The minimum support is got from the user and the association rules are generated accordingly.
5.3 GENERATING ITEMSETS

Table 1. Attributes with support

<table>
<thead>
<tr>
<th>Attributes</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>17</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>81</td>
</tr>
<tr>
<td>Smoke</td>
<td>65</td>
</tr>
<tr>
<td>History</td>
<td>62</td>
</tr>
</tbody>
</table>

Support is calculated by,

$$\text{Support} (x \Rightarrow y,t) = \frac{|(x \land y,t)|}{|t|}$$

i.e. support of items $x$ with $y$ in the transaction $t$ is, ratio of number of records that contain $x$ and $y$ in the transaction $t$ to the total number of records in transaction. In simple, support of an itemset is percentage of transaction that contains the itemset. We have taken 121 records to explain the calculations,
Table 2. Calculation of support

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Count of yes</th>
<th>Count of no</th>
<th>Total count</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>21</td>
<td>100</td>
<td>121</td>
<td>( \frac{21}{121} \times 100 = 17 )</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>98</td>
<td>23</td>
<td>121</td>
<td>( \frac{98}{121} \times 100 = 81 )</td>
</tr>
<tr>
<td>Smoke</td>
<td>79</td>
<td>42</td>
<td>121</td>
<td>( \frac{79}{121} \times 100 = 65 )</td>
</tr>
<tr>
<td>History</td>
<td>75</td>
<td>46</td>
<td>121</td>
<td>( \frac{75}{121} \times 100 = 62 )</td>
</tr>
</tbody>
</table>

Let the minimum support be 30, so diabetes would not be present in 1-frequent itemset.

Table 3. One-frequent itemset

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol</td>
<td>81</td>
</tr>
<tr>
<td>Smoke</td>
<td>65</td>
</tr>
<tr>
<td>History</td>
<td>62</td>
</tr>
</tbody>
</table>

For 2-itemset generation, ‘AND’ operation is done among the itemsets. So for 2-itemset, the attributes are,

- Cholesterol and smoke
- Cholesterol and history
- Smoke and history

Sample calculation for 2-itemset of Cholesterol and smoke:

Table 4. Sample representation of calculation for 2-frequent itemset

<table>
<thead>
<tr>
<th>Cholesterol</th>
<th>Smoke</th>
<th>Cholesterol and smoke</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5. Sample calculation of support for 2-frequent itemset

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Count of yes</th>
<th>Count of no</th>
<th>Count of transaction</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesteral and smoke</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>( (3/5) \times 100 = 60 )</td>
</tr>
</tbody>
</table>

2-frequent itemset of the XML file:

Table 6. Two-frequent itemset

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Count of ‘1’</th>
<th>Count of ‘0’</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol and Smoke</td>
<td>62</td>
<td>59</td>
<td>51</td>
</tr>
<tr>
<td>Cholesterol and History</td>
<td>61</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>History and Smoke</td>
<td>46</td>
<td>75</td>
<td>38</td>
</tr>
</tbody>
</table>

3-frequent itemset of the XML file:

Table 7. Three-frequent itemset

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Count of ‘1’</th>
<th>Count of ‘0’</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol, Smoke and History</td>
<td>37</td>
<td>84</td>
<td>31</td>
</tr>
</tbody>
</table>

GENERATION OF ASSOCIATION RULE

The frequent itemsets which passes the minimum support value is taken and confidence for the frequent itemsets were found.

Expression: 3 Confidence

\[
\text{Confidence (x} \Rightarrow \text{y, t)} = \frac{|\forall(x \lor y), t|}{|\forall(x)|}
\]
Confidence is, ratio of number of transactions that contain \( X \cup Y \) to the number that contain \( X \).

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Numerator</th>
<th>Denominator</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke =&gt; Cholesterol</td>
<td>S U C = 62</td>
<td>C = 98</td>
<td>63</td>
</tr>
<tr>
<td>=&gt; Smoke</td>
<td>S U H = 62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>History =&gt;</td>
<td>S U H = 46</td>
<td>H = 75</td>
<td>61</td>
</tr>
<tr>
<td>Smoke</td>
<td>C U H = 46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>=&gt; History</td>
<td>C U H = 61</td>
<td>H = 75</td>
<td>81</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>S U C = 62</td>
<td></td>
<td>59</td>
</tr>
<tr>
<td>=&gt; Heart diseases</td>
<td>U H = 37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heart diseases =&gt;</td>
<td>S U C = 46</td>
<td>U H = 60</td>
<td>80</td>
</tr>
<tr>
<td>History =&gt;</td>
<td>U H = 37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heart diseases</td>
<td>S U C = 61</td>
<td>C U H = 60</td>
<td>60</td>
</tr>
<tr>
<td>=&gt; Heart diseases</td>
<td>U H =</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To generate association rule, minimum support value need to be specified. Let minimum value be 70. Then the association rule generated is

<table>
<thead>
<tr>
<th>Association Rule</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke =&gt; Cholesterol</td>
<td>78</td>
</tr>
<tr>
<td>History =&gt; Cholesterol</td>
<td>81</td>
</tr>
<tr>
<td>Smoke, History =&gt; Cholesterol</td>
<td>80</td>
</tr>
</tbody>
</table>

The result is interpreted as,

- 78% of heart pain patients who smokes also have cholesterol.
- 81% of heart pain patients, who has history of heart diseases, also have Cholesterol.
- 80% of heart pain patients who smokes and have history of heart diseases, also have cholesterol.

6. CONCLUSION AND FUTURE WORK

We have proposed a framework which helps in better mining of the items. Since the data items are converted to binary format, space occupancy is less, faster access of data items and also helps in securing the data items. Ontology is used for better understanding of the data obtained. Security is provided by authentication; role based access and privatized data. Privacy of the data items is done by swapping, a hiding technique. Thus the framework is user friendly and intelligent. Thus efficient technologies are integrated together to form a high performance, intelligent and user friendly system. My future works is concerned with implementing more privacy preserving techniques and allow an intelligent miner in choosing the most efficient mining technique.

7. REFERENCES

2. Fan W., Zheng Qin, Xiao-Ling Jia, “Data Mining Application Issues


