Intelligence Ingrained Data Mining Engine Architecture

S. P. Deshpande*1, Dr. V. M. Thakare*2
*1 P.G. Department of Computer Science and Technology, D.C.P.E., Amravati (MS), India.
1 Shrinivasdeshpande68@gmail.com
*2 Department of Computer Science, S.G.B. Amravati University, Amravati (MS), India.
2 vilthakare@yahoo.co.in

Abstract: Information plays vital role in every field. To take complete advantage of data; it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. ‘Data mining’ is a tool to accomplish the above mentioned needs. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. The domain experts play important role in the different stages of data mining. The decisions at different stages are influenced by the factors like domain and data details, aim of the data mining, and the context parameters. This paper present the architecture of a Data mining engine that work for specific problem domain. The engine is developed with some prerequisite domain knowledge therefore a non expert can use the system effectively for data mining.

Key Words: Data Mining, Engine Architecture, problem specific data mining engine.

I. Introduction

Data mining is the process of extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [1,2,3,4]. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining is one of the tasks in the process of knowledge discovery from the data. The data stored in the database is used to discover the patterns of data, which then interpreted by applying the domain knowledge. The data mining applications can be generic or domain specific. The generic application is required to be an intelligent system that by its own can takes certain decisions like: selection of data, selection of data mining method, presentation and interpretation of the result. Some generic data mining applications cannot take its own these decisions but guide users for selection of data, selection of data mining method and for the interpretation of the results.

The domain specific applications are focused to use the domain specific data and data mining algorithm that targeted for specific objective. The data generating sources generate different type of data. Data can be from a simple text, numbers to more complex audio-video data. To mine the patterns and thus knowledge from this data, different types of data mining algorithms are used. The collection and selection of context specific data and applying the data mining algorithm to generate the context specific knowledge is thus a skillful job. In many domains specific data mining applications the domain experts plays vital role to mine useful knowledge.

II. The Problem Definition

Most of the previous studies on data mining applications in various fields like: Language Science [11,12,11], E-learning[34], Medical science and Health care [12,13,15,18], Education [14,33], Banking [29], Market research [26], E-commerce [31], Engineering [15], Sports Science [24,28,36], Detection of terrorist activities [17,27,30,35] Intrusion detection [21], Quality Control [39], Software Engineering [19,20], Library Science [32], Bio-informatics [22,23] etc. uses the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic
[6,7,8,9,10]. Thus, for every domain the domain expert’s assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge [5]. The domain experts are required to select specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns for knowledge generation.

Most of the domain specific data mining applications show accuracy above 90%. The generic data mining applications are having the limitations. From the study of various data mining applications it is observed that, no application called generic application is 100% generic. The intelligent interfaces and intelligent agents up to some extent make the application generic but have limitations. The domain experts play an important role in the different stages of data mining. The decisions at different stages are influenced by the factors like domain and data details, aim of the data mining, and the context parameters [37]. The domain specific applications are aimed to extract specific knowledge. The domain experts by considering the user’s interest and other context parameters guide the system to select the data, preprocess the data, apply the data mining algorithm and to discover the knowledge. The results yields from the domain specific applications are more accurate and useful. Therefore it is concluded that the domain specific applications are more specific for data mining. From above study it seems very difficult to design and develop a data mining system, which can work dynamically for any domain. Designing a domain specific system is also a challenging task. Data mining engine that work for specific problem domain and can be used by the user do not have much knowledge of the domain is thus the central idea of this study.

III. The Proposed System Architecture

In this research work a new architectural model of a data-mining engine is proposed. As the problem domain is known and well defined, the environment in which the system run is also well defined. The user’s requirements are clearly defined and target processing on the data is also known, in this situation system design become easy. The essential domain knowledge is maintained in the knowledgebase. The system has four major components: The user Interface, The data store, Data mining engine and Knowledgebase.

The problem related data selected from the data repository is preprocessed and then copied in to the data mart. The user interface for selection of specific data mining algorithm enables user in the selection of specific algorithm. The domain knowledge required for actual mining purpose shall be referenced from the domain knowledge component. The discovered knowledge shall be presented to the users through the user interfaces. The generated knowledge is also useful for the future DM applications and the knowledge discovery process therefore it shall be stored in the knowledge base also. The Proposed Data Mining Engine Architecture is as given below:

IV. Implementation of Proposed Architecture

The proposed architecture is implemented for generating players profile from sports related text. In the pilot application ‘Profile Generator’ the system contains four main components: The user interface, Data mining engine, data store and knowledge base. The data store is simple folder contain text of the interest. This text is preprocessed; the interfaces are available to accomplish the preprocessing work. The
preprocessing includes conversion of text in lower case letters, removal of punctuation marks, removal of commonly used words using stop-list of words, and tokenization of the text. The knowledge base contain prerequisite knowledge like: Name of countries, Name of the sports/games, commonly used words (stop words) etc. Rules like: Word presided with, word followed by, if_word_is etc.

Ex.

<table>
<thead>
<tr>
<th>IF_PRESIDED_WITH</th>
<th>IF_TEXT_IS</th>
<th>THEN_VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr or Ms or Miss or Mrs or Master or Shri or Smt or Ku.</td>
<td>Not Name_of_Country Not Commonly_used_Word Not Numeric string Not date string</td>
<td>Name</td>
</tr>
<tr>
<td>Any word</td>
<td>Not Name_of_Country Not Commonly_used_Word Not Numeric string Not date string</td>
<td>May be Name</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IF_TEXT_IS</th>
<th>IF_FOLLOWED_BY</th>
<th>THEN_VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Name&gt;</td>
<td>Of &lt;Country&gt;</td>
<td>Country He/She representing</td>
</tr>
</tbody>
</table>

The rule base for other attributes like: Date of Birth, Country he/she representing, Playing hobby, Favorite food, Favorite Movie, Favorite clothing and Sports Record etc. is created and made available with the system. The knowledge base used with the data mining engine is lace with domain and problem specific knowledge. The training phase also fulfills the knowledge base with relevant knowledge and therefore a user can use data mining tool without domain expertise.

This knowledge base is used for the extraction of values for the above-mentioned variables. The concept of use of knowledge base in the data mining application is new. The domain experts are required to create the initial knowledge base. The derived values for different attributes are also stored in the knowledge base. This table is used to train the system and to use in the next pass of the system.

The Data mining engine is design to apply different data mining algorithm for the

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Attribute</th>
<th>Value retrieve Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Full Name of Player</td>
<td>86 %</td>
</tr>
<tr>
<td>2</td>
<td>Date of Birth</td>
<td>92 %</td>
</tr>
<tr>
<td>3</td>
<td>Game He/She Plays</td>
<td>99%</td>
</tr>
<tr>
<td>4</td>
<td>Country He/She Representing</td>
<td>93%</td>
</tr>
<tr>
<td>5</td>
<td>Other Sports He/She Likes</td>
<td>84%</td>
</tr>
<tr>
<td>6</td>
<td>Favorite Food</td>
<td>82%</td>
</tr>
<tr>
<td>7</td>
<td>Favorite Movie</td>
<td>84%</td>
</tr>
<tr>
<td>8</td>
<td>Favorite Clothing</td>
<td>79%</td>
</tr>
<tr>
<td>9</td>
<td>Sports Record</td>
<td>61%</td>
</tr>
</tbody>
</table>
For the attributes for which success rate is above 90% the algorithm works good and very rare retrieves false value. For the attributes for which success rate is between 80% to 90% the system prompts the user with few options and asks for selection of most correct value. In the extraction of sports record, the text document plays important role, the algorithm requires modification to extract sport record as it has different parameters in different games. This architecture of a data mining engine is using domain and problem relevant knowledge. This knowledge is partially loaded by domain experts and partially generated in the training phase of the algorithm. Due to the use of knowledge base this system can be used by novice users.

VI. Conclusion
The application developed using proposed architecture shown success rate 84.44 % in average therefore it can be comfortably used for data mining. The major advantage of this architecture is the presence of domain specific knowledge along with data. This facilitates the non domain-expert users in data mining. Existing data mining tools are required domain expertise and also training of using the tool. They are limited to extraction of patterns of data of features which need further interpretation by the domain experts to generate knowledge. Using this architecture system can generate the knowledge but developing knowledgebase is the main issue in this context.

References
[7] Ralf Rantzau and Holger Schwarz, A Multi-Tier Architecture for High-Performance Data Mining, A Technical Project Report of ESPRIT project, The consortium of CRITIKAL project, Attar Software Ltd. (UK), Gehe AG (Denmark); Lloyds TSB Group (UK), Parallel Applications Centre, University of Southampton (UK), BWL, University of Stuttgart (Denmark), IPVR, University of Stuttgart (Denmark).
[11] Identification of foreign-accented French using data mining techniques, By Bianca Vieru-Dimulescu, Philippe Boula de...


