Extracting Peculiar Data from Multidatabases Using Agent Mining

S.Shahar Banu, V.Saravanan, R. Shriram

Abstract- Data mining is a broad term that describes the search to extract some meaningful information from data that is unformatted and either unstructured or partially structured Similarly, Fayyad et. al. described it as "The nontrivial process identifying valid, novel, potentially useful, and ultimately understandable patterns in data" . Data mining is also known as knowledge discovery, knowledge extraction, information harvesting, data archeology, and data pattern processing. Although most algorithms provide some unique implementation of each phase, there are several common steps to achieve the goal of identifying patterns in data. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. This paper discusses the peculiar data mining and agent mining. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Keywords: Data mining peculiar mining, agent based system, multi agent .

I. INTRODUCTION

Data mining refers to the process of finding interesting patterns in data that are not explicitly part of the data (Witten & Frank, 2005, p. xxiii). The interesting patterns can be used to tell us something new and to make predictions. The process of data mining is composed of several steps including selecting data to analyze, preparing the data, applying the data mining algorithms, and then interpreting and evaluating the results. Sometimes the term data mining refers to the step in which the data mining algorithms are applied. The first step in data mining is data cleaning, or pre-processing. All input data must meet certain conditions to ensure optimal performance including:

1. The data must be in a usable form.
2. There must be sufficient data to derive a solution.

Different levels of analysis Exist:

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.

Rule induction:

The extraction of useful if-then rules from data based on statistical significance.

Datavisualization:

The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

This research mainly deals with three proposed approaches [as rules] for peculiarity mining. They are

- Exception rules
- Peculiarity rules
- Multi agent system

The underlying objective of this work is to develop effective and efficient data mining technique that can have good accuracy and performance.

II. LITERATURE SURVEY

Ribeiro, Kaufman and Kerschberg,[1995] have described a way of multi-database mining by incorporating primary and foreign keys, as well as developing and processing knowledge segments[1]. Wrobel[1997], has extended the concept of foreign keys to include foreign links, since multi-database mining also involves accessing non-key attributes. Aronis *et al.* introduced a system called WoRLD that uses spreading activation to enable inductive learning from multiple tables in multiple databases spread across the network.

Liu, Lu and Yao [1998],have proposed an alternative multi-database mining technique that selects relevant databases and searches only the set of all relevant databases. Their work has focused on the first step in multi-database mining, which is the identification of databases that are most relevant to an application. A relevance measure was thus proposed to identify relevant databases for mining with an objective to find patterns or regularity within certain attributes. This can overcome the drawbacks that are the result of joining all databases into a single huge database upon which existing data mining techniques or tools are applied. The approach is effective in reducing search costs for a given application.

Revised Manuscript Received on 30 March 2013.

* Correspondence Author
S.Shahar Banu, Assistant Prof (Sel.Gr),B.S.Abdur Rahman University,Chennai-48, India.
V.Saravanan, Professor & Director, Department of Computer Applications, Sri Venkateswara College of Computer Applications and Management, India.
Dr. R. Shriram, Professor, Dept of CSE, B.S.Abdur Rahman University, Chennai-48, India.

Zhong *et al.*[1999] have proposed a way of mining peculiarity patterns from multiple statistical and transaction databases based on previous work. A peculiarity pattern is discovered from the peculiar data by searching the relevance among the peculiar data. A data item is peculiar if it represents a peculiar case described by a relatively small number of objects and is very different from other objects in a data set. Although it looks like an **exception pattern** from the viewpoint of describing a relatively small number of objects, the peculiarity pattern represents a well-known fact with common sense, which is a feature of the general pattern.

Wu and Zhang[2001] have advocated an approach for identifying patterns in multi-database by weighting .Kargupta [2001], have built a collective mining technique for distributed data. Grossman have built a system, known as Papyrus, for distributed data mining. Existing parallel mining techniques can also be used to deal with multi-databases.

A notion related to peculiarity is noise, which is an unavoidable problem in real-world databases. Although noise may appear as peculiar data, one may identify noise based on domain knowledge or Meta knowledge of the database. In this paper, we concentrate on defined peculiar data, which are characterized by attribute values and the distribution of values in a database. The task of differentiating actual peculiar data and noise is left to domain experts.

III. PROPOSED SYSTEM

- ***Exception Rules:***

Association rules are generated from frequent item sets satisfying high confidence constraint. The confidence calculation is a straightforward procedure after all frequent item sets have been generated. We do not consider the confidence calculation as it is easy and conceptually proven correct. The input of the exception rules mining algorithm are frequent 1-itemsets. The output of the algorithm is exceptional item sets. Exceptional item sets will become exception rules after the confidence of association rules has been checked. We generate frequent item sets and on each step k (k is the length of the item set). Exception rules have been previously defined as rules with low interest and high confidence. Interconnection between exception and negative association rules will be considered. Based on the knowledge about negative association rules in the database, the candidate exception rules will be generated. A novel *exceptionality* measure will be proposed to evaluate the candidate exception rules.

- ***Peculiarity Rules:***

Peculiarity rules are a new class of rules which can be discovered by searching relevance among a relatively small number of peculiar data. Peculiarity oriented mining in multiple data sources is different from, and complementary to, existing approaches for discovering new, surprising, and interesting patterns hidden in data. A theoretical framework for peculiarity oriented mining is presented. Within the proposed framework, we give a formal interpretation and comparison of three classes of rules, namely, association rules, exception rules, and peculiarity rules, as well as describe how to mine interesting peculiarity rules in multiple databases. Peculiarity represents a new interpretation of interestingness, an important notion long identified in data mining. Peculiarity,unexpected relationships/rules, may be hidden in a relatively small number of data. Peculiarity rules are a typical regularity hidden in many scientific, statistical, and transaction databases. They may be difficult to find by applying the standard association rule mining method due to the requirement of large support. In contrast, peculiarity oriented mining focuses on some interesting data (peculiar data) in order to find novel and interesting rules (peculiarity rules). The second keyword is multiple databases, which are the objects of discovery and learning. Mainstream KDD (Knowledge Discovery and Data Mining) research is limited to rule discovery in a single universal relation or an information table. Multidatabase mining is to mine knowledge in multiple related information sources.

By considering the two related issues of peculiarity and multiple databases, we propose a framework of peculiarity oriented mining in multi databases. The identification of peculiarity rules, as well as algorithms of mining peculiarity rules, will enhance the effectiveness of data mining and extend its domain of applications. Studies on peculiarity oriented mining can be divided into three phases:

- ***Agent-Based Distributed Data Mining:***

Applications of distributed data mining include credit card fraud detection system, intrusion detection system, health insurance, security related applications, distributed clustering, market segmentation, sensor networks, customer profiling, evaluation of retail promotions, credit risk analysis, etc. These DDM applications can be further enhanced with agents. ADDM takes data mining as a basis foundation and is enhanced with agents; therefore, this novel data mining technique inherits all powerful properties of agents and, as a result, yields desirable characteristics.

In general, constructing an ADDM system concerns three key characteristics: interoperability, dynamic system configuration, and performance aspects, discussed as follows. Interoperability concerns, not only collaboration of agents in the system, but also external interaction which allow new agents to enter the system seamlessly. The architecture of the system must be open and flexible so that it can support the interaction including communication protocol, integration policy, and service directory.

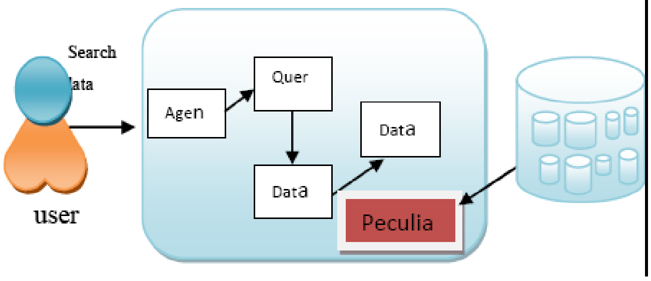


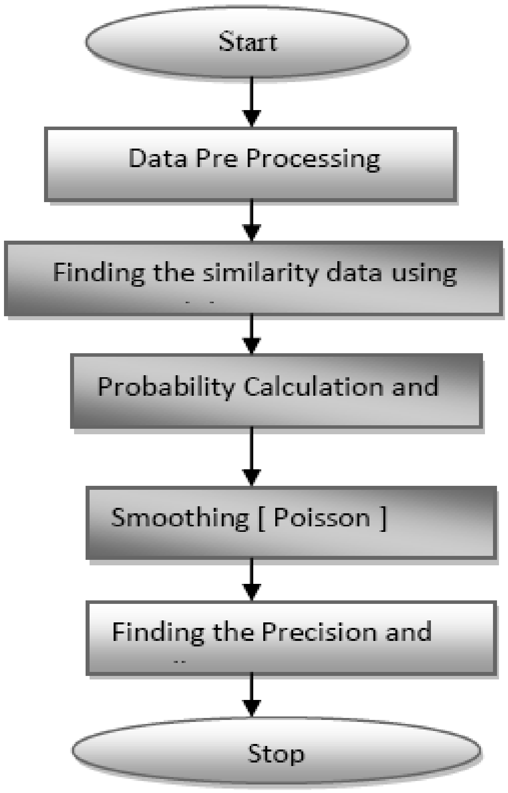
Figure 3.1 posed system architecture

A mining task may involve several agents and data sources, in which agents are configured to equip with an algorithm and deal with given data sets. Change in data affects the mining task as an agent may be still executing the algorithm. Lastly, performance can be either improved or impaired because the distribution of data is a major constraint.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

In distributed environment, tasks can be executed in parallel, in exchange, concurrency issues arise.

Quality of service control in performance of data mining and system perspectives is desired; however it can be derived from both data mining and agents fields. We may generalize activities of the system into request and response, each of which involves a different set of components


Start
Data Pre Processing
Finding the similarity data using
Probability Calculation and
Smoothing [Poisson]
Finding the Precision and
Stop

3.2 Data Mining Agent's Functionality Flow Diagram Finding the Similarity Measurement using Euclidian Distance.

The above figure 3.2 shows the flow of data mining functionality. There are many ways of finding the similar and dissimilar data in a dataset. One of the method is Euclidean Distance method. If we consider the data set D1 and D2, where each data from both set attribute will be noted as $a_1,a_2,a_3...a_n$. and the $b_1,b_2,b_3....b_n$ for first set and the second set respectively. Now the formula ,

$Ed = \text{sum}\sqrt{}(b_i\text{-}a_i)$.

There is no possibility for probability is zero. So we go for smoothing technique. The best technique in smoothing is Poisson. The formula is where the

$P(q \mid D) = P(q) + (1\text{-} (P(qi \mid Di)$,

where 0> <1,

From this formula the score of the probability becomes non-zero and we go for the precision and recall for comparing the IR results from the existing system,

Precision = |{ relevant data's from the Di}|

∩ |{ retrieved data's from the Di }|/

|{ retrieved data's from the Di }|

Recall = |{ relevant data's from the Di}|

∩ |{ retrieved data's from the Di }|

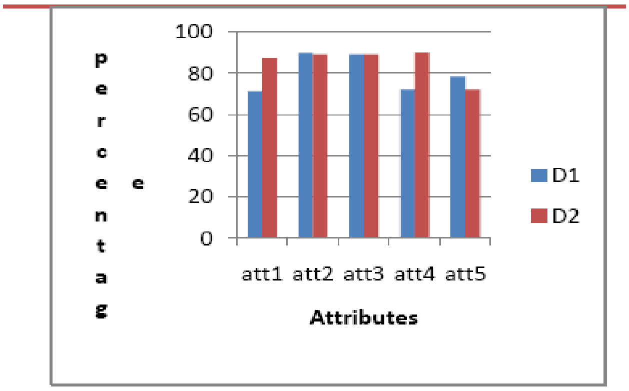
/|{ total relevant data's from the Di }|

The following table shows the sample datasets,

	att1	att2	att3	att4	att5
D1	70.8	89.7	89	72	78
D2	87	89	89	90	72

The Euclidean Distance is 17.90.

From the above table we can conclude that the similarity of the data sets is 82%



IV. CONCLUSION

Multi-agent systems are fundamentally designed for collaborative problem solving in distributed environments. Many of these application environments deal with empirical analysis and mining of data. This paper suggests that traditional centralized data mining techniques may not work well in many distributed environments where data centralization may be difficult because of limited bandwidth, privacy issues and/or the demand on response time.

This paper pointed out that distributed data mining algorithms may offer a better solution since they are designed to work in a distributed environment by paying careful attention to the computing and communication resources.

REFERENCES

1. Ribeiro, K. Kaufman, 1995,knowledge "Discovery From Multiple databases." In: Proceedings of KDD95. 240-245.
2. Wrobel,1997, "An algorithm for multi Relational discovery of subgroups Principle of Data Mining and Knowledge Discovery, 367-375.
3. J.Yao and H. Liu, 1997,"Searching Multiple Databases for Interesting complexes" Proc. of PAKDD, 198-210.
4. H. Lu, and J.Yao, 1998,"Identifying Relevant Databases for Multidata base mining"Proceedings of pacific Asia conf on Knowledge discover and Data mining 210–221.
5. N.Zhong,Y.Yao, and S. Ohsuga 1999 "Peculiarity Oriented mining in multi Database mining "Proceeding PKDD,136-146.
6. H. Kargupta, K.Sivakumar,B.Park and S.Wang, 2000, "Collective Principal Component Analysis from Distributed Heterogeneous Data." Principles of Data Mining and knowledge discovery, 452-457.
7. S.Zhang, 2001, "Knowledge discovery Multi- databases by analyzing Local Instances". PhD Thesis, Deakin University,
8. Kargupta,W. Huang,K. Sivakumar, And E.Johnson, 2001, "Distributed clustering Using collective principalcomponent analysis." Knowledge and Information Systems, 3(4) : 4 22-448.

*Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)*

AUTHOR PROFILE

Mrs.S.Shahar Banu , working as Assist Prof(Sel.Gr) , in Department of Computer Applications at B.S.Abdur Rahman University, Chennai. She is doing research in data mining . She has published papers in National , International conferences and journals.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)