International Journal of Advanced and Innovative Research (2278-844)/

Volume 7 Issue 1

Data Mining Technique for Temporal Association Mining

Vaishali Sahu[#1], Asst. Prof. Anubhav Sharma[#2]

[#]*IES College of Technology, Bhopal, India.*

[1]vaishalisahu44@gmail.com

ABSTRACT: ***Data mining and feature extraction from the multilevel available text is being studying in past. There are techniques which take part of different data resource and process them according to requirement and research analysis. In recent research extraction of temporal information that too in specific medical domain came into significance, where the different research performed in this segment. Medical data and clinical text is rich with various source of data which take part in data processing and further suggested analysis. In existing work paper CRF based technique which is conditional random field's model is used. They achieved best f-measure, accuracy and precision parameters while comparing with other approach such as Baseline, CRF+ Lexical is used. The future work remain by the research is developing of semi-supervised scheme for the temporal extraction and also working with un-annotated data text to make it annotating and thus obtaining better precision, recall, accuracy and F-Measure values.***

Keywords–Rule Mining, Classification, Data Mining Algorithms, K-Theory.

I. INTRODUCTION

Data mining is the process of analyzing data from different aspects and summarizing it into useful information. Data mining software is a tool for analyzing data, in this user can analyze data from many different dimensions, also categorize it, and summarize the relationships identified. Data mining is technically the process of finding interrelations or patterns among dozens of fields in large relational databases. Nowadays, several administrations together with popular hospitals are capable of generating and accumulating a large quantity of data. The accumulation of digital data by governments, companies, and people has made a domain that encourages large-scale data mining and data analysis. This volatile growth of data needs an automatic way to extract helpful knowledge and there is a demand for data sharing among various parties.

Association Rule Mining

Mining of association rules from a database consists of finding all rules that meet the user-specified threshold support and confidence.

Although multimedia databases have become one of the most promising research areas in the database community, discovering association rules in multimedia databases has not received much attention. Many relational and object oriented databases have been using multimedia objects more frequently than the previous year's [Zaiane1998]. These multimedia objects include photo, video, audio, etc. Tremendous use of the global internet has increased the demand of multimedia objects. It is a necessity to extract these data and find associations among them.

Temporal And Spatial Association Rules

Spatial databases contain location information concerning the data being stored. This may be in the form of latitude-longitude pairs, street addresses, zip codes, or other geographic data. While spatial data mining examines the same types of problems as traditional data mining, problem statements and potential solutions may be tailored to the fact that spatial data is involved. For example, spatial operations (within, near, next to, etc.) can be used to describe relationships among tuples in the database. A *spatial association rule*, $X \Rightarrow Y$, is an association rule where both X and Y is sets of predicates, some of which are spatial [Koperski1995]. A spatial association rule holds for a tuple, T, in a database if both predicates, X and Y, are true for T. Definitions for confidence and support are identical to those for regular association rules. Suppose that a database contains information about public schools in a particular county. This database contains data about public facilities (parks, schools, municipal buildings, etc.), geographic features (rivers, lakes, etc.), private buildings, and public infrastructure (roads, bridges, etc). The following is a spatial association rule:

Elementary(T) ∧ Near(T,housing development) ⇒ Adjacent to (T,park)

This rule indicates that an elementary school which is near a housing development is also adjacent to a park. Unlike market basket data, we may need to look at other tuples outside the one being examined, to determine the validity of a spatial association rule. We only need to look at a tuple T to see if it is an elementary school as this will be shown in the value for some attributes. However interpreting the truth of the spatial predicates (near and adjacent to in this case) may require looking at other tuples in the database (or other databases). Thus determining the truth of a spatial predicate may be quite difficult and expensive. One approach to improve the efficiency of mining spatial association rule is a two-step technique where the first step examines approximate satisfaction of spatial predicates by using a coarse interpretation of the spatial relationships [Koperski1995]. This step serves as a filtering process which can drastically improve the second step which examines an exact matching of the predicate. The use of dedicated spatial data structures including R-trees and MBR representations of the spatial features also improves performance.

II. EXISTING BASE PAPER WORK

In the existing system there are following approach is being discussed and perform to execute their work –

1. A temporal rule mining over the data and algorithm is being performed by the proposed algorithm in current paper.
2. They have worked with CRF based approach along with the clinical dataset text such that a processing occupies better performance.
3. They have worked with Annotated dataset which is taken from I2B2 dataset available over the internet resources.
4. Existing another paper also proposed ARTAR approach for temporal data mining and extraction.
5. Association rule mining approach with the data mining dataset approach is being simulated in existing work.
6. T-Priori Approach algorithm is simulated in the approach and comparison is defined by author.

Further Enhancement

A further enhancement needs to be described and performed in the approach with following scenario:

1. An approach to make it semi-supervised and learning is going to perform in our further work.
2. Proposed further enhancement is going to perform with I2B2 dataset as well as an annotation determination approach to work with Un-annotated data is going to perform in enhanced work.
3. A further comparison is going to perform with precision, recall and accuracy parameter.

Related Terms (Brief Introduction)

While performing rule mining and various approaches over the dataset, there are terms which associate while working with the Data mining approach over dataset.

Briefly, the stages are noted as follows:

i. Obtain data from various sources
ii. Data preprocessing
iii. Pattern Discovery
iv. Pattern Analysis

Data Mining Process Diagram Temporal

A. Data Pre-Processing

The data should be pre-processed to improve the efficiency and ease of the mining process. The main task of data pre-processing is to prune noisy and irrelevant data, and to reduce data volume for the pattern discovery phase. Field Extraction and data cleaning algorithms parse the web log records separating the fields and purging. Covering

B. Pattern Discovery

Few techniques to discover patterns from pre-processed data are listed like converting IP addresses to domain names, filtering, dynamic site analysis, cookies, path analysis, association rules, sequential patterns, clustering, decision trees etc.

C. Pattern Analysis

Following statistics are a few listed ones which are the end products of analysis such as the frequency of visits per document, most recent visit per document, who is visiting which documents, frequency of use of each hyperlink, and most recent use of each hyperlink. The common techniques used for pattern analysis are visualization techniques, OLAP techniques, Data & Knowledge Querying, Usability Analysis.

III. LITERATURE REVIEW

A. Jiancheng Ni [1] 2016 approach is based on the concept of outlier detection ARTAR which is rule mining algorithm over the large dataset.

1. A single node experiment which they have compared with T-Apriori and PPT algorithm. Author works on ARM (Association rule mining) approach to solve the rules over dataset.
2. They have stated the advantage of their proposed approach in terms of relationship between the number of cluster nodes and running time of the algorithm.
3. Further they have stated the execution of their algorithm using computation time comparing with existing algorithm.

B. Huiping Peng [2] 2010 stated the use of FP-growth algorithm for processing the web log records, obtaining a set of frequent access patterns, then using the combination of browse interestingness and site topology interestingness of association rules. Sanjay Kumar Malik [3] explains about the web mining that shows there are three general classes of information that can be discovered by web mining: Web activity, from server logs and Web browser activity tracking.

1. Analysis of data mining by using Log analyzer tool, "Log Expert".
2. They focused on the development of Ontology for an intelligent or efficient data and it's relation with data usage mining.
3. Finally, they also summarize some other research challenges towards an intelligent machine and web environment.

C. Hao Yan [4] 2010 proposed a two-step K-means clustering algorithm to search user groups in realistic data collected from WAN.

1. They gave some useful practical conclusions to facilitate design of targeting and recommending applications.
2. K-mean is efficient algorithm which is already proven for the best clustering approach , further more K-mean is extended to K-Medoit and other scheme to make further more efficient clustering from the available data and document from the user.

D. Jiawei Han in [5] 2004 proposed a novel frequent-pattern tree structure.

1. In which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns.
2. They develop an efficient FP-tree based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth .

E. Rakesh Agrawal and Ramakrishan Srikant in [6] 2010

1. They consider the problem of discovering association rules between items in a large database of sales transactions.
2. They present two new algorithms for solving this problem that are fundamentally different from the known algorithms.
3. They also show how the best features of the two proposed algorithms can be combined into a hybrid algorithm, called AprioriHybrid

F. Mohd helmyAbdWahab in [7] 2008 describes the pre-processing techniques on IIS Web Server Logs ranging from the raw log file until before mining process can be performed.

G. C.P Sumathiin [8] 2011 presented an overview of the various steps involved in the preprocessing stage.

H. Renata I vancsy in [9] 2006 investigated three pattern mining approaches from the web usage mining point of view.

1. Also author has done the in-depth analysis of Web Log Data of NASA website to find information about a web site, top errors, potential visitors of the site etc.
2. which help system administrator and Web designer to improve their system by determining occurred systems errors, corrupted and broken links by using web using mining

I. Vaibhav Kant singh in [10] 2008 shows how the different approaches achieve the objective of frequent mining.

1. They also look for hardware approach of cache coherence to improve efficiency of the above process.

IV. PROBLEM FORMULATION

Today the World Wide Web is popular and interactive medium to distribute information. The web is huge, diverse, dynamic and unstructured nature of web data, web data research encountered lot of challenges for web mining. Information user could encounter following challenges when interacting with web.

1. Finding Relevant Information- People either browse or use the search service when they want to find specific information on the web. Today's search tools have problems like low precision which is due to irrelevance of many of the search results. This results in a difficulty in finding the relevant information. Another problem is low recall which is due to inability to index all the information available on the web.
2. Creating new knowledge out of the information available on the web- This problem is basically sub problem of the above problem. Above problem is query triggered process (retrieval oriented) but this problem is data triggered process that presumes that already have collection of web data and extract potentially useful knowledge out of it
2. Personalization of information- When people interact with the web they differ in the contents and presentations they prefer.
3. Learning about Consumers or individual users- This problem is about what the customer do and want. Inside this problem there are sub problem such as customizing the information to the intended consumers or even to personalize it to individual user, problem related to web site design and management and marketing.
4. Most of the algorithm worked only with annotated data thus a proper approach also need to determine for un-annotated data.

V. CONCLUSION

Data mining and its different category of rule mining find its multiple applications in real-time scenario. Mining approach was given in different scenario and extraction over the dataset. In this synopsis our work is executed and discussed to perform the approach with temporal mining outperform on clinical text which deals with the large dataset which is medical dataset can give multiple direction of research. Our work also demonstrates the multiple other work done for proper mining approach. Thus based on current study the further a real-time scenario is required to monitor with CRF approach and further extraction using the Tree based Data annotation with spatial real-time dataset need to be investigated.

Further parameters such as precision, recall, Accuracy and detection rate need to be calculated to outperform the result over existing approach.

REFERENCES

[1]. Jiancheng Ni, "ARTAR: Temporal Association Rule Mining Algorithm Based on Attribute Reduction", IEEE, 2016.

[2]. Huiping Peng "Discovery of Interesting Association Rules Based on Web Usage Mining" 2010 International Conference.

[3]. Sanjay Kumar Malik, Nupur Prakash, S.A.M. Rizvi" Ontology and Web Usage Mining towards an Intelligent Web focusing web logs" 2010 International Conference.

[4]. Hao Yan, Bo Zhang, Yibo Zhang, Fang Liu, Zhenming Lei "Web usage mining based on WAN users' behaviours" 2010 International Conference.

[5]. Han J., Pei J., Yin Y. and Mao R., "Mining frequent patterns without candidate generation: A frequent-pattern tree approach" Data Mining and Knowledge Discovery, 2004.

[6]. Agrawal R, Srikant R., "Fast Algorithms for Mining Association Rules", VLDB. Sep 12-15 1994, Chile, 487-99, ISBN 1-55860-153-8.

[7]. Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd ,Mohamad Mohsin "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm" 2008.

[8]. C.P. Sumathi, r. padmajavalli,"An overview of preprocessing of web log files for web usage mining" 2011.
Gandhimathi Moharasan, Tu Bao Ho, "A Semi-Supervised Approach for Temporal Information Extraction from Clinical Text",2016, IEEE.

[9]. RenataIvancsy, IstvanVajk "Frequent Pattern Mining in Web Log Data"2006.

[10]. Vaibhav Kant Singh, Vijay Shah, Yogendra Kumar Jain, "Proposing an Efficient Method for Frequent Pattern Mining" 2008.