INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES
ISSN: 2322-0821(O) VOLUME 3 ISSUE 4
ISSN: 2394-9910(P) Oct 2015 - Dec 2015

A REVIEW ON CLOSED PATTERN GRAPH MINING USING TWO LAYERED ARCHITECTURE OF BIG DATA

Ankit Kumar Gupta[1], Garbita Gupta[2]

[1]*M. Tech Scholar, Department of Computer Science Engineering, BIST, Bhopal, India, 462021*
[2]*Assistant Professor, Department of Computer Science Engineering, BIST, Bhopal, India, 462021*

Abstract - The aim of this paper is to analysis and decide the value of a neighborhood preference based on the previously access web pages by users. Here two layers maintain authorization and authentication by users. It also maintains neighborhood profile similarity for next prediction of users. The main achievement of this research is to provide decision tree based of Horizontal Partition with Genetics Algorithm. This analysis is important for the prediction of different close value of website. In this paper we combine two techniques of classification and genetic algorithm to increase the efficiency of the website performance. The classification of items in website is used to provide classification among different values of the user profile and then genetic algorithm provide a close among these values.

Index Terms— Closed Pattern, Sequential Pattern Mining, Association Rule Mining, SOM clustering.

I. INTRODUCTION

Data mining is primarily used today by many companies with a strong consumer focus just like retail, financial, communication, and marketing organizations areas. It is used to determine the relationships of internal factors such as product positioning, price, or staff skills in the company. It is also determine external factors just like economic indicators, customer interest and the market competition strategy as in the store company. The entire factors are used to make company profits, increase the sales and customer satisfaction etc. It also shows the summary information to view details of transactional data. A retailer store is per day purchase by customer in database for finding some targeted promotions. It is based on an individual's purchase history by mining demographic data. This analysis help to retailer for generation of products and different promotions offer to specific customer segments.
The enormous amount of data normally stored in files, databases, and other repositories. It is used to extract of interesting knowledge for analysis and interpretation of data that help in decision making. Data mining and knowledge discovery (or KDD) are frequently treated as synonyms in databases. It is actually part of the knowledge discovery process which having some steps in an iterative knowledge discovery process.

A. Neural Networks/Pattern Recognition

Neural network is a set of connected input/output units. Each connection has a weight present with it. It predicts the correct class labels of the input tuples during the learning phase of network learns by adjusting weights. It has the remarkable ability to derive meaning from complicated or imprecise data. It can be used to extract patterns and detect trends that are too complex. These are well suited for continuous valued inputs and outputs e.g. handwritten character reorganization, for training a computer to pronounce English text and many real world business problems. It is identifying the patterns or trends in data which well suited for prediction or forecasting needs.

B. Clustering Analysis

Clustering analysis is unsupervised classification technique which group the items based on similarity basis. It groups the web users according to web page access pattern in website. It also provides personalized web content to the individual user for market segmentation in e-commerce application. Web page clustering is useful for Internet search engines and Web service providers. There are many clustering approaches which are based on the maximizing the similarity between objects in a same class and minimizing the similarity between objects of different classes. The different types of clustering methods are - a) Partitioning Methods, b) Density based methods, c) Hierarchical Agglomerative (divisive) methods, d) Model-based methods, and e) Grid-based methods etc.

C. Genetic Algorithms

GA is a technique which performs like bacteria growing in a petri dish. The data set gives ability to do different things for whether a direction or outcome is favorable. It optimizes the final result which is used mostly for process optimization, such as scheduling, workflow, batching, and process re-engineering.

D. Decision Tree/Rule Induction

Decision trees use real data mining algorithms. It helps with classification and split out information that is very descriptive, helping users to understand their data. A decision tree process will generate the rules followed in a process. For example, a lender at a bank goes through a set of rules when approving a loan. Based on the loan data a bank has, the outcomes of the loans (default or paid), and limits of acceptable levels of default, the decision tree can

set up the guidelines for the lending institution. These decision trees are very similar to the first decision support (or expert) systems.

Web Mining is the extraction of interesting with potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. It automatically discovers and extracts information from website. It categorize into three areas of interest based on which part of the web to mine -

a. Web Content Mining
b. Web Structure Mining
c. Web Usage Mining.

II. OBJECTIVE AND SCOPE

In recent time data mining on big data is very tedious task in current as well as future scenario because huge data cannot be handling in memory with different type. Big data is nothing but just huge data in different format. After five years internet having terra bytes to petta bytes data on server so that not possible to handle into memory. Now Hadoop Distributed Foundation Class (HDFC) handled different data format using Map-Reduce layer by mapping the object.

Each web page is link to each other so its access by user is used to find the behavior in website. We know that we scan database three times, first at initial phase for pruning, second for frequent and third for finding for next item. By using link structure it create efficient prefix graph structure and use variable length bit vectors to present the relationship between the database and its item. If we put node and node count with binary positional code so that we can reduce third database scan.

Big data produce large graph so it can't put into the memory. So graph partition is needed so we used Self Organizing Map Algorithm of Neural Network for clustering of large graph and also used some conditional parameter (eg. Age, Experience, Research Area, Sex etc.) based on registered user profile. Here we used two functionality Similarity and Expert of Profile, we used combination of both functions which support as conditional parameter for partition of graph. So only user related information is gathered in web.

Generally we count of node for pruning but not used importance of node if occur more than one in any transaction. So we used Minimum Support and Maximum Support for every page. Initial we set the min. and max. weight page every page but after that web service set by monthly according to use by user. Previously we used only support and confidence. If support of page is greater than or equal to support then its frequent but now we add additional condition for infrequent items if infrequent items support between minimum and maximum support then also its frequent item otherwise finally its infrequent items. As per importance of item its scope will maintain.

Next problem is to find next item in advance in memory when user browsing the web. Here first search the item in tree and put only partial or sub graph in memory and also remove duplicate pattern. Now we also used closed sequential pattern so that less candidate pattern is generated in FP-tree (Frequent Pattern Tree). Suppose we having five items so 2^5=32 candidate set is generated generally but by using closed sequential pattern only 7 or 8 candidate set is generated. It reduced memory spaces and increases response time. So less memory is used as a scope in future.

III. RELATED WORK

The web mining research is a converging area from several research communities, such as Databases, Information Retrieval, Machine Learning and Natural Language Processing. Due to the widespread computerization and affordable storage facilities, enormous wealth of information is embedded in huge database belonging to different enterprise or scientific experiment. It provides a tremendous interest in the areas of Knowledge Discovery and Data Mining. These areas have motivated allowed statistician and data miners to develop faster analysis tools that can help to analyze the stockpiles of data, turning up in to valuable and often surprising information.

Omar Zaarour et al. (2013) proposed an improvement the web log mining procedure for the prediction of online navigational pattern. Their contribution contains three different components. First they proposed for session identification, a refined time-out based heuristic. Secondly, suggested the practice for navigational pattern detection by using a specific density based algorithm. Finally, a new method for efficient online prediction is also recommended to improve the applicability and effectiveness of the website.

Rahul Moriwal et al. (2013) presented a method for Finding Frequent Sequential Traversal Patterns from Web Logs which is based on Dynamic Weight Constraint, where various frequent sequential pattern mining algorithms have been proposed that mines the set of frequent subsequences pattern which satisfying a min-support constraint in a particular session database. Though, previously sequential pattern mining algorithms gives equal weightage to sequential traversal patterns whereas the pages in sequential patterns have different importance and also have different weightage. Other problem in most of the frequent sequential pattern mining algorithms is that a large number of sequential patterns is generates, when min-support is lowered and here they do not have any alternative ways for adjusting the number of sequential patterns other than increment in the minimum support. The proposed frequent sequential pattern mining algorithm with weights constraint main purpose is to append the weight constraints in to the sequential pattern while maintaining the downward closure property. In this a weight range is defined for maintaining the downward closure property. The pages are given dissimilar weights and traversal sequences assign a minimum and maximum weight. For scanning a session database maximum and minimum weight in the session database is utilized to cut infrequent sequential subsequence and by this downward closure property is maintained.

Chitraa et al. (2012) proposed method, presented, analyzed, and evaluated is to automatically give the actual value of k and select the right initial points based on the datasets objects. The algorithm enhances the k- means clustering algorithm by finding initial points and optimize for accurate results. This algorithm selecting initial points is more complex than the random methods, but this algorithm is stable, running it in different times, the clustering results obtained are the same, the random algorithms cannot ensure this, and different initial points lead to different running time on random algorithms, compared with proposed algorithm, its running time is uncertain and more long.

Nayana Mariya Varghese et al. (2012) are proposed cluster optimization technique using fuzzy logic. Web page access pattern is collected from web log file as input and then eliminate irrelevant data items. The cleaned web log is used for pattern discovery. The web page personalization is used for clustering of web pages. It is based on similar usage of web access patterns by users. Some clustering algorithms have some drawbacks when the number of web user is increased, because the size of cluster also increases. The proposed algorithm is used for eliminating the redundancies occur in data based on fuzzy logic after clustering optimization methodology.

Nanhay Singh et al. (2013) proposed a new framework to improve the performance of web proxy server through cluster (k-means algorithm) based prefetching schemes (LRU and LFU) and Apriori algorithm is applied to generate rules for web pages. Web caching is used to minimize the network traffic at the proxy server level by caching web pages. There is demand to improve the cache performance by using the prefetching technique. It fetches the objects from database and store in advance that are likely to be accessed in the near future. This will result reduction of the response time of the user request.

Ketki Muzumdar et al. (2013) proposed a method to discover useful knowledge by obtaining secondary data from the access pattern of the web users. The proposed method uses Self Organizing Map (SOM), which is a kind of neural network approach to detect user's patterns. It shows the comparison between the traditional K-Means with SOM algorithm. This process describes the transformations necessaries to modify the data storage in the Web Servers Log files to an input of SOM. Neural Network based method has shown that the trend analysis performance depends on the number of requested cluster.

Srishti Taneja et al. (2014) introduced a novel algorithm proposed which uses some features of algorithm of Univariate tree and if noise remains then that can be removed by implementing some features of Multivariate algorithm with some additional features of new algorithm to be designed.

Javeriya Naaz (2015) proposed MinRPset algorithm for locating minimum frequent pattern sets is introduced. It is used for the analysis to get the frequent item sets/ pattern sets. Because some frequent pattern mining often produces a large number of frequent patterns, which imposes a great challenge on visualizing. So that by understanding and further analysis of the generated patterns it emerges the need for finding small number frequent occurring patterns. Here time needed for generating frequent pattern sets plays associate important role. Some algorithms are designed, considering solely the time issue.

Mrs. Neha Chaure (2015) uses a divide-and-conquer strategy to project and partition a large database recursively into a smaller set of patterns. RPglobal is time-consuming and space-consuming. RPlocal is very efficient, but it produces more representative patterns than RPglobal. FP-tree is used to Preserve complete information for frequent pattern mining. For mining frequent closed itemsets, FPclose algorithm is developed. FreeSpan and PrefixSpan are projection based approaches. They reduce the candidate sequence generation. For finding minimum representative pattern sets two algorithms, MinRPset and FlexRPset are developed. Both these algorithms mines frequent patterns first and then find representative patterns. They use CFP-tree to store and retrieve frequent patterns.

REFERENCES

[1] Chun-Chieh Chen, Kuan-Wei Lee, Ming-Syan Chen, "Efficient Large Graph Pattern Mining for Big Data in the Cloud", IEEE International Conference on Big Data, 2013, pp. 531-536.

[2] Ms. N. Preethi, Dr. T. Devi, "New Integrated Case And Relation Based Page Rank Algorithm", ICCCI, Jan. 2013.

[3] Hideaki Ishii, Roberto Tempo, and Er-Wei Bai, "A New Approach for Aggregated PageRank Computation via Distributed Randomized Algorithms", 50th IEEE Conference on Decision and Control and European Control Conference, USA, Dec.2011, pp. 6421-6426.

[4] Yi Pan, HongYan Du, "A Novel Prefix Graph Based Closed Frequent Itemsets Mining Algorithm", The 14th IEEE International Conference on Computational Science and Engineering, 2011, pp. 627-631.

[5] Hideaki Ishii, Roberto Tempo, Er wei Bai, "A Web Aggregation Approach for Distributed Randomized Page Rank Algorithms", IEEE Transactions on Automatic Control, Vol. 57, No. 11, November 2012, pp. 2703-2717.

[6] V. Bonnici, R. Giugno, A. Pulvirenti, D. Shasha, and A. Ferro,"A subgraph isomorphism algorithm and its application to biochemical data," BMC Bioinformatics, vol. 14, no. Suppl 7, p. S13, 2013.

[7] F. Zhu, Q. Qu, D. Lo, X. Yan, J. Han, and P. Yu, "Mining topk large structural patterns in a massive network," Proceedings of the VLDB Endowment, vol. 4, no. 11, 2011.

[8] P. Anchuri, M. J. Zaki, O. Barkol, R. Bergman, Y. Felder, S. Golan, and A. Sityon, "Graph mining for discovering infrastructure patterns in configuration management databases," Knowledge and information systems, vol. 33, no. 3, pp. 491–522, 2012.

[9] U. Kang, D. H. Chau, and C. Faloutsos, "Pegasus: Mining billion-scale graphs in the cloud," in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference, pp. 5341–5344.

[10] Z. Khayyat, K. Awara, A. Alonazi, H. Jamjoom, D. Williams, and P. Kalnis, “Mizan: a system for dynamic load balancing in large-scale graph processing,” in Proceedings of the 8th ACM European Conference on Computer Systems. ACM, 2013, pp. 169–182.
[11] H. Ishii, R. Tempo, E.-W. Bai, and F. Dabbene, “Distributed randomized PageRank computation based on web aggregation,” in Proc. 48th IEEE Conf. Decision Control and Chinese Control Conf., 2009, pp. 3026–3031.
[12] H. Ishii and R. Tempo, “Distributed randomized algorithms for the PageRank computation,” IEEE Trans. Autom. Control, vol. 55, no. 9, pp. 1987–2002, Sep. 2010.
[13] Amit Mittal, Ashutosh Nagar, "Comparative Study of Various Frequent Pattern Mining Algorithms", Internatonal Jornal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol.4, Issue 4, pp. 550-553, April 2015.
[14] Rahul Anil Ghatage, "Frequent Pattern Mining Over Data Stream Using Compact Sliding Window Tree & Sliding Window Model" International Research Journal of Engineering and Technology (IRJET)", Vol.2, Issue-4, pp. 217-223, July 2015.
[15] V. Purushothama Raju, G.P. Saradi Varma, "Mining Closed Sequential Patterns in Large Sequence Databases", International Journal of Database Management Systems (IJDMS), Vol.7, No.1, pp. 29-39, Feb 2015.
[16] Mrs. Neha Chaure, Prof. Dr. S. S. Sane, "A Survey on Finding Representative Pattern Sets in Frequent Pattern Mining", International Journal for Science And Research In Technology (IJSART), Vol 1, Issue 1, pp. 50-53, Jan. 2015.
[17] Prabha S., Shanmugapriya S., Duraiswamy K., "A Survey on Closed Frequent Pattern Mining", International Journal of Computer Applications, Vol. 63, No. 14, pp 48-52, Feb. 2013.
[18] Javeriya Naaz I. Syed, Rajeshri R. Shelke, "Efficient Analysis of Closed Frequent Pattern Set Mining Approach", International Journal of Science and Research (IJSR), Vol. 4, Issue 5, pp. 366-369, May 2015.
[19] Sushila S. Shelke, Suhasini A. Itkar, "A Review on Sequential Pattern Mining Algorithms", International Journal of Electrical, Electronics and Computer Engineering (IJEECE), Vol. 4, Issue 1, pp. 14-19, 2015.