Global Journal of Advanced Engineering Technologies Volume 5, Issue 4- 2016
ISSN (Online): 2277-6370 & ISSN (Print):2394-0921

REVIEW ON PATTERN MINING IN LINKED DATA

[1]Vishakha Manohar Warke , [2]A. S. Vaidya
[1]ME(CSE)Student, Gokhale Education Society R.H.Sapat College of Engineering management studies and Research , Savitribai Phule Pune University,Nashik, Maharashtra, India.
[2]Professor, Gokhale Education Society R.H.Sapat College of Engineering management studies and Research, Savitribai Phule Pune University, Nashik, Maharashtra, India.

Abstract: *As the growth of online linked data, problem was how to analyze the data from linked data. For this problem an important knowledge was obtained the linked patterns among objects which are helpful for characterizing, analyzing, and understanding the linked data. Mining link patterns in large-scale linked data has two main challenges computational complexity of mining algorithms and memory limitations. They introduced a partitioning strategy algorithm mainly used in pattern mining. Partitioning is one of the techniques used by designers to improve the performance of database access. The algorithms like primary partitioning, Bi-partitioning used for analyzing the data in database. In this paper, present a novel approach for the mining linked patterns in pattern mining that means a typed object graph (TOG) was proposed as the graph data model for mining linked patterns [1]. With these algorithms, the various case studies had been analyzed for making subjective and objective studies in pattern mining*

Keywords—partitioning, bi-partitioning, mining-link-patterns, linked-data, frequent links, Patterns, TOG, pattern mining.

I. INTRODUCTION

Frequent pattern mining was one of the fundamental problems in data mining: In that most general form, one was given a database of transactions, each of includes a set of items [2]. A frequent pattern includes a set of items that co-occur in the same transactions more often than a predefined frequency threshold.

Frequent pattern mining was a heart of important problems such as association rule mining.

II. RELATED WORK

In this chapter, a detailed survey of pattern mining approaches and techniques presented in the literature to mine data from large database.

A) Partitioning strategy

In this paper, purpose of study was to the effect of partitioning on query response time using three partitioning strategies are the zero partitioning, list partitioning and range partitioning. There are the three major benefits acquired from partitioning viz. the high performance, manageability and availability. Furthermore operations are backup and recovery can be done more efficiently and effectively with partitioning strategies. The dataset used in this study was extracted from the student information system (SIS) at Yarmouk University (YU)[3]. As compared the results confirmed that partitioning improves query response time over non-partitioning. Norton & McCloskey are the concentrated on the activity was needed to be performed with the 4 kinds of fraction operations: unitizing, partitioning, disembedding and iterating for the effective fraction learning. The flexibility of system is compared with the various partitioning strategies. The best partitioning strategy was not fixed. Therefore, if they select the partitioning strategies by themselves, through that can learn the fraction more positively.

Yin-Fu Huang, Chen-Ju Lai studied the various algorithms for finding an optimal database partition. The main algorithms viz. Apriori algorithm and cosine similarity are used to determine weighted frequent patterns. On the basis of the weighted frequent patterns, they developed two methods for partitioning a database: the candidate method and the optimal method [4]. The optimal method involves using a branch-and-bound algorithm and considering costs in each step of combined the attributes until an optimal solution is reached. Finally, they concluded the experimental results show that the optimal method was the highest performance among all examined methods. And the refined method was considerably more efficient than the original method. Sabeur Aridhi , Laurent d'Orazio ,et.al, the graph mining are now a day's very popular approaches in the various domains. The frequent sub graph discovery task had been highly motivated by the tremendously increased size of existing graph databases [5]. Due to this fact, there was an urgent need of efficient and scaling approaches for the frequent sub graph discovery. In this paper, they study a novel approach for large-scale sub graph mining by means of a density- based partitioning technique used the Map-Reduce framework. This partitioning aims to balance computational load on a collection of machines. They experimentally show that this approach decreases

significantly the execution time and scales up the sub graph discovery process to large graph databases [6].

B) Pattern mining

Faisal Orakzai, Thomas Devogele, Toon Calders, the distributed algorithm for pattern mining can be divided into three stages: partitioning, local convoy pattern mining and merging to produce the global result. The primary step in the development of a distributed mining algorithms was to found an efficient data partitioning strategy with the following properties: Data exchange, Data redundancy, Partitioning costs, Disk seeks, Data ordering[7].This algorithm was computationally expensive and the existing algorithms are not scale up to huge amounts of the movement of data. For that they used partitioning strategies to improve the performance of the system.

Ning Zhong, Senior Member, Yuefeng Li,et.al,Graph mining was a specific kind of frequent pattern mining, the task of enumerating patterns which occur frequently in a dataset. A first class of pattern mining was unstructured mining, such as item set mining, where the pattern was a set of items without any additional structural relation between the different items[8]. Yuqiu Kong, Lijun Wang, Xiuping Liu,et.al, To remove error outputs and preserve accurate predictions, develop a pattern mining based saliency seeds selection method. The given initial saliency maps, this method can effectively recognize discriminative and representative saliency patterns (features), which are robust to the noise in initial maps and more accurately distinguish foreground from background. They studied the saliency labels of saliency seeds to other image regions, an Extended Random Walk (ERW) algorithm was proposed. Compared with prior methods, the proposed ERW regularized by a quadratic Laplacian term ensures the diffusion of seeds information to more distant areas and allows the incorporation of external classifiers. This method was able to improve the performance of existing algorithms and performs favourably against the state-of-the-arts[8].

Several algorithms were introduced to find frequent patterns discover all the combinations of frequent item sets for a given minimum support threshold. But sometimes, it was needed to discovered the frequency of specified few frequent item sets found in the last dataset to check its existence in current dataset to improved the strategy of future business. From the various techniques, Apriori and FP-tree are the most common used techniques for discovering frequent item sets. Apriori finds all significant frequent item sets using candidate generation with minimum support threshold and several number of database scans. FP-tree finds all the significant frequent item sets using specified minimum support threshold with two database scans. This method proposed SIFPMM (Selective Item sets Frequent Pattern Mining Method) finds frequency of selective item sets using Existence Count Table (ECT) with one database scan[9]. Finally they concluded the experimental results of SIFPMM shows that this method outperforms than Apriori and FP-tree.

KU Leuven, Inria Rennes, There are various worked done on the frequent item set mining, they had been considerable effort on mining more structured patterns such as the sequences or graphs. A number of papers had been logically proposed to extend the declarative approached to structured pattern mining problems. In this paper, they introduced a framework that defines the core components of item set, sequence and graph mining tasks, and used to compare existing specialised algorithms to their declarative counterpart [10].

C) Linked Data

Marijn Janssen, Jeroen van den Hoven, they studied the results in new opportunities and had been the potential to transformed government and its interactions with the public [11]. The transparency and privacy would be conceptualized as complex, non-dichotomous constructs interrelated with other factors. Only by conceptualizing

These values in this way, the nature and impact of BOLD (Big open linked data) on privacy and transparency can be understood,and their levels can be balanced with security, safety, openness and other socially-desirable values[12].

Elena Simperl, Maribel Acosta, Marin Dimitrov,et.al, EUCLID had a major contribution to this by providing a comprehensive educational curriculum, supported by multi-modal training materials and state-of-the-art eLearning distribution channels to the real needs of data practitioners[13]. Amrapali Zaveri, Anisa Rula b, Andrea Maurino,et.al, observed widely varying data quality ranging from extensively curate datasets to crowd sourced and extracted data of relatively low quality. Data quality was commonly conceived as fitness for used. In this article, they present the results of a systematic review of approaches for assessing the quality of LOD (linked open data). They gather the existing approaches and compare, group them under a common classification scheme. In particular, they unify and formalize commonly used terminologies across papers related to data quality and provide a comprehensive list of the dimensions and metrics [14].

III. CONCLUSION

In this review paper, had made a survey of different algorithms that can be used for analyzing and understanding the linked data. It can be concluded that the using partitioning strategies they analyzed the mining link patterns in linked data. Using the partitioning strategy results may be the feasible or efficient in mining the link patterns.

IV. ACKNOWLEDGMENT

We wish to express my sincere thanks and the deep sense of gratitude to respected guide Prof. A. S. Vaidya in Computer Department of Gokhale education society's R. H. Sapat College of Engineering and Research Center, Nasik for the technical advice, encouragement and constructive criticism which motivated to strive harder for excellence. We also wish acknowledgment to the people who gives support direct or indirectly to the paper writing.

REFERENCES

[1] Xiang Zhang, Wenyao Cheng, "Pattern Mining in Linked Data by Edge-Labeling", Tsinghua Science And Technology, Vol. 21, no2, April 2016, pp. 168-175

[2] Mahito Sugiyama, Karsten M. Borgwardt ,"Fast and MemoryEfficient Significant Pattern Mining via Permutation Testing", 2015.

[3] Salam H. Matalqa & Suleiman H. Mustafa,"The Effect of Horizontal Database Table Partitioning on Query Performance", Faculty of Information Technology and Computer Sciences Department of Computer Information Systems,2016.

[4] Yeon Joo Lim, Kwangho Lee, "The Analysis of 5th Graders' Partitioning Strategies", Advanced Science and Technology Letters, Vol. 127, 2016.

[5] Yin-Fu Huang, Chen-Ju Lai," Integrating frequent pattern clustering and branch-and-bound approaches for data partitioning", Elsevier, pp 288-301,2015.

[6] Sabeur Aridhi , Laurent dOrazio, "Density-based data partitioning strategy to approximate large-scale subgraph mining", Elsevier, pp 213-223,2013.

[7] Faisal Orakzai, Thomas Devogele, Toon Calders," Towards Distributed Convoy Pattern Mining", Dec 2015.

[8] Ning Zhong, Yuefeng Li, Sheng-Tang Wu," Effective Pattern Discovery for Text Mining", IEEE Transactions On Knowledge And Data Engineering, pp 30-44,2010.

[9]Saravanan Suba, Dr. Christopher. T, " An Efficient Frequent Pattern Mining Algorithm to Find the existence of K-Selective Interesting Patterns in Large Dataset Using SIFPMM", International Journal of Applied Engineering Research, Volume 11, Number 7, pp 5038-5045 ,2016.

[10]Yuqiu Kong, Lijun Wang, Xiuping Liu,et.al," Pattern Mining Saliency", National Natural Science Foundation,researchgate,pp 583-598,2016

[11] Marijn Janssen, Jeroen van den Hoven," Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy?", Elsevier, pp 263-268, 2015.

[12] Elena Simperl, Maribel Acosta, Marin Dimitrov, et.al," EUCLID: EdUcational Curriculum for the usage of LInked Data",2012.

[13] Pascal Hitzler, Krzysztof Janowicz, "Linked Data, Big Data, and the 4th Paradigm", National Science Foundation, 2013.

[14] Amrapali Zaveri, Anisa Rula, Andrea Maurino, et.al," Quality Assessment for Linked Open Data: A Survey",2012.