International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

Mining Linked Patterns Using Parallelization Technique for Improving Efficiency

Vishakha M. Warke[1], A. S. Vaidya[2]
[1,2]*Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering Management Studies and Research, Nashik-5, India*

***Abstract*— Mining link patterns in linked data has been inefficient due to the computational complexity of mining algorithms and memory limitations.For that to improve efficiency for pattern mining partitioning approachs are used. A partitioning strategy for mining link patterns in linked data, in which linked data is partitioned according to edge-labeling rules. A primary step is multi-partition according to edge labels. In this system, an edge-labeling partition approach for efficient mining of link patterns are used. There are the two edge-labeling rules for pattern mining. A primary multi-partitioning step produces a set of partitions. A quick miner process is to provide feedback to multi-partitioning and a secondary bi-partitioning step will further improve mining efficiency of the linked patterns. In this system we apply the TOG(Typed object graph)for classifying the dataset using the pattern mining in linked data. Parallel Mining approach divides the dataset and gives them as input to multiple systems at the same time. The results obtained here are finally merged to obtain the result.**

***Keywords*—Edge labeling, Link pattern, Linked data, Partitioning, Scalability evaluation.**

I. INTRODUCTION

To improve the scalability and efficiency of mining,a partition strategy is needed. Partitioning is one of the technique used to improve the performance of linked database access.

The edge-labeling rules are used for mining link patterns in pattern mining. There are the two edge labeling rules first one is multi-partitioning and second is bi-partitioning. In this system apply the TOG(Typed object graph)for classifying the dataset using the pattern mining in linked data. Parallel Mining approach divides the dataset and gives them as input to multiple systems at the same time. The results obtained here are finally merged to obtain the result. The system will improve the efficiency in parallel mining using pattern mining. The multi-partitioning and bi-partitioning are used for analyzing the linked data in database[1].

II. LITERATURE REVIEW

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

The basic idea for this system,two edge-labeling rules that is multi-partitioning and bi-partitioning.The TOG is derived from the RDF graph.In this system, RDF graph is used for sharing the information.But problem was that those were unclassified documents discarded.For this system SWDF dataset and DBpedia dataset's used[1].

In this system, the Mining of subgraphs in a currently beyond the scope of these algorithms. To bridge this gap,first bring up a partition-based approach called PartMiner for mining the graphs. The PartMiner algorithm finds the frequent subgraphs by dividing the database into minor units, mining frequent subgraphs on these units and finally combining the results of these units to recover the complete set of subgraphs in the database.A partition-based algorithm PartMiner for discovering the set of frequent graphs.PartMiner can reduce the number of candidate generation graphs by exploring the chain information of the units. It also have the present IncPartMiner, an extended version of PartMiner to handle renew in the graph databases[2].

In this system,a now frequent subgraph mining algorithm: FFSM (Fast Frequent Subgraph Mining) which employs a search scheme within an algebraic graphical framework have been developed to reduce the number of redundant candidates considered. FFSM achieves a steady performance gain over the current start-of-the-art subgraph mining algorithm gSpan. In this system, a data structure CAM tree are used for all connected subgraphs of a single connected undirected graph.A single CAM tree can be built for a graph database. This method clearly suffers from the number of available subgraphs in a graph database and very rare scale to large databases[3].

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

In this system, a algorithm used which is called as gSpan (graph-based, Substructure pattern mining), which determines frequent substructures without candidate generation. gSpan builds a graphs and maps each graph to a unique DFS code as its authorize label. Based on this lexicographic order, gSpan assents the depth-first search approach to mine frequent connected subgraphs dexterously. A global performance study has been conducted in experiments on both synthetic and real world datasets. The real dataset tested a chemical compound dataset[4].

The main objective of mining semi-structured data, symbolic sequences and ordered trees is to extract patterns from structured data.In this system, the patterns mined are characterized by frequency and information entropy are mined. The classes of the patterns are handled in the multirelational data mining are more indicative than mentioned data structures.In mathematics, one of the most generic topological structures are used by graphs. Semi-structure represented by text tags, symbolic sequence and tree including ordered and unordered trees are subclasses of general graphs[5].

The greedy search method is divided into two forms depth first search (DFS) and breadth-first search (BFS). DFS can save the memory consumption. A drawback of DFS approach is that an some part of the identical subgraphs can be found when the search must be stopped due to the search time compulsions if the search space is very wide in size.

In this system, The given dataset is divided into a fixed number of fragments, associated with the number of the graphs in each dataset. Then each of the fragment is mined individually using a graph mining algorithm that is FSG either gSpan and the results are combined to generate global results. A major problem in fragmenting graphs is concerning on similarity or dissimilarity of them. Graph-based data mining is defined as the extraction of useful knowledge or information from a graph representation of data. One of the most important hypothesis in graph mining is to find a frequent subgraph, that is to be find a graph that is repeated in the main graph. There are the two real datasets applied to the system. The first dataset, Chemical 340, is a Predictive Toxicology database containing 340 molecules which is equivalent that it has 340 graphs.

This chemical dataset is sparse. The second dataset, Compound 422, is a confirmed active compound from AIDS antiviral screen and contains 422 graphs and is dense[6].

In this system, represents a computationally efficient algorithm called FSG(Frequent Sub Graph), for finding all frequent subgraphs in large graph datasets. Experimentally evaluated the performance of FSG using a variety of real and synthetic datasets. The results are show that the underlying complexity associated with frequent subgraph discovery. FSG is effective in finding all frequently occurring subgraphs in datasets containing over 200,000 graph transactions and scales linearly with respect to the size of the dataset[21].

Vishakha M. Warke and Prof A. S. Vaidya are discusses a review on pattern mining in linked data. The study of mining the linked data in database with the help of various techniques are discussed. The techniques are viz. partitioning strategy, pattern mining, linked data. These techniques are used for mine the data from large linked database. The detailed survey of these techniques[20].

To improve the efficiency our contribution, we propose the system that executed in parallel mining approach. Unlike the existing system, where TOG was applied for single system and for single dataset, we introduce parallel mining where single dataset is divided and provided to different systems simultaneously. This reduces the time complexity for data classification pattern-wise and improves the efficiency of the system.

III. Proposed System

This chapter includes the all information related to the system block diagram, algorithms and the dataset.

A. Problem Statement

Pattern mining for linked pattern in linked data using parallelization improves the efficiency of system.

B. System Architecture

In fig the architecture of the pattern mining for linked data pattern system is depicted.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

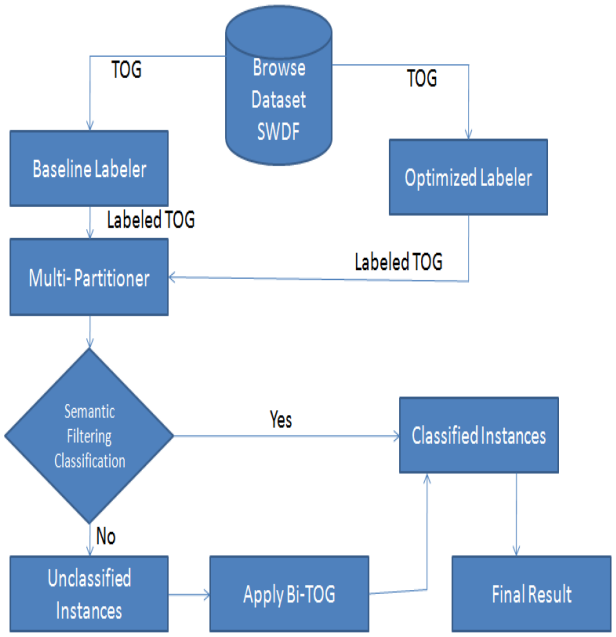

Fig 1: Block Diagram of Pattern mining System

Initial step is to browse the dataset(SWDF).A TOG derived from the browsed dataset(SWDF) is transferred to the labellers namely Baseline Labeler and Optimized Labeler. These two labelers assign a labels or number to each in the linked data that is labeled TOG from unlabeled linked data,using the different edge labeling rules. A multi-partitioner evaluates only the labeled TOGs, and aggregates all labelled edges into primary partitions.

The multi-partitioner sends its output to semantic filtering classification which classifies the given linked data input. Some instances may be classified and some may not.

The classified instances are shown as the final results and for the unclassified instances bi-TOG is applied. Again it is checked whether the instances are classified or not. Bi-TOG is applied on unclassified instances and then updated in classified instances. Conclusively showing the final result.

C. Algorithm and its analysis

Algorithm 1: Multi-partitioning

Input: Labeled TOGs derived from the linked dataset using edge labeling rules

1) The edges in linked data are labeled by using edge labeling rule and then divided into edge sets.

2) From edge set, sort the edge set in descending order.

3) If linked data in dataset is success case of edge labeling rule, then prepare partitions.

Output: Partitions

Algorithm 2: Bi-partitioning

Input: Partitions generated from multi-partitioning.

1) Linked data is divided into partitions

2) Call gSpan to locally design the link patterns in each partition.

3) obtain the result of quick mining in that each partition is a set of local link patterns

Output: Partitions

Algorithm 3: Merging

Input: Partitions

1) Linked dataset is divided into two partitions

2) Call gSpan to locally design link patterns in the two partitions and obtain set of local patterns

3) Get the results with sum of all the local patterns

Output: Global Patterns

Analysis:

In proposed system, these three algorithms are used for mining the link patterns based on edge labeling partition rules. Analysis of the proposed system is an NP-Hard problem. NP-Hard problems have not solved in polynomial time.

D. Dataset

- For this system we use the SWDF(Semantic Web Dog Food) dataset with the number of triples 166083 and number of TOG's 148.
- SWDF is an most popular dataset for mining the linked data.

Name of Dataset	Size of Dataset	URL
SWDF	49.3 KB	http://data.semanticweb.org/dumps/other/pref_labels.rdf

Fig. 2 : showing details of SWDF dataset

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

IV. System Analysis

A. Mathematical Model

Let S be a pattern mining system,
such that,

S = {S0, S4, Fs, D, T, C, R, L| φs}

where,
D represents set of Datasets
T represents Type of graph
C represents set of Classes
R represents set of Results
L represents set of Labels

Initial State(S0)
User browse the dataset for mining link patterns in linked data.

End State(S4)
User obtained the results.

Input
Dataset(D)={d0,d1,.....,dn}

Output
The relevant results in the form of graphs.

Functions(Fs)={f1,f2,f3,f4,f5}
f1 - Parsing of system
f2 - Generate graph
f3 - Generate labels
f4 - Generate classes
f5 - obtain results

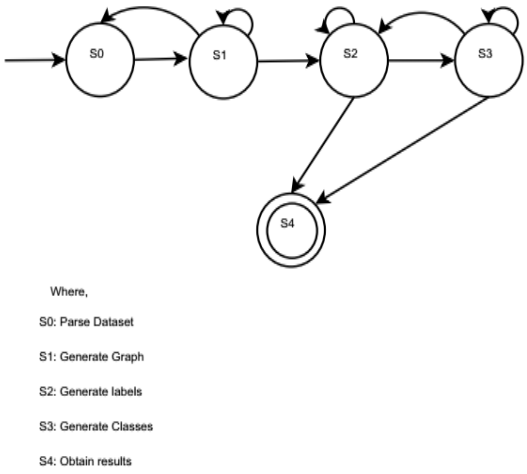

Fig. 3 : State transition diagram of Pattern mining System

B. Implementations Datails

Hardware Requirements

There is the new functionality will run on all standards hardware platform like Intel and Mac.These systems consist of standard and upgraded Windows,Apple,and Mac operating systems.Hardware interfaces include optimal for PC with P4 and AMD 64 processor.The minimum configuration is required for proposed system 2.4 GHZ,80 GB HDD for installation and 512 MB memory.

Software Requirements

There are the various service providers will have different software interfaces to access the authentication services provided by the system.they can perform their services independently as long as they adhere with the policies and standard agreed upon.The proposed system uses the software for implementation as JDK 1.7

V. Result And Discussion

Dataset	No. of Processors	Proposed System Time in ms	Existing System Time in ms
SWDF	1	35.8	35.8
	2	28.6	
	3	22.9	
	4	16.2	
	5	11.7	
	6	8.4	

Fig. 4 : Table showing time required in processing SWDF dataset

A. *Comparison between Existing system and Proposed system*

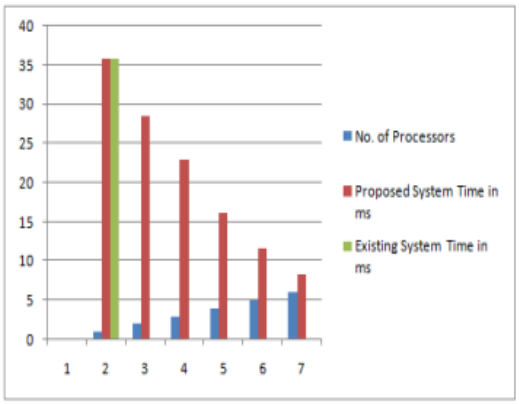

Fig. 5 : Graph showing comparison between Existing and Proposed system

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

In the existing system, we are using a single dataset (SWDF) to be parsed into our system as input. For this system, we are using a single processor and the time required to process using a single process is 35.8 ms as shown in Table. In the above Graph fig.6, a comparison is done between the existing and the proposed system. We can see that for the existing system, by using a single processor, it takes 35.8 ms to process the dataset.

But in our implementation, we would be using different number of processors. The result may like this shown in the graph. As described in table, when the number of processors are 2 the time taken as processing is 28.6 ms and again when we increase the processor number to 3, the time taken is reduced to 22.9 and so on as the different number of processors. The results show that as the number of processors increases, the time taken as processing gradually decreases. This concludes that the processing time and the number of processors are inversely proportional to each other.

B. Results of proposed system

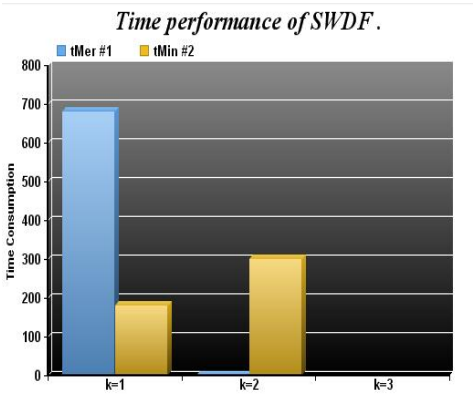

Fig. 6 : Time performance of SWDF

The columns in represent the time consumption for labeling and partitioning; for both the values of k=1 and k=2. Each part of the time consumption is also presented. Time consumption of labeling and partitioning is too small to clearly display. The mining time efficiency is greatly improved by using edge-labeling partitioning. For SWDF dataset, mining time is reduced to half or one-third of the non-partitioned dataset. As shown in the graph, the value for time consumption is 680 for k=1.

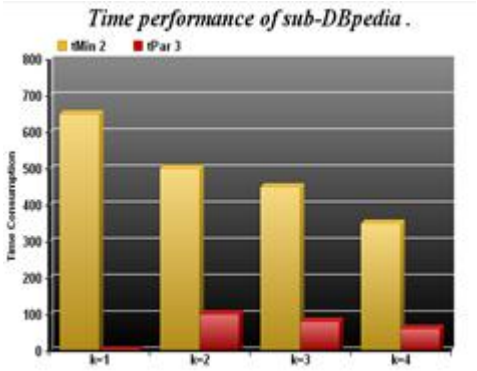

Fig. 7: Time performance of sub-DBpedia

Fig. Shows the graphical representation of Time consumed Vs k for each constant value ranging from 1 to 4. Time consumed for mining is shown in yellow while time consumed for partitioning is shown in red color. For k=1, the time consumed for mining is 650 ms while time consumed for partition for partition is 2 ms. Similarly other constant values of k, is shown in the graph.

VI. Conclusion

In this system, have made a detailed analysis of different technologies that can be used for analysing and understanding the linked data. It can be concluded that the using partitioning strategies they resolved the mining link patterns in linked data for obtaining the relevant information with TOG and BI-TOG. Using the partitioning strategy results are the feasible or efficient in mining the link patterns and reduces the time complexity.

Acknowledgment

I have a tremendous pleasure in presenting the project "Mining linked patterns using parallelization technique for improving efficiency" under the guidance of Prof. A. S. Vaidya and PG coordinator Prof. A. S. Vaidya. I am truly indebted and thankful to Head of the Department Dr. D. V. Patil for their valuable guidance and encouragement. I would also like to thank the Gokhale Education Society's R. H. Sapat College of Engineering, Management Studies Research, Nashik-5 for providing the required facilities, Internet access and important books. At last I must express my sincere heartfelt gratitude to all the Teaching Non-teaching Staff members of Computer Department of GESRHSCOE who helped me for their valuable time, support, remarks, and ideas.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

REFERENCES

[1] Xiang Zhang and WenyaoCheng,Pattern Mining in Linked Data by EdgeLabeling,IEEE transactions on knowledge and data engineering,Volume 21, Number 2, April 2016,ISSNll1007-0214ll05/10ll, pp168-175.

[2] J. Wang, W. Hsu, M. L. Lee, and C. Sheng, A partition based approach to graph mining, presented at the 22nd International Conference on Data Engineering, Atlanta, GA, USA, 2006.

[3] J. Huan, W. Wang, and J. Prins, Efficient mining of frequent subgraph in the presence of isomorphism, presented at the 3rd International Conference on Data Mining, Melbourne, FL, USA, 2003.

[4] X. F. Yan and J.W. Han, Gspan: Graph-based substructure pattern mining, presented at the 2002 IEEE International Conference on Data Mining, Maebashi, Japan, 2002.

[5] A. Inokuchi, T. Washio, and H. Motoda, An apriori-based algorithm for mining frequent substructures from graph data, presented at the 4th European Symposium on the Principle of Data Mining and Knowledge Discovery, Lyon, France, 2000.

[6] M. Gholami and M. Norouzi ,A Fixed-Size Fragment Approach to Graph Mining,IEEE transactions on knowledge and data engineering, July 19, 2014.

[7] A. Sheth, B. Aleman-Meza, B. Arpinar, C. Bertram, Y. S. Warke, and C. Ramakrishnan, Semantic association identification and knowledge discovery for national security applications, Journal of Database Management, vol. 16, no. 1, pp. 3353, 2005.

[8] A. Basse, F. Gandon, I. Mirbel, M. Lo," DFSbased frequent graph pattern extraction to characterize the content of RDF triple stores", presented at the Web Science Conference 2010: Extending the Frontiers of Society Online, Raleigh, USA, 2010.

[9] S. N. Nguyen, M. E. Orlowska, and X. Li, "Graph mining based on a data partitioning approach", presented at the 19th Conference on Australasian Databases, Wollongong, Australia, 2008.

[10] J. W. Han, X. F. Yan ," CloseGraph: Mining closed frequent graph patterns", presented at the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC, USA, 2003.

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

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

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

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

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

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

[17] 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 , 2016, pp 5038-5045.

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

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

[20] Vishakha Manohar Warke, Prof. A. S. Vaidya, "Review on Pattern mining in Linked data",Gokhale Education Society's R. H. Sapat College of Engineering Management Studies and Research,Nashik-5, Global Journal of Advanced Engineering Technologies(GJAET), volume 5,issue 4,ISSN:2277-6370, 2016, pp 423-425.

[21] Michihiro Kuramochi , George Karypis, "An Efficient Algorithm for Discovering Frequent Subgraphs", IEEE Transactions on Knowledge and Data Engineering, pp 1-13.

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

AUTHOR'S PROFILE

Miss. Vishakha M. Warke has received her BE Degree in Computer in 2015 from University of Pune. Presently, pursuing ME (Computer) from University of Pune.PG student in Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering Management Studies and Research,Nashik-5.

Prof. A. S. Vaidya, Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering Management Studies and Research,Nashik-5.