Research Article

Ensemble Distributed Search-FSGM-CRD Compressed Cache Algorithm for Large Datasets

M. Sailaja[1], Dr. CVPR Prasad[2]

[1]Research Scholar, Acharya Nagarjuna University
[2]Supervisor, Acharya Nagarjuna Universit

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Frequent sub-graph mining (FSM) is a alternative of frequent pattern mining where patterns are graphs. Among the entities, graph based representation is utilized to effectively represent the complex relationships. Various graph mining techniques are developed from the past many years, most the challenging tasks in graph mining is frequent sub-graph mining (FSM). In FSM many of the existing algorithms consider only graph based structure, the relationships based on entities involved and strength is not considered. It is very important to handle the complex and huge data. There is very huge demand in distributed computational approaches. In this paper, An Ensemble Distributed Search-FSGM-CRD Compressed Cache Algorithm is developed and implemented to find frequent sub graphs.

Keywords: FSM, CRD, Compressed Cache.

1. Introduction

Data mining is the process which is used to extract the information which is meaningful and knowledge from huge databases [1]. Large amount of data is generating day by day and it is very need to process the interesting information from this large volume of data. For this reason, data mining becomes alternative for many researchers. Graph mining becomes more important to process the complicated structures such as social networks, indexing of video, retrieval of text etc. Graphs represents the relationships in various types such as users are called as (nodes) and the relationship in social networks called as (edges), maintaining the relationship between the atoms (nodes) and bonds (edges) are represented with chemical structures, and in biological network proteins (nodes) and protein interactions (edges) and connections between the computer (nodes) within the computer network [2]. The sub-domain of the data mining is graph mining to represent the data in the form of graph [3].

In general FSM algorithms are having two phases: generation of candidates and calculation of frequency. Generation of candidates is done with breadth first technique or depth first technique. One of the most influence factor that effects the performance of the algorithm is the same candidate is generated for more than once. If the data increases, the candidates generation is also increases. In this candidate generation, duplicate and redundant candidates should be avoided during candidate generation for an efficient algorithm. In the second phase, the frequency should be calculated for the candidate's generation and to specify which are most frequent among them. The frequency of the sub-graph is calculated, if it is required to the number of graphs that are isomorphic to this sub-graph in a database. Testing of sub-graph isomorphism is the basic problem of these algorithms since this problem is NP-complete [4]. If the size of graph is increases the cost of identifying isomorphic graphs increases gradually.

In this paper, An Ensemble Distributed Search-FSGM-CRD Compressed Cache Algorithm is introduced to find the improved frequent sub graphs for complex databases or datasets. The proposed methodology consists of three phases. 1.) Searching Phase: In this phase, the efficient searching is done for the keywords, which is called Ensemble distributed Search. 2.) Assigning Phase: In this phase, the tasks are assigned for the workers based on partitioning of data. Efficient partitioning is done in this phase. 3.) Compressed Cache: In this phase, the enhanced compressed technique is utilized to process the data effectively within the less time. This will also reduce the computation time.

Rest of this paper is organized as follows. Section 2 explains the existing frequent graph mining techniques. Section 3 introduces the new proposed methodology. Section 4 experimental results and implementation. Section 5 describes the conclusion.

2. Literature Survey

Many researchers have been developed the various types of algorithms on graph mining. Some of the algorithms are explained in this section. The author Ullmann [5] introduced a new algorithm for sub-graph isomorphism. This method is initialized by means of brute-force tree search procedure. This calculation achieves talent via way of means of inferentially putting off replacement's nodes inside the tree search. The author Agarwal and Srikant [6] has taken into consideration the issue of finding association policies among matters at some point of a large facts base of offers exchange. They added new algorithms for handling this hard which are on a certainly fundamental stage now no longer an equal due to the fact the recognized algorithm. Cook and Holder [7] observed some other variation in their SUBDUE basis disclosure framework relies upon on least depiction period rule. Holder, Cook and Djoko [8] described about the SUBDUE framework which the minimum description length (MDL) rule is observed foundations that % the understanding base and communicate to number one thoughts inside the facts. at some point of this paper they depicted using SUBDUE and moreover tested the lowest portrayal period rule and basic facts used by SUBDUE can control base revelation in an collection of space.

Blockeel and Raedt [9] provided a initial-request system for top-down enlistment of decision tree (DT) which is logically proved. Top-down enlistment of DT's is that the better and first-class ML method. It is been applied to address various diverse available issues. It makes use of a separation and vanquishes approach, and at some point of this it contrasts from its popular prepare contenders which might be primarily based totally with regards to overlaying methodologies. Chakrabarti, Dom and Indyk [10] developed some other approach for obviously ordering hypertext into a given difficulty progression, using an iterative unwinding calculation. After bootstrapping off a content material primarily based totally classifier, they applied each close by messages at some point of a report while the dispersion of the assessed training of diverse information in its area, to refine the class movement of document being organized. They pointed out 3 territory of examination: textual content and hypertext statistics recovery, AI in putting different content material or hypertext, and PC imaginative and prescient and instance acknowledgment.

The author M.Sailaja and C.V.P.R. Prasad [11] introduced a new algorithm for finding shortest path for Community Structures with Deep Learning. The author M.Sailaja and C.V.P.R. Prasad [12] research on frequent sub graph mining from Distributed Database,

The author in [13] proposed the high efficiency algorithm which finds all the frequent sub-graphs from huge graph databases. The experiments are done with synthetic datasets and also with chemical compound dataset. The new methodologies for non-stop diagram primarily based totally instance mining in chart datasets and proposed a totally particular calculation referred to as gSpan. gSpan can be a diagram primarily based totally base instance mining. This observed normal bases without competitor age.

CPAR has done excessive precision and talent that have several treasured highlights. CPAR speaks to a unique technique toward gifted and excellent order. it is charming to extra enhance the productiveness and versatility of this gadget and evaluation it and different entrenched grouping plans. Additionally, the electricity of the decided prescient requirements likewise rouses us to play out a pinnacle to backside research on optionally available methodologies toward possible association rule mining.

3. An Ensemble Distributed Search-FSGM-CRD Compressed Cache Algorithm

The proposed methodology is very strength and powerful algorithm which process the results very effectively and efficiently. Firstly the searching is done with Greedy algorithm and this follows:

Phase 1: Searching

```
set Greedy (Set Contender)
  infusion=new Set();
  while (Contender.isNotEmpty()){
          next = Contender.select(); //utilize selection criteria,
               //remove from Contender and return value
          if(infusion.isFeasible(next))
//constraints satisfied
           infusion.union(next);
          if(infusion.solves()) return infusion}
   return null
}
```

- The select () command selects a candidate based on a local selection criteria, removes it from Candidate, and the value is returned.

- isFeasible()checks whether the selected value is added to the present solution or not and this results in a feasible solution.

- solves() checks whether the problem is solved.

Phase 2:

Scalar Fitness To evaluate the better performance that one can achieve all the tasks. Scalar fitness φi is defined as the best factorial rank of individual pi among all the tasks that can be expressed as =

$$\varphi i = \frac{1}{min_{j\in\{1,2...k\}} r^{i}_{j}}$$

Phase 3:

A-Optimized Compression: A compression approach includes a compression algorithm to reduce the size of the result. Let A be a piece of data. The code of A, denote as code (A), is the compressed form of A. A can be restored from code (A) if the compression is lossless. The compression ratio on A is defined as

$$\rho(A) = \frac{|code(A)|}{|D|}$$

4. Dataset Description

DBLP originally stood for Database systems and Logic Programming. DBLP is a bibliographic database for computer sciences. The main problem in DBLP is the assignment of papers to author entities. This dataset provides bibliographical information about computer science journals and proceedings. It includes 50,000 objects.

- Provides efficient algorithms for locating hidden patterns in information.
- Finds marginal sets of knowledge
- Evaluates significance of knowledge,
- it's simple to know,
- Offers easy interpretation of obtained results,
- Most algorithms supported the rough pure mathematics are significantly fitted to data processing.

5. Experimental Results

The experimental results are conducted on synthetic dataset and Netbeans 8.0.2 as IDE, Java as programming language. Ram 8 GB and 1TB hard drive is required to process this proposed algorithm.

The overall comparative results are shown in table 1.

Algorithm	**No of Results of One keyword**	**Processing Time (MS)**
Disk based Technique (DBT)	1461	3.155
Partition Based Technique (PBT)	2621	4.248
FSGM-CRD	2621	3.352
CRD-PPA	2621	2.970

Ensemble Distributed Search-FSGM-CRD Compressed Cache	2621	1.580

Table 1 shows the performance of the proposed system with total no of results from the DBLPL dataset for one keyword (network) and total processing time (ms).

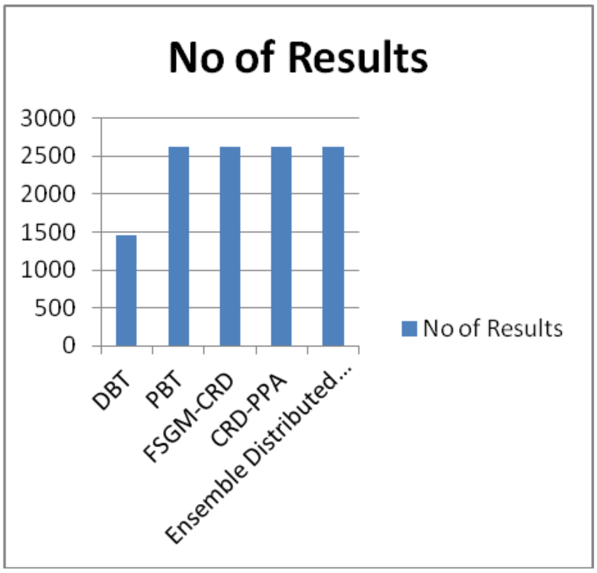

Figure 2: Comparative performance in terms of number of frequent graphs of various FSGM Algorithms

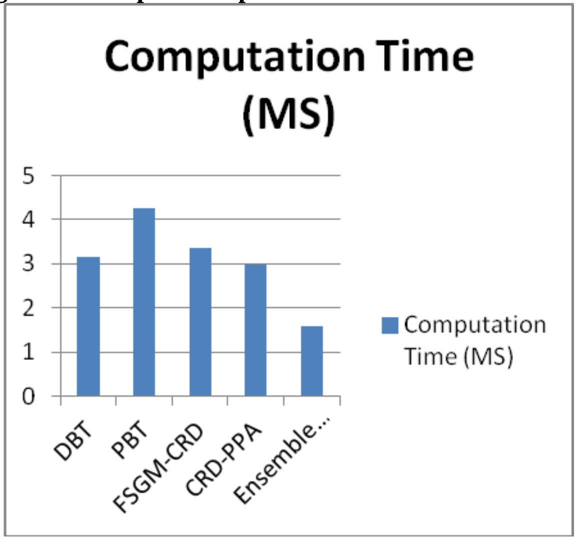

Figure 3: Comparative performance in terms of time taken to complete one keyword frequent graphs of various FSGM Algorithms

6. Conclusion

In this paper, An Ensemble Distributed Search-FSGM-CRD Compressed Cache Algorithm is implemented to solve various issues such as reducing computation time by maintaining the compression cache algorithm. The algorithm focuses on maintaining the cache results within the system and reduces the computation time. This also finds the best frequent sub-graphs that distributed. An efficient allocation of task is done in this research.

References

1. Han J., Pei J., Kamber M., Data mining: Concepts and Techniques, Elsevier, 2011
2. Chakrabarti D., Faloutsos C., Graph mining: Laws, generators,and algorithms, ACM Computing Surveys (CSUR), 2006, 38(1), 2
3. Rehman S. U., Khan A. U., Fong S., Graph mining: A survey of graph mining techniques, Seventh International Conference on Digital Information Management (ICDIM 2012), IEEE, 2012, 88-92.
4. D. J. Cook and L. B. Holder, “Substructure discovery using minimum description length and background knowledge” Journal of Artificial intelligence Research, 1, 1994, 231-255.
5. H. Blockeel, L.D. Raedt, “Top-down induction of first-order logic decision trees”, Artificial Intelligence, 101, 1998, pp. 285-297.
6. S. Chakrabarti, B. Dom, P. Indyk, “Enhanced hypertext categorization using hyperlinks” ACM, (SIGMOD’98), 1998, pp. 307-318.
7. A. Inokuchi, T. Washio, H. Motoda, “An Apriori-based Algorithm for Mining Frequent substructures from Graph Data. In proc. 2000 European Symp. Principle of Data mining and knowledge Discovery (PKDD’00), 1998, pp. 13-23.
8. T. Calders, J. Wijsen, “On Monotone mining Languages”, In proc. Of international workshop on database programming Languages(DBPL), 2001, pp. 119-132.
9. S. Kramer, L.D. Raedt, C. Helma, “Molecular feature mining in HIV data”, In Proc.ational conf. on of the 7th ACM SIGKDD International conf. on knowledge discovery and data mining, 2001, pp. 136-143
10. M. Kuramochi, G. Karypis, “ Freovequent Subgraph Discovery “, In Proc ICDM’01.
11. M.Sailaja, Dr.C.V.P.R Prasad, "Finding shortest path for Community Structures with Deep Learning", International Journal of Pure and Applied Mathematics, Volume 118 No. 14 2018, 167-174.
12. M.Sailaja, Dr.C.V.P.R. Prasad, "A Research on Frequent Sub Graph Mining From Distributed Database", Jour of Adv Research in Dynamical & Control Systems, Vol. 11, No. 8, 2019.
13. M.Sailaja,Dr.C.V.P.R Prasad, "An Improved CRD (Close, Reach& Degree) algorithm to Process the Complex Graph Datasets",, Solid State Technology, Volume:63,Issue:6,Publication Year:2020.