International Journal of Latest Trends in Engineering and Technology
Vol.(9)Issue(3), pp.246-250
DOI: http://dx.doi.org/10.21172/1.93.41
e-ISSN:2278-621X

A STUDY ON GRAPH MINING ALGORITHMS TO DISCOVER FREQUENT SUBGRAPH PATTERNS FROM EXACT GRAPH DATA AND UNCERTAIN GRAPH DATABASE

Saroj A. Shambharkar[1], D.Anantha Reddy[2] & Mohammed Jabeed Rihaz[3]

Abstract- For knowledge discovery from complex objects we require mining algorithms,they extract frequent subgraphs from graph database,required to know the relationship between the object. In this paper the different approaches/algorithms,techniques and methods for finding the frequent graph patterns from the given dataset is specified. The dataset can be certain graph data or Uncertain graph data.The paper discusses about the approaches,techniques and methods used to find the frequent graph patterns on protein structure and genes. There are also several algorithms for finding the frequent subgraph. One of the novel frequent subgraph mining algorithm is used to solve one of the important problem in Bioinformatics is finding recurring residue packing patterns and spatial motifs. There are different frequent subgraph mining algorithms such as gSpan, FFSM,SPIN,SUBDUE .This paper presented the information about the algorithms for frequent subgraph mining algorithms,techniques of graph mining,domain where graph mining is used and also the creation of subgraphs from the graph database. In this paper it is also mentioned that we can use mining algorithms to extract the frequent subgraph patterns from certain graph data as well as we can extract frequent subgraph patterns from uncertain graph data / databases. To discover the frequent subgraph patterns from the uncertain graph database is a very challenging job as uncertainties may occur due imprecision data.

Keywords – Expected support, Graph mining, spatial motifs,supervised learning,uncertain data,unsupervised learning.

1. INTRODUCTION

In Bioinformatics interaction between the proteins can be represented using the data structure called Graph.The objective behind going for data analysis is to obtain the various patterns from the available data sets in huge amount of data. It is considered as one of the way of mining the data. These extracted data patterns can be represented in the form of graph. There are various applications of graph mining.

Graph Mining is helpful in mining the data of web browsing performed by net users,it is helpful in mining subsequences of DNA,in inferencing diagnostic rules from the stored patient history records. There are three graph mining techniques and each is having different way of mining the data from the database. They are graph clustering,graph classification and subgraph mining[2].

A graph is used in data mining to represent data and used in many applications of Bioinformatics. The task of mining frequent patterns from the graph database is challenging as operation related to graph such as subgraph testing requires more time or high complexity task as comparison to the operations related to trees,sequences and so on[5].

Graph mining is used to extract the information on the social media websites or social network. The communication on social network can be done through grouping,messaging, writing comments or by some other means like audio messaging ,emotion icons,and so on.

2. LITERATURE SURVEY

Karsten Borgwardt and Oliver Stegle,in their research mentioned the use of graphs are in Co-expression network,social network,program flow,chemical compound,protein structure,and so on. The pattern graph mining exist in frequent graph patterns,pattern summarization,Optimal graph patterns,Graph patterns with constraints ,Approximate graph patterns .The graph classification given by them is Pattern-based approach ,Decision tree and Decision stumps. The application of graph patterns mentioned by them are Mining biochemical structures ,Finding biological conserved subnetworks,Finding functional modules,Program control flow analysis,Intrusion network analysis,Mining communication networks,Anomaly detection,Mining XML structures and building blocks for graph classification,clustering,compression,comparison,correlation analysis and indexing[1].

In paper “Mining Protein Family Specific Residue Packing Patterns from Protein Structure Graphs”, Jun Haun ,Wein Wang, Deepak Bangyopadhyay,Jack Snoeyink,Jan Prins,Alexander Tropsha,the important problem in Bioinformatics is mentioned as

[1] Department of Information Technology, Kavikulguru Institute of Technology And Science, Ramtek,Nagpur, Maharastra, India
[2] Department of Information Technology, Kavikulguru Institute of Technology And Science, Ramtek,Nagpur, Maharastra, India
[3] Department of Information Technology, Kavikulguru Institute of Technology And Science, Ramtek,Nagpur, Maharastra, India

finding recurring residue packing patterns or spatial motifs. This algorithm is applied to three dimensional 3D protein structure. In this paper the protein graph vertices's are represented by amino acid and vertex-residues are connected by edges. Three approaches mentioned in this for connected vertices's with edges,first is distance threshold between contact residues , second is delaunay tesselation from computational geometry and third is recently developed delaunay tesselation approach. These approaches can be applied to SCOP database. The SCOP is structural Classification of Proteins. And applying the techniques ,obtaining several hundred common subgraphs[4].

In paper entitled as “ Efficient Mining of frequent subgraph in the presence of isomorphism”, a novel mining algorithm named as FFSM,where they applied a vertical search scheme which reduces the number of redundant candidates. Also,they mentioned their algorithm achieved substantial performance as compared to subgraph mining algorithm gSpan[5].

In paper entitled as “A Comparative study of frequent subgraph Mining Algorithms” three aspects were involved in frequent subgraph mining has been mentioned. They were graph representation,subgraph enumeration, frequency counting and also described all three aspects.

The authors of paper on “Frequent Subgraph Mining from uncertain data”,they mentioned that the algorithm used can be scalable ,efficient and can also give accurate results to get the frequent subgraph patterns even from uncertain Graph data.

The objectives of approximate mining algorithm are to determine as efficiently as possible whether a subgraph patter can be output or not and to examine the subgraph patterns as efficiently as possible to obtain the frequent patterns[11].The researchers mentioned that the analytical and experimental results of the approximate mining algorithm shows that it is very efficient, accurate, and scalable for large uncertain graph databases[16].

3. GRAPH MINING ALGORITHMS,TECHNIQUES AND METHODS

3.1 Functions,Techniques and methods finding graph patterns –

The objective functions or measures useful in finding graph patterns are

i) Frequency function for finding frequent graph pattern
ii) Discriminative function for getting information gain, fisher score.
iii) Significance measure using G-test.

The demonstration of first function that is frequency function to find the frequent graph pattern is given below:-

Let us given is a graph dataset D1 ,and we have to find the subgraph GS1 of it,such that fre(GS1) greater than a threshold value theta, where fre(GS1) is the percentage of graphs in D1 that contains subgraphs, DS1.consider an example of find the the frequent graph pattern in chemical bonds of caffeine,diurobromine and viagra,their structure are given in figure 1[1].

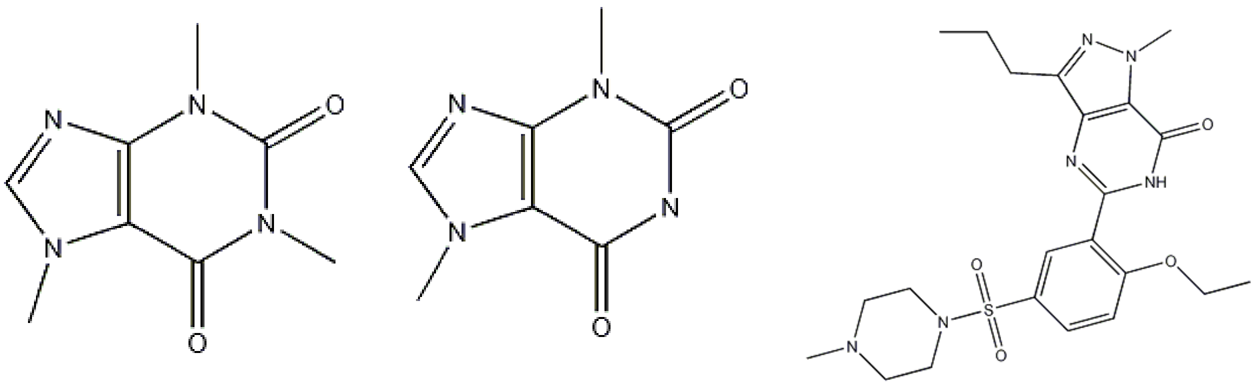

Figure 1. dataset of caffeine,diurobromine and Viagra

Taking the above datasets as input then the frequent subgraph generated after applying the frequency function is given in figure 2.

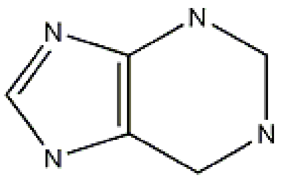

Figure 2. Frequent subgraph

3.2 Graph Mining Techniques –

Graph Mining Techniques are divided into following types:-

i) Clustering or graph clustering or decision tree ii) Graph Classification iii)Sub Graph Mining

In first technique that is graph clustering the vertices's are grouped together to form clusters. While making clusters one has keep in mind the number of edges should be large in a cluster and number of clusters generated should be less. Graph clustering is based on unsupervised learning,here the we don’t know the data is belongs to which classes ,it is unknown before clustering. The classes were formed based on similarity measure done on each data present in the datasets.

The second technique is graph classification which illustrate the classification of the individual graph taken from the database and classifying them into different classes. The first technique is based on unsupervised leaning ,but the second technique is based on supervised or it may belongs to semi supervised learning. In supervised or semi-supervised learning the classes are prior to clustering. The third technique is sub graph mining where the subgraph is considered as set of edges and vertices's taken from the graph[2].The third technique can also be called as frequent sub graph mining.

The frequent subgraph mining process involves three things first is graph representation,second is subgraph enumeration and third is frequency or support counting. The simplest way to represent a graph is using adjacency matrix and adjacency list. The subgraph enumeration is classified into two categories one is FSG and another is AGM [6].

3.3 Related terms to graph and frequent subgraph patterns

The data can be from databases, data warehouses, it can be ordered/sequence data, graph data, text data, and so on. The data can be categorized as federated data, high dimensional data, longitudinal data, streaming data, web data, numeric, categorical, or text data[7].

A graph structure can be represented in different ways either using adjacency matrix,adjacency list, adjacency multi list and so on. The subgraph pattern or graph pattern is one of the important application of data mining and used in social network analysis, Bioinformatics.

Labeled graph is a undirected graph consisting of 5 tuples V, vertices's of a graph G,E edges of the vertices's,∑V set of vertex labels, ∑E set of edge labels and fifth tuple is labeling function denoted by δ. It is mapping of set of edge labels and set of vertex labels .The problem of frequent graph mining is to identify all connected subgraphs from graph database. A set of graphs is called as graph database[4].

An attributed graph with labels on nodes and edges ,labels are called as attributes[6].The labels can be in pair as {attribute_name, attribute_value}.

There can be exact graph and uncertain graph data .The exact graph data containing precise and correct data. There may be graphs with uncertainties found due to imprecision,noise and inaccuracy,such graphs are called as uncertain graphs and to extract frequent subgraph patterns from these graphs is a challenging.

A subgraph is a subset of given graph, and every subset is also a graph. Frequent subgraphs are subgraphs that are occurring frequently in the database[9].

The frequent subgraph patterns are those patterns that occur frequently in the given graph database or in set of graphs or in large graphs.

In Bioinformatics,the PPI(Protein-Protein Interaction) network is a graph consisting of vertices and edges. The vertices are also called as nodes of the network are proteins ,and the protein to protein interaction is represented as edges.

Graph Mining Techniques are used to discover outliers ,patterns from structured data can be represented in the form graph or frequent subgraph patterns. Using graph mining understand the relationships,represented as links between nodes in the graph,contents can be text,images ,numbers depending on the structure data collected with reference to the application [13].

3.4 Finding frequent sub graphs

A Graph G is defines as set of edges and vertices's, G={V,E}.The edges of the graph can labeled dependent on the application for which the graph is taken into consideration,sometimes it can labeled by parameter cost,sometimes as weight,and so on. A set of graphs is called as graph database. Consider the datasets or graph for chemical structures given in G1,G2 and G3[3].

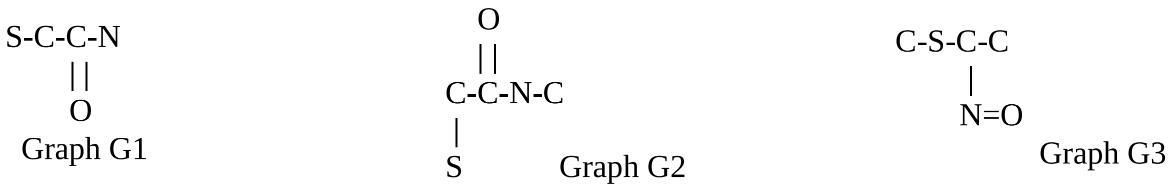

Figure 4. graphs G1,G2 and G3 of some chemical structures

For the above graphs G1,G2, G3 the frequent subgraphs obtained can be

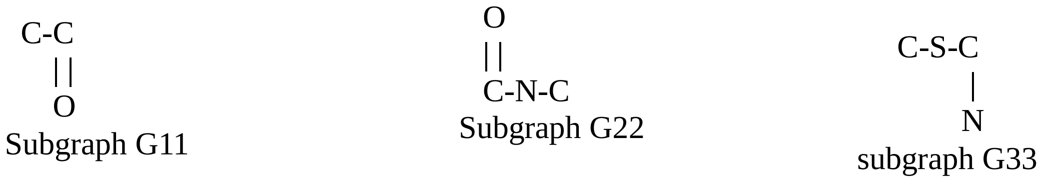

Figure 5. Frequent Subgraphs G11,G22 and G33 of G1,G2 and G3

The method of obtaining the frequent subgraphs from the given set of graphs is called as mining frequent subgraph. For this two parameters as input are to be provided first is set of graphs or graph database and second is minimum threshold support. These two parameters are taken as input parameter to generate the frequent subgraphs.

C-C C-C-N

Subgraph G1-1 Subgraph G2-2

The second subgraph G2-2 is obtained from the input graph G1 and G2,this subgraph is frequent and have support or frequency as 2,as it appears in two input graphsG1 and G2.In this way by giving input parameter support or frequency and input graphs or graph database we can obtain the frequent subgraphs.

In the above example we have obtain the frequent subgraph ,by taking more than one input graph. We can also obtain the frequent subgraphs if only one or single graph is given as input parameter but in such case there will different way in computation of support or frequency parameter.

3.5 Frequent subgraph pattern algorithms

Frequent sub graph mining is a sub section of graph mining domain which is used in mostly for graph classification, it also helps in building in dices and also used for the purpose of graph clustering. The frequent sub graph mining is addressed from various perspectives and viewed in different directions based upon the domain expectations[10].The data obtained from social network websites like twitter, Facebook, where data is in large amount and for mining such voluminous data,stream mining algorithms were evolved and for social networks graph mining algorithms are exist. The data can be obtained from websites social networks by user's e-mail or name or account, search terms, locations on map, companies, IP addresses, books, films, music, and products and data mining techniques can be applied to data collected from social network data may be old or emerging data[7].

For mining frequent subgraph patterns the following algorithms can be used are gPLS algorithm, graphSig, gSpan, Rightmost path extensions,subgraph isomorphism enumeration algorithm,Canonical checking algorithm,GREW,SPIN,SUBDUE, FFSM. Extracting knowledge or discovering knowledge is important in domain of Bioinformatics,web mining,drug discovery,adverse drug events .There are exist different ways for check whether the given two graphs are similar or not. There are two approaches,one of them is based on the comparison done either between the nodes or between the edges,any two pairs of nodes or edges in two networks/ Graphs and calculating an overall similarity score of the two networks. This approach takes time quadratic in the number of nodes and edges, having feasibility even for large graphs[14].

There are also apriori-based frequent subgraph mining algorithms exist,they are AGM,FSG and Edge-disjoint Path-join Algorithms. The AGM algorithm outputs candidate graphs and at instant of time any two candidate graphs were merged and gives a graph ,then this resultant graph is checked to see whether it is subgraph of given graph or a graph database or not[10].The size of the graph is equal to the number of vertices available in that graph. Suppose if we choose any 2 graphs from the candidate set say number of vertices of one is m and number of vertices of second graph is m,then if we merge them the resultant graph will be of size (m+1) using apriori-based algorithm. In we apply the algorithm in every iteration the same process is repeated and the size of the resultant graph will get increase by 1.Therefore,this algorithm is also called as vertex-based candidate generation algorithm[10].Similarly we are having FSG,which is edge based apriori algorithm,in vertex-based candidate generation algorithm the size of the resultant graph is dependent on number of vertices ,here the resultant graph size is dependent on number of edges. In every iteration just like vector-based candidate generation algorithm the number of edges are exactly increased by one as compared the previous frequent subgraphs.

The FSG is an efficient mining algorithm to discover the frequent subgraph patterns from the large graph dataset and gives good performance when transaction data sets increases. The performance evaluated on synthetic datasets. The researchers [12] discovered efficiently frequent Subgraphs in in a datasets over 200,000 graph transactions using FSG. They did the experiment on three types of dataset,out of which 2 datasets are from chemical compounds containing graph transaction 200,000 in number and one for market- basket transaction. When we apply FSG algorithm for mining frequent subgraphs from large datasets,in each iteration ,candidate subgraphs generated , and number of edges present in the subgraph will be one more than arts by enumerating all frequent single- and double-edge subgraphs. Then, it enters its main computational phase, which consists of a main iteration loop. During each iteration, FSG first generates all candidate subgraphs whose size is greater than the previous frequent ones by one edge, and then counts the frequency for each of these candidates and prunes subgraphs that do no satisfy the support constraint. FSG stops when no frequent subgraphs are generated for a particular iteration.

4. FREQUENT SUBGRAPH MINING ALGORITHMS FOR UNCERTAIN DATA

There are some applications were mining algorithms are required for mining the frequent subgraph patterns from uncertain Graph data. As compared to mining frequent subgraph patterns from certain Graph data ,mining the frequent subgraph patterns from Uncertain Graph data is a challenging task. We can can have Uncertain graph database as a collection of uncertain Graph data sets. In previous sections we have seen mining algorithms for certain graph data[11].In Bioinformatics, the PPI network is represented as uncertain graph and also in biological networks consisting of uncertainties. For a given uncertain graph database U’, a connected exact graph that is subgraph isomorphic to at least one implicated graph in some implicated graph database of U’ is a subgraph pattern [11].For wide range of applications the structured and unstructured data collected ,topologies of wireless sensor network ,social networks can have uncertain graph data. There is a theoretical and practical significance of mining frequent subgraph patterns from the uncertain data [15].Also,there are many real time applications where uncertainties will be found because of incompleteness and imprecision of data. For uncertain data we can use a parameter or a measure, called expected support. An approximate mining algorithm is proposed to find a set of approximately frequent subgraph patterns which tolerate the errors on expected supports measured for discovered subgraph patterns. The algorithm uses efficient methods to determine whether a subgraph pattern can be output or not and a new pruning method to reduce the complexity of examining subgraph patterns[16].

5. CONCLUSION

Graph mining algorithms are important in field of Bioinformatics and Chem-informatics. The mining algorithms are useful to perform sentiment analysis on data collected from social networks like twitter. There are some applications where the challenging task is to find the frequent subgraph patterns from uncertain Graph data or databases as there were presence of uncertainties and imprecision in it. It can be formalized using expected support. There are mining algorithms for certain Graph data and also for uncertain Graph data. It is proved to be an NP- hard problem. For uncertain graph data ,the efficient algorithm is approximate mining algorithm.

6. REFERENCES

[1] Karsten Borgwardt and Oliver Stegle, “An Introduction to Graph Mining”,Machine Learning andComputational Biology Research Group,Max Planck Institute for Biological Cybernetics and Max Planck Institute for Developmental Biology, Tübingen based upon K. Borgwardt and X. Yan: Graph Kernels and Graph Mining. KDD 2008, with permission from Xifeng Yan.

[2] Saif Rehman,Asmat Ullah Khan,Simon Fong, "Graph Mining:A survey of Graph Mining Techniques”," Conference Paper · Augus 2012 DOI:10.1109/ICDIM.2012 IEEE.

[3] “Graph Mining, Social Network Analysis, and Multirelational Data Mining”, chapter 9, pp. 535-589.

[4] Jun Huan,Wein Wang,Deepak Bandyopadhyay,Jack Snoeyink,Jan Prins,Alexander Tropsha, “Mining Protein Family Specific Residue Packing Patterns From Protein Structure Graphs”, ‘RECOMB’04, March 27–31, 2004, San Diego, California, USA,Copyright 2004 ACM 1-58113-755-9/04/0003 .

[5] Jun Huan, Wei Wang, Jan Prins, “Efficient Mining of frequent subgraph in the presence of isomorphism", 3rd IEEE international conference on Data Mining, ICDM-Melborne,FL,United States, 2003.

[6] K. Lakshmi and Dr. T. Meyyappan, “A comparative study of frequent subgraph mining Algorithms", International Journal of Information Technology Convergence and Services (IJITCS) vol. 2, No. , April 2012.

[7] Bater Makhhabel, “Learning data mining with R” , January 2005.

[8] Shilpa Arora,Elijah Mayfield, carolyn Penstein-Rose and Eric Nyberg, “Sentiment Classification using Automatically Extracted Subgrah Features,Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and generation of Emotion in Text,pages 131-139,LOS Angeles,California,June 2010.

[9] S.V.S.Shanthi,Dr. P. Padmaja, “A Survey of frequent subgraph Mining Algorithms for Uncertain Graph Data,” International Research Journal of Engineering and Technology(IRJET), vol. 02,Issue 02 , e-ISSN 2395-0056,p-ISSN 2395-0072, May 2015.

[10] T. Ramraj, R. Prabhakar, “Frequent Subgraph Mining Algorithms-A Survey”, Science Direct,Published by Elsevier , 2015, E 1877-0509.

[11] Zhaonian Zou, Jianzhong Li, Hong Gao, and Shuo Zhang, "Frequent Subgraph Pattern Mining on Uncertain Data", CIKM’09, November 2–6, 2009, Hong Kong, China. ACM 978-1-60558-512-3/09/11, pp. II.258 – II.261, 2002.

[12] MichiHero KuraMochi and George Korypis, "An efficient Algorithm for Discovering Frequent Subgraphs", IEEE Transactions on Knowlege and Data Engineering.

[13] Mary McGlohon, Christos,Faloutsos, "Graph Mining Techniques for Social Media Analysis",Carnegie Mellon,International Conference on Weblogs and Social media,2008.

[14] K.Lakshmi and Dr. T.Meyyapan,“Frequent Subgraph Mining Algorithms-A Survey and Framework for Classification”,Natarajan Meghanathan, et al. (Eds): ITCS, SIP, JSE-2012, CS & IT 04, pp. 189–202, 2012.

[15] JianZhong Li, “Algorithms for mining uncertain data”, KDD’12 proceeding of 18th ACM SIGKDD international conference on knowledge discovery and data mining,page 813,ISBN: 978-1-4503-1462-6.

[16] Zhaonian zou,Jianzhong Li,Hong Gao,Shuo Zhang,“Mining frequent subgraph patterns from Uncertain Graph data”,Issue No. 09- September 2010 vol. 22,ISSN: 1041-4347,pp: 1203-1218.