Graph Mining Approaches

Ankita V. Raiyani
Alpha College of Engineering and Technology,
Gandhinagar,
Gujarat

Abstract

Graph mining is one of the novel approaches for mining the data set represented in graph data structure. In the scientific and commercial applications, graph as a structure has become important for modeling sophisticated structure especially the interactions within them. In this paper, we propose different frequent pattern mining algorithm based on labeling as well as different structures. In our future work, we will provide our own graph mining approach which will efficiently perform mining on the graph data structure.

I. Introduction

There are number of research work based on data mining for to mine the data from dataset in seeking for better performance and innovation. One innovation includes mining from structure data, which is new challenge. This structure is represented as a graph data. For example Biological cells can be represented as biological network, which include various molecules and relationship between molecules. Here molecules are represented as entity and relationship represented as edges.

Graph mining has been a popular area of research in recent years because of numerous applications in computational biology, software bug localization and computer networking. In addition, many new kinds of data such as semi structured data, DBLP and XML can typically be represented as graphs. In the graph domain, the requirement of different applications is not very uniform. Thus, graph mining algorithms which work well in one domain may not work well in another. Consider the example that Chemical data is often represented as graphs in which the nodes correspond to atoms, and the links correspond to bonds between the atoms. In some cases, substructures of the data may also be used as individual nodes. In this case, the individual graphs are quite small, though there are significant repetitions among the different nodes. This leads to isomorphism challenges in applications such as graph matching. The isomorphism challenge is that the nodes in a given pair of graphs may match in a variety of ways. The number of possible matches may be exponential in terms of the number of the nodes. In general, the problem of isomorphism is an issue in many applications such as frequent pattern mining, graph matching, and classification. In this paper, we have provided comprehensive summary details of the different graph mining techniques. Here, we outline the Framework for graph mining, graph matching, indexing in different graph structure, frequent pattern mining in graph data set.

The rest of this paper is organized as follows: In Section II, the terminologies of graph theory is provided. In Section III a detailed literature review is provided on graph mining techniques. This study will end with the conclusion of our work with some future direction in section IV.

II. Graph Terminologies

A graph G (V,E) is made of two sets ,V is set of vertices, E is set of edges. In undirected, labeled graphs, Lv is set of vertex label, L_E: set of edge label and Labels need not be unique like element names in a molecule. An Undirected graphs are normally considered as bi-directional graphs. In a directed graph, each edges are order pairs (u,w) means w is followed by u. The u and w are called endpoints of the edge. In a weighted graph, a weight function w is used to represent the label on each edge.

A graph is said to be connected if there is path between every pair of vertices. A graph G_s (V_s, E_s) is a sub graph of another graph G (V, E) if V_s is subset of V and E_s is subset of E. Two graphs G_1 (V_1, E_1) and G_2 (V_2, E_2) are isomorphic if they are topologically identical. There is a mapping from V_1 to V_2 such that each edge in E_1 is mapped to a single edge in E_2 and vice-versa.

Graphs can be represented in various forms like adjacency matrix, adjacency list, and incidence matrix and incidence list. Based on topology, graphs can be further classified as labeled and unlabelled. Labeled graphs can be further classified as only vertex labeled; only edge labeled and both vertex and edge labeled. If graph vertices or edges have more than one label or attribute, a graph is called multi-labeled.

A complete graph is a graph where every pair of distinct vertices is adjacent. A complete graph on n vertices is denoted by K_n .The complete graph K_n of order n is a simple graph with n vertices in which every vertex is adjacent to every other is called clique.

III. Literature Review

This section summarizes the different graph mining algorithm with their major research contribution and limitation.

In [1],swapnil *et.al.* have proposed new graph mining framework. This can capture entities and relation between entities from different data sources into graph data base, extraction of dense substructure, frequent substructure discovery and also provide direct interactive visualization of graph. This Framework includes five modules like graph preprocessing, graph data base, dense substructure extraction, frequent substructure extraction and graph visualization. The graph preprocessing should integrate the data from various data sources, transform them into graph format. This module improves the quality of graph. Frequent substructure discovery module is used to find commonalities or associations among related entities in the graph dataset. Dense substructure is a collection of vertices such that many or sometimes all has edges between them. Graph visualization module should receive as input a description of graph and return the drawing of the graph as output. This framework cannot find the graph classification techniques.

In [2], hakan *et.al.* have propose a new structural indexing approach. For indexing, we specify a set of common graph structures such as star, bipartite, triangle and cliques. These structures are used in biological, chemical and social networks. Structural graph indexing lists all substructure matching structure formulation and other graph structure can be indentified and added to the list. This technique takes anonymous routers on sampled network initials data. Here, we resolve the anonymous nodes between two known nodes. Then, use the SGI algorithm to resolve nodes. Starting from maximum cliques, resolve all nodes within the cliques and triangle structures.SGI resolve most of all anonymous routers.

In [3], Yuhua Li *et. al.* have proposed new algorithm mSpan for directed labeled graph frequent pattern mining. This algorithm is based on FP-growth gets minimum edge code and abstract node code sequence to identify a direct graph pattern uniquely through minimum extension. This algorithm can also solve the graph pattern isomorphism problem and redundant extension problem. This algorithm can be work on directed labeled graph. This techniques use financial transaction data with minimum threshold value. The dataset has 23 graphs; average graph scale value is 10 nodes. This algorithm use most right extension, which extends each possible forward or backward edges .This algorithm, includes two major parts, the first is graph extension and other is the judgment of graph isomorphism. Graph extension is based on Depth first search, which time complexity is $O(2^n)$.For second step time complexity is also $O(2^n)$.So entire time complexity is $O(2^n*2^n)$.

In [4], ChangHau *et al.* have proposed a graph mining algorithm for dynamic network using graph rewriting rule. A dynamic graph means that graph which structure change over the time. That contains sequences of graph. This paper proposed algorithm which use graph rewriting rule based on compression. That introduce algorithm in two step .For the analysis of dynamic graph, first we determine how one graph is changed into another and how much. That algorithm use temporal patterns which describe how and which graph rewriting rules are applied to generate sequences of graphs over time. There are two step algorithms: learning graph rewriting rule and learning description rule. The first algorithms represent how two sequential graphs are different. The other algorithm describes how one graph changes over time. This techniques has been implemented foe biological networks which are change over time. Here that graph cannot be change that structure according to timeline but that structure can be changes according to threshold value. That algorithm has several issues. First is to synchronize the temporal patterns with structural patterns. Next issue is to evaluate the predication time when the sub graph will appear.

In [5],Winnie *et.al.* have proposed a new techniques, which is MIGDAG (mining graph data for classification) applies on graph theory and to discover sub graph .That algorithm uses an interesting threshold and find exact pattern from frequent sub graph of each class. MIGDAC first calculates measure for each frequent sub graph and then using threshold, distinguishes between interesting and no interesting sub graph. The interesting sub graph consists of patterns that can uniquely characterize the class. That can also define class specific patterns according to their ability to characterize a class and graph sample across multiple classes. MIGDAC [5] compares CS patterns with unseen web page structure by graph mining. Finally that calculates weight of evidence and also classifies the unseen graph into class. This algorithm has several benefits like CS pattern reduce no of interesting patterns, which speed up to classification of graph. This technique uses 50 web pages from

local site. The training data set consists of 40 web pages with the indication of popular or no popular for adverting. The remaining 10 web pages are used to test the accuracy .This algorithm applied on web data and to discover the patterns of layout .This algorithm cannot properly handle more graphical features.

In [6], varun *et. al*. have proposed multi labeled graph matching approach. Graph matching is the process of finding the occurrences of a given pattern with respect to the reference graph. Here graph can be classified as a multiple label on edges as well as vertices. This technique enhancing the indexing method which can provide fast process. Graph matching performance depends upon the nature of input, labeling, data structure and indexing and matching process. This technique use query matching in three steps includes vertex matching, degree centrally based spanning tree generation and NCC based graph matching. This technique is evaluated using three data sets. One is PPI biological networks, which has 2361 vertices and 7182 edges. This can be carried out on 64 bit Linux system with 2 GB memory and Intel Xeon processor. The MuGRAM[6] is focusing on neighborhood connectivity check for graph matching with an enhanced indexing process. Multi labeled graphs are relevant in the context of communication graphs. For example, in a telephone call graph, two callers are represented as two vertices with telephone number, addresses consider as vertex labels, average duration of all calls, frequency of calls, and total duration of call consider as edge labels. The main issue related to multi labeled graph matching is used in social network analysis, molecular chemistry in which graph is static. Social networks and communication networks are dynamic in nature.

IV .Conclusion

In this study, we have presented the summary information of the different graph mining techniques. These graph mining algorithm are based on different graph frequent pattern mining, classification, graph indexing, graph matching. The spatiality of this work is that it reveals literature review of different graph mining techniques and provides a vast amount of information under a single paper. In our future work, we have planned to propose a new graph matching, classification method based on graph mining techniques, provide its implementation.

REFERENCES

[1] Swapnil Shrivastava and Supriya N.Pal: Graph Mining Framework for Finding and Visualizing Substructures Using Graph Database. In*: Advance in Social Network Analysis and Mining* (2009) 379-380

[2] Hakan Kardes , Mehmet Hadi Gunes: Structural Graph Indexing for Mining Complex Networks. In *IEEE 30th International Conference on Distributed Computing Systems* (2010) 99-104

[3]Yuhua Li, Quan Lin, Gang Zhong, Dongsheng Duan,Yanan Jin: A Directed Labeled Graph Frequent Pattern Mining Algorithm Based on Minimum Code. In: *3rd International Conference on Multimedia and Ubiquitous Engineering* (2009) 353-359

[4]Chang Hau You, Lawerence Holder, Diane Cook, Graph Based Data Mining in Dynamic Networks: Empirical Comparison of Compression based and frequency based sub graph Mining .In: *IEEE International Conference on data mining* (2008) 929-938

[5] Winnei W.M. Lam,Keith C.C. Chan, Analysing Web Layout Structures using Graph Mining. In: *IEEE* (2008) 361-366

[6] Varun Krichna, NNR Ranga Suri ,G Athithan, MuGRAM: An approach for Multi Labeled Graph Matching .In: *IEEE International Conference on Recent Advances in Computing and Software Systems* (2012)19-26