# Graph Theoretic Approach for Data Mining

Sabiha Firdaus, Md. Mahbubur Rahman

**Abstract**— The need for mining structured data was apparent to the research community and one such approach focused on the topological view of data structures. Since the graph has a generic topological structure and is one of the most thoroughly researched data structures in Computer Science and Discrete Mathematics, state of- the-art techniques in graph-based data mining (GDM) have had profound influence. GDM has tremendous utility because graph-structured data occur widely in practical fields like biology, chemistry, material science and communication networking. Graph-based data mining represents a collection of techniques for mining the relational aspects of data represented as a graph. Two major approaches to graph based data mining are frequent sub graph mining and graph-based relational learning. This article will focus on one particular approach embodied in the Subdue system, along with recent advances in graph-based supervised learning, graph-based hierarchical conceptual clustering, and graph-grammar induction. The need for mining structured data has increased rapidly. One of the best studied data structures in computer science and discrete mathematics are graphs. Graph based data mining has become quite popular in the last few years. This study introduces the theoretical basis of graph based data mining and surveys the state of the art of graph-based data mining. Brief descriptions of some representative approaches are provided as well.

**Index Terms**— Graph theory, data mining, knowledge discovery, clustering, greedy search, kernel function, inductive logic programming.

## 1 INTRODUCTION

DATA mining is a part of the process of Knowledge Discovery in Databases (KDD) which is the search for meaningful and useful patterns and relationships in large databases. Data mining algorithms include machine learning and statistical techniques that are designed especially to work with large amounts of multidimensional data. Clustering is a form of data mining called unsupervised learning. The computer is given an unclassified dataset, and attempts to group records with similar attributes together. During the few years the field of data mining has introduce as an important field of research, investigating interesting research issues and developing challenging real-life applications. The objective data formats in the beginning of the field were limited to relational tables and transactions where each in-stance is represented by one row in a table or one transaction represented as a set. However, the studies within the last several[2] years began to extend the classes of considered data to semi-structured data such as HTML and XML texts symbolic sequences, ordered trees and relations represented by advanced logics.

Graph based data mining or graph mining has a strong relation with the afore mentioned Multi-relational data mining. However, the main objective of graph mining is to provide new principles and efficient algorithms to mine topological substructures embedded in graph data, while the main objective of multi-relational data mining is to provide principles to mine and/or learn the relational patterns, represented by the expressive logical languages. The former is more geometry oriented and the latter more logic and relation oriented in this study, the theoretical basis of graph-based data mining is explained in the following section. Second the approaches to graph-based data mining are reviewed and some representative approaches are briefly described.

- *Sabiha Firdaus is currently working as a lecturer at the department of Computer Science and Engineering at Bangladesh University of Business and Technology, Bangladesh*
  *E-mail: sabiha_firdaus@yahool.com*
- *Md. Mahbubur Rahman is currently working as an assistant professor at the department of Computer Science and Engineering at Bangladesh University of Business and Technology, Bangladesh.*
  *E-mail: mahabub.cse.buet@gmaill.com*

## 2 GRAPH-BASED DATA MINING (GDM)

### 2.1 Theoretical Bases of Graph Data Mining (GDM)

Graph Based data Mining (GDM) has five theoretical bases. Here the five theoretical bases of graph-based data mining approaches are reviewed. They are sub-graph categories, sub-graph isomorphism, graph invariants, mining measures and solution methods.

(1) Sub-graph Categories: Sub-graphs are categorized into various classes (namely general sub-graphs, induced sub-graphs, connected sub-graphs, ordered trees, unordered trees and paths) and the approaches of graph-based data mining strongly depend on the targeted class.

(2) Sub-graph Isomorphism: Sub-graph isomorphism is the mathematical basis of substructure matching and/or counting. In graph-based data mining, the sub-graph isomorphism problem is further extended to cover multiple graphs.

(3) Graph Invariants: Graph invariants are the quantities (like the number of vertices, the degree of each vertex and the number of cyclic loops) to characterize the topological structure of a graph and they help to efficiently reduce the search space of the targeted graph structures. If two graphs are topologically identical, i.e., isomorphic, Isomorphic graphs always have identical values of all graph invariants, while the identical values of given graph invariants does not imply the isomorphism of the graphs. Accordingly the use of graph invariants is not equivalent to complete isomorphic sub-graph matching and counting. However, graph invariants can be

used to reduce the seach space to solve the sub-graph isomorphism problem. If any of the graph invariants show different values between two sub-graphs, the sub-graphs are not isomorphic. Divide and conquer approach based on the graph invariants significantly enhances the computational efficiency in most of the practical problems. One of the most generic and important graph invariants is canonical label and canonical form by which a graph can be represented Graph Theoretic Approach for Data Mining 3 by multiple forms. Graph invariants used for the construction of a high dimensional feature space characterizing a graph has been proposed. Various machine learning, data mining and statistical approaches can be applied if the graph is transformed into a feature vector.

(4) Mining Measures: These are various measures, similar to those in conventional data mining, to mine substructures in the graph, whose selection depends on the objective and the constraints of the mining approach. Some popular mining measures are support, information entropy, information gain, gin index and minimum description length (MDL).

(5) Solution Methods: Approximately five types of search methods are used to solve the sub-graph isomorphism problem amidst a number of graphs. They are categorized into (1) heuristic search methods and (2) complete search methods, in terms of the completeness of the search. They are also classified under (1) direct and (2) indirect matching methods, in terms of the sub-graph isomorphism matching problem. The five types of search methods are: (1) conventional greedy search, (2) inductive logic programming, (3) inductive database, (4) complete level-wise search and (5) support vector machine (SVM).

### 2.2 Recent Developments Carried Out On Graph Based Data Mining

Researchers have proposed a variety of unsupervised-discovery approaches for structural data. One approach is to use a knowledge base of concepts to classify the structural data. Systems using this approach learn concepts from examples and then categorize observed data. Such systems represent examples as distinct objects and process individual objects one at a time. In contrast, subdue stores the entire database as one graph and processes the graph as a whole. Scientific discovery systems that use domain knowledge have also been developed, but they target a single application domain. An example is Mechem, which relies on domain knowledge to discover chemistry hypotheses. In contrast, Subdue performs general-purpose, automated discovery with or without domain knowledge and hence can be applied to many structural domains. Logic-based systems have dominated relational concept learning, especially inductive logic programming (ILP) systems. However, first-order logic can also be represented as a graph and, in fact, is a subset of what graphs can represent. Therefore, learning systems using graphical representations potentially can learn richer concepts if they can handle the larger hypothesis space. FOIL, the ILP system discussed in this article, executes a top-down approach to learning relational concepts (theories) represented as an ordered sequence of function-free definite clauses. Given extensional background knowledge including relations and examples of the target concept relation, FOIL begins with the most general theory. Then it follows a set-covering approach, repeatedly adding a clause that covers some positive examples and few negative examples. Then FOIL removes the positive examples covered by the clause and iterates the process on the reduced set of positive [4], examples and all negative examples until the theory covers all the positive examples. To avoid over complex clauses, FOIL ensures that a clauses description length does not exceed the description length of the examples the clause covers. In addition to the applications discussed here, as well as applications in numerous recursive and no recursive logical domains, FOIL has been applied to learning search-control rules and patterns in hypertext.

## 3 APPROACHES OF GRAPH BASED DATA MINING

The approaches to graph-based data mining are categorized into five groups. They are greedy search based approach, inductive logic programming (ILP) based approach, inductive database based approach, mathematical graph theory based approach and kernel function based approach.

### 3.1 Greedy Search Based Approach

Two pioneering works appeared in around 1994, both of which were in the framework of greedy search based graph mining. Interestingly both were originated to discover concepts from graph representations of some structure, e.g. a conceptual graph similar to semantic network and a physical system such as electric circuits. One is called SUBDUE. SUBDUE deals with conceptual graphs which belong to a class of connected graph. The vertex set V (G) is R UC where R and C are the sets of labeled vertices representing relations and concepts respectively. The edge set E (G) is U which is a set of labeled edges. Though the original SUBDUE targeted the discovery of repeatedly appearing connected sub-graphs in this specie type of graph data, i.e., concept graph data, the principle can be applied to generic connected graphs. SUBDUE starts looking for a sub-graph which can best com press an input graph G based on Minimum Description Length (MDL) principle. The found sub-graph can be considered a concept. This algorithm is based on a computationally constrained beam search. It begins with a sub-graph comprising only a single vertex in the input graph G, and grows it incrementally expanding a node in it. At each expansion it evaluates the total description length (DL), I (Gs) +I (Gj jGs), of the input graph G which is defined as the sum of the two: DL of the sub-graph, I (Gs), and DL of

the input graph, I (Gj jGs), in which all the instances of the sub-graph are replaced by single nodes. It stops when the sub-graph that minimizes the total description length is found.

### 3.2 ILP Based Approach

The first system to search for the wider class of frequent sub-structure in graphs named WARMR was proposed in 1998. They combined ILP method with Apriori like level wise search to a problem of carcinogenesis prediction of chemical compounds. The structures of chemical compounds are represented by the first order predicates such as atomel(C;A1; c), bond(C;A1;A2;BT), aromatic ring(C;S1) and alcohol(C;S2). The first two state that A1 which is a carbon atom bond Graph Theoretic Approach for Data Mining 5 to A2 where the bond type is BT in a chemical compound C. The third represents that substructure S1 is an aromatic ring in a chemical compound C, and the last represents that S2 is an alcohol base in C. Because this approach allows variables to be introduced in the arguments of the predicates, the class of structures which can be searched is more general than graphs. However, this approach easily faces the high computational complexity due to the equivalence checking under subsumption on clauses and the generality of the problem class to be solved. To alleviate this difficulty, a new system called FARMAR has recently been proposed. It also uses the level wise search, but applied less strict equivalence relation under substitution to reduced atom sets. FARMAR runs two orders of magnitudes faster. However, its result includes some propositions having different forms but equivalent in the sense of the -subsumption due to the weaker equivalence criterion. A major advantage of these two systems is that they can discover frequent structures in high level descriptions. These approaches are expected to address many problems, because many context dependent data in the real-world can be represented as a set of grounded first order predicates which is represented by graphs.

### 3.3 Inductive Database Based Approach

A work in the framework of inductive database having practical computational efficiency is MolFea system based on the level-wise version space algorithm. This method performs the complete search of the paths embedded in a graph data set where the paths satisfy monotonic and anti-monotonic measures in the version space. The version space is a search subspace in a lattice structure. The monotonic and anti- monotonic mining measures described in define borders in the version space. To define the borders, the minimal a maximal elements of a set in terms of generality are introduced.

### 3.4 Mathematical Graph Theory Based Approach

The mathematical graph theory based approach mines a complete set of sub-graphs under mainly support measure. The initial work is AGM (Apriori-based Graph Mining) system. The basic principle of AGM is similar to the Apriori algorithm for basket analysis. Starting from frequent graphs where each graph is a single vertex, the frequent graphs having larger sizes are searched in bottom up manner by generating candidates having an extra vertex. An edge should be added between the extra vertex and some of the vertices in the smaller frequent graph when searching for the connected graphs. One graph constitutes one transaction. The graph structured data is transformed without much computational effort into an adjacency matrix mentioned. Let the number of vertices contained in a graph be its size, an adjacency matrix of a graph whose size is k be Xk, the ijth element of Xk, xij and its graph, G(Xk). AGM can handle the graphs consisting of labeled vertices and labeled edges [6].

### 3.5 Kernel Function Based Approach

A kernel function K defines a similarity between two graphs Gx and Gy. For the application to graph-based data mining, the key issue is to find the good combinations of the feature vector XG and the mapping defines appropriate similarity under abstracted inner product. A recent study proposed a composition of a kernel function characterizing the similarity between two graphs Gx and Gy based on the feature vectors consisting of graph invariants of vertex labels and edge labels in the certain neighbor area of each vertex. This is used to classify the graphs into binary classes by SVM mentioned in subsection. Given training data consisting of graphs having binary class, the SVM is trained to classify each graph. Though the similarity is not complete and sound in terms of the graph isomorphism, the graphs are classified properly based on the similarity defined by the kernel function. Another framework of kernel function related with graph structures is called diffusion kernel. Though this is not dedicated to graph-based data mining, each instance is assigned to a vertex in a graph structure, and the similarity between instances is evaluated under the diffusion process along the edges of the graph. Some experiments report that the similarity evaluation in the structure characterizing the relations among the instances provides better performance in classification and clustering tasks than the distance based similarity evaluation. This type of work is supposed to have some theoretical relation with graph-based data mining.

## 4 RELATED RESEARCH WORK ON GRAPH BASED DATA MINING

For Graph based data mining given a database of graph objects, the goal of Graph based data mining is to find all the commonly occurring sub-graph patterns. Some of the early work in Graph based data mining include Cook and Holder (1994), Yoshida and Motoda (1995), and Dehaspe et al. (1998). Many recent methods have been proposed which improve the efficiency of mining, these include Inokuchi et al. (2000), Kramer et al. (2001), Kuramochi and Karypis (2001), Yan and Han (2002a), Huan et al. (2003a), and Nijssen and Kok (2004).

Closed Graph based data mining methods have also been proposed (Yan and Han 2003).Graph based data mining is used in different field of data mining a recent work of mukharjee et. al. shows a structural mining system of social network data she shows the application of the concept of N-clique .Since the strict definition of a clique may be too strong to capture the meaning of the concept, it may be relaxed a bit to include an actor as a member of a clique if he/she is connected to every other member of a group at a distance greater than one. Usually the path distance of two is used. This corresponds to the Friend of a friend (FoaF) concept. This approach to defining substructures is called N-clique, where N stands for the length of the path allowed to make a connection to all other members. The problem with the N-clique approach is that it tends to find long and stringy groupings rather than the tight and discrete Graph Theoretic Approach for Data Mining 7 ones of the maximal approaches. The research of Complete Mining of Frequent Patterns from Graphs by Takashi Washio and Hiroshi Motoda et al. ( 2003) propose an algorithm can derive all frequent induced sub-graphs from both directed and undirected graph structured data having loops (including self-loops) with labeled or unlabeled nodes and links. Its performance is evaluated through the applications to Web browsing pattern analysis and chemical carcinogenesis analysis. Episode mining is one of the data mining methods for time-related data introduced by Mannila et al. in 1997. The purpose of episode mining is to extract all frequent episodes from input event sequences. Here, the episode is formulated as an acyclic labeled digraph in which labels correspond to events and edges represent temporal precedent-subsequent relations in an event sequence. Then an episode gives a richer representation of temporal relationship than a subsequence, which represents just a linearly ordered relation in sequential pattern mining. Episode mining uses serial episode (Mannila et al.), parallel episode (Mannila etal.), sectorial episode (Katoh et al.), diamond episode, elliptic episode (Katoh et al.) bipartite episode , and k-partite episode for data mining.

## 5 RESULTS OF THIS STUDY

This study results indicate that Subdue CL, the graph-based relational concept learner that is competitive with logic-based relational concept learners on a variety of domains. This comparison has identified a number of avenues for enhancements. Subdue CL would benefit from the ability to identify ranges of numbers. We could accomplish this by utilizing the systems existing capability to find similar but not exact matches of a substructure in the input graph.

Numeric values within the instances could be generalized to the encompassing range. A graph-based learner also needs the ability to represent recursion, which plays a central part in many logic-based concepts. More research is needed to identify representational enhancements for describing recursive structures for example, graph grammars.US Geological Survey earthquake data, and software call graphs. Subdue has discovered several interesting patterns in the ASRS database. Burke Burkart of UTAs Department of Geology evaluated Subdues results on the geology data and found that Subdue correctly identified patterns dependent on earthquake depth, often the distinguishing factor among earthquake types. These and other results show that Subdue discovers relevant knowledge in structural data and that it scales to large databases. There also have some complexity to ranking the retrieved data from different sources that contains same class of information. This problem can solve by applying graph base structuring, ordering and weighting. My future work will enhance this approach by applying graph theoretic techniques.

## 6 CONCLUSION

There are many other studies related to graph based data mining. An approach is proposed to derive induced subgraphs of graph data and to use the induced [8], sub-graphs as attributes on decision tree approaches. Geibel and Wysotzki proposed a method can be used to find frequent induced sub-graphs in the set of graph data. A method to completely search homomorphically equivalent sub-graphs which are the least general over a given set of graphs and do not include any identical triplet of the labels of two vertices and the edge direction between the vertices within each subgraph. Liquiere and Sallantin proposed a method to completely search homomorphically equivalent sub-graphs which are the least general over a given set of graphs and do not include any identical triplet of the labels of two vertices and the edge direction between the vertices within each sub-graph. In addition this approach may miss many interesting and useful sub-graph patterns since the homomorphically equivalent sub-graph is a small subclass of the general sub-graph. This study focuses the theoretical basis of the graph-based data mining was explained from multiple points of views such as sub-graph types sub-graph isomorphism problem, graph invariants, mining measures and search algorithms. Thus representative graph-based data mining approaches were shown in the latter half of this article. Even from theoretical perspective, many open questions on the graph characteristics and the isomorphism complexity remain.

## REFERENCES

[1] MRDM'01:. Workshop multi-relational data mining. In conjunction with PKDD'01 and ECML'01, 2002.http://www.kiminkii.com/mrdm/.

[2] R. A grawal and R. Srikant. Fast algorithms for mining association rules. In VLDB'94: Twentieth Very Large Data Base Conference, pages 487- 499, 1994.

[3] Takashi Washio and Hiroshi Motoda. State of the Art of Graphbased Data Mining SIGKDD Explorations Special Issue on Multi- Relational Data Mining,Volume 5, Issue 1, 2003.

[4] L. Dehaspe and H. Toivonen Discovery of frequent data log patterns Data Mining and Knowledge Discovery, 3(1):736, 1999

[5] K. Yoshida, H. Motoda, and N. Indurkhya.. Graph- based induction as a united learning framework. J. of Applied Intel, 4:297328, 1994.

[6] ] D. J. Cook and L. B. Holder. Graph-Based Data Mining. IEEE Intelligent Systems, 15(2), 32-41, 2000.

[7] ] Maitrayee Mukherjee and Lawrence B. Holder. Graph-based Data Mining on Social Networks. KDD04, 2004.

[8] ] Takashi Katoh . E_cient Algorithms for Extracting Frequent Episodes from Event Sequences.

[9] Vineet Chaoji , Mohammad Al Hasan , Saeed Salem and Mohammed J. Zaki. An integrated, generic approach to pattern mining: data mining template library. Springer Science+Business Media LLC 2008.

[10] ] William Eberle and Lawrence Holder. Mining for Structural Anomalies in Graph-based Data. IEEE. Graph Theoretic Approach for Data Mining 9

[11] Akihiro Inokuchi,Takashi Washio And Hiroshi Motoda. Complete Mining of Frequent Patterns from Graphs: Mining Graph Data. Machine Learning, 50, 321354, 2003.

IJSER