Expert Systems with Applications 41 (2014) 1847–1863

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Mining frequent correlated graphs with a new measure

Md. Samiullah [a], Chowdhury Farhan Ahmed [b,*], Anna Fariha [a], Md. Rafiqul Islam [c], Nicolas Lachiche [b]

[a] Department of Computer Science and Engineering, University of Dhaka, Bangladesh
[b] ICube Laboratory, University of Strasbourg, France
[c] School of Computing and Mathematics, Charles Sturt University, Australia

ARTICLE INFO

Keywords:
Data mining
Knowledge discovery
Correlated patterns
Graph mining

ABSTRACT

Correlation mining is recognized as one of the most important data mining tasks for its capability to identify underlying dependencies between objects. On the other hand, graph-based data mining techniques are increasingly applied to handle large datasets due to their capability of modeling various non-traditional domains representing real-life complex scenarios such as social/computer networks, map/spatial databases, chemical-informatics domain, bio-informatics, image processing and machine learning. To extract useful knowledge from large amount of spurious patterns, correlation measures are used. Nonetheless, existing graph based correlation mining approaches are unable to capture effective correlations in graph databases. Hence, we have concentrated on graph correlation mining and proposed a new graph correlation measure, *gConfidence*, to discover more useful graph patterns. Moreover, we have developed an efficient algorithm, *CGM* (*C*orrelated *G*raph *M*ining), to find the correlated graphs in graph databases. The performance of our scheme was extensively analyzed in several real-life and synthetic databases based on runtime and memory consumption, then compared with existing graph correlation mining algorithms, which proved that *CGM* is scalable with respect to required processing time and memory consumption and outperforms existing approaches by a factor of two in speed of mining correlations.

1. Introduction

Data mining extracts useful knowledge from databases. It also discovers patterns (Agrawal & Srikant, 1994; Ahmed, Tanbeer, Jeong, & Lee, 2009; Ahmed, Tanbeer, Jeong, & Choi, 2011; Ahmed, Tanbeer, Jeong, & Choi, 2012; Han, Pei, & Yin, 2000; Hu, Huang, & Kao, 2013; Inokuchi, Washio, & Motoda, 2000; Inokuchi, Washio, & Motoda, 2005; Kuramochi & Karypis, 2001; Nishi, Ahmed, Samiullah, & Jeong, 2013; Nori, Deypir, & Sadreddini, 2013; Tanbeer, Ahmed, Jeong, & Lee, 2009) hidden in data and useful correlations/affinities between the patterns. Correlation mining is a very interesting and important area of data mining which finds the underlying dependencies/affinities among the patterns/objects (Lee, Kim, Cai, & Han, 2003; Omiecinski, 2003; Tan, Kumar, & Srivastava, 2002; Xiong, Tan, & Kumar, 2003). The implicit information within databases, and mainly the interesting association relationships among sets of objects, those lead to association rules, may disclose useful patterns for decision support, financial forecast, marketing policies, even medical diagnosis and many other applications. Nowadays, as data mining techniques are being increasingly applied to non-traditional domains, existing approaches for finding frequent patterns cannot be used as they cannot model the requirements of these domains.

Graphs can be used as an alternate way of modeling the objects in datasets (Inokuchi et al., 2000; Kuramochi & Karypis, 2001; Lahiri & Berger-Wolf, 2008; Lahiri & Berger-Wolf, 2010; Yan & Han, 2003). Within that model, the problem of finding frequent patterns becomes that of discovering subgraphs that occur frequently over the entire set of graphs (Kuramochi & Karypis, 2001). In particular, each vertex of the graph will correspond to an entity and each edge will correspond to a relation between two entities. In this model, both vertices and edges may have labels associated with them which are not required to be unique. The graph structured data mining to derive frequent subgraphs from a graph dataset is difficult because the search for subgraphs is combinatorially explosive and includes subgraph isomorphism matching (Kramer, Pfahringer, & Helma, 1997) which is an NP-complete problem. Power of using graphs, to model complex datasets, has been recognized by many researchers in chemical informatics (Chittimoori, Holder, & Cook, 1999; Dehaspe, Toivonen, & King, 1998; Srinivasan, King, Muggleton, & Sternberg, 1997a; Srinivasan, King, Muggleton, & Sternberg, 1997b), computer vision (Klviinen & Oja, 1990; Piriyakumar, Murthy, & Levi, 1998), image and object retrieval (Cicirello, 1999; Dupplaw & Lewis, 2000), and machine learning (Chen & Yun, 2003; Holder, Cook, & Djoko, 1994; Yoshida & Motoda, 1995) domain.

* Corresponding author. Mob.: +33 629 271 568.
E-mail addresses: samiullah@cse.univdhaka.edu (Md. Samiullah), cfahmed@unistra.fr, farhan@cse.univdhaka.edu (C.F. Ahmed), purpleblueanna@gmail.com (A. Fariha), mislam@csu.edu.au (Md. Rafiqul Islam), nicolas.lachiche@unistra.fr (N. Lachiche).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.082

Correlation mining in graph databases is a very important graph mining task due to its wide range of application domains. Existing works (He & Singh, 2006; Holder et al., 1994; Raymond, Gardiner, & Willett, 2002; Williams, Huan, & Wang, 2007; Yan, Zhu, Yu, & Han, 2006) mainly focus on structural similarity search, which aim to find graphs those are similar in structure. However, in many applications, two structurally similar graphs do not imply that they are correlated or similar in characteristics. For example, in chemistry, isomers refer to molecules with the same chemical formula and similar structures. The chemical properties of isomers can be quite different due to different positions of atoms and functional groups. Consider the case that a chemist needs to find some molecules those share similar chemical properties with a given molecule. Structural similarity search is not relevant, since it mostly returns isomers of the given molecule that have similar structures but different chemical properties, which is undesirable.

In Ke, Cheng, and Ng (2008), the authors proposed an algorithm of mining graph correlation based on statistical similarity, that is, *CGS* (Correlated Graph Search) algorithm, which is able to obtain the molecules that share similar chemical properties but may or may not have similar structures to the given molecule. In particular, *CGS* is a searching algorithm, which works for searching correlation of a specific query graph with the database. Therefore, it has limitations in describing inherent correlation within graphs of graph databases and the domain knowledge is mandatory in using *CGS*, otherwise lots of queries would be meaningless.

Consider a scenario shown in Fig. 1, where two frequent graphs are found from a set of graphs representing a group of people in a social network. Each graph in the set represents friend circle of an individual where nodes represent individuals and edges represent interaction among individual pairs. The circles around graphs represent the interaction of a group of people all together (sub-group). Moreover, nodes in the frequent graphs represent the most interactive individuals and edges are their mutual interaction in various friend circle representing graphs. Integer values beside the edges and circles represent the frequencies of the edges and frequent graphs within the set of graphs respectively. The frequency of edges indicates the number of graphs where such interactions occur in the graph database (network) and the frequency of circles represents the number of graphs where such sub-grouping of individuals with their interactions occurred.

In order to determine the most correlated group between the two groups in Fig. 1, so that we can perform various operations on the most correlated group (as example, target group for task assignment, social/cyber crime investigation, common notification sending etc.), frequencies of the frequent graphs cannot provide any hint due to the tie in frequency of both graphs.

In this circumstance, our measure, which is proposed in Section 3.1, will suggest that people of G_2 are more correlated. Because, G_1's people interact together in 30 events and the maximum number of interactions between any pair in G_1 is 100. Therefore, according to our approach, $Correlation(G_1) = \frac{30}{100} = 0.3$. The second group's people interact together in 30 events and the maximum number of interactions between any pair in G_2 is 60, hence $Correlation(G_2) = \frac{30}{60} = 0.5$. Indeed, such correlation measure helps in mining more useful/meaningful graph patterns and knowledge, since it can discover inherent correlation among the elements of a graph. As a consequence, if the graph database, from which the two frequent subgraphs of Fig. 1 were extracted, can be used for mining correlated graphs with the correlation threshold value 0.4, then the first frequent subgraph will be pruned and the second one will be selected as a strong correlated subgraph among the graphs of the database.

These facts motivated us in developing such a new measure and to the best of our knowledge, such correlation mining in graph databases has not been proposed yet. The key contributions of this paper are as follows:

1. A new measure, *gConfidence*, is proposed to capture more interesting inherent correlation in graph databases.
2. Our proposed measure satisfies the downward closure property, consequently, allows to prune a large number of candidate patterns.
3. We have proposed an algorithm, *CGM* (Correlated Graph Mining), which uses the proposed measure and efficiently mines correlation by constructing a hierarchical reduced search space in large graph databases.
4. Elaborate descriptions with examples of real-life applications are given to explain the realistic usefulness of our approach. Advantages of *CGM* over existing graph correlation mining algorithms as well as the relationship with them are discussed and comprehensively analyzed.
5. An extensive performance study was conducted to show the efficiency, scalability, correctness and effectiveness of our approach. Real-life and complex-large synthetic graph datasets were used to compare our method with existing approaches with respect to runtime and memory consumption.

Our proposed algorithm can be applied in various real-life domains where data can be represented by graphs such as chemical informatics domain, gene sequence databases, bio-informatics, image processing, machine learning, neural networks and lot more.

Rest of the paper is organized as follows: Section 2 contains Related Works, our proposed scheme is presented in Section 3, where Section 4 focuses on the performance analysis of our proposed algorithm. We have discussed the applicability of our scheme in real-life scenarios in Section 5 and finally, we conclude our work in Section 6.

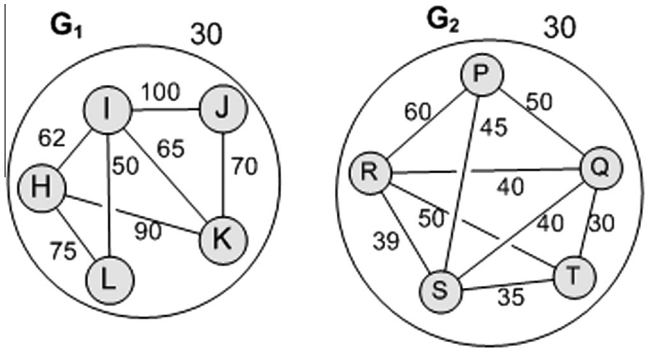

Fig. 1. An example scenario.

2. Related works

Data mining focuses on frequent data values in structured data, but in semi-structured and graph data, the emphasis is on frequent labels and common topologies. Difficulties arise in the discovery task from the complexity of some of the required sub-tasks, such as subgraph isomorphism. In any data mining algorithm which uses an Apriori-based approach, two issues arise: (1) the basic building block from which frequent patterns are composed; (2) making sure that at each step of the algorithm, all frequent patterns for that step are found (Inokuchi et al., 2000).

In graph mining domain, most of the graphs are considered labeled, that is, either the vertices or the edges or both contain a specific value. Transaction graphs can be represented by