Knowledge-Based Systems 85 (2015) 265–276

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Approximating the maximum common subgraph isomorphism problem with a weighted graph

Alan Chia-Lung Chen [a], Ahmed Elhajj [b], Shang Gao [c,*], Abdullah Sarhan [a], Salim Afra [a], Ahmad Kassem [d], Reda Alhajj [a]

[a] Dept of Computer Science, University of Calgary, Calgary, Alberta, Canada
[b] University of Bradford, Bradford, Yorkshire, United Kingdom
[c] College of Computer Science and Technology, Jilin University, Changchun, China
[d] Dept of Management Information Systems, Global University, Beirut, Lebanon

ARTICLE INFO

Article history:
Received 19 May 2014
Received in revised form 7 February 2015
Accepted 9 May 2015
Available online 18 May 2015

Keywords:
Maximum common subgraph
Isomorphism
Weighted graph

ABSTRACT

The maximum common subgraph isomorphism problem is a difficult graph problem, and the problem of finding the maximum common subgraph isomorphism problem is NP-hard. This means there is likely no algorithm that will be able to find the maximal isomorphic common subgraph in polynomial time because as the size of the graphs grows the search space for the solution will grow exponentially. This research provides a method that will approximate the maximum common subgraph isomorphism problem by producing a weighted graph, where the weights will give an indication of the probability that the associated link will be in the maximum common subgraph of the two input graphs. The experimental results show that the method can effectively generate a weighted graph containing most of the expected links that would exist in the maximum common subgraph of some given graphs with similar structures.
© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Graph theory is an important branch of mathematics which studies mathematical structures that contain nodes and links. Graph theory helps in constructing models for many fields of research such as social network analysis, gene expression data analysis, computer networks, and the World Wide Web. A graph is a mathematical model that contains a set of nodes, visualized as circles or dots, and a set of links that connect pairs of nodes; each link is visualized as a line between two associated nodes (see Fig. 1). There are many different types of graphs, and this work incorporates weighted graphs. Weighted graphs are graphs where each link is associated with a numerical value set as the weight of the link to reflect a degree of association between the two nodes connected by the link.

In graph theory, a graph G_N is said to be a subgraph of another graph G_M if the set of nodes in G_N form a subset of the set of nodes in G_M, and the set of links in G_N is a subset of the set of links in G_M (see Fig. 2). Isomorphism is another important concept in graph theory, which is a mapping of the nodes between two graphs, such that if there is a link between two nodes in one graph, then there is also a link between the two corresponding nodes in the other graph (see Fig. 3). Furthermore in graph theory, a modular product can be determined for two graphs G_1 and G_2 to find the modular product graph (see Fig. 4). The modular product graph contains a set of nodes that are the cartesian-product of the two sets of nodes in G_1 and G_2 . Two nodes (N_1, N_2) and (N_A, N_B) in the modular product graph are linked if N_1 and N_A are linked nodes in G_1, and additionally N_2 and N_B are linked nodes in G_2. Alternatively, the two nodes are linked if N_1 and N_A are not linked nodes in G_1, and additionally N_2 and N_B are also not linked nodes in G_2. The concept of a clique in graph theory refers to a set of nodes such that every node in the set is linked to every other node in the set (see Fig. 5).

1.1. Overview of the problem

In this section, we will try to describe carefully the problem tackled in this paper. We concentrate on how the problem is NP-hard and hence justify and motivate the need for the approximate solution described in this paper. Indeed, the maximum common subgraph isomorphism problem is a difficult graph problem. In this problem, two input graphs G_1 and G_2 are examined to find the largest subgraph of G_1 that is isomorphic to any subgraph of G_2. One solution to this problem is to create a modular product graph

* Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun, China.
E-mail addresses: shanggao@jlu.edu.cn (S. Gao), alhajj@ucalgary.ca (R. Alhajj).

http://dx.doi.org/10.1016/j.knosys.2015.05.012
0950-7051/© 2015 Elsevier B.V. All rights reserved.

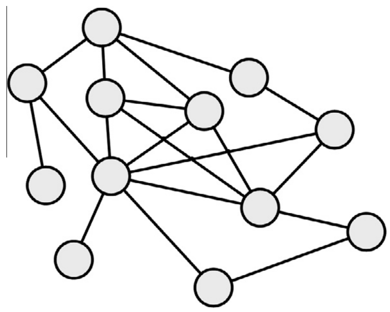

Fig. 1. Example of a graph.

of G_1 and G_2, and find the largest clique in the modular product graph to determine the maximum common subgraph between G_1 and G_2.

The modular product graph solution will essentially compare every combination of nodes between the two input graphs to determine what links according to the combinations appear in both graphs. Finding the maximal clique in the modular product graph will determine the maximum common subgraph because links are present in the modular product graph to show whether there are no links in either input graphs between the corresponding nodes. Therefore, the maximal clique in one graph will solve the maximum common subgraph isomorphism problem after removing the links that represent no links in both G_1 and G_2, as well as removing the node references to the second graph in the maximal clique.

The main issue is that finding the maximal clique of a graph is an NP-hard problem [18]. This makes the maximum common subgraph isomorphism problem NP-hard [13]. In other words, there is likely no algorithm that will be able to find the maximal isomorphic common subgraph in polynomial time because as the size of the graphs grows, the search space for the solution will grow exponentially.

Maximum common subgraph isomorphism solutions are applicable to many different fields, such as biology [11], database management [15], image recognition [23], and social network analysis [16], among others. This problem is also important in bioinformatics and cheminformatics because solutions are used to find similarities in the interactions of biological molecules or the structures of chemical compounds [19]. The problem with the maximum common subgraph isomorphism algorithms is that they can take considerable time to execute especially on large graph and even larger graphs. This is more apparent and challenging with the advances in science and technology to collect and store evermore data.

1.2. Solution

This research provides a method that will approximate the maximum common subgraph isomorphism problem by producing a weighted graph, where the weights will give an indication of the probability that the associated link will be in the maximum common subgraph of two input graphs G_1 and G_2. The first phase in the proposed method will map similar nodes in G_1 to the counterpart nodes in G_2. This is accomplished by distinguishing the nodes in each graph based on their betweenness, closeness, degree, and eigenvector centrality measures, and adjusting the measurements of the nodes in G_2 to map to the measurements in G_1. Lastly, every combination of a pair of nodes from G_1 and G_2, respectively, are compared to determine their similarity, and pairs of nodes with a similarity measure lower than a user specified threshold will be excluded from the next phase of the solution. The final phase of the proposed solution uses the pairings of the mapped nodes between G_1 and G_2 to create the weighted graph of G_1. The weights in the weighted graph of G_1 are calculated by assuming every pairing of nodes returned is applied to the weighted graph of G_1, and checking the number of times a link exists in both graphs against the number of times that the link only appears in one graph, according to the node mappings. The links of the weighted graph of G_1 will then have a weight that signifies the probability that the link will be a part of the maximum common subgraph of G_1 and G_2.

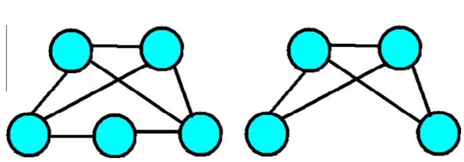

Fig. 2. Subgraph example.

The proposed method is similar to the modular product graph solution, as the modular product graph solution basically maps all the nodes between G_1 and G_2, and considers whether a link exists in both G_1 and G_2 based on each pairing of the mapped nodes; therefore, the detection of the maximal clique in the modular product graph will reveal the maximum common subgraph of G_1 and G_2. This is similar to the proposed method in that both solutions map the nodes of both input graphs to determine what links between the two graphs to consider as part of the solution. However, the modular product graph solution will consider every possible combination of nodes in the mapping to check over all the possible combinations of links that will yield the maximum common subgraph; whereas the proposed solution will consider the mappings of nodes with high similarity to help determine the likelihood that the links would be part of the maximum common subgraph. By generating a weighted graph to offer inferences on the structure of the maximum common subgraph instead of searching for a rigid maximum common subgraph, the proposed solution is characterized by polynomial time. This way larger input graphs will not exponentially increase the runtime of the solution. The weighted graph itself can also allow for some flexibility on how to determine a fitting solution to the problem, following further modifications to the weighted graph such as elimination of nodes or links based on the weights.

After the above introduction which provided a brief background for understanding the research presented, and an overview of the problem and solution being addressed and proposed in this research (Section 1), the rest of the paper is organized as follows. Section 2 discusses other works which have tried to approximate a solution to the maximum common subgraph isomorphism problem. The algorithm for the proposed method is presented in Section 3. The results of the proposed solution and experiments are discussed in Section 4. Section 5 is the conclusion and possible future research.

2. Related work

The subgraph isomorphism problem and the subtree isomorphism problem are similar problems to the maximum common subgraph isomorphism problem. In the subgraph isomorphism

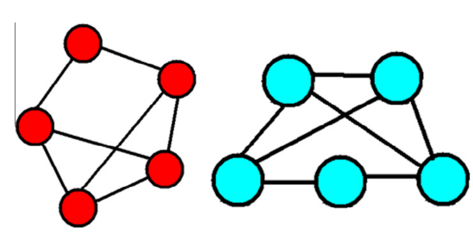

Fig. 3. Isomorphism example.

Download English Version:

https://daneshyari.com/en/article/402267

Download Persian Version:

https://daneshyari.com/article/402267

Daneshyari.com