J Inf Process Syst, Vol.13, No.5, pp.1158~1167, October 2017
https://doi.org/10.3745/JIPS.04.0048

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Similarity Evaluation between Graphs: A Formal Concept Analysis Approach

Fei Hao*, Dae-Soo Sim**, Doo-Soon Park**, and Hyung-Seok Seo***

Abstract

Many real-world applications information are organized and represented with graph structure which is often used for representing various ubiquitous networks, such as World Wide Web, social networks, and protein-protein interactive networks. In particular, similarity evaluation between graphs is a challenging issue in many fields such as graph searching, pattern discovery, neuroscience, chemical compounds exploration and so forth. There exist some algorithms which are based on vertices or edges properties, are proposed for addressing this issue. However, these algorithms do not take both vertices and edges similarities into account. Towards this end, this paper pioneers a novel approach for similarity evaluation between graphs based on formal concept analysis. The feature of this approach is able to characterize the relationships between nodes and further reveal the similarity between graphs. Therefore, the highlight of our approach is to take vertices and edges into account simultaneously. The proposed algorithm is evaluated using a case study for validating the effectiveness of the proposed approach on detecting and measuring the similarity between graphs.

Keywords

Formal Concept Analysis, Graph, Social Networks, Similarity Evaluation

1. Introduction

As the rapid development of big data techniques and powerful ubiquitous computing abilities, the research on massive graph analysis and mining are opening another new door for complex networks systems. On the basis of the promotion of massive graph analysis and mining, various real-world applications are emerging in biological science, social media, and transportation fields recently [1-3]. Therefore, more interesting points and knowledge are hidden in the internal topological structure of massive graphs.

Among the existing massive graphs analysis and mining techniques, subgraph matching technology is to detect the isomorphic subgraph structures in terms of similarity between graphs. The working principle of subgraph matching technology is described as: for a given two subgraphs $g1$ and $g2$, the similarity degree denoted as $sim(g1,g2)$ between $g1$ and $g2$ is evaluated.

Fig. 1 shows a motivating example on functions identification of a certain newly produced medicine.

Manuscript received May 19, 2017; first revision June 9, 2017; accepted July 24, 2017.
Corresponding Author: Doo-Soon Park (parkds@sch.ac.kr)
* School of Computer Science, Shaanxi Normal University, Xi'an, China (feehao@gmail.com)
** Dept. of Computer Software Engineering, Soonchunhyang University, Asan, Korea (parkds@sch.ac.kr)
*** Dept. of Science, Konyang University, Nonsan, Korea (hsseo@konyang.ac.kr)

To explore the functions of this new medicine, a traditional clinic medical approach is to test it in both animal and human. Unfortunately, it often consumes a long time for identifying the functions of the medicine. Fortunately, graph similarity search is becoming a main technical solution for addressing this problem and saving much time for us. As shown in Fig. 1, the molecular structure of our targeted medicine is in the left-most side. It is regarded as a query q in graph similarity search problem, then the graph similarity search algorithm will evaluate the similarity between q and the existing graphs, i.e., $g1$ and $g2$ (molecular structures) in medicine database. Therefore, the essence of this problem is similarity evaluation between graphs.

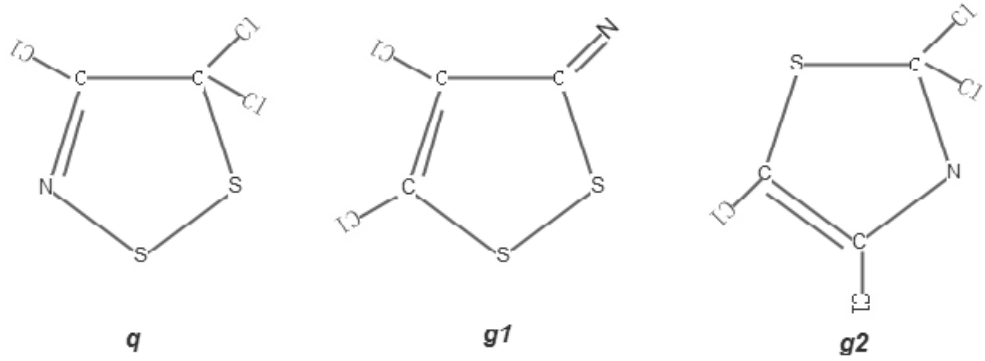


Fig. 1. A motivating example on functions identification of certain newly produced medicine.

Regarding the above problem on query-oriented similarity evaluation between graphs, the existing typical works mainly focus on the kernel function based approach and feature-based approach for the similarity evaluation between graphs. In [4,5], the authors formulated the similarity between two graphs with a kernel function. The basic idea of work [6] is to incorporate the domain knowledge and obtain the features by distilling the topological structures. Then, they further calculate the similarity between two graphs based on their common topological structures. However, these two types of evaluation approaches have not considered the global structural connectivity information of graphs. To address this disadvantage, this work provides a formal concept analysis based evaluation approach for obtaining the similarity between graphs. Differs from most of existing approaches, the formal concept analysis methodology, as an efficient mathematical tool for describing the objects and attributes, is storing the global and local information of graph. It firstly represents the given graphs as the formal contexts. Then, the concept lattices are correspondingly established. With the constructed concept lattices, a similarity evaluation approach is devised. The major contributions are summarized as follows:

(1) (**Graph Representation as Formal Context**) since formal concept analysis methodology is adopted in this paper, we need represent the graph as formal context initially. Technically, the formal context for given graphs is easily constructed via Modified Adjacency Matrix presented in [7-9];

(2) (**Similarity Evaluation Feature Construction**) generally, features are usually needed for evaluation of similarity between graph. In this paper, the formal concept lattices are correspondingly constructed as the feature for further evaluation of similarity between graphs.

(3) (**Equivalent Theorem**) based on our research and proof, it is easily to obtain an equivalent theorem between the similarity on graphs and on the generated formal concept lattices.

The rest of this paper is structured as follows. The related work is over-viewed in Section 2. Section 3 formally describes the addressed problem. The proposed approach is detailed in Section 4. Section 5 shows a case study. Eventually, Section 6 concludes this research.

2. Related Work

There have been a lot of existing approaches that are often used to characterize graph properties but not for similarity evaluation between graphs. From the literature review, two main categories can be summarized: (1) algorithms that are used for detecting vertex similarities; (2) algorithms that are used for evaluating edge similarities. As for the first category, Bunke [10] firstly connected the graph edit distance problem [11] with the one of maximum common subgraphs. Elzinga and Wang [12] obtained the number of the common paths by using inner product of the devised kernel function. This approach can be applied into large graph dataset with a high accuracy. But, its computational complexity is high, i.e, $O(n^3)$. Within the second category, those can be evaluated using Levenshtein distance [13] or the DeltaCon framework [14]. Specifically, Vishwanathan et al. [15] proposed random walk graph kernel (RWGK) for measuring two graph similarities by computing their common paths. But the approach proposed in [15] led to a big kernel function value in the process of measuring in a small range of graph; also the computational complexity is high. Borgwardt and Kriegel [16] proposed shortest-path graph kernel (SPGK) for computing the shortest path, and then further avoiding the phenomenon of tottering. By evaluation the similarity between shortest path between vertices and the similarity between graphs is obtained. Tian and Patel [17] investigated the approximate matching of query graph in a graph database. They assumed the graphs are labeled. Inspired by B-tree index, they devised a hybrid index for preserving the structural information. They firstly matched the important vertices of a query graph then extended it gradually. Zheng et al. [5] presented a minimum edit distance based graph similarity search algorithm.

3. Problem Statement

This section focuses on describing the problem of similarity evaluation between graphs. The problem statement is formally presented as follows.

Problem Statement (Graph Similarity based on Formal Concept Analysis) [18] For two given graphs $G_1(V_1, E_1)$ (including the $|V_1|=n_1$ vertices, $|E_1|=e_1$ edges) and $G_2(V_2, E_2)$ (including the $|V_2|=n_2$ vertices, $|E_2|=e_2$ edges), and the links between the nodes, respectively. The objective of this problem is to put forward an algorithm for evaluating the similarity of two graphs, i.e., $sim(G_1,G_2)$. Since this paper attempts to solve the evaluation of similarity between graphs by Formal Concept Analysis, this problem is further formulated as: for a two given $G_1(n_1, e_1)$ and $G_2(n_2, e_2)$, how to represent them by using formal context $K=(O,A,I)$, and then investigate the similarity $sim(L_A, L_B)$ between the concepts which formed from the formal concept lattice L. Table 1 lists the major variables used throughout this paper.

Table 1. Important variables used in the paper

Notation	Description
$G_1(V_1, E_1)$	A graph G, with n_1 vertices and e_1 edges
$G_2(V_2, E_2)$	A graph G, with n_2 vertices and e_2 edges
C	Formal context
O	Object
A	Attribute
I	Binary relationship between object and attribute
L	Concept lattice
$sim(L_A, L_B)$	Similarity between concept lattice L_A, L_B

4. The Proposed Approach

This section is devoted on presenting the working process of our approach for evaluating the similarity degree between graphs via the methodology of formal concept analysis.

4.1 Similarity Evaluation between Formal Concept Lattices

Similarity evaluation between formal concept lattices is a key technique for our paper. Hence, this section focuses on elaborating how to calculate the similarity between formal concept lattices generated from two given graphs.

Formal Concept Analysis, as a powerful methodology for describing the binary relationships between object and attribute, has been applied into many areas. Formally, a formal context is formulated as $C=(O,A,I)$ where O indicates the object set and A denotes the attribute set respectively, and the relation $I \subseteq O \times A$ is a binary relation between object and attribute. Generally, $o \in O$ and $a \in A$, $(o, a) \in I$ is explained that the objective o has the attribute a.

For better explanations of formal concept lattice and its generated formal concepts, the following two operators are given.

(**Operator for extracting the common attribute of objects subset X**) [18] For $X \subseteq O$, we define a set of common attributes of X,

$$X^{\uparrow} = \{a \in A \mid (x,a) \in I, \forall x \in X\};$$

(**Operator for extracting the common objects of attributes subset X**) [18] For $Y \subseteq A$, we also define a set of common objects of Y,

$$Y^{\downarrow} = \{o \in O \mid (o,y) \in I, \forall y \in Y\}.$$

Generally, for a given formal context $C=(O,A,I)$, a pair (X,Y) is called as a concept if $X^{\uparrow\downarrow} = Y$. Note that, X and Y are called as the extent and intent of the concept, respectively. With the above operators, a concept lattice $L(O,A,I)$ can be generated including the concepts that can be organized according to a special hierarchical partial order.

(Similarity Degree Function) [18,19] Let L_A, L_B be the concept lattices, the similarity degree between the nodes in L_A, L_B is formalized as follows,

$$sim(L_A, L_B) = \frac{\sum_{C_i \in L_A} sim(C_i, L_B)}{n}$$

where $sim(C_i, L_B) = \max(\frac{\sum_{l \in R_i} sim(C_i, l)}{n})$, R_i refers to the set of path which describes the concept C_i.

4.2 Similarity Evaluation between Graphs based on Formal Concept Analysis

Basically, the proposed approach is composed of the following steps (as shown in Fig. 2). Clearly, the two graphs are the input of our approach. Then, it goes into the step 1 for constructing the formal context, after that, the formal concept lattices are correspondingly generated as shown in step 2; the similarity between the generated formal concept lattices are evaluated (step 3) for assisting the evaluation of similarity between graphs.

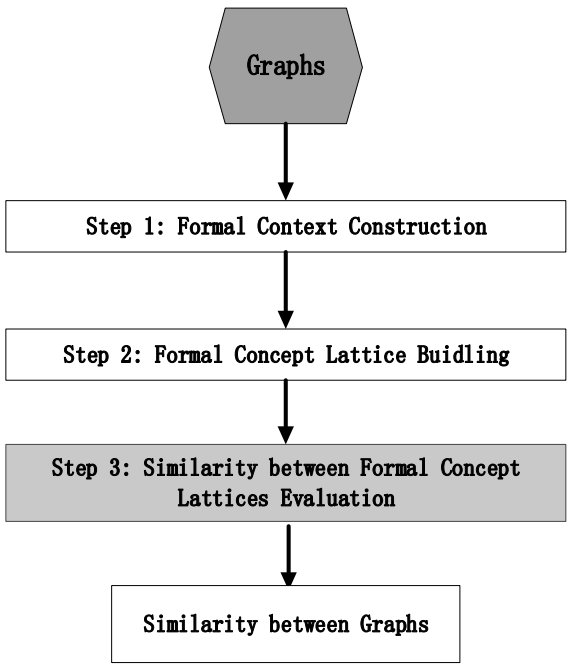


Fig. 2. The work flow chart of the proposed approach.

4.2.1 Formal contexts construction

The formal contexts are easily obtained according to our previous work [7,8,18]. The basic idea of the construction approach is to take the vertices as the both objects and attributes. Then, the modified adjacency matrix is adopted for the construction of formal contexts for two given graphs $g1$ and $g2$. Formally, the formal context is represented as

$$C = (V, V, I)$$

where V is the vertex in the graph, thus the formal context C is a special context compared to the traditional one. We denote $C(g1)$ and $C(g2)$ are the constructed formal contexts of $g1$ and $g2$.

4.2.2 Generating the formal concept lattice

According to the generation algorithm of formal concept lattice [7,9], the lattices of two graphs $g1$, $g2$ are separately generated as $L(C(g1))$, $L(C(g2))$.

4.2.3 Evaluating the similarity degree between concept lattices

Up to now, we can evaluate the similarity degree between lattices according to the similarity degree function as defined in Section 4.1. Hence, the similarity between two lattices can be calculated as

$$sim(L(C(g_1)), L(C(g_2))) = \frac{\sum_{C_i \in L(g_1)} sim(C_i, L(C(g_2)))}{n}$$

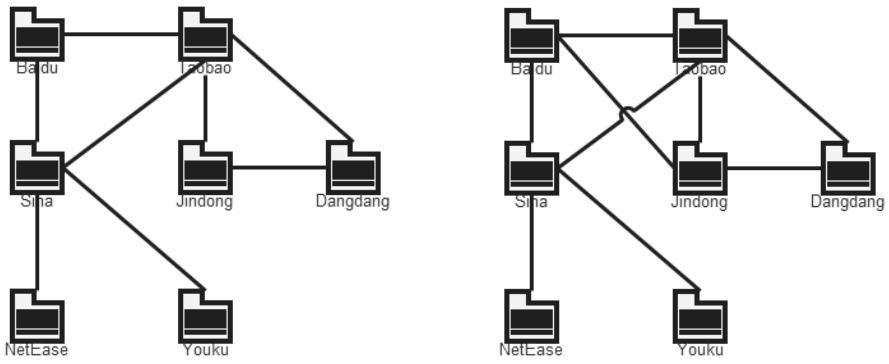
Currently, once we obtain the similarity degree between concept lattices, the similarity between graphs are equivalently obtained. That is to say, for the similarity between $g1$ and $g2$, denoted as $sim(g1,g2)$, then the following equivalence relation holds.

$$sim(g1, g2) \equiv sim(L((C(g1)), L(C(g2)))$$

5. Case Study

In this section, we adopt a useful case about high click rate websites of China given by [20]. This case is described as follows: some high click rate websites in China are given, such as Baidu, Google, Sina, NetEase, Youku, Taobao, Jingdong, and dangdang. Formally, each website in this case study is viewed as a node, and each link is regarded as an edge of graph. We can establish the following two graphs $g1$ and $g2$ shown in Fig. 3.

We can easily obtain the following two formal concept lattices respective to $g1$ and $g2$ by using the above proposed approach.



(a) (b)

Fig. 3. The establishment of (a) $g1$ and (b) $g2$.

$$sim(L(g_1), L(g_2)) = \frac{\sum_{C_i \in L(g_1)} sim(C_i, L(g_2))}{9} = 8/9 * 1 = 0.889$$

Clearly, we obtain the similarity between $g1$ and $g2$ is 0.889.

Throughout this case study, the proposed approach has been validated further from the aspects of feasibility and effectiveness. The proposed approach will be utilized for various large complex graph related applications, such as social networking analysis, Web mining. Additionally, by incorporating with the richer domain knowledge, graph objects classification, subgraph searching are becoming more meaningful.

6. Conclusions

Graph similarity evaluation is becoming a promising technology in the fields of pattern searching, objects tracking and biological complex identification. In order to evaluate the similarity between two graphs, this paper presents a novel formal concept analysis based approach. First of all, the proposed approach constructs the formal contexts for given two graphs, respectively; then the formal concept lattices of them are correspondingly generated; finally, we defined a similarity degree function for concept lattice in order to evaluate the similarity of graphs. The case study on the networks of high click rate websites of China is investigated for performance evaluation of the proposed approach. It is clearly to conclude that our proposed approach can efficiently characterize the relationship between nodes and further obtain the similarity between graphs by calculating the similarity between nodes which appear in the formal concept lattices of the given graphs.

Acknowledgement

This research was supported by the Ministry of Science, ICT and Future Planning (MSIP), Korea, under the Information Technology Research Center (ITRC) support program (No. IITP-2017-2014-0-

00720-002) supervised by the Institute for Information & communications Technology Promotion (IITP) and the National Research Foundation of Korea (No. NRF-2017R1A2B1008421) and it was also supported by the National Natural Science Foundation of China (No. 61702317), and the Fundamental Research Funds for the Central Universities, China (No. GK201703059).

References

[1] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, et al, "Topological structure analysis of the protein–protein interaction network in budding yeast," *Nucleic Acids Research*, vol. 31, no. 9, pp. 2443-2450, 2003.

[2] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social network," *in Link Mining: Models, Algorithms, and Applications*. New York, NY: Springer, 2010, pp. 337-357.

[3] L. Tong, X. Zhou, and H. J. Miller, "Transportation network design for maximizing space–time accessibility," *Transportation Research Part B: Methodological*, vol. 81(Part 2), pp. 555-576, 2015.

[4] Z. Zeng, A. K. H. Tung, J. Wang, J. Feng, and L. Zhou, "Comparing stars: on approximating graph edit distance," in *Proceedings of the VLDB Endowment*, vol. 2, no., 1, pp. 25-36, 2009.

[5] W. Zheng, L. Zou, X. Lian, D. Wang, and D. Zhao, "Graph similarity search with edit distance constraint in large graph databases," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, San Francisco, CA, 2013, pp. 1595-1600.

[6] X. Yan, F. Zhu, P. S. Yu, and J. Han, "Feature-based similarity search in graph structures," *ACM Transactions on Database Systems (TODS)*, vol. 31, no. 4, pp. 1418-1453, 2006.

[7] F. Hao, G. Min, Z. Pei, D. S. Park, and L. T. Yang, "k-Clique community detection in social networks based on formal concept analysis," *IEEE Systems Journal*, vol. 11, no. 1, pp. 250-259, 2017.

[8] F. Hao, D. S. Park, G. Min, Y. S. Jeong, and J. H. Park, "k-Cliques mining in dynamic social networks based on triadic formal concept analysis," *Neurocomputing*, vol. 209, pp. 57-66, 2016.

[9] F. Hao, S. S. Yau, G. Min, and L. T. Yang, "Detecting k-balanced trusted cliques in signed social networks," *IEEE Internet Computing*, vol. 18, no. 2, pp. 24-31, 2014.

[10] H. Bunke, "On a relation between graph edit distance and maximum common subgraph," *Pattern Recognition Letters*, vol. 18, no. 8, pp. 689-694, 1997.

[11] X. Gao, B. Xiao, D. Tao, and X. Li, "A survey of graph edit distance," *Pattern Analysis and Applications*, vol. 13, no. 1, pp. 113-129, 2010.

[12] C. H. Elzinga and H. Wang, "Kernels for acyclic digraphs," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2239-2244, 2012.

[13] B. Cao, Y. Li, and J. Yin, "Measuring similarity between graphs based on the Levenshtein distance," *Applied Mathematics and Information Sciences*, vol. 7, no. 1, pp. 169-175, 2013.

[14] D. Koutra, J. T. Vogelstein, and C. Faloutsos, "DeltaCon: a principled massive-graph similarity function," in *Proceedings of the 2013 SIAM International Conference on Data Mining*, Austin, TX, 2013, pp. 162-170.

[15] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt, "Graph kernels," *Journal of Machine Learning Research*, vol. 11, pp. 1201-1242, 2010.

[16] K. M. Borgwardt and H. P. Kriegel, "Shortest-path kernels on graphs," in *Proceedings of the 5th IEEE International Conference on Data Mining*, Houston, TX, 2005, pp. 74-81.

[17] Y. Tian and J. M. Patel, "Tale: a tool for approximate large graph matching," in *Proceedings of the 24th IEEE International Conference on Data Engineering*, Cancun, Mexico, 2008, pp. 963-972.

[18] F. Hao, D. S. Sim, and D. S. Park, "Measuring similarity between graphs based on formal concept analysis," in *Proceedings of the 11th International Conference on Ubiquitous Information Technologies and Applications (CUTE 2016)*, Bangkok, Thailand, 2016, pp. 730-735.

[19] F. Hao and S. Zhong, "Tag recommendation based on user interest lattice matching," in *Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology*, Chengdu, China, 2010, pp. 276-280.

[20] X. Wang and J. Ouyang, "A novel method to measure graph similarity," in *Proceedings of the IEEE 12th International Conference on e-Business Engineering*, Beijing, China, 2015, pp.180-185.

Fei Hao

He received the B.Sc. degree in information and computing science and the M.Sc. degree in computer software and theory from Xihua University, China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science and engineering from Soonchunhyang University, Korea, in 2016. He is currently an associate professor with the School of Computer Science, Shaanxi Normal University, China. He has authored over 60 papers in international journals and conferences. He received five best paper awards from KISM 2012, GreenCom 2013, MUE 2015, UCAWSN 2015, and CUTE 2016. He is a recipient of the IEEE Outstanding Leadership Award at CPSCom 2013 and the 2015 Chinese Government Award for Outstanding Self-Financed Students Abroad. His research interests include social computing, ubiquitous computing, big data analysis and processing and mobile cloud computing. He is a member of ACM, KIPS and CCF.

Dae-Soo Sim

He is currently an undergraduate student with the Department of Computer Software Engineering, Soonchunhyang University, Korea. His interests include data mining, complex network analysis.

Doo-Soon Park

He received his Ph.D. in Computer Science from Korea University in 1988. Currently, he is a professor in the Department of Computer Software Engineering at Soonchunhyang University, Korea. He is Director of Wellness Service Coaching Center at Soonchunhyang University. He was President of KIPS (Korea Information Processing Society) from 15 to 2015, and Director of Central Library at Soonchunhyang University from 2014 to 2015. He was editor in chief of *Journal of Information Processing Systems* (JIPS) at KIPS from 2009 to 2012, and Dean of the Engineering College at Soonchunhyang University from 2002 to 2003. He has served as an organizing committee member of international conferences including FutureTech 2016, MUE 2016, WORLDIT 2016, GLOBAL IT 2016, CUTE 2015, CSA 2015. His research interests include data mining, big data processing and parallel processing. He is a member of IEEE, ACM, KIPS, KMS, and KIISE.

Hyung-Seok Seo

He received his Ph.D. in Microbiology from Soonchunhyang University in 2008. He also received MPH at Graduate School of Public Health, Yonsei University in 2005. He received the M.D. license in 1986 and became a urologic specialist in 1990 in South Korea. Currently he is a professor in the department of Sports Medicine and Director of Primary Healthcare Center at Konyang University, Korea. He served as a Medical officer at Western Pacific Regional Office (WPRO) of WHO in Manila, Philippines in 2008.