《智能信息处理》课程作业

基于形式概念分析的图相似性度量

董静阳

作业	分数[20]
得分	

2021 年 11 月 29 日

基于形式概念分析的图相似性度量

董静阳

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘　要　图是一种重要的信息组织结构，通常用于表示社交网络、互联网和其他互联网应用程序。本文研究了图之间相似性度量的一个基本问题，这是图搜索、匹配、模型发掘的基本步骤。为了有效地度量图之间的相似性，提出了一种基于形式概念分析的图之间相似性度量方法。我们还提供了一个案例研究，以证明建议方法的可行性。本文介绍了图相似的基本原理，介绍了形式概念格的相似性，并对用形式概念分析度量图相似的可行性进行了证明。

关键词　形式概念分析　图相似度　社交网络

中图法分类号　TP311.20　　DOI 号　10.3970/j.issn.1001-3695.2019.11.031

Graph Similarity Metric Based on Formal Concept Analysis

Dong Jingyang

(Computer science and technology，Dalian maritime university, LiaoningDalian,116026,China)

Abstract　Graph, an important information organizational structure, is commonly used for representing the social networks, web, and other internet applications. This paper tackles a fundamental problem on measuring similarity between graphs that is the essential step for graph searching, matching, pattern discovery. To efficiently measure the similarity between graphs, this paper pioneers a novel approach for measurement of similarity between graphs by using formal concept analysis that can clearly describe the relationships between nodes. A case study is provided for demonstrating the feasibility of the proposed approach. This paper introduces the basic principles of graph similarity, describes the similarity of formal concept lattices, and demonstrates the feasibility of using formal concept analysis to measure graph similarity.

Key words　Formal concept analysis; Graph similarity; Social networks;

1 介绍

大规模图论技术和普适计算范式的最新进展丰富了图数据和其他复杂网络系统的挖掘和分析，如蛋白质相互作用网络、社会网络和运输网络等。因此，了解网络的内部拓扑结构有助于从图中获得深刻的见解和知识。

相似子图匹配是指在同构的基础上寻找具有相似拓扑结构的子图结构。在许多实际应用中，一个给定的应用问题通常被转换成一个图，然后度量图之间的相似度。本文主要研究图之间的相似度评价问题。关于图形相似性度量的研究已经有很多。一种被广泛使用的方法，称为 GED，用于计算基本操作的开销：节点替换、节点插入。然而，这种方法通常是 np 难的，它的主要缺点是按照图编辑顶点数的指数计算复杂度。Yan 等人 [9]提出了一种基于特征的方法来处理图结构中的最近邻搜索。他们使用图数据库中的索引特征来过滤图形，而不进行成对相似性计算。

本文将利用形式概念分析方法对图之间的相似性度量进行了研究，这是一个开拓性的研究，可以弥补图挖掘和软计算之间的差距。本文的重点在于提出了一种有效的图之间相似性度量方法。首

先，我们根据修正的邻接矩阵构造给定图的形式背景；然后生成相应的形式概念格；由于所生成的形式概念格之间的相似性等价于图之间的相似性，因此获得图之间的相似性。

这篇文章的其余部分是这样组织的。第二部分会介绍已解决的问题，并提供解决方案的详细介绍。第三部分给出了形式概念格之间相似性的定义。第四部分通过一个案例阐述了计算图之间相似度的详细方法。第五节为结论。

2 问题的定义和解决思路

在提出问题陈述之前，首先给出了图、相似度等几个基本定义。然后，给出了所处理问题的形式。

定义 1（图）： 在图论中，一个图可以被数学化为一对 G = (V，E) ，其中 V 表示顶点集，E 表示由顶点对组成的边集。顶点 V 和边 E 的个数称为 V 和 E 的基数，通常用 n 和 m 或 |V| 和 |E|来表示。

邻接定义为顶点和边对： 两个顶点 u 和 v 是邻接的当且仅当 e_{uv}是图的边，而两条边 e_a和 e_b是邻接的当且仅当它们有一个共同的端点。

问题陈述（图相似性）： 给出两个图 $G_1(n_1, e_1)$和 $G_2(n_2, e_2)$，它们可能有不同数量的节点和边，以及图的节点之间的映射，问题是找到一个算法来计算两个图的相似性，表示为 sim(G_1, G_2)，并返回一个能很好捕捉直觉的相似度量。

解决方案：与现有的利用结构特征计算图形相似度的方法不同，本文的解决思路是基于形式概念分析来评价图形之间的相似度。具体的技术步骤如下：

1）根据以前的工作，用形式背景表示给定的图
2）为上述构造的形式背景建立形式概念格
3）计算形式概念格之间的相似度
4）将上述得到的相似性返回给图之间得到的相似性

3 形式概念格的相似性

这一部分主要介绍了形式概念格之间相似性的评估方法。让我们回顾一下形式概念分析的方法论。

定义 2（形式背景）:一个正式的背景可以组织为三元组 *K=(O,A,I)* ，其中 *O* 和 *A* 分别表示为对象集和属性集，并且 I⊆O×A 表示对象和属性之间的二元关系。例如，$o \in O, a \in A, (o,a) \in I$可以解释为对象 o具有属性 a。

定义 3 :在形式概念分析方法论中，定义了两个关键操作符，对于 $X \subseteq O$，我们给出了 X 的一组常见属性,$X^{\uparrow} = \{a \in A | (x,a) \in I, \forall x \in X\}$;并且对于 $Y \subseteq A$，我们还定义了一组 Y的共同对象，$Y^{\downarrow} = \{o \in O | (o,y) \in I, \forall y \in Y\}$。

定义 4 （概念）:在形式背景 *K=(O,A,I)*，对于 $X \subseteq O$，$Y \subseteq A$，如果$X^{\uparrow\downarrow}$=Y，则（*X,Y*）可以被称为概念。

定义 5 （概念格）:在一个形式背景 *K=(O,A,I)* 中，概念格 *L(O,A,I)* 定义为按照一个特殊的层次偏序组织的概念。

在详细介绍了形式概念分析方法学的基本知识之后，在下面定义概念格之间的相似度函数，

定义 6（相似度函数）: L_{A}，L_{B}表示为概念格，相似度定义为 L_{A}，L_{B}中节点间相似度的平均值，因此，它的形式如下，

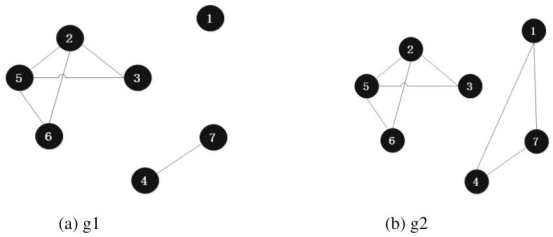
$$sim(L_{\mathrm{A}}, L_{\mathrm{B}}) = \frac{\sum_{C_i \in L_A} sim(C_{\mathrm{i}}, L_{\mathrm{B}})}{n}$$

其中$sim(C_{\mathrm{i}}, L_{\mathrm{B}}) = \max\left(\frac{\sum_{l \in R_i} sim(C_i, l)}{n}\right)$，$R_i$ 表示描述概念C_i的路径集。

4 提出的方法以及案例研究

在这一部分，我们从技术上阐述了提出的基于形式概念分析的图之间相似度的度量方法。正如前面在解决方案的想法中提到的，一个案例研究被用来说明所提出的方法的工作原理。

给出两个图g_1和g_2,如图 1 所示。每个图都包含 7 个节点，但是它们的拓扑结构是不同的。这个例子的目的是返回g_1和g_2的相关度$sim(g_1, g_2)$。



(a) g1　　(b) g2

图 1　g_1和g_2的可视化结构

4.1 构建图的形式背景

通过使用[4]中的构造方法，很容易获得形式背景，结果如图 2 和图 3 表示。

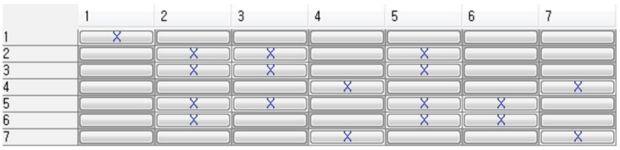


图 2　g_1的形式背景

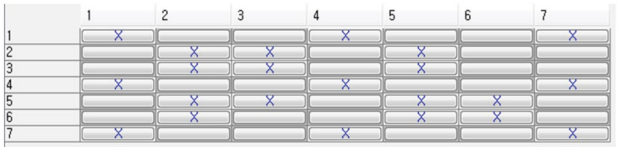


图 3　g_2的形式背景

4.2 构建形式概念格

根据文献[4]中提出的形式概念格生成算法，两个图的形式概念格分别显示在图 4 中。

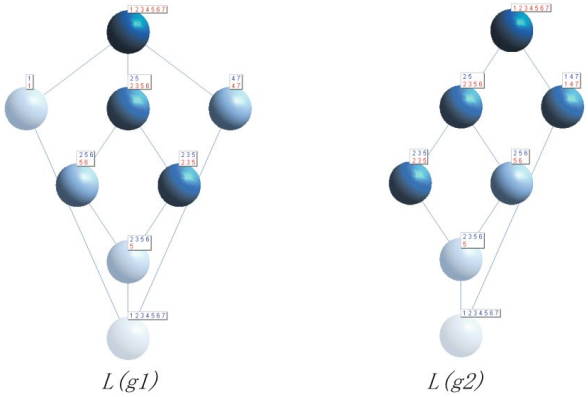


图 4　g_1和g_2的形式概念格

4.3 计算概念格之间的相似度

根据定义 6，可以很容易地计算概念格之间的相似度。从图 4 中可以看到，概念格g_1和g_2的区别主要在于两个位置，分别为g_1 中的$<\{1,2,3,4,5,6,7\},\{\ \}>-><\{1\},\{1\}>-><\{\},\{1,2,3,4,5,6,7\}>$和 g_2中的$<\{1,2,3,4,5,6,7\},\{\ \}>-><\{1,\ 4,\ 7\},\{1,\ 4,\ 7\}>-><\{\},\{1,2,3,4,5,6,7\}>$。因此，两个概念格的相似性是

$$sim(L(g_1),L(g_2)) = \max\left(\frac{\sum_{C_i\in(Lg_1)} sim(C_i,L(g_2))}{n}\right)=0.667$$

得到概念格之间的相似度后，我们可以说它等价于图之间的相似度。换句话说，g_1和g_2之间的相似度$sim(g_1,g_2)$为 0.667。

5 结论

基于形式概念分析的评价方法是一种新的基于形式概念的评价方法。该方法首先为给定的两个图构造形式背景，然后构造相应的形式概念格，最后定义概念格的相似度函数，并采用相似度函数来度量图的相似性。我们亦会进行个案研究，以证明建议方法的可行性。重要的是，我们的方法可以清晰地刻画节点之间的关系，并通过计算节点之间的相似度进一步返回图之间的相似度。

结束语

随着数据库系统的广泛应用和网络技术的高速发展，数据库技术也进入一个全新的阶段。我们面临着称为的“信息丰富而知识贫乏”窘境。数据库在给我们提供丰富信息的同时，也体现出明显的海量信息特征。信息爆炸时代的当下，图作为一种用于表示社交网络和其他互联网应用程序重要的信息组织结构受到越来越多的关注 ，对图相似相关问题的研究也变得至关重要。本文开创性的将形式概念分析运用到图的相似度量中，并用实验结论证明了其可行性。将来随着形式概念分析和图相关理论的进一步完善和发展，该方法必定能够发挥它应有的作用。

参 考 文 献

[1]. Bu, D. , Zhao, Y. , Cai, L. , Xue, H. , Zhu, X. , Lu, H. , Zhang, J. , Sun, S. , Ling, L. , Zhang, N. , Li, G. , & Chen, R. . (2003). Topological structure analysis of the protein-protein interaction network in budding yeast.. 31(9), 2443-2450.

[2]. Kumar, Ravi, et al. "Structure and evolution of online social networks." ACM Knowledge Discovery and Data Mining (2006): 611-617.

[3]. tong, lu, et al. "Transportation network design for maximizing space–time accessibility." 81 (2015): 555-576.

[4]. Hao, Fei, et al. "K-Clique Community Detection in Social Networks Based on Formal Concept Analysis.." IEEE Systems Journal 11 (2017): 250-259.

[5]. Hao, F. , Park, D. , Min, G. , Jeong, Y. , & Park, J. . (2016). k-Cliques mining in dynamic social networks based on triadic formal concept analysis.. 209(C), 57-66.

[6]. Hao, F. , Yau, S. , Min, G. , & Yang, L. . (2014). Detecting k-Balanced Trusted Cliques in Signed Social Networks. 18(2), 24-31.

[7]. hao, fei, and shengtong zhong. "Tag recommendation based on user

interest lattice matching." (2010).

[8]. Zeng, Zhiping, et al. "Comparing stars: on approximating graph edit distance." 2 (2009): 25-36.

[9]. Yan, Xifeng, et al. "Feature-based similarity search in graph structures." 31 (2006): 1418-1453.

[10]. Girvan, M. , & Newman, M. . (2002). Community structure in social and biological networks.. 99(12), 7821-7826.

[11]. Hu Xuegang,Wang DeYing，Liu Xiaoping，Guo Jun，Wang Hao，The Analysis on Model of Association Rules Mining Based on Concept Lattice and Apriori Algorithm[C].In: IEEE The Third International Conference on Machine Learning and Cybernetices， Shanghai，2002，8:1620-1624.