《智能信息处理》课程作业

基于形式概念分析方法的图相似

贾康

作业	分数[20]
得分	

2021 年 11 月 30 日

基于形式概念分析方法的图相似

贾康

（大连海事大学 信息科学与技术学院 大连 116026）

摘 要 许多现实世界的应用信息都是用图形结构组织和表示的，图形结构通常用于表示各种无处不在的网络，如万维网、社交网络和蛋白质-蛋白质交互网络。特别是，图之间的相似性评估在图搜索、模式发现、神经科学、化合物探索等许多领域都是一个具有挑战性的问题。有一些基于顶点或边属性的算法被提出来解决这个问题。然而，这些算法不同时考虑顶点和边的相似性。为此，本文开创了一种基于形式概念分析的图间相似性评价新方法。这种方法的特点是能够描述节点之间的关系，并进一步揭示图之间的相似性。因此，我们方法的重点是同时考虑顶点和边。通过一个实例对该算法进行了评估，验证了该算法在检测和度量图间相似性方面的有效性。

关键词 形式概念分析，图结构，相似性评估

Graph similarity based on formal concept analysis

Kang Jia

(Department of Information Science Technology, Dalian Maritime University, Dalian 116026)

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 structure, Similarity Evaluation

1 介绍

随着大数据技术的飞速发展和强大的普适计算能力，海量图分析与挖掘的研究为复杂网络系统打开了又一扇新的大门。在大规模图形分析和挖掘的推动下，各种现实世界的应用正在生物科学、社交媒体和交通领域涌现[1-3]。因此，在海量图的内部拓扑结构中隐藏着更多有趣的点和知识。

在现有的海量图分析和挖掘技术中，子图匹配技术是根据图之间的相似性来检测同构子图结构。子图匹配技术的工作原理是：对于给定的两个子图 g_1 和 g_2，计算 g_1 和 g_2 之间的相似度 $sim(g_1,g_2)$

图1显示了某一新生产药品的功能识别激励示例。

为了探索这种新药的功能，传统的临床医学方法是在动物和人类身上进行试验。不幸的是，识别药物的功能通常需要很长时间。幸运的是，图形相似性搜索正在成为解决这个问题并为我们节省大量时间的主要技术解决方案。如图 1 所示，我们的靶向药物的分子结构位于最左侧。在图相似性搜索问题中，它被看作是一个查询 q，然后图相似性

搜索算法将评估q与医学数据库中现有的图，即 g_1 和 g_2（分子结构）之间的相似性。因此，这个问题的实质是图之间的相似性评估。

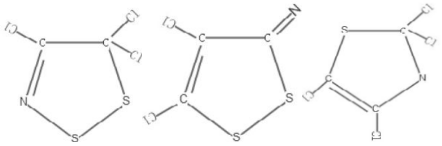

图 1　某新药功能鉴定的激励性示例

针对上述面向查询的图间相似度评估问题，现有的典型工作主要集中在基于核函数的图间相似度评估方法和基于特征的图间相似度评估方法。在[4,5]中，作者用核函数描述了两个图间的相似度。工作的基本思想[6]就是将领域知识结合起来，通过提取拓扑结构来获取特征。然后，他们根据两个图的公共拓扑结构进一步计算两个图之间的相似度，但是这两种评估方法没有考虑图的全局结构连通性信息。为了解决这个缺点，这项工作提供了一种基于形式化概念分析的评估方法来获取图之间的相似性。与现有的大多数方法不同，形式概念分析方法作为描述对象和属性的有效数学工具，存储了图形的全局和局部信息。它首先将给定的图表示为形式上下文。在此基础上，建立了相应的概念格，并设计了一种相似性评价方法。主要贡献总结如下：

（1）（图表示为形式上下文）由于本文采用了形式概念分析方法，我们首先需要将图表示为形式上下文。从技术上讲，给定图的形式上下文很容易通过[7-9]中提出的改进邻接矩阵构造；

（2）（相似性评价特征构造）通常，图之间的相似性评价通常需要特征。本文构造了相应的形式概念格，作为进一步评价图之间相似性的特征。

（3）（等价定理）基于我们的研究和证明，很容易得到图上的相似性与生成的形式概念格上的相似性之间的等价定理。

本文的其余部分结构如下。第 2 节正式描述了解决的问题。第 3 节详细介绍了建议的方法。第4节展示了一个案例研究。最后，第 5 节总结了本研究。

2 问题陈述

本节重点描述图之间的相似性评估问题。问题陈述正式呈现如下。

问题陈述（基于形式概念分析的图相似性）[18]分别针对两个给定图 $G_1(V_1, E_1)$（包括$|V_1|=n_1$ 个顶点，$|E_1|=e_1$ 边）和 $G_2(V_2, E_2)$(包括$|V_2|=n_2$ 个顶点，$|E_1|=e_2$ 边）以及节点之间的链接。该问题的目的是提出一种评估两个图的相似性的算法，即 $sim(G_1,G_2)$。由于本文试图通过形式概念分析来解决图之间的相似性评估问题，因此该问题进一步表述为：对于两个给定的 $G_1(n_1, e_1)$ 和 $G_2(n_2, e_2)$，如何用形式上下文 K=(O, A, I)来表示它们，然后研究由形式概念格 L 形成的概念之间的相似性 sim（L_A,L_B）。

表 1 论文中使用的重要变量

Notation	Description
$G1(V_1, E_1)$	A graph G, with n_1 vertices and e_1 edges
$G2(V_2, E_2)$	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

3 方法

本节介绍了通过形式概念分析方法评估图之间相似度的方法的工作过程。

3.1 形式概念格之间的相似性评估

形式概念格之间的相似性评估是本文的一项关键技术。因此，本节重点阐述如何计算由两个给定图生成的形式概念格之间的相似性。

形式概念分析作为描述对象与属性之间二元关系的一种强有力的方法，已被应用于许多领域。形式上，形式上下文表示为 C=（O，A，I），其中 O 表示对象集，a 分别表示属性集，关系 $I \subseteq O \times A$ 是对象和属性之间的二进制关系。一般来说，$o \in O$ 和 $a \in A$，$(o, a) \in I$ 被解释为目标 o 具有属性 a。

为了更好地解释形式概念格及其生成

的形式概念，给出了以下两个算子。

（用于提取对象子集 X 的公共属性的运算符）[18]用于 X ⊆ O、 我们定义了一组 X 的公共属性,

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

（用于提取属性子集 X 的公共对象的运算符）[18]对于 Y C A，我们还定义了一组 Y 的公共对象，

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

通常，对于给定的形式上下文 C=(O，A，I)，如果 $X^{\uparrow\downarrow}$=Y，则一对（X，Y）被称为概念。注意，X 和 Y 分别被称为概念的范围和意图。使用上述运算符，可以生成概念格 L(O，A，I)，其中包括可以根据特殊的层次偏序组织的概念。

（相似度函数）[18,19]设 L_A，Ls 为概念格，L_A，Ls 中节点之间的相似度形式化如下，

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

其中,

$$\text{sim}\,(C_i,L_B)=max\left(\frac{\sum_{l\in R_i}\text{sim}\,(C_i,l)}{n}\right)$$

R_i指描述概念 C 的路径集。

3. 2 基于形式概念分析的图间相似性评估

基本上，建议的方法由以下步骤组成（如图 2 所示）。显然，这两个图是我们方法的输入。然后，进入步骤 1 构建形式上下文，然后，如步骤 2 所示，相应地生成形式概念格；评估生成的形式概念格之间的相似性（步骤 3），以帮助评估图之间的相似性。

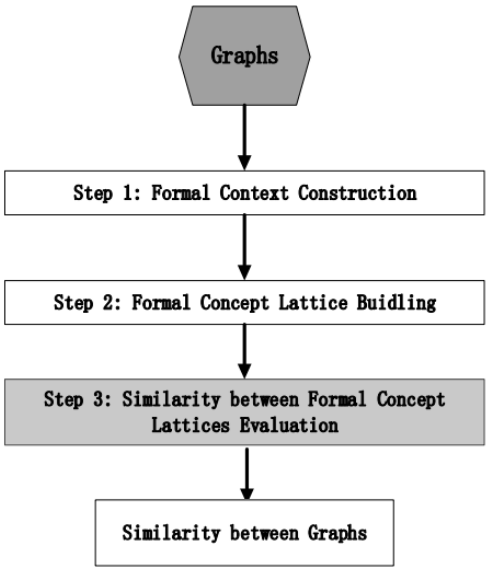

图 2 工作流程图

3.2.1 正式语境构建

根据我们以前的工作[7,8,18]，形式上下文很容易获得。构造方法的基本思想是将顶点作为对象和属性。然后，对两个给定的图 g_1 和 g_2 采用改进的邻接矩阵构造形式上下文。形式上，形式上下文表示为

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

其中 V 是图中的顶点，因此形式上下文 C 与传统上下文相比是一个特殊的上下文。我们表示 C（g_1）和 C（g_2）是 g_1 和 g_2 构造的形式上下文。

3.2.2 生成形式概念格

根据形式概念格的生成算法[7,9]，两个图 g_1、g_2 的格分别生成为 L（C（g_1））, L（C（g_2））。

3.2.3 评估概念格之间的相似度

到目前为止，我们可以根据第 3.1 节中定义的相似度函数来评估格之间的相似度。因此，两个格之间的相似度可以计算为

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

目前，一旦我们得到概念格之间的相似度，就可以等价地得到图之间的相似度。也就是说，对于 g_1 和 g_2 之间的相似性，表示为 sim(g_1，g_2)，则以下等价关系成立。

$$\text{sim}\,(g1,g2) \equiv \text{sim}\,(L((C(g1)),L(C(g2)))$$

4 研究

在本节中，我们采用了[20]给出的一个关于中国高点击率网站的有用案例.本案例描述如下：给出了中国一些点击率较高的网站，如百度、谷歌、新浪、网易、优酷、淘宝、京东、当当等。从形式上讲，本案例研究中的每个网站都被视为一个节点，每个链接都被视为一个图的边。我们可以建立以下两个图 g_1 和 g_2，如图 3 所示。

通过使用上述方法，我们可以很容易地得到以下两个形式概念格，分别对应于 g_1 和 g_2。

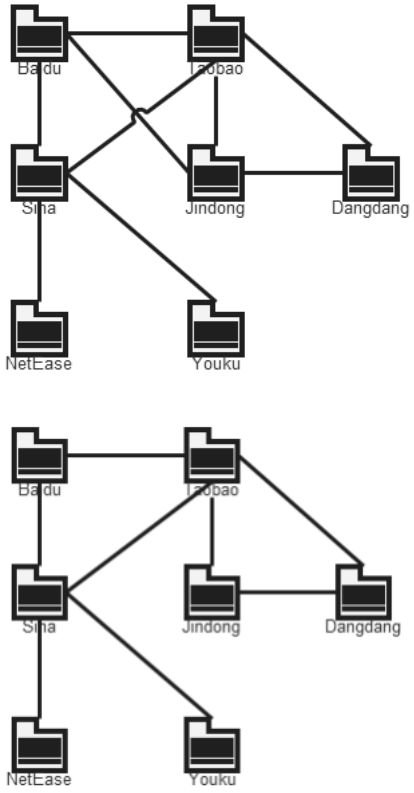

图 3　g_1 和 g_2 的构建

显然，我们得到 g_1 和 g_2 之间的相似性为 0.889。

在整个案例研究中，从可行性和有效性方面进一步验证了所提出的方法。所提出的方法将用于各种大型复杂图形相关应用，如社交网络分析、Web 挖掘。此外，随着领域知识的不断丰富，图对象分类、子图搜索变得越来越有意义。

5 结论

在模式搜索、目标跟踪和生物复合体识别等领域，图形相似性评价技术正成为一种很有前途的技术。为了评估两个图之间的相似性，本文提出了一种新的基于形式概念分析的方法。首先，该方法分别为给定的两个图构造形式上下文；然后相应地生成它们的形式概念格；最后，我们为概念格定义了一个相似度函数来评价图的相似度。以中国高点击率网站网络为例，对该方法进行了性能评估。可以清楚地得出结论，我们提出的方法可以有效地描述节点之间的关系，并通过计算给定图的形式概念格中出现的节点之间的相似度来进一步获得图之间的相似度。

参考文献

[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 (TOD

S), 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-BusinessEngineering,Beijing,China,2015,pp.180-185.