



基于特征子图的异构信息网络节点相似性度量^{*}

张彪^{1,2,3}, 李川^{1,2,3}, 徐洪宇¹, 李艳梅¹, 杨宁¹, 罗谦⁴

(1. 四川大学计算机学院 成都 610065; 2. 国家空管自动化系统技术重点实验室 成都 610065;
3. 武汉大学软件工程国家重点实验室 武汉 430072; 4. 中国民用航空总局第二研究所 成都 610065)

摘要:为解决异构信息网络相似性度量的问题,提出了基于节点特征子图的节点相似性度量算法,通过节点特征子图的最大公共子图与最小公共超图之间的差异性,进行节点间的相似性度量。该算法以图理论为基础,根据连边的不同类型设定不同权值,在考虑节点信息相似的同时,加入节点在网络中的结构信息,最大程度地利用了异构信息网络所富含的信息。实验结果表明,提出的算法具有较好的性能和有效性。

关键词:异构信息网络;图相似;相似性度量;特征子图

doi: 10.3969/j.issn.1000-0801.2014.11.012

Heterogeneous Information Networks Node Similarity Measurement Based on Feature Sub-Graph

Zhang Biao^{1,2,3}, Li Chuan^{1,2,3}, Xu Hongyu¹, Li Yanmei¹, Yang Ning¹, Luo Qian⁴

(1. College of Computer Science, Sichuan University, Chengdu 610065, China;
2. National Key Laboratory of Air Control Automation System Technology, Chengdu 610065, China;
3. State Key Laboratory of Software Engineering of Wuhan University, Wuhan 430072, China;
4. The Second Research Institute of CACC, Chengdu 610065, China)

Abstract: To solve the problem in measuring the similarity of heterogeneous information networks, a similarity measuring algorithm was proposed. It calculates the difference between the maximum common sub-graph and minimum common hyper-graph, based on feature sub-graph of the current node. The algorithm takes graph theory as its foundation, set different weight to different kinds of edges, considers nodes information as well as graph to topological information, and makes full use of the information in heterogeneous network. The result shows that the proposed algorithm has wonderful effectiveness and efficiency.

Key words: heterogeneous information network, graph similarity, similarity measurement, feature sub-graph

1 引言

随着信息技术的不断发展,对基于图论的复杂网络进行深入研究已成为可能,复杂网络的相关研究也逐渐成为学术界关注的热点,信息网络(info-network)作为复杂网络的重

要特殊形式,为复杂网络问题求解提供了一个有效途径。

相比信息网络中的同构信息网络,异构信息网络包含更为丰富和复杂的信息。学术网络是典型的异构信息网络实例,网络中包含作者、论文、会议或刊物、研究领域、关键词等节点类型,这些节点连接在一起,构成了错综复杂的

^{*} 国家自然科学基金资助项目(No.61103043, No.61173099, No.U1233118),国家“十二五”科技支撑计划基金资助项目(No.2012BAG04B0),武汉大学软件工程国家重点实验室开放基金资助项目(No.SKLSSE2012-09-26)

关系,形成了包括作者—论文、论文—会议或刊物、作者—研究领域、论文—关键词等多种类型的边。

节点相似性度量在数据挖掘中应用非常广泛,如分类、聚类、节点排名等问题,通过计算节点之间的相似性,可以进行 Top-K 的相似性节点查询,利用节点的相似性完成协同过滤等数据挖掘任务。然而,目前大部分的节点相似性度量都是基于同构网络,在计算节点的相似度时并没有考虑不同路径所具有的不同语义,具有一定局限性,因而,不能直接采用针对同构网络节点的相似性度量方法来解决异构信息网络的节点相似性度量问题。

本文提出一种基于节点特征子图的节点相似性(feature sub-graph based node similarity, SUBFG-SIM) 度量算法,通过计算节点特征子图的最大公共子图与最小公共超图之间的差异性来度量节点间的相似性。该方法以图理论为基础,根据连边的不同类型设定不同权值,并且在考虑节点信息相似的同时,加入节点在网络中的结构信息,最大程度地利用异构信息网络所富含的信息。

2 相关工作

Han 和 Yu 等人在参考文献[1,2]中正式提出和倡导信息网络这一新概念,给出了信息网络的形式化定义和性质。参考文献[3]首次提出了基于图数据集的信息网络在线分析处理问题,提出并定义了信息维、拓扑维的概念以及信息网络探索性的设计和框架。参考文献[4]提出了基于 Graph-OLAP 建模的双星模型,该模型较好地解决了基于 Graph-OLAP 的数据立方体建模问题。

对于同构图中节点相似度的研究已经有很长的历史,计算节点相似度的常用度量有 Jacard 距离^[9]、cosine 相似度^[5]等。最近的相关研究中,参考文献[6]提出了 PPR(personalized page rank)算法,PPR 算法是一种非对称的度量方法,通过度量从定点 A 出发通过随机游走模型到定点 B 的概率来表示两个定点之间的相似性。参考文献[7]提出 SimRank 算法,SimRank 是基于图中节点的结构上下文相似度的算法,其中,结构上下文相似度通过迭代计算两个节点邻居的相似度得到。然而这些方法都是基于同构网络提出的度量,并不能直接用于异构信息网络。

针对异构网络,参考文献[8]提出了 ObjectRank 算法,并首次表明了网络中的异构链接会对随机游走模型结果产生影响。然而 ObjectRank 算法在为各种不同的异构边分配权重时,采用了固定的方式,不能体现用户对不同类型

连边的喜好程度。参考文献[5]提出了一种基于元路径(meta-path)的异构信息网络节点相似性度量算法,该算法基于对节点在网络中路径的分析,首次提出元路径的概念,通过计算节点之间元路径的相似性来度量两个节点的相似度。然而该方法路径形式过于单一,未能有效利用图论处理网络问题,并且该方法对不同连边赋予相同的权重,与现实网络中不同连边具有不同影响力的特征不符。

3 异构信息网络的图相关定义

异构信息网络的图相关定义如下。

(1) 异构信息网络

设网络 $G(V, E, T, R, \varphi, \psi)$ 的所有节点属于有限集合 V , 所有边属于有限集合 E , 节点类型有限集合为 T , 边类型有限集合为 R , 若对所有的节点和节点类型存在函数 φ 满足 $\varphi: V \rightarrow T$, 且对于所有边和边类型存在函数 ψ 满足 $\psi: E \rightarrow R$ 。当 $|T| > 1$ 或者 $|R| > 1$ 时,称网络 G 为异构信息网络;当 $|T| = 1$ 并且 $|R| = 1$ 时, G 为同构信息网络。为了便于描述,如无特殊说明,本文后续提到的图网络均指异构信息网络图。

(2) 异构信息网络中的子图

对于图 G_1 和 G_2 , G_1 是 G_2 的子图,表示为 $G_1 \in G_2$, 必须满足: $V_1 \in V_2, E_1 \in E_2, T_1 \in T_2, R_1 \in R_2, \varphi_1(x) = \varphi_2(x) \quad \forall x \in V_1, \psi_1(x) = \psi_2(x) \quad \forall x \in E_1$ 。对于 G_1 和 G_2 , 若 G_1 是 G_2 的子图,则 G_2 被称为 G_1 的超图。

(3) 异构信息网络中的公共子图

如果图 $\tilde{G}$ 是 G_1 与 G_2 的公共子图,需要满足: $\tilde{G} \in G_1$ 并且 $\tilde{G} \in G_2$ 。若不存在比 $\tilde{G}$ 更大的 G_1 与 G_2 的公共子图,则 $\tilde{G}$ 为 G_1 与 G_2 的最大公共子图。

(4) 异构信息网络中的公共超图

如果图 $\tilde{G}$ 是 G_1 与 G_2 的公共超图,需要满足: $G_1 \in \tilde{G}$ 并且 $G_2 \in \tilde{G}$ 。若不存在比 $\tilde{G}$ 更小的 G_1 与 G_2 的公共超图,则 $\tilde{G}$ 为 G_1 与 G_2 的最小公共超图。

(5) 异构信息网络中的图同构

图 G_1 与 G_2 同构,可表示为 $G_1 \cong G_2$,必须存在一个双射函数 $f: V_1 \rightarrow V_2$,使得 G_1 和 G_2 满足: $f(v'_1) = v'_2, f(e'_1) = f((v'_1, u'_1)) = (v'_2, u'_2) = e'_2, T_1 = T_2, R_1 = R_2, f(\varphi_1) = \varphi_2, f(\psi_1) = \psi_2$ 。

异构信息网络图同构的条件中,未对节点或者边所带的信息做任何限定,但要求节点类型和边类型一一对应,这与传统图同构有一定区别,传统图同构只要求结构上相同即满足条件,异构信息网络由于存在多种类型的节点和



边,除在结构上有要求外,对节点和边类型也进行了限定。

(6) 异构信息网络中的公共共同构子图

如果图 $\hat{G}$ 是 G_1 与 G_2 的公共共同构子图,需要满足: $\hat{G}_1 \in G_1, \hat{G}_2 \in G_2, \hat{G}_1 \cong \hat{G}_2 \cong \hat{G}$ 。若不存在比 $\hat{G}$ 更大的 G_1 与 G_2 的公共共同构子图,则 $\hat{G}$ 为 G_1 与 G_2 的最大公共共同构子图。

(7) 异构信息网络中的公共共同构超图

如果图 $\tilde{G}$ 是 G_1 与 G_2 的公共共同构超图,需要满足: $G_1 \in \tilde{G}_1, G_2 \in \tilde{G}_2, \tilde{G}_1 \cong \tilde{G}_2 \cong \tilde{G}$ 。若不存在比 $\tilde{G}$ 更小的 G_1 与 G_2 的公共共同构超图,则 $\tilde{G}$ 为 G_1 与 G_2 的最小公共共同构超图。

4 异构信息网络图相似性度量

本节提出了基于超图和子图差异性比较的异构信息网络图相似性度量。将异构信息网络图相似性度量分为信息相似性度量与结构相似性度量两类,其中,信息相似性度量要求考虑节点和边所属的类型以及其带有的信息,而结构相似性度量只关注网络的拓扑结构,即只考虑节点和边的所属类型,而不关注其带有的相关信息。

对于图 G_1 和 G_2 , G_1 与 G_2 的最小公共超图 $\tilde{G}$ 与最大公共子图 $\hat{G}$ 的差异性被定义为异构信息网络中图之间的信息相似度(graph information similarity, GISIM),因此, G_1 与 G_2 的信息相似度可用式(1)表示。

$$GISIM(G_1, G_2) = Diff(\tilde{G}, \hat{G}) = \sum_{i=1}^{|\tilde{T}|} \frac{|t'_i|}{|t_i|} m_i + \sum_{j=1}^{|\tilde{R}|} \frac{|r'_j|}{|r_j|} n_j \quad (1)$$

其中, $|\tilde{T}|$ 表示 $\tilde{G}$ 中的节点类型数, $t_i \in T$, $|t_i|$ 表示 $\tilde{G}$ 中类型为 t_i 的节点个数, $|\tilde{R}|$ 表示 $\tilde{G}$ 中边的类型数, $r_j \in R$, $|r_j|$ 表示图 $\tilde{G}$ 中类型为 r_j 的边数, $|r'_j|$ 表示 $\hat{G}$ 中类型为 r'_j 的边数, m_i 表示类型为 t_i 的节点在计算相似性度量时的权重, n_j 表示类型为 r_j 的边在计算相似性度量时的权重, m_i 和 n_j 满足 $0 \leq m_i \leq 1, 0 \leq n_j \leq 1$ 和 $\sum_{i=1}^{|\tilde{T}|} m_i + \sum_{j=1}^{|\tilde{R}|} n_j = 1$ 。

对于图 G_1 和 G_2 , G_1 与 G_2 的最小公共共同构超图 $\tilde{G}$ 与最大公共共同构子图 $\hat{G}$ 的差异性被定义为异构信息网络中图之间的结构相似度(graph structural similarity, GSI),因此, G_1 与 G_2 的结构相似度可用式(2)表示。

$$GSI(G_1, G_2) = Diff(\tilde{G}, \hat{G}) = \sum_{i=1}^{|\tilde{T}|} \frac{|t'_i|}{|t_i|} m_i + \sum_{j=1}^{|\tilde{R}|} \frac{|r'_j|}{|r_j|} n_j \quad (2)$$

式(2)和式(1)具有相同的表述形式,但式(2)所用的超图和子图为同构公共超图与同构公共子图。

5 基于节点特征子图的异构信息网络相似性度量算法

异构信息网络中存在多种类型的节点,节点间连接着错综复杂的边,确定一个节点在网络中的意义以及找出一类节点类型与网络中其他节点类型之间的关系,对于研究异构信息网络非常重要。传统图论中节点间的度、中心性等都是衡量节点间关系的重要指标,其主要依据是节点间的结构信息,而在异构信息网络中存在很多类型节点通过其他节点建立联系的情况,传统单一的度量方法不适用于异构信息网络。本节针对异构信息网络的特点,提出基于节点特征子图的 SUBFG-SIM 度量算法。

在网络 G 中能够连接两种节点类型 t_1 和 t_{L+1} 的模式 $M(t_1, t_{L+1}) = t_1 \xrightarrow{r_1} t_2 \xrightarrow{r_2} \cdots \xrightarrow{r_L} t_{L+1}$,若满足条件: $r_i \in R$ 且 $1 \leq i \leq L$; $t_i \in T$ 且 $1 \leq i \leq L+1$; $\forall i, j$ 满足 $t_i \neq t_j$,且 $1 \leq i \leq L+1, 1 \leq j \leq L+1$,则模式 M 被称为异构信息网络节点间关联模式。其中,连接节点的边类型数量 L 表示该模式的长度,连接两种类型的所有关联模式中, L 最小的模式为最短关联模式,表示异构信息网络中不同类型节点间最密切的联系程度。

在网络 G 中连接两种不同类型节点 v^{t_1} 和 v^{t_2} 的路径 $P = v_1 \xrightarrow{r_1} v_2 \xrightarrow{r_2} \cdots v_l \xrightarrow{r_l} v_{l+1}$,若路径 P 满足节点类型 t_1 和 t_2 的关联模式,则路径 P 被称为节点 v^{t_1} 和 v^{t_2} 的关联路径;若路径 P 符合节点类型 t_1 和 t_2 间的最短关联模式,则路径 P 为节点间的最短关联路径,两个节点可能不存在最短关联路径。

在网络 G 中,与节点 v 之间存在最短关联路径的所有节点及最短关联路径中的边所构成的网络为异构信息网络节点特征子图(feature sub-graph, SUBFG),即 $G_F = \bigcup_{i=1}^n P_i$,其中 $\{P_1, P_2, \dots, P_n\}$ 表示节点 v 与其他类型节点间存在的所有最短关联路径的集合。

设网络 G 中同种类型节点 x 和 y 的特征子图分别为 G_x 和 G_y ,则基于特征子图的节点相似性(SUBFG-SIM)度量可表示为:

$$\begin{aligned} SUBFG-SIM &= GISIM(G_x, G_y) = Diff(\tilde{G}, \hat{G}) \\ &= \sum_{i=1}^{|\tilde{T}|} \frac{|t'_i|}{|t_i|} m_i + \sum_{j=1}^{|\tilde{R}|} \frac{|r'_j|}{|r_j|} n_j \end{aligned} \quad (3)$$

其中, $\tilde{G}$ 为 G_x 和 G_y 的最小公共超图, $\hat{G}$ 为 G_x 和 G_y 的最

大公共子图。相关因子计算方法与式(1)相同,虽然在计算 SUBFG-SIM 时,没有利用结构相似度计算图的结构相似度,但是在选取相关的特征子图时,已经利用了图的结构信息。

6 异构信息网络节点相似性度量相关算法

本节提出了基于 SUBFG-SIM 的异构网络 Top-K 相似节点查询算法以及基于 SUBFG-SIM 的社区发现算法。

网络中的节点相似性度量对网络节点间关系的研究以及对整个网络的研究都非常重要,Top-K 相似节点查询能够找出指定节点在网络中最相似的 K 个节点且给出排名。异构信息网络中 Top-K 相似节点查询算法如算法 1 所示。Top-K 相似查询算法的时间复杂度取决于查询类型节点的数量 n 以及节点特征子图中节点和边类型的数量 m , 时间复杂度为 $O(m \times n)$, 在异构信息网络中节点和边的类型远小于节点和边的数量,因此,相似查询算法的时间复杂度接近于 $O(n)$ 。

算法 1 异构信息网络 Top-K 相似节点查询

输入:异构信息网络 G , t 类型节点 V_0 , 查询数量 k , 节点类型与边类型的权重比 W ;

输出:最相似的 K 个节点及排名列表 $TopList$;

List = $\emptyset$;

vPathSet = getPathSet(G, V_0); //获取 V_0 所有最短关联路径

vFSG = getFSG(vPathSet, V_0); //获取节点 V_0 的特征子图

For each V_i in V_t Do

pathSet = getPathSet(G, V_i); //获取 V_i 的所有最短关联路径

FSG = getFSG(pathSet, V_i); //获取节点 V_i 的特征子图

minSuperG = getMinSuperG(vFSG, FSG); //获取最小公共超图

maxSubG = getMaxSubG(vFSG, FSG); //获取最大公共子图

FSGSim = getSim(minSuperG, maxSubG, W); //计算相似度

List.put(V_i , FSGSim); //将 V_i 和相似度放入列表

End For

TopList = $\emptyset$;

TopList = getTopList(List, k); //获取相似度前 k 的节点顺序列表

随着复杂网络研究的不断深入,针对复杂网络社区发

现的研究也备受关注。传统复杂网络的社区发现算法非常多,包括基于图分割、基于模块度、基于谱聚类、基于层次聚类的算法等^[9]。根据本文提出的异构信息网络节点相似性度量理论,可以将传统复杂网络的部分社区发现算法扩展到异构信息网络的社区发现中。本文采用基于层次聚类的社区发现算法。

对异构信息网络中同种类型但没有边相连的节点进行社区发现。基于层次聚类的社区发现算法在聚类过程中未限定聚类数量 k , 而是将聚类过程保存, 这样可以根据需要调整 k 值, 得到更符合要求的聚类结果, 具体过程见算法 2。

算法 2 基于层次聚类的社区发现算法

输入:异构信息网络 G , 社区发现节点类型 t , 社区发现个数 k , 节点类型与边类型权重比 W ;

输出:社区发现结果 Communities;

$V_t = \text{getV}(G, t)$; //获取信息网络 G 中类型为 t 的所有节点

simMatrix = initializeSimMatrix(V_t); //初始化相似矩阵

clusters = initializeClusters(simMatrix); //初始化类簇

clusterSet = $\emptyset$; //层次聚类过程

repeat

minClusters(i, j) = getMinClusters(); //获取最相近的两个类簇 i 和 j

cluster(k) = mergeClusters(i, j); //合并簇

clusters.add(cluster(k)); //合并的新簇加入类簇中

clusters.remove(cluster(i)); //移除已合并簇

clusters.remove(cluster(j)); //移除已合并簇

clusterSet.add(i, j, k); //记录层次聚类过程

until clusters.size.equal(1); //直到所有簇合并结束

Communities = getClusters(clusterSet, k); //根据层次聚类获取 k 个社区

7 实验分析

7.1 实验数据及实验平台

本文实验平台中的处理器采用 Intel 的酷睿 i5 4258U, CPU 主频为 2.4 GHz, 操作系统为 MacOS 10.9, 内存为 8 GB, 硬盘为 256 GB 固态硬盘, 编程语言为 Java, 数据库为 MySQL 5.1.7。

实验采用 DBLP 数据集, 得到的异构学术信息网络包含 4 种类型的节点以及 3 种类型的边, 相关的统计数据见表 1 和表 2, 其中关键词由论文题目中出现的单词去除停



表 1 学术网络中节点类型及数量

	作者	论文	会议	关键词
fullDBLP	913 312	1 348 545	6 485	129 226
CCF DBLP	304 742	358 118	280	45 551

表 2 学术网络中边类型及数量

	作者—论文	论文—会议	论文—关键字
fullDBLP	4 246 559	1 348 545	9 518 924
CCF DBLP	1 113 476	358 118	2 501 663

用词操作后得到。本文采用全部数据(fullDBLP)和抽取中国计算机学会(CCF)推荐的计算机相关国际学术会议所涉及的数据集(CCF DBLP)分别进行多组实验。

7.2 基于特征子图的 Top-K 查询实验

本实验从 DBLP 数据集中选取具有代表性的作者,从全数据集中找出这些作者与各节点最相似的 10 个节点。

表 3 为相似作者的查询实验结果,本实验给出作者 Tang Changjie 和 Han Jiawei 在 fullDBLP 数据集中最相似的 10 个作者。为验证结果的正确性,在学术搜索引擎上分别搜索了排名前 10 的与 Tang Changjie 最相似的作者,结果表明这些研究者都是他的博士学生,属于同一组织机构,且以合作者的方式共同发表了诸多相似领域的论文。表 3 中与 Han Jiawei 最相似的 Philip S Yu 则是来自不同学校的数据挖掘领域的知名教授,他们与自己的学生以及相关学者在相似领域的会议上合作发表了诸多论文。

表 3 的右半部分列举了利用本文所提方法与 PathSim 方法分别计算出的与 Christos Faloutsos 相似的 Top-10 作者的结果,由于该方法作者所做实验数据集截至 2009 年,为保证对比效果,本文也将查询范围缩小到 2009 年。分析实验结果发现,二者找出的相似度 Top-10 作者集合有相当数量是重合的,由此证明了本文所提方法确实能有效度量两

个节点的相似程度。

7.3 社区发现算法实验

本文选用 CCF DBLP 数据集中的会议作为社区发现算法关注的节点类型,CCF DBLP 中共包含 10 个领域共 280 个会议。实验的评价方法为将社区发现结果与 CCF 对计算机相关会议所属领域的分类进行对比,以评价社区发现方法的有效性。本文将聚类所得到类簇中出现次数最多的领域作为该类簇的类别,类簇中的其他领域被认为是错误的划分,最后计算所有类簇中正确划分的会议数与会议总数的比值,得到社区发现算法的正确率。

基于层次聚类的聚类树如图 1 所示,聚类的正确率在图 2 中给出。由图 2 可知,类簇数量越多,社区发现的准确性越高,但当类簇数量达到 19 个时,会出现一个会议属于一类的情况,本文以此作为社区发现的临界点,选取 18 个类簇作为社区发现结果,此时社区发现准确率为 74.64%。通过对社区发现结果的分析可知,影响聚类准确率的因素主要是部分会议领域直接存在交叉甚至非常接近,例如 CCF 所给会议推荐目录将数据挖掘领域与人工智能两个领域分别列出,但实际上这两个领域非常接近,很多研究人员会同时活跃在这两个领域发表论文,因而,这两类会议在社区发现中区分难度较大。

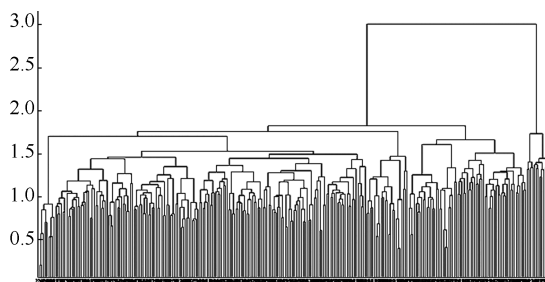


图 1 基于层次聚类的聚类树

表 3 前 10 名相似作者的查询实验结果

排名	Tang Changjie	Han Jiawei	SubFG-Sim	PathSim-AP	PathSim-PAP
1	Qiao Shaojie	Philip S Yu	Christos Faloutsos	Christos Faloutsos	Christos Faloutsos
2	Li Chuan	Pei Jian	Sun Jimeng	Spiros Papadimitriou	Han Jiawei
3	Duan Lei	Yan Xifeng	Spiros Papadimitriou	Sun Jimeng	Rakesh Agrawal
4	Zuo Jie	Sun Yizhou	Agma J M Traina	Pan Jiayu	Pei Jian
5	Zhang Tianqing	Aggarwal Charu C	Caetano Traina Jr	Agma J M Traina	Charu C Aggarwal
6	Qiu Jiangtao	Gao Jing	Pan Jiayu	Jure Leskovec	H V Jagadish
7	Peng Jing	Fan Wei	Jure Leskovec	Caetano Traina Jr	Raghu Ramakrishnan
8	Dai Shucheng	Wang Ke	Philip S Yu	Tong Hanghang	Nick Koudas
9	Liu Yintian	Jin Xin	Tong Hanghang	Deepayan Chakrabarti	Surajit Chaudhuri
10	Yang Ning	Wang Haixun	Dimitrios Gunopulos	Flip Korn	Divesh Srivastava

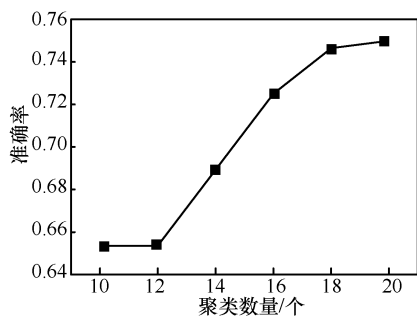


图2 不同聚类数时的聚类准确率

上述实验结果表明,本文提出的相似度模型能够有效地对异构信息网络节点间的相似程度进行度量。

7.4 实验性能分析

图3和图4给出了读取数据库时查询范围对 Top-K 查询时间消耗的影响。图3表明 Top-K 作者相似度查询时间随着查询范围的增大基本呈线性增长,而图4表明 Top-K 会议相似度查询速度会越来越快,并不是呈现线性增长,其原因在于读取部分数据后,已经将大部分作者等其他相关信息读入内存中,因而所需时间变少。图5和图6给出将所有信息读入内存时查询范围的变化对 Top-K 相似度查询时间消耗的影响。从图5和图6可以看出,查询时间随着查询范围的增加基本呈线性增长。实验表明,本文提出的 Top-K 查询能够适应大规模数据集的应用。

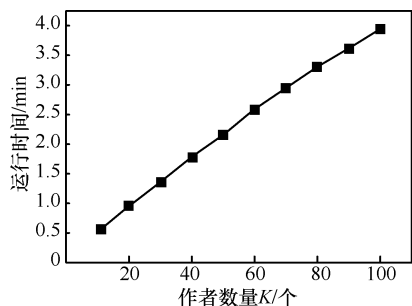


图3 读取数据库查询 Top-K 作者时间消耗情况

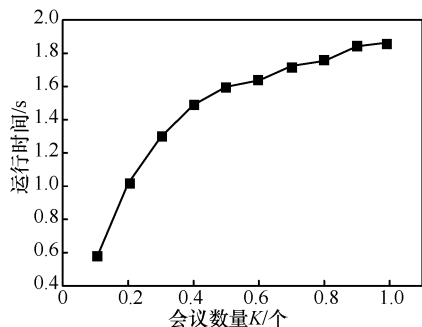


图4 读取数据库查询 Top-K 会议时间消耗情况

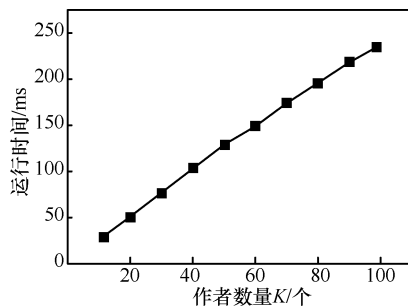


图5 读取内存查询 Top-K 作者时间消耗情况

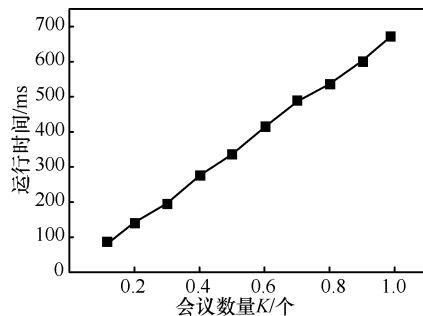


图6 读取内存查询 Top-K 会议时间消耗情况

8 结束语

本文首先提出了异构信息网络节点特征子图的概念,节点特征子图用来衡量节点在网络中与其他节点的关联以及结构;然后,基于异构信息网络图相似模型对节点特征子图进行相似度计算;最后,提出基于节点相似度的异构信息网络 Top-K 相似节点查询算法及基于层次聚类的异构信息网络社区发现算法。实验表明,基于节点特征子图的相似性度量能较好地反映节点间的相似程度。由于异构信息网络的图匹配问题与传统图匹配问题具有本质区别,并具有重要的研究意义,因此,将提出的异构信息网络中图相似度模型用于解决异构信息网络中的图匹配问题,将是未来工作的研究重点。

参考文献

- 1 Han J W, Yan X F, Yu P S. Scalable OLAP and mining of information networks. Proceedings of the 12th International Conference on Extending Database Technology, Saint Petersburg, Russia, 2009
- 2 Chen M S, Han J W, Yu P S. Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 866-883
- 3 Chen C, Yan X F, Zhu F D, et al. Graph OLAP: towards online analytical processing on graphs. Proceedings of the 8th IEEE International Conference on Data Mining, Pisa, Italy, 2008: 103-112
- 4 李川, 赵磊, 唐常杰等. Graph OLAPing 的建模、设计与实现. 软件学报, 2011, 22(2): 258-268

- 5 Sun Y Z, Han J W, Yan X F, *et al.* Pathsim: meta path-based top- k similarity search in heterogeneous information networks. Proceedings of the VLDB Conference, Seattle, Washington, USA, 2011
- 6 Glen J, Widom J. Scaling personalized web search. Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary, 2003: 1~35
- 7 Jeh G, Widom J. SimRank: a measure of structural-context similarity. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, 2002: 1~11
- 8 Balmin A, Hristidis V, Papakonstantinou Y. Objectrank: authority-based keyword search in databases. Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004: 564~575
- 9 Fortunato S. Community detection in graphs. Physics Reports, 2010, 486(3~5): 75~174
- 10 Biggs N, Lloyd E K, Wilson R J. Graph Theory 1736-1936. New York: Oxford University Press, 1976
- 11 Hamilton W R. Memorandum respecting a new system of roots of unity. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1856, 12(81):446
- 12 Han J, Sun Y, Yan X, *et al.* Mining knowledge from databases: an information network analysis approach. Proceedings of the ACM SIGMOD International Conference on Management of Data, Indianapolis, Indiana, USA, 2010: 1251~1252
- 13 Ji M, Han J, Danilevsky M. Ranking-based classification of heterogeneous information networks. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 2011: 1298~1306
- 14 Mashaghi A, Ramezanpour A, Karimipour V. Investigation of

protein complex network. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 41(1): 113~121

[作者简介]



张彪,男,四川大学计算机学院硕士生,主要研究方向为数据挖掘、信息网络、社会网络。

李川[通信作者],男,博士,四川大学副教授、硕士生导师,主要研究方向为数据库、数据挖掘等。近年来发表学术论文 50 余篇,主持或参与过多项国家及省部级项目。

徐洪宇, 男, 四川大学计算机学院硕士生, 主要研究方向为数据挖掘、信息网络、社会网络。

李艳梅,女,四川大学计算机学院硕士生,主要研究方向为数据挖掘、信息网络、社会网络。

杨宁,男,博士,四川大学讲师、硕士生导师,主要研究方向为数据库、数据挖掘。

罗谦,男,博士,现就职于中国民用航空总局第二研究所,主要研究方向为数据挖掘进化计算、企业智能计算。

(收稿日期:2014-08-12)

· 简讯 ·

2014 年 9 月通信业主要指标完成情况(一)

指标名称	单位	本年本月止累计到达	比上年同期累计(±%)	本月
电信营业收入	亿元	10 175.0	1.2	1 091.8
其中:电信主营业务收入	亿元	8 864.8	1.8	940.7
电信固定资产投资完成额	亿元	2 246.0	8.1	404.3
固定本地电话通话时长合计	万分钟	19 861 009.7	-13.5	2 199 682.7
区间电话通话时长	万分钟	19 90 101.2	-17.4	221 836.6
区内电话及拨号上网通话时长	万分钟	17 870 908.5	-13.1	1 977 846.2
固定长途电话通话时长合计	万分钟	4 011 236.0	-10.0	448 216.7
国内长途电话通话时长	万分钟	3 935 400.3	-9.6	439 974.2
国际电话通话时长	万分钟	43 320.0	-32.0	4 565.0
港澳台电话通话时长	万分钟	32 515.7	-20.0	3 677.5
移动电话通话时长合计(含本地)	万分钟	219 894 124.6	1.8	245 54 444.7
移动电话国内长途通话时长	万分钟	51 238 828.2	2.9	5 841 490.8
移动电话国际电话通话时长	万分钟	97 683.3	0.7	10 948.6
移动电话港澳台电话通话时长	万分钟	69 445.2	-4.5	7 663.7
移动短信业务量	万条	57 255 955.3	-17.9	6 640 191.1
移动互联网接入流量	万吉比特	139 936.6	57.2	18 593.0

注: (1) 收入增长率按可比口径计算; (2) 固定长途电话通话时长和移动电话通话时长均包含相应的 IP 电话通话时长; (3) 通话时长各项指标均为去话通话时长。

(来源:工业和信息化部运行监测协调局)