Journal of Computer Applications
计算机应用,2014,34(9):2604 - 2607,2638
ISSN 1001-9081
CODEN JYIIDU
2014-09-10
http://www.joca.cn

文章编号:1001-9081(2014)09-2604-04
doi:10.11772/j.issn.1001-9081.2014.09.2604

# 异构信息网中基于元路径的动态相似性搜索

陈湘涛,丁平尖*,王　晶
(湖南大学 信息科学与工程学院,长沙 410082)
(*通信作者电子邮箱 dingpingjian@sina.cn)

**摘　要**:现有的相似性搜索算法通常没有考虑时间因素,为此,提出一种异构信息网中基于元路径的动态相似性搜索算法 PDSim。PDSim 算法首先计算给定元路径下实体的链接矩阵,得到实体之间的元路径实例数比值,同时基于建立时间的不同,计算其时间差异度;在此基础上针对给定的元路径,获得异构信息网中动态相似性的度量。在多个相似性搜索实例中,PDSim 能够捕获到实体随时间变化而产生的兴趣的变化;应用于聚类时,相对于 PathSim 和 PCRW 方法,其标准互信息聚类精度可以提高 0.17% ~9.24%。实验结果表明,PDSim 方法与传统的基于链接的相似性搜索算法相比,显著提高了异构信息网中动态相似性搜索的效率和用户满意度,是一种研究实体随时间而发生动态变化的相似性搜索方法。

**关键词**:异构信息网;元路径;动态相似性搜索;链接
**中图分类号**:TP391.1　　**文献标志码**:A

## Meta path-based dynamic similarity search in heterogeneous information network

CHEN Xiangtao, DING Pingjian*, WANG Jing
(*College of Computer Science and Electronic Engineering, Hunan University, Changsha Hunan* 410082, *China*)

**Abstract:** The existing similarity search algorithms do not consider the time factor. To address this problem, a meta path-based dynamic similarity search algorithm named PDSim was proposed for the heterogeneous information network. Firstly, PDSim calculated the link matrix of object under the given meta-path, thus obtained the instances ratio of meta-path between different objects. Meanwhile, the differences of establishing time were calculated. Finally, the dynamic similarity was measured under the given meta-path. In multiple instances of the similarity search, PDSim kept up with the interest variation of object which dynamically changed with time. Compared with the PathSim (Meta Path-Based Similarity) and PCRW (Path-Constrained Random Walks) methods, the clustering accuracy of Normalized Mutual Information (NMI) could be increased by 0.17% to 9.24% when applied to clustering. The experimental results show that, compared to the traditional similarity search algorithm based on link, the efficiency of dynamic similarity search and the satisfaction of user of PDSim are significantly improved, and it is a dynamic similarity search algorithm for object changes with time.

**Key words:** heterogeneous information network; meta-path; dynamic similarity search; link

## 0　引言

人们通常把数据库作为存储大量数据的仓库,并支持索引、检索、更新和查询处理。然而数据库中的实体不是孤立的元组,它们包含丰富的内在关联语义信息,通过多种关系相互链接形成庞大的信息网络。信息网络分析能够为面向深入的网络数据挖掘和分析提供方法,而相似性搜索是数据库中的重要操作。因此如何利用信息网络中实体提供的链接信息进行相似性搜索,成为异构信息网研究的重要方向。

目前,相似性搜索算法大致可以分为两类:基于特征的方法和基于链接的方法。基于特征的度量方法利用对象的特征值进行相似性度量,比如:余弦相似度[1]、欧氏距离[2]和 Jaccard 系数等;但这类方法不考虑实体之间的链接关系,无法满足网络数据的相似性搜索要求。为此,有研究人员提出各种基于链接的度量方法。Jeh 等[3]提出一种对称的相似性度量方法,利用两实体的邻居之间的相似度来度量它们的相似性;但该算法计算很复杂,后期有很多的改进研究以期加速计算。在网络结构聚类算法(Structural Clustering Algorithm for Network, SCAN)[4]中利用两对象的直接邻居集来度量其相似性。Jeh 等[5]提出一种不对称的相似性搜索算法,在网络中,利用实体 $x$ 随机走到实体 $y$ 的概率来度量其相似性。但上述这些算法在进行相似性度量时都没有考虑对象之间链接的差异性。2011 年,Sun 等[6]基于不同类型实体之间构建的元路径包含不同语义的事实,提出一种异构信息网中基于对称元路径的相似性搜索方法(Meta Path-Based Similarity, PathSim),实现相同类型实体之间的相似性度量。Lao 等[7]提出一种限制路径随机移动(Path-Constrained Random Walks, PCRW)算法,在由丰富数据构建的有向图上进行相似性搜索。

但是,上述相似性搜索算法着重关注实体与实体之间的相关性,忽略了实体之间的相关性会随着时间变化。若考虑实体与实体之间的相互关系随时间变化时,上述相似性搜索算法将难以计算或性能较差。据此,本文提出一种基于元路径的动态相似性搜索算法 PDSim,用来分析在异构信息网中

收稿日期:2014-03-21;修回日期:2014-06-04。
作者简介:陈湘涛(1974 - ),男,湖南邵阳人,副教授,博士,CCF 会员,主要研究方向:数据挖掘;　丁平尖(1990 - ),男,湖南衡阳人,硕士研究生,主要研究方向:数据挖掘;　王晶(1989 - ),女,湖南邵阳人,硕士研究生,主要研究方向:数据挖掘。

实体与实体的关系随时间变化的规律。

## 1　相关概念

信息网络可以抽象为一个有向图 $G=(V,E)$,其中含有实体类型映射函数 $\phi:V\to A$ 和链接类型映射函数 $\psi:E\to L$,每个实体 $v\in V$ 属于一种特定实体类型 $\phi(v)\in A$,每条链接 $e\in E$ 属于一种特定的关系类型 $\psi(e)\in L$,且实体和链接都具有属性的特征。若实体类型数 $|A|>1$ 或者关系类型数 $|L|>1$,则该信息网络即为异构信息网。

**定义 1**　网络模式[8]。网络模式是异构信息网络 $G=(V,E)$ 的模板,表示为 $T_G=(A,L)$,其中:有向图顶点表示实体类型 $A$,边表示关系 $L$。

**定义 2**　对称元路径[9]。对称元路径 $P$ 是定义在网络模式 $T_G=(A,L)$ 上的一条路径,表示为 $A_1\xrightarrow{L_1}\cdots A_{(c-1)}\xrightarrow{L_{(c-1)}}A_c\xrightarrow{L_{(c-1)}}A_{(c-1)}\xrightarrow{L_{(c-2)}}\cdots\xrightarrow{L_1}A_1$。其中:$A_1$ 为起点类型,$A_c$ 为中间类型,$L_1$ 表示 $A_1$ 与 $A_2$ 之间的关系类型,$L_i$ 表示 $A_i$ 与 $A_{(i+1)}$ 之间的关系类型,$P$ 的长度即为 $P$ 中关系的数目。

在网络 $G$ 中,给定元路径 $P$,若存在一条路径 $p=(a_1\cdots a_{c-1}a_ca_{c-1}\cdots a_1)$,使得,对于 $\forall i$,有 $\phi(a_i)=A_i$,且每一条链接 $(e_i=\langle a_ia_{i+1}\rangle)\in L_i$,则称这一路径 $p$ 为元路径 $P$ 的元路径实例。

## 2　动态相似性搜索算法

传统的基于链接的相似性搜索算法[10]仅仅利用实体之间的链接数进行相似性搜索,忽略了链接建立的时间差异。随着数据的快速增长和异构信息网络的发展,在异构信息网络中考虑时间因素的动态相似性搜索已经变得越来越重要,所以提出引入时间因素的元路径实例数比值。然而,用户对时间因素的重视程度不一,所以时间差异度的考虑是十分必要的。本章将介绍引入时间因素进行基于元路径的动态相似性搜索的具体方法 PDSim。

### 2.1　元路径实例数比值

**定义 3**　元路径实例数比值。假设 $\boldsymbol{V}_{x\to A_c}$ 表示实体 $x$ 与类型 $A_c$ 各实体之间的链接数向量,$\boldsymbol{V}_{y\to A_c}$ 表示实体 $y$ 与类型 $A_c$ 各实体之间的链接数向量,则实体 $x$ 与实体 $y$ 的元路径实例数比值为:

$$Ratio_{(x,y|P)}=\frac{2*\boldsymbol{V}_{x\to A_c}*\boldsymbol{V}_{y\to A_c}^{\mathrm{T}}}{\boldsymbol{V}_{x\to A_c}*\boldsymbol{V}_{x\to A_c}^{\mathrm{T}}+\boldsymbol{V}_{y\to A_c}*\boldsymbol{V}_{y\to A_c}^{\mathrm{T}}}\tag{1}$$

实体 $x$ 与实体 $y$ 之间的元路径实例数比值由实体 $x$ 和实体 $y$ 之间的元路径实例数与它们自身元路径实例数之间的比值来衡量。其中 $\boldsymbol{V}_{x\to A_c}$ 可以通过关系矩阵 $\boldsymbol{W}_x|P$(具体参考定义 4)计算得到。

若已知元路径 $P=(A_1\cdots A_{c-1}A_cA_{c-1}\cdots A_1)$ 和被搜索实体 $x\in A_1$,假设 $\boldsymbol{Q}_x|P=[Q_{u\in U,(a_2\in A_2)}]_{N_1*M_1}$ 表示实体 $x$ 对实体类型 $A_2$ 的实体邻接矩阵。其中:$Q_{u\in U,(a_2\in A_2)}$ 表示实体 $x$ 与实体 $a_2\in A_2$ 在 $u\in U$ 时间段建立的链接数,且 $N_1$ 表示 $U$ 中时间段的个数,$M_1$ 表示类型 $A_2$ 中含有实体 $a_2$ 的个数。在本文中,为了解决元路径长度大于 2 的情况,提出了类型邻接矩阵 $\boldsymbol{R}_{A_i,A_{(i+1)}}|u\in U=[R_{a_i\in A_i,a_{(i+1)}\in A_{(i+1)}}]_{N_2*M_2}$ 表示类型 $A_i$ 各实体和 $A_{(i+1)}$ 各实体在时间段 $u\in U$ 建立的链接数。其中:$R_{a_i\in A_i,a_{(i+1)}\in A_{(i+1)}}$ 表示对象 $a_i\in A_i$ 与对象 $a_{(i+1)}\in A_{(i+1)}$ 在时间段 $u\in U$ 建立的链接数,且 $N_2$ 和 $M_2$ 分别表示 $A_i$ 和 $A_{(i+1)}$ 中实体的个数。

**定义 4**　关系矩阵。在异构信息网络中,给定一条元路径 $P=(A_1\cdots A_{c-1}A_cA_{c-1}\cdots A_1)$,$x\in A_1$ 表示被搜索实体,实体 $x$ 对类型 $A_c$ 的关系矩阵表示为 $\boldsymbol{W}_x|P=[W_{u\in U,a_c\in A_c}]_{N_3*M_3}$。其中:$W_{u\in U,a_c\in A_c}$ 表示实体 $x$ 与 $A_c$ 中的实体 $a_c$ 在 $u\in U$ 时间段建立的元路径实例数,且 $N_3$ 表示 $U$ 中时间段的个数,$M_3$ 表示类型 $A_c$ 中含有实体 $a_c$ 的个数。给定元路径通过以下计算模型可以得到 $\boldsymbol{W}_x|P$:

$$\begin{aligned}\boldsymbol{W}_x|P&=\boldsymbol{W}_x(A_1A_2\cdots A_{c-1})\circ\{\boldsymbol{R}_{c-1}\}=\\&\boldsymbol{W}_x(A_1A_2\cdots A_{c-2})\circ\{\boldsymbol{R}_{c-1}\}\circ\{\boldsymbol{R}_{c-2}\}=\cdots=\\&\boldsymbol{W}_x(A_1A_2\cdots A_i)\circ\{\boldsymbol{R}_{c-1}\}\circ\cdots\circ\{\boldsymbol{R}_{i+1}\}=\cdots=\\&\boldsymbol{W}_x(A_1)\circ\{\boldsymbol{R}_{c-1}\}\circ\cdots\circ\{\boldsymbol{R}_2\}\end{aligned}\tag{2}$$

其中 $\boldsymbol{W}_x(A_1)$ 为 $\boldsymbol{Q}_x|P=[Q_{u\in U,(a_2\in A_2)}]_{N_1*M_1}$ 实体邻接矩阵,即当元路径长度为 2 时,关系矩阵即为实体邻接矩阵,$\{\boldsymbol{R}_2\}$ 表示为 $\{\boldsymbol{R}_{A_2,A_3}|u=[R_{a_2\in A_2,a_3\in A_3}]_{N_2*M_2},u\in U\}$,$\boldsymbol{W}_i$ 表示矩阵 $\boldsymbol{W}$ 的第 $i$ 行,算子"$\circ$"表示为:

$$\boldsymbol{W}\circ\{\boldsymbol{R}\}=\begin{pmatrix}\boldsymbol{W}_1\\\boldsymbol{W}_2\\\vdots\\\boldsymbol{W}_n\end{pmatrix}\circ\begin{pmatrix}\boldsymbol{R}|u_1\\\boldsymbol{R}|u_2\\\vdots\\\boldsymbol{R}|u_n\end{pmatrix}=\begin{pmatrix}\boldsymbol{W}_1*\boldsymbol{R}|u_1\\\boldsymbol{W}_2*\boldsymbol{R}|u_2\\\vdots\\\boldsymbol{W}_n*\boldsymbol{R}|u_n\end{pmatrix}\tag{3}$$

利用关系矩阵,式(1)可以表示为:

$$\begin{aligned}Ratio_{(x,y|P)}=2*&\sum_j\Big(\sum_i\boldsymbol{W}_x(i,j).*\sum_i\boldsymbol{W}_y(i,j)\Big)\Big/\\&\Big(\sum_j\Big(\sum_i\boldsymbol{W}_x(i,j).*\sum_i\boldsymbol{W}_x(i,j)\Big)+\\&\sum_j\Big(\sum_i\boldsymbol{W}_y(i,j).*\sum_i\boldsymbol{W}_y(i,j)\Big)\Big)\end{aligned}\tag{4}$$

其中 $.*$ 表示为矩阵的 Hadamard 乘积。

### 2.2　时间差异度

**定义 5**　时间差异度。在异构信息网 $G$ 中,记实体 $x$ 与 $a_{cj}\in A_c$ 之间的链接 $p_i$ 的时间属性值为 $T_{xp_i\to a_{cj}}$,表示实体 $x$ 与 $a_{cj}\in A_c$ 之间的链接 $p_i$ 的建立时间;记实体 $y$ 与 $a_{cj}\in A_c$ 之间对应的链接 $q_i$ 的时间属性值为 $T_{yq_i\to a_{cj}}$,表示实体 $y$ 与 $a_{cj}\in A_c$ 之间链接 $q_i$ 的建立时间。若实体 $y$ 与实体 $a_{cj}\in A_c$ 之间没有相应的链接,则令 $T_{yq_i\to a_{cj}}$ 的取值为 $\min(U)-1$,故定义实体 $x$ 与 $y$ 的时间差异度 $A_{(x,y)}$ 如下:

$$A_{(x,y)}=a^{\left(-\sum_{j=1}^{m}\sum_{i=1}^{n}|T_{xp_i\to a_{cj}}-T_{yq_i\to a_{cj}}|\right)/(n*m)}\tag{5}$$

其中 $a$ 是一常数,其值反映了链接建立时间在动态相似性搜索中所占权重,而且:

$$n=\max(N_{x\to a_{cj}},N_{y\to a_{cj}})\tag{6}$$

$$m=count(A_{cx}\cup A_{cy})\tag{7}$$

其中:$N_{x\to a_{cj}}$ 表示实体 $x$ 与实体 $a_{cj}\in A_c$ 之间的链接数,$N_{y\to a_{cj}}$ 表示实体 $y$ 与实体 $a_{cj}\in A_c$ 之间的链接数,$A_{cx}$ 表示与实体 $x$ 相连的实体 $a_c\in A_c$ 所构成的集合,$A_{cy}$ 表示与实体 $y$ 相连的实体 $a_c\in A_c$ 所构成的集合,函数 $count()$ 用来计算集合中元素的个数。

本文给出时间差异度 $A_{(x,y)}$,用来表示异构信息网中实体 $x$ 与 $a_c\in A_c$ 之间的链接和实体 $y$ 与 $a_c\in A_c$ 之间的链接的建立时间的差异程度,其取值范围为 $(0,1]$。时间差异度的大小取决于两实体与中间类型的相同实体链接建立时间的相隔时间长短,值为 1 时表示相同链接建立的时间没有任何差异。

### 2.3 动态相似性搜索算法 PDSim

在文献[6]提出的基于链接的相似性度量方法的基础上,结合时间差异度,给定元路径 $P$,衡量实体 $x$ 和实体 $y$ 之间的 PDSim 相似度为:

$$S_{(x,y|P)} = A_{(x,y)} * Ratio_{(x,y|P)} \tag{8}$$

其中:$A_{(x,y)}$ 为时间差异度;$Ratio_{(x,y|P)}$ 为元路径实例数比值;$S_{(x,y|P)}$ 表示在给定元路径 $P$ 下,实体 $x$ 和实体 $y$ 之间的动态相似性。算法 PDSim 的具体实现步骤如下所示:

1) 依据 $x,y$ 与 $a_{2i} \in A_2$ 的邻接关系得到实体邻接矩阵 $\boldsymbol{Q}_x|P$ 和 $\boldsymbol{Q}_y|P$,并读取 $x,y$ 相关链接的时间属性值为 $T_{xp_i \to a_{ej}}$,$T_{yq_i \to a_{ej}}$。

2) 如果元路径 $P$ 的长度大于2,则执行步骤3);否则执行步骤4)。

3) 计算元路径 $P$ 下的类型邻接矩阵 $\boldsymbol{R}_{A_i,A_{(i+1)}}|u$,并通过式(2)、(3) 得到关系矩阵 $\boldsymbol{W}_x|P$ 和 $\boldsymbol{W}_y|P$。

4) 关系矩阵 $\boldsymbol{W}_x|P$ 和 $\boldsymbol{W}_y|P$ 分别为:$W_x|P = \boldsymbol{Q}_x|P, W_y|P = \boldsymbol{Q}_y|P$。

5) 通过式(4) 计算得到 $Ratio_{(x,y|P)}$。

6) 根据定义5将 $T_{xp_i \to a_{ej}}, T_{yq_i \to a_{ej}}$ 进行相应处理,然后利用式(5) 计算得到时间差异度 $A_{(x,y)}$。

7) 通过式(8) 计算得到实体 $x$ 与实体 $y$ 之间的 PDSim 相似度 $S_{(x,y|P)}$。

本文中动态相似性度量方法结合了链接数和链接建立的时间因素,因此可以很好地表述实体的兴趣动态变化和整体兴趣。

## 3 实验与结果分析

本章将在数字参考书目和图书馆项目(Digital Bibliography & Library Project, DBLP)[11] 数据集中进行相关实验,对本文提出的动态相似性搜索算法 PDSim 与传统的基于链接的相似性搜索方法进行实验比较,验证 PDSim 的有效性。

### 3.1 实验环境与数据集描述

实验环境:处理器为 Intel Pentium CPU G630 @ 2.70 GHz,RAM 2 GB,操作系统为 Windows XP。

在实验中,使用的数据集是 DBLP 网络中选择的子集,包括4个研究领域的主要会议:数据库和数据挖掘、计算机图形图像、计算机网络和计算机体系结构。在该数据集中,共有4个研究领域,2510 个作者,20 个会议;同时考虑了各实体之间的链接及其链接的时间属性。

### 3.2 动态相似性搜索算法的有效性

假设给定对称元路径 $P$,对 PathSim 和 PDSim 算法分别进行分析。例如:对于元路径 ACA(即作者-会议-作者),找到共享相同会议的作者。依据 ACA 计算得到的关系矩阵可以应用到 PathSim 和 PDSim 上。在实验中,PDSim 的参数 $a$ 设置为 1.2。下面通过两个实例说明该算法的有效性。

在实验数据集中分别采用 PDSim 和 PathSim 搜索作者 Shmuel T. Klein,得到5个兴趣相似的作者。图1为这5个作者和被搜索作者的研究兴趣变化趋势示意图。从图1(a)中可以看出,2000 年以前 Shmuel T. Klein 对数据库与数据挖掘领域比较感兴趣,2000 年以后对计算机图形图像领域更感兴趣。使用 PDSim 搜索作者 Shmuel T. Klein 时,得到前三个相似度最高的作者为 Hui Fang、Vo Ngoc Anh 和 Gonzalo Navarro,其兴趣变化趋势分别如图1(b) ~ (d)所示,可以看出在 2000 年左右这三个作者都具有从数据库和数据挖掘领域向图形图像领域转变的趋势。然而通过 PathSim 算法搜索得到的相似作者有 Andrew Turpin 和 Jamie Callan 等,其兴趣变化趋势分别如图1(e)、(f)所示,用 PathSim 搜索得到的结果更倾向于发表文章数相同的作者,而忽视了作者研究领域的变化情况。

在另外一个研究实例中,使用 PDSim 查找作者 Subrata Banerjee,得到在相似时间段相同会议发表相同文章数的作者,如:F. Borgonovo、N. Shacham 和 S. Liew 等,从表1可以看出这些作者在 INFOCOM 会议上都发表了6篇文章且集中在20世纪90年代发表。而使用 PathSim 得到的前三个作者分别是 D. Du、D. Kandlur 和 E. Coyle,这些作者虽然在 INFOCOM 会议上也都发表了6篇文章,但与被搜索作者的文章发表时间相差较大。所以说,通过元路径 ACA,即通过会议文章来衡量作者的相似性,作者 F. Borgonovo、N. Shacham 和 S. Liew 比 D. Du、D. Kandlur 和 E. Coyle 与被搜索作者更相似,这是更为合理的结果,因为这些作者都是20世纪90年代活跃在计算机网络领域的研究人员。

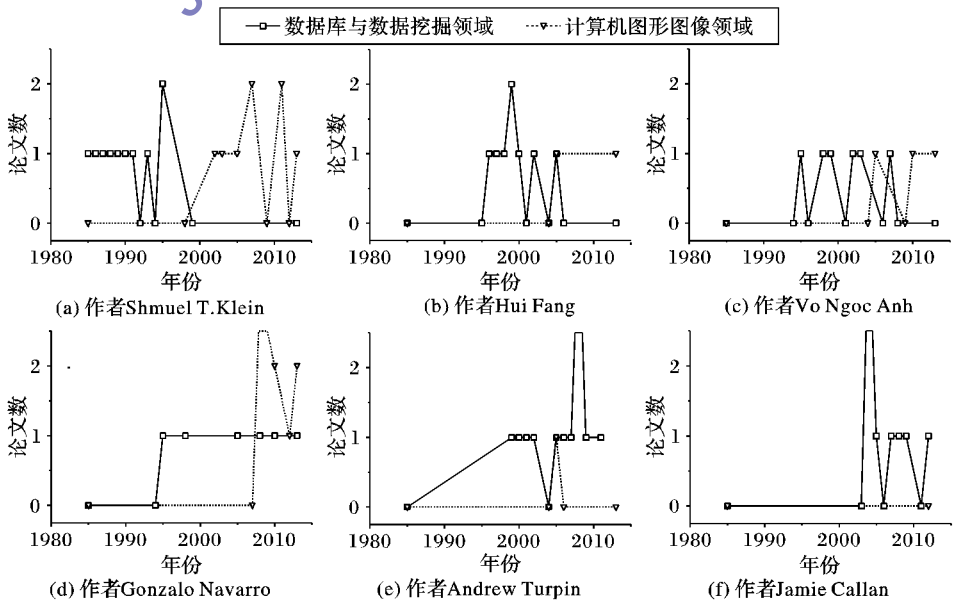


图1 作者的研究兴趣变化趋势示意图

表 1　作者文章数分布表

| 作者 | 1989—1993 年 | 1994—1998 年 | 1999—2003 年 | 2004—2008 年 | 2009—2013 年 |
|---|---|---|---|---|---|
| Subrata Banerjee | 3 | 3 | 0 | 0 | 0 |
| F. Borgonovo | 4 | 2 | 0 | 0 | 0 |
| N. Shacham | 3 | 3 | 0 | 0 | 0 |
| S. Liew | 2 | 4 | 0 | 0 | 0 |
| D. Du | 0 | 0 | 1 | 2 | 3 |
| D. Kandlur | 1 | 0 | 4 | 0 | 1 |
| E. Coyle | 2 | 0 | 1 | 2 | 1 |

上述实验结果表明,PDSim 在异构信息网中进行相似性搜索时,考虑了链接建立的时间,证明了 PDSim 算法在动态相似性搜索中的有效性。

### 3.3　PDSim 算法在聚类中的应用

在本实验中采用标准互信息(Normalized Mutual Information, NMI)[12]来计算聚类精度。给定 $N$ 个实体,簇数目为 $K$,两个聚类结果,假设第一个聚类标签为 $i$ 第二个聚类标签为 $j$ 的实体的个数为 $n(i,j)(i,j=1,2\cdots,K)$,则定义联合分布为:

$$d(i,j)=n(i,j)/N \tag{9}$$

行分布和列分布分别表示为:

$$d_1(j)=\sum_{i=1}^{K}d(i,j) \tag{10}$$

$$d_2(i)=\sum_{j=1}^{K}d(i,j) \tag{11}$$

则 *NMI* 定义为:

$$NMI=\sum_{i=1}^{K}\sum_{j=1}^{K}d(i,j)\mathrm{lb}\left(\frac{d(i,j)}{d_1(j)d_2(i)}\right)\Bigg/\sqrt{\sum_{i=1}^{K}d_1(j)\mathrm{lb}(d_1(j))\sum_{i=1}^{K}d_2(i)\mathrm{lb}(d_2(i))} \tag{12}$$

PDSim 也能应用到聚类任务,在实验数据集中进行两个聚类任务:在元路径 APCPA(即作者-文章-会议-文章-作者)下对作者进行聚类,以及在元路径 CPAPC(即会议-文章-作者-文章-会议)下对会议进行聚类。利用 PDSim、PathSim 和 PCRW 得到不同相似度矩阵,应用 Normalized Cut[13]进行聚类,其中簇的数目设置为 16;然后利用 *NMI* 来评价作者、会议聚类结果,其中第一个聚类结果为人工聚类结果,第二个聚类结果是分别采用 PCRW、PathSim、PDSim 算法得到的聚类结果。表 2 给出了各算法聚类得到的聚类精度。可以看出,作者聚类中 PDSim 的 *NMI* 聚类精度最高,而会议研究领域随时间变化不大,所以会议 *NMI* 聚类精度相差无几。这说明 PDSim 能应用于聚类中。

表 2　*NMI* 聚类精度

| 算法 | 作者 *NMI* | 会议 *NMI* |
|---|---|---|
| PCRW | 0.6364 | 0.8108 |
| PathSim | 0.6735 | 0.8162 |
| PDSim | 0.7288 | 0.8125 |

### 3.4　PDSim 中参数 $a$ 的作用

在 PDSim 中参数 $a$ 取不同值时,表示链接建立数和链接建立时间之间的权重不同。若 $a=1$ 且元路径长度为 2,表示相似性搜索退化为不考虑动态变化。在给定元路径 ACA,表 3 给出了 $a$ 在三种取值下搜索作者 Christos Faloutsos 得到的 5 个相似性作者。其中:1982 年至今,Philip S. Yu 与 Christos Faloutsos 都在各大会议上发表文章;1984 年至今,Jiawei Han 也在各会议上发表文章。然而在文章发表数方面,Christos Faloutsos 与 Jiawei Han 更为接近。当 $a=1$ 时,Jiawei Han 与被搜索作者 Christos Faloutsos 更为相似;当 $a=1.2$ 或 $a=2$ 时,Philip S. Yu 与 Christos Faloutsos 更为相似。从表 3 的结果排序中可以看出,随着 $a$ 的值增大,链接建立时间的权重越大,而链接建立数的权重越小。

表 3　PDSim 中参数 $a$ 不同取值搜索"Christos Faloutsos"的相似作者

| Rank | $a=1$ | $a=1.2$ | $a=2$ |
|---|---|---|---|
| 1 | Christos Faloutsos | Christos Faloutsos | Christos Faloutsos |
| 2 | Jiawei Han | Philip S. Yu | Philip S. Yu |
| 3 | Philip S. Yu | Jian Pei | Hector Garcia-Molina |
| 4 | Rakesh Agrawal | Hector Garcia-Molina | Vagelis Hristidis |
| 5 | Jian Pei | Charu C. Aggarwal | Gautam Das |

注:Rank 表示搜索结果排序,其中:"1"表示相似度最高,"5"表示相似度最低。

## 4　结语

针对异构网络实体的兴趣变化这一现象,本文提出一种基于元路径实现动态相似性搜索的算法.引入链接建立的时间因素,从而考虑实体的动态变化,如引导用户快速了解相关兴趣的用户,跟踪某人的兴趣变化,影迷及时找到相同爱好的人,实现异构信息网中的动态相似性搜索,从而提高相似性搜索的准确度.实验结果验证了该算法的有效性.然而,本文仅研究了单条对称元路径下的动态相似性搜索算法,只捕获到实体之间的一种语义信息.下一步研究可以考虑多条对称元路径下的动态相似性搜索算法;另一方面也可以考虑在不对称的元路径下求解不同类型实体之间的动态相关性算法.

**参考文献:**


[1] MOHAMMADZADEH H, GOTTRON T, SCHWEIGGERT F, *et al*. TitleFinder: extracting the headline of news Web pages based on cosine similarity and overlap scoring similarity [C]// Proceedings of the 12th International Workshop on Web Information and Data Management. New York: ACM, 2012: 65-71.

[2] QIAN G, SURAL S, GU Y, *et al*. Similarity between Euclidean and cosine angle distance for nearest neighbor queries [C]// Proceedings of the 2004 ACM Symposium on Applied Computing. New York: ACM, 2004: 1232-1237.

[3] JEH G, WIDOM J. SimRank: a measure of structural-context similarity [C]// Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 538-543.

[4] XU X, YURUK N, FENG Z, *et al*. SCAN: an structural clustering algorithm for networks [C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007: 824-833.

[5] JEH G, WIDOM J. Scaling personalized Web search [C]// Proceedings of the 12th International Conference World Wide Web. New York: ACM, 2003: 271-279.

[6] SUN Y, HAN J, YAN X, *et al*. PathSim: Meta path-based top-*k* similarity search in heterogeneous information networks [C]// Proceedings of the 37th International Conference on Very Large Data Bases. New York: ACM, 2011: 992-1003.


(下转第 2638 页)

方案。该方案基于 ElGamal 签名,结合 1-out-*n* 不经意传输协议,可用来保护电子交易动态环境中的授权隐私;方案中用户对自己的选择使用的是序号,具有制约性和隐蔽性,用序号作幂运算,加密得到密钥,完成密钥协商;商家运用这种对序号的制约,使得用户不能以没有选择的序号打开消息,用户可以得到且只能得到自己订购的数字商品;并对方案进行了正确性证明和安全性分析,结果表明方案具有不经意性、接收方选择的无条件安全性和在 CDH 假设下的安全性。方案和现有的网上交易方案相比不拘泥于价格的限定,依据用户选择的自选序号以及所拥有的加密签名,可以有效防止商家以次充好的恶意欺诈行为;采用 ElGamal 签名,生成的签名文件相对较短,签名时计算量较小;密钥协商时利用序号幂运算,实现密钥动态变化,整个方案在电子交易动态环境中适应性更好,安全性更强。方案目前的研究仍有不足之处:整个协议中没能考虑签名和用户对应的唯一性,也就是说,用户如果将拥有的签名给予第三方,协议无法识别拥有签名的第三方是否用户本身,这也是下一步研究的方向。

**参考文献:**


[1] TONG Y, TAO Y, TANG S, *et al*. Identity-reserved anonymity in privacy preserving data publishing [J]. Journal of Software, 2010, 21(4): 771 - 781. (童云海, 陶有东, 唐世渭, 等. 隐私保护数据发布中身份保持的匿名方法[J]. 软件学报, 2010, 21(4): 771 - 781.)

[2] LINDELL Y, PINKAS B. Secure two-party computation via cut-and-choose oblivious transfer [C]// TCC 2011: Proceedings of the 2011Theory of Cryptography Conference, LNCS 6597. Berlin: Springer-Verlag, 2011: 329 - 346.

[3] BAO F, DENG R H, FENG P. An efficient and practical scheme for privacy protection in the e-commerce of digital goods [C]// ICISC 2001: Proceedings of the 3rd International Conference on Information Security and Cryptology. London: Springer-Verlag, 2001: 162 - 170.

[4] MAO J, YANG B, WANG Y. A new scheme for privacy protection in the e-commerce of digital goods [J]. Acta Electronica Sinica, 2005, 33(6): 1053 - 1055. (毛剑, 杨波, 王育民. 保护隐私的数字产品网上交易方案[J]. 电子学报, 2005, 33(6): 1053 - 1055.)

[5] JIANG Y, ZHANG M, YANG B, *et al*. An online ordering scheme for privacy protection [J]. Journal of Human University: Natural Science Edition, 2010, 37(3): 77 - 79. (蒋亚军, 张明武, 杨波, 等. 一个具有隐私保护的网上订购方案[J]. 湖南大学学报: 自然科学版, 2010, 37(3): 77 - 79.)

[6] SHEN Y, LIN X. Schnorr signature-based privacy protection online ordering scheme [J]. Application Research of Computers, 2013, 30(3): 882 - 884. (申艳光, 林祥龙. 基于 Schnorr 签名的隐私保护网上订购方案[J]. 计算机应用研究, 2013, 30(3): 882 - 884.)

[7] ZHANG Y, ZHU Y. Price oblivious transfer based privacy protection transaction scheme [J]. Computer Application and Software, 2012, 29(5): 35 - 37. (张云鹤, 朱艳琴. 基于价格不经意传输的隐私保护交易方案[J]. 计算机应用与软件, 2012, 29(5): 35 - 37.)

[8] HUANG N, GUI X, YU S, *et al*. Privacy-preserving computable encryption scheme of cloud computing [J]. Chinese Journal of Computers, 2011, 34(12): 2391 - 2402. (黄汝雄, 桂小林, 余思, 等. 云环境中支持隐私保护的可计算加密方法[J]. 计算机学报, 2011, 34 (12): 2391 - 2402)

[9] XU J, HUANG X, GUO M, *et al*. Location privacy through anonymous chain in dynamic P2P network[J]. Journal of Zhejiang University: Engineering Science Edtion, 2012, 46(4): 712 - 718. (徐建, 黄孝喜, 郭鸣, 等. 动态 P2P 网络中基于匿名链的位置隐私保护[J]. 浙江大学学报: 工学版, 2012, 46(4): 712 - 718)

[10] ElGAMAL T. A public-key cryptosystem and a signature scheme based on discrete logarithms [J]. IEEE Transactions on Information Theory, 1985, 31(4): 469 - 472.

[11] TZENG W. Efficient 1-out-of-n oblivious transfer schemes with universally usable parameters [J]. IEEE Transactions on Computers, 2004, 53(2): 232 - 240.

[12] LI N, DU W, BONEH D. Oblivious signature - based envelope [C]// PODC 2003: Proceedings of the 22nd ACM Symposium on Principles of Distributed Computing. New York: ACM, 2003: 182 - 189.

[13] OGAWA K, HANAOKA G, KOBARA K, *et al*. Anonymous pay-TV system with secure revenue sharing [C]// KES 2007: Proceedings of the 11th International Conference on Knowledge-based Intelligent Information and Engineering Systems, LNCS 4694. Berlin: Springer-Verlag, 2007: 984 - 991.

[14] JIAO Y, FU D. A sequential multi-signature scheme based on ElGamal [J]. Journal of Sichuan University: Natural Science Edition, 2013, 50(4): 757:759. (焦阳, 傅德胜. 基于 ElGamal 的有序多重数字签名方案[J]. 四川大学学报: 自然科学版, 2013, 50(4): 757 - 759.)


(上接第 2607 页)


[7] LAO N, COHEN W. Fast query execution for retrieval models based on path constrained random walks [C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010: 881 - 888.

[8] SUN Y, HAN J, AGGARWAL C, *et al*. When will it happen? — Relationship prediction in heterogeneous information network [C]// Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. New York: ACM, 2012: 663 - 672.

[9] YU X, SUN Y, NORICK B, *et al*. User guided entity similarity search using meta-path selection in heterogeneous information network [C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 2025 - 2029.

[10] SHI C, KONG X, YU P, *et al*. Relevance search in heterogeneous networks [C]// Proceedings of the 15th International Conference on Extending Database Technology. New York: ACM, 2012: 180 - 191.

[11] JI M, SUN Y, DANILEVSKY M, *et al*. Graph regularized transductive classification on heterogeneous information networks [C]// Proceedings of the 21th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Berlin: Springer-Verlag, 2010: 570 - 586.

[12] SUN Y, HAN J, ZHAO P, *et al*. RankClus: integrating clustering with ranking for heterogeneous information network analysis [C]// Proceedings of the 12th International Conference on Extending Database Technology. New York: ACM, 2009: 565 - 576.

[13] SHI J, MALIK J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888 - 905.