基于标签相似度计算的学术圈构建方法

傅城州[1] 汤 庸[1] 贺超波[2] 王津凌[1] 袁成哲[1]

(华南师范大学计算机学院 广州 510631)[1] (仲恺农业工程学院信息科学与技术学院 广州 510225)[2]

摘 要 为面向学者的社交网络系统中的用户构建学术圈,对促进学者之间的交流具有重要的应用价值。根据学者之间的共同属性进行相似度计算,形成学术领域相似和研究课题相近的学术圈,能让学者们更加紧密和频繁地协同合作。提出了利用学者的学术信息提取代表个人特征的学术标签,并对不同类别标签的权重进行衡量,再通过相似度计算和聚类算法构建学术圈的方法。通过抓取学者社交网络平台SCHOLAT公开的学者信息进行实验,进而验证所提方法的可靠性和实用性。

关键词 社交网络,标签,相似度计算,聚类算法,学术圈

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.009

Construction Method of Academic Circle Based on Label Similarity Computation

FU Cheng-zhou[1] TANG Yong[1] HE Chao-bo[2] WANG Jin-ling[1] YUAN Cheng-zhe[1]

(School of Computer Science, South China Normal University, Guangzhou 510631, China)[1]

(School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China)[2]

Abstract Constructing academic circles for users in the scholar-oriented social network system has important application values for promoting exchanges among scholars. Similarity computation is done based on the common properties of scholars, constituting academic circles with similar academic field and research subject, allowing scholars to collaborate more closely and frequently. This paper proposed a method which uses scholars of the academic information to extract the personal characteristics of academic labels, and measures the weight of different class labels. Then through the similarity computation and clustering algorithm, the academic circle can be constructed. By crawling the public information on the academic social network platform named SCHOLAT to perform experiments, the reliability and usefulness of the proposed method are verified.

Keywords Social network, Label, Similarity computation, Clustering algorithm, Academic circle

1 引言

在Web 2.0技术快速发展的时代[1],在线社交网络(Online Social Network, OSN)如Facebook、Twitter、Qzone和Weibo等聚集了大量用户注册使用。随着学术活动的开展,在线社交网络同时也成为国内外各领域学者交流研讨的重要途径。诸如ResearchGate、学者网和学术圈等都是目前非常流行的学术社交网络系统,吸引了众多国内外知名学者、学术机构和在校学生注册使用。当学者用户量日益庞大、用户类型趋向复杂化时,如何在系统中构建学术圈,提高学者用户对系统的粘性并促进他们的交流分享,成为一个具有实际应用意义的重要研究课题。

Wang等人提出了基于社交圈的在线社交网络朋友推荐算法,他们首先提出社交圈检测算法,并定义用户间的社交圈相似性,基于社交圈相似度为用户推荐新朋友[2]。Lin等人提出了一种边稳定系数模型和表达个体之间关系紧密度的完全信息图模型,在此基础上设计并实现了一种有效的社区发现算法。他们提出的完全信息图模型具有较高通用性,适用于需要融合个体和链接属性的社区发现算法[3]。Julian等人建立了一个模型,根据用户信息在自我网络中进行社区发现[4]。

与传统社区发现的相关研究相比,本文基于学术信息的挖掘,针对学者这一群体的学术社区发现进行研究。信息是社交网络用户特征标识的重要依据,如何提取和过滤用户信息为我们构建社交圈提供基础,是一个重要的步骤。对于学者社交圈而言,学者的研究领域、工作单位、个人简历、科研成果和学术活动都是重要的学术社交属性,如何在大量且无规则的用户信息中提取标签也是重要的过程[5]。

本文通过提取近年发表学术文献的关键词,构建用于实验的学术标签库,并且利用标签对学者的社交信息文本进行分析,提出结合相似度计算和聚类算法构建学术圈的方法。

本文第2节提出构建标签库的方法,并通过实例推导出

到稿日期:2015-08-01 返修日期:2015-09-28 本文受863高技术研究发展计划项目基金(2013AA01A212),国家自然科学基金(61272067,61502180),广东省公益研究与能力建设专项(2015A020209178)资助。

傅城州(1986—),男,博士生,主要研究方向为社交网络与大数据;汤 庸(1964—),男,教授,博士生导师,主要研究方向为社交网络与大数据;贺超波(1981—),男,博士,副教授,主要研究方向为社会网络;王津凌(1991—),男,硕士生,主要研究方向为移动互联网;袁成哲(1991—),男,硕士生,主要研究方向为社交网络。

相似度计算的最终模型；第 3 节在计算模型的基础上，提出运用聚类算法构建学术圈的方法；第 4 节利用学术社交网络的数据进行学术圈构建实验，验证所提方法的可靠性和准确性；最后总结目前取得的成果并对下一步的研究工作进行展望。

2 学者相似度计算

2.1 标签提取方法

用户之间的相似度计算方法有余弦相似度[6]、皮尔森系数[7]和修正的余弦相似度[8]3 种。这 3 种相似度计算方法都是基于“用户-项目”评分矩阵的数据结构进行计算的。社交圈构建的过程通常可以认为是好友的分组过程，而这个过程也等同于自我网络(Ego Network)社团发现问题[9]。通过标签库的构建，形成集合为 L_{sam} 的标签信息库，对于每位学者，通过提取标签，构成学者的标签信息库，用集合 L 表示，则学者 S 的标签集为：

$$U_s=\{l_s \mid l_s \in L_{sam}\} \tag{1}$$

其中，U_N 表示学者 N 的标签信息集合，l_n 表示学者 N 拥有的标签，对于学者 A 和学者 B 的共同标签集表示为：

$$f_{common}(A,B)=U_A \cap U_B=\{l_i \mid l_i \in L_A, \text{且 } l_i \in L_B\} \tag{2}$$

在标签集合图中，对学者 A 与其他学者之间的标签共同域进行连结，构成的标签关系如图 1 所示。

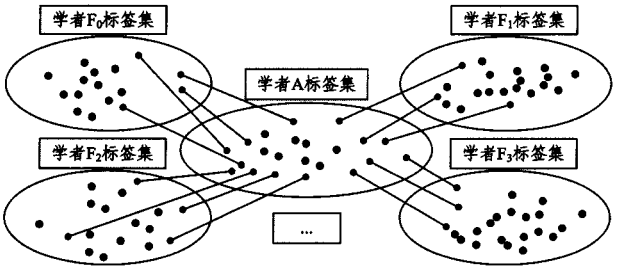


图 1 通过标签构建学者 A 的关系

定义 1 对于学者标签集合图，其中连结点表示标签词汇；两条连线之间表示学者之间拥有的共同标签；部分重叠的连结点表示这几位学者具有共同标签；其它孤立的节点表示该学者独有的标签。学者间的标签连线数量越多，说明他们之间的相似度越高，将他们划分在同一个学术圈的可能性也越高。

本文对学者的信息进行提取，将学者的各类信息进行归并，假设学者的专业背景、研究领域、科研项目、学术论文和个人简介等信息分别为 i_0,i_1,i_2,i_3,i_4，将其重新合并成为新的信息集合 $I=\{i_0,i_1,i_2,i_3,i_4\}$。从文本内容集合 I 中提取出学者的标签需要解决以下两个问题：1)确定能够作为标签的词汇；2)如何从文本分离出符合条件的词汇。

为了解决上述两个问题，本文提出利用学术文献的关键词作为词汇库，用于提取文本的标签。基于这种方法，从学者网的学术论文库中抓取大约 300000 篇学术水平比较高的文献，从中提取词汇近 1000000 个，并通过去重过滤，获得无重复词汇 91009 个。依照上述构建的标签库对学术信息文本中的词汇进行高效的统计，采用 KMP 算法[10]进行字符串匹配。

算法 1 提取文本中的所有标签

步骤 1 在串 S 和串 T 中，分别假设比较的起始下标为 i,j；

步骤 2 循环直到 S 中剩余字符长度小于 T 中的长度或者 T 中所有字符均比较完毕：

步骤 2.1 如果 S[i]=S[j]，继续比较 S 和 T 的下一个字符；否则转至步骤 2.2；

步骤 2.2 将 j 向右滑动到 next[j]的位置，即 j=next[j]；

步骤 2.3 如果 j=0，则将 i 和 j 分别加 1，准备下一次比较；

步骤 3 如果 T 中所有字符均比较完毕，则返回匹配的起始下标；否则返回 0。

由于在此过程提取的标签数据，需要进行大量的运算和处理，以及查询排序的过程，对计算机的运算能力要求比较高。为解决单机运算能力较低的瓶颈，本文在实验的过程中搭建 Hadoop[11]的云计算平台进行运算。Hadoop 是一个开源成熟的云平台，实现了一个分布式文件系统(Hadoop Distributed File System，HDFS)。HDFS 有高容错性的优点，能够部署在比较廉价的硬件上，而且它能够提供高吞吐量且用于访问应用程序的数据，适合超大数据集的应用程序。Hadoop 是基于 MapReduce 模式开发的系统，最基本的 MapReduce 应用程序至少包含 3 部分：一个 Map 函数，一个 Reduce 函数和一个 main 函数。main 函数将作业控制和文件 Input/Output 结合起来。Hadoop 提供了大量的接口和抽象类，为 Hadoop 应用程序开发人员提供许多工具，便于调试和性能度量等。

2.2 文本相似度计算方法

2.1 节描述了 3 种关于用户相似度计算的方法，但是这 3 种方法在实际的计算过程中都存在一些缺陷。

余弦相似度(Cosine Similarity)：该算法是以两个向量的内积空间的夹角的余弦值作为衡量它们之间相似度的标准。当夹角为 0 时，余弦值是 1；而其他任何角度的余弦值都不大于 1；并且其最小值是−1。从而根据两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。因此当两个向量有相同的方向时，余弦相似度的值为 1；两个向量夹角为 90°时，余弦相似度的值为 0；两个向量指向完全相反的方向时，余弦相似度的值为−1。在比较过程中，向量的规模大小不予考虑，仅考虑向量的指向。余弦相似度通常用于两个夹角小于 90°的向量，所以余弦相似度的值域为[0,1]。

皮尔森相关系数(Pearsoncorrelation Coefficient)也称皮尔森积矩相关系数(Pearson Product-moment Correlation Coefficient)，是一种线性相关系数。皮尔森相关系数是两个变量线性相关度的统计量。相关系数用 r 表示，其中 n 为样本容量，分别为两个变量的观测值和相似度均值。r 描述的是两个变量间线性相关的强弱程度。r 的绝对值越大表明它们相关性越强。

修正的余弦相似度(Adjusted cosine)：余弦相似度未考虑到用户评分度量问题，如在评分区间[1,5]的情况下，对用户 A 来说，评分 3.5 以上就是感兴趣的；而对于用户 B，评分 3.8 以上才是感兴趣的。通过减去用户对项的平均评分，修正的余弦相似度度量方法对以上问题进行改善。

在数据样本比较稀疏的情况下，这 3 种方法均存在以下问题：余弦相似性和修正余弦相似度都是基于用户没有评价、项目评分为 0 的假设；皮尔森系数中用户共同评分项目集合可能很小。基于前人的研究工作，本文在计算用户相似度时，选择余弦相似度算法。

2.3 基于标签提取的学者相似度计算

通过计算学者之间的标签相似度，近似地计算学者之间的相似度，再通过聚类算法对学者进行无监督分类，从而达到构建学术圈的目标。假设：$l_{A0},l_{A1},l_{A2},\cdots,l_{Am}$ 表示学者 A 所拥有的标签，并且对应的标签数量为 $A_0,A_1,A_2,\cdots,A_m$；$l_{B0},l_{B1},l_{B2},\cdots,l_{Bn}$ 表示学者 B 所拥有的标签，并且对应的标签数

量为 $B_0,B_1,B_2,\cdots,B_n$；$l_{s0},l_{s1},l_{s2},\cdots,l_{si}$ 表示学者 A 与学者 B 共同拥有的标签，其中 $i=\min(m,n)$，学者之间独有的标签记为 0，则他们的标签相似度 $Sim(A,B)$ 为：

$$Sim(A,B)=\begin{cases}\sum_{k=0}^{m}\dfrac{l_{ik}\times l_{jk}}{\sqrt{\sum_{k=0}^{m}l_{ik}^2}\sqrt{\sum_{k=0}^{m}l_{jk}^2}}, & m<n\\ \sum_{k=0}^{n}\dfrac{l_{ik}\times l_{jk}}{\sqrt{\sum_{k=0}^{n}l_{ik}^2}\sqrt{\sum_{k=0}^{n}l_{jk}^2}}, & m\geqslant n\end{cases}\tag{3}$$

由于学者的学术信息文本可能比较冗长，提取得到的标签数量也比较大，因此需要将提取的标签按数量进行降序排列，以截取前 p 个标签进行相似度计算。设初始学者集合为 $U=\{U_0,U_1,U_2,\cdots,U_n\}$，对应学者两两之间的相似度为 $U(0,1),U(0,2),U(0,3),\cdots,U(n-1,n)$。将这些相似度作为聚类计算的质心距离，从而进一步通过计算构建学术圈内。构成的学者社交加权图[12]如图 2 所示，图中没有连线表示这两位学者没有共同标签，此时相似度记为 0，但是他们能够通过间接的相似度被聚合在同一个学术圈内。

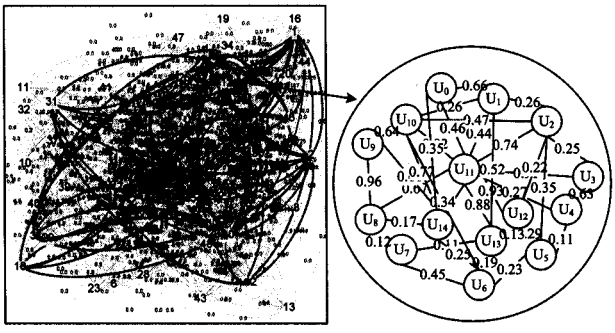


图 2　学者相似度关系图

此算法中有一个问题需要解决：当两学者的标签集合元素数量都比较少时，会造成标签越少而相似度值越大的可能。因此，需要对参与计算的学者用户信息的文本质量进行约束。

例 1　学者 A 的标签集只有一个元素“计算机”，即 $U_A=$ {“计算机:1”}；学者 B 的标签集也只有一个元素“计算机”，即 $U_B=$ {“计算机:1”}，此时他们的相似度为：

$$Sim_0=Sim(A,B)=\frac{1\times1}{\sqrt{1^2\times1^2}}=1\tag{4}$$

基于存在此种情况，为避免数据的运算结果与现实不符，作以下条件约束：对于所有参与运算的用户数据，用户标签集合的元素数量 L 必须满足 $L\geqslant N$，N 为预设的阈值。

2.4　标签类别加权相似度计算模型

上文介绍了基于标签提取的学者相似度计算，上述的计算方法是基于如下假设建立的：所有的文本对学者的相似度计算权重是等同的。但是显然在某些情况下，需要提取学者信息时，针对不同类别的信息，在参与相似度运算时，应当对其赋予不同的权重。

例 2　假设学者 A、学者 B 和学者 C 的信息文本集合分别表示为 I_A,I_B,I_C；标签库表示为 L_{sum}。

$I_A=$ {“‘研究领域’:‘大数据与云计算，信息搜索，社交网络，服务计算，移动互联网应用’”,“‘个人简介:湖南张家界人，武汉大学学士和硕士、中国科学技术大学博士，教授，博士生导师，中国计算机学会(CCF)杰出会员。’”}

$I_B=$ {“‘研究领域’:‘信息搜索，社交网络，服务计算，移动互联网’”,“‘个人简介’:‘广东汕尾人，华南师范大学计算机学院，博士研究生，中国计算机学会(CCF)学生会员。’”}

$I_C=$ {“‘研究领域’:‘网络与教育技术’”,“‘个人简介’:‘湖南张家界人，武汉大学教育科学学院，博士，教授，博士生导师。’”}

$L_{sum}=$ {“服务计算”,“技术”,“武汉大学”,“博士生导师”,“湖南”,“教授”}

观察 I_A,I_B,I_C 的信息文本，根据常识，显然可以认为学者 A 和学者 B 的相似度比学者 A 和学者 C 的相似度高，因为从他们的研究领域了解到学者 A 和学者 B 的专业领域都是计算机，更应该聚合在同一个学术圈；而学者 C 则为教育学专业的教授，与前两位学者的学术领域相差较大。然而，我们进行以下运算，利用标签库 L_{sum} 对以上 3 位学者提取标签，分别用集合 L_A,L_B,L_C 表示，则：

$L_A=$ {“服务计算:2”,“技术:2”,“武汉大学:1”,“博士生导师:1”,“湖南:1”,“教授:1”}

$L_B=$ {“服务计算:1”,“技术:0”,“武汉大学:0”,“博士生导师:0”,“湖南:0”,“教授:0”}

$L_C=$ {“服务计算:0”,“技术:1”,“武汉大学:1”,“博士生导师:1”,“湖南:1”,“教授:1”}

利用式(3)，求解学者 A 与学者 B、学者 C 的相似度分别记为 $Sim(A,B),Sim(A,C)$，则：

$$Sim(A,B)=\frac{2\times1+2\times0+1\times0+1\times0+1\times0+1\times0}{\sqrt{(2^2+2^2+1^2+1^2+1^2+1^2)(1^2+0^2+0^2+0^2+0^2+0^2)}}=0.5774\tag{5}$$

$$Sim(A,C)=\frac{2\times0+2\times1+1\times1+1\times1+1\times1+1\times1}{\sqrt{(2^2+2^2+1^2+1^2+1^2+1^2)(0^2+1^2+1^2+1^2+1^2+1^2)}}=0.7746\tag{6}$$

对比运算结果发现 $Sim(A,B)<Sim(A,C)$，因此，根据前面阐述的学者相似度理论(定义 1)，学者 A 和学者 C 的相似度比学者 A 和学者 B 的相似度高，显然，求解的结果与现实矛盾。

因此，在此基础上，我们对学者相似度计算理论进行进一步研究，引入标签类别加权相似度计算模型，即在计算相似度的过程中，同时考虑信息的类别，不同信息在计算的过程中权重应当不同。如何对不同类别的权重进行衡量，是一个非常重要的问题。

定义 2　对于学者的信息文本，所有类别的集合为 $IK=\{ik_0,ik_1,ik_2,\cdots,ik_n\}$，如果信息文本类别元素 ik_n 在所有学者的平均长度越短，其权重就越大，否则反之。

结合式(3)，假设学者各类别信息的平均长度依次为 $\overline{len_0},\overline{len_1},\overline{len_2},\cdots,\overline{len_n}$，它们的总和为 $\overline{Sum}$；学者之间各类别信息的标签相似度依次为 $Sim_0,Sim_1,Sim_2,\cdots,Sim_n$，则标签类别加权相似度计算的最终模型表示为：

$$\overline{Sim}=\sum_{k=0}^{n}\frac{\overline{Sum}-\overline{len_k}}{\overline{Sum}}Sim_k\tag{7}$$

3　学术圈聚类构建

聚类是数据挖掘中用来发现数据分布和隐含模式的一项重要技术[13]，也用于发现数据中未知的分类。聚类分析已经有很长的研究历史，其重要性越来越受到人们的肯定。聚类算法是机器学习、模式识别和数据挖掘等研究领域的重要研

究课题之一，在识别数据对象的内在联系方面具有极其重要的作用。

聚类算法主要应用于模式识别中的语音识别和字符识别等，机器学习中的聚类算法应用于图像分割时在数字图像处理中主要用于数据压缩和信息检索等。另一方面，聚类算法的重要应用是数据挖掘、时态数据库应用、序列和数据清洗等。此外，聚类还应用于统计科学，同时在生物学、地质学、地理学以及市场营销等方面也发挥着重要的作用。

本文通过聚类分析，构建学术情况相近和研究领域相似的学术圈，促进同领域学者之间的交流，对学术水平的提高具有重要的现实意义，并且该算法能够帮助研发团队构建高水平的学术交流平台。在大数据时代，通过搭建云计算平台，能够对大量的数据进行复杂且快速的运算，有利于支撑具有海量数据系统的部署。学术圈构建的基本流程如图 3 所示。

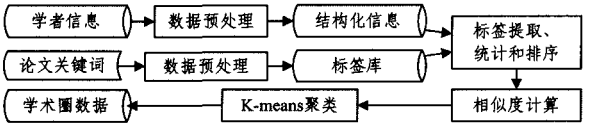


图 3　学术圈构建基本流程

K-means[14-18]是目前常用的文本聚类算法，是典型的基于原型的目标函数聚类方法的代表，它将数据点到原型的某种距离作为优化的目标函数，利用对函数求极值的方法得到迭代运算的调优规则。K-means 算法以欧氏距离作为相似度的测度，它求解对应某一初始聚类的中心向量最优分类，使得评价指标最小，采用误差平方和准则函数作为聚类的准则函数。

算法 2　K-means 算法

步骤 1　从 n 个文档中随机选取 k 个作为初始质心；

步骤 2　对剩余的其它每个文档，测量其到每个质心的距离，并把它归到最近的质心的类；

步骤 3　重新计算已经得到的各个类的质心；

步骤 4　迭代步骤 2、步骤 3 直至满足既定条件，算法结束。

4　实验结果与分析

本节用大量实验来验证所提出的学术圈构建方法的有效性，并且运用学者网的部分有效用户的数据进行实际运算。

由于学者网是基于实名制构建的学术社交网络系统，可以认为学者的信息基本上都具有真实性，由此得到的结果也是可信的。学术文献中的关键词绝大部分为专业词汇，因此这些词汇也可以作为标识学者的依据。本实验采用的数据均为真实的学者用户公开的学术信息，包括学者的专业背景、研究领域、科研项目、学术论文和个人简介等。实验中，将专业背景、研究领域、科研项目、学术论文和个人简介合并为“科研简历”作为一个信息类别，而将“研究领域”作为一个信息类别。在该实验环境中，选用 4 台普通台式计算机（基本配置为 Debian5.0，Intel（R）Core（TM）i7-2600 CPU @ 3.4GHz，8.0GB RAM）用于部署 Hadoop 平台。另外，使用其中 1 台主机进行提取标签实验。

4.1　标签提取实验

对学者 A 和学者 B 进行标签提取，结果如表 1－表 4 所列。实验统计得到“科研简历”的平均长度为 3422，“研究领域”的平均长度为 27。

表 1　学者 A“科研简历”标签提取结果

标签	频数	标签	频数
研究	91	服务	34
计算	91	科技	32
软件	80	模型	29
数据	79	本体	29
广东	75	网络	26
技术	72	教育	25
计算机	71	小平	24
时态	67	计划	23
广东省	64	中山	23
大学	57	国家	23
数据库	53	技术研究	22
科学	49	博士	22
系统	49	基金	21
信息	43	中山大学	21
协同	39	中心	21
学报	37	课程	20
项目	37	出版	20

表 2　学者 A“研究领域”标签提取结果

标签	频数	标签	频数
信息	3	时态	1
数据	2	互联网	1
计算	1	移动	1

表 3　学者 B“科研简历”标签提取结果

标签	频数	标签	频数
计算	94	数据库	17
计算机	90	项目	17
算法	59	分布	16
工程	54	分布式	16
研究	49	方法	16
应用	47	事务	15
广西	42	学会	15
设计	41	本体	15
系统	31	软件	15
构架	26	机制	14
技术	25	安全	13
网络	24	信息	12
路由	24	协同	12
数据	21	发展	12
科学	20	管理	12
遗传	20	中国	11
遗传算法	18	大学	11

表 4　学者 B“研究领域”标签提取结果

标签	频数	标签	频数
数据	2	搜索	1
计算	1	科学	1
语义	1	分析	1

通过实验，我们注意到提取标签的过程的耗时是比较理想的，整个实验中，从单个学者的信息中提取标签总耗时不超过 1000ms。虽然如此，当社交系统用户群庞大时，进行实时的运算仍然不太现实，因此在实际的应用中，可以采用离线计算的方式定期对系统的社交关系数据进行更新，例如在系统空闲时（通常为夜间时段）进行运算，这样不会影响用户的使用，同时也兼顾数据的更新周期在用户的可接受范围之内。

4.2　学术圈构建实验

例 3　根据标签相似度，求解学者 A 和学者 B 的相似度。由于他们的标签数量相等，依据上文描述，兼顾实验结果和运算速度，标签数量的阈值统一取值为 34，因此 $m=n=34$，他们的加权相似度为：

$$\overline{Sim_0}=\overline{Sim(A,B)}$$
$$=\frac{27}{3422+27}\times\frac{91\times49+91\times94+\cdots+20\times0}{\sqrt{(91^2+91^2+\cdots+20^2)(49^2+91^2+\cdots+3^2)}}+\frac{3422}{3422+27}\times\frac{3\times0+2\times2+1\times1+1\times0+1\times0+1\times0}{\sqrt{(3^2+2^2+1^2+1^2+1^2+1^2)(0^2+2^2+1^2+0^2+0^2+0^2)}}$$
$$=0.4037 \tag{8}$$

不失一般性，假设样本学者总量为 n，如果两两求解相似度，则可以得到的相似度解集元素个数为 C_n^2，对所有样本学者进行计算，加权相似度依次为：

$$\overline{Sim_0},\overline{Sim_1},\overline{Sim_2},\overline{Sim_3},\cdots,\overline{Sim_{C_n^2-1}} \tag{9}$$

因为此实验的数据集比较大，这里只列举部分数据。在运算过程中，运用 Hadoop 平台对 key-value pair 数据集进行如下设计：假设学者 A 和学者 B 的标识符分别为 U_A 和 U_B，将(U_A，U_B)整体作为 *key* 值；而他们的共同标签对应的数量分别记为 L_A，L_B，将(L_A，L_B)整体作为 *value* 值，在整体赋值时均需在中间用特殊符号“，”隔开，在 Mapper 和 Reducer 的过程中，对 *key* 和 *value* 按照特殊符号分割(split)进行还原。

例 4　学者 A 和学者 B 的标识为 001 和 002，并且他们具有同一标签为“服务计算”，数量分别为 3 和 4，则此时 *key*=“001，002”，*value*=“3，4”。

表 5 列出了学者 A 与其他学者的相似度。根据相似度计算结果，可以直观认为结果数值比较大的为最有可能与学者 A 在同一个学术圈的学者。

表 5　学者相似度实验结果

Sim(A,X)	学者标识	相似度	Sim(A,X)	学者标识	相似度
Sim_{00}	0000	0.1206	Sim_{15}	0015	0.1512
Sim_{01}	0001	0.1524	Sim_{16}	0016	0.2186
Sim_{02}	0002	0.1948	Sim_{17}	0017	0.2518
Sim_{03}	0003	0.3211	Sim_{18}	0018	0.0000
Sim_{04}	0004	0.0000	Sim_{19}	0019	0.0000
Sim_{05}	0005	0.1853	Sim_{20}	0020	0.0103
Sim_{06}	0006	0.1698	Sim_{21}	0021	0.0000
Sim_{07}	0007	0.2121	Sim_{22}	0022	0.3430
Sim_{08}	0008	0.0000	Sim_{23}	0023	0.2121
Sim_{09}	0009	0.2667	Sim_{24}	0024	0.0000
Sim_{10}	0010	0.1109	Sim_{25}	0025	0.0000
Sim_{11}	0011	0.2037	Sim_{26}	0026	0.2739
Sim_{12}	0012	0.0000	Sim_{27}	0027	0.1023
Sim_{13}	0013	0.1700	Sim_{28}	0028	0.0000
Sim_{14}	0014	0.0694	Sim_{29}	0029	0.0187

继续抽取学者网 500 位注册用户的学术信息进行相似度计算，得到他们两两之间的相似度值 124750 个。如图 4 所示，再从学者 A(标识为 17)的节点出发，提取相似度值域为[0.3500，0.9999]的学者，得到其中与学者 A 相似度最大的标识为 129、317 和 367 的 3 位学者，分别记为学者 X_1、学者 X_2 和学者 X_3。为进一步考察提出的相似度计算方法的效果，查阅了他们公开的个人学者主页，并列出以下关键信息。

学者 A：教授，博士生导师，武汉大学计算机学士和硕士，中国科学技术大学博士。研究领域为大数据与云计算、时态数据库、服务计算和社交网络等。

学者 X_1：副教授，硕士生导师，学者 A 指导的硕士和博士研究生。研究领域为时态数据库。

学者 X_2：计算机博士，副教授，硕士生导师，学者 A 指导的博士研究生。研究领域为语义 Web(描述逻辑)、数据挖掘(机器学习)、个性化推荐等。

学者 X_3：副教授，武汉大学计算机学士和硕士，耶鲁大学毕业的计算机博士，学者 A 的武汉大学同专业的校友。研究领域为生物信息学和计算复杂性。

观察学者 A、学者 X_1、学者 X_2 和学者 X_3，发现他们的研究领域和学术简历都有相似之处，因此将他们构建在同一个学术圈事实上也是合理的，同时也证明了提出的相似度计算方法是可靠的。

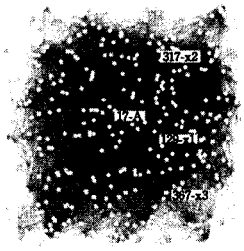
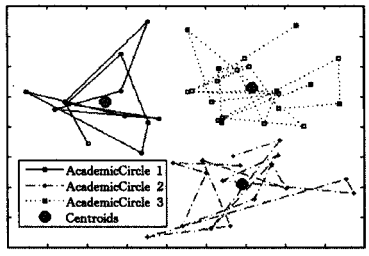
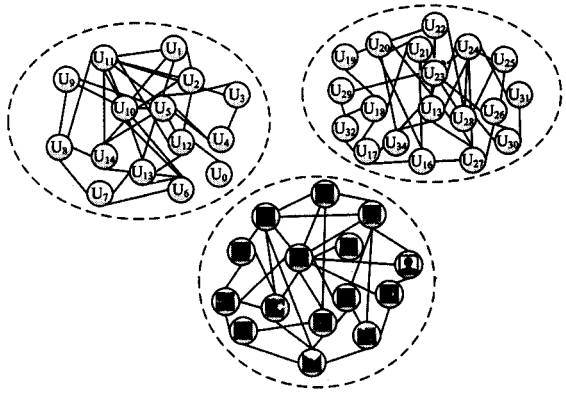


图 4　与学者 A 学术信息相似度较高的学者

最后，从上述实验样本中选取 10%的信息更为完整的用户，将得到的数据导入 Matlab[20] 软件中，利用 K-means 算法(在 Matlab 中利用系统内置 kmeans 函数实现)，当 $k=3$ 时，构建的学术圈结果如图 5(a)所示。为直观表示学术圈，将其中一个学术圈贴上对应学者的头像，如图 5(b)所示。



(a)Matlab 聚类结果



(b)学术圈构建结果

图 5　当 $k=3$ 时的学术圈构建结果

结束语　本文将学术圈看作用户标签相似关系的集合，提出了与人际关系不相关的基于标签相似关系的学术圈构建方法，该方法结合用户特征信息与社交网络拓扑结构，定义了关系的相似度，通过聚类相似的关系，快速识别学术圈所属。定义了学者间的学术标签相似度，设计的方法能够对已经完善个人资料的新近注册学者所属的社交圈给出建议，实现学者分组的自动更新。最后使用真实数据验证学术圈构建方法的准确性，实验结果表明本文方法能够快速检测学术圈并准

(下转第 76 页)

commender Systems. ACM, 2014: 265-272

[11] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2000, 29(5): 1189-1232

[12] Riccardi A, Fernandez-Navarro F, Carloni S. Cost-sensitive AdaBoost algorithm for ordinal regression based on extreme learning machine[J]. Cybernetics IEEE Transactions on, 2014, 44(10): 1898-1909

[13] Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution[C]// Proceedings of International Conferences on Machine Learning. 2003: 856-863

[14] H P, F L, C D. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(8): 1226-1238

[15] Steffen R. Learning recommender systems with adaptive regularization[C]// Fifth ACM International Conference on Web Search & Data Mining. 2012: 133-142

[16] Steffen R, Zeno G, Christoph F, et al. Fast context-aware recommendations with factorization machines[C]// Proceedings of the 34th ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011: 635-644

[17] Steffen R. Scaling Factorization Machines to Relational Data[C]// Proceedings of the 39th International Conference on Very Large Data Bases (VLDB 2013). Trento, Italy, 2013: 337-348

[18] Christoph F, Lars S T, Steffen R. Bayesian Factorization Machines[C]// Workshop on Sparse Representation and Low-rank Approximation, Neural Information Processing Systems (NIPS-WS). Granada, Spain, 2011

[19] Cover T, Hart P. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27

[20] Steffen R. Factorization Machines with libFM[J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(3): 451-458

(上接第 56 页)

确预测学术关系。以后的研究工作将重点考虑社交关系方向问题，将构建方法应用到有向网络中。

参 考 文 献

[1] Zhang H P, Zhang R Q, Zhao Y P, et al. Big data modeling and analysis of microblog ecosystem[J]. International Journal of Automation and Computing, 2014, 11(2): 119-127

[2] Wang Y, Gao L. Social circle-based algorithm for friend recommendation in online social networks[J]. Chinese Journal of Computer, 2014(4): 801-808(in Chinese)
王珏，高琳. 基于社交圈的在线社交网络朋友推荐算法[J]. 计算机学报，2014(4): 801-808

[3] Lin Y F, Wang T Y, Tang R, et al. An effective model and algorithm for community detection in social networks[J]. Journal of Computer Research and Development, 2012, 49(2): 337-345(in Chinese)
林友芳，王天宇，唐锐，等. 一种有效的社会网络社区发现模型和算法[J]. 计算机研究与发展，2012，49(2)：337-345

[4] McAuley J, Leskovec J. Discovering social circles in ego networks[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2014, 8(1): 73-100

[5] Wang J, Lochovsky F H. Data extraction and label assignment for web databases[C]// Proceedings of the 12th International Conference on World Wide Web. New York: ACM Press, 2003: 187-196

[6] Huang A N. Similarity measures for text document clustering [C]// Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008). Christchurch, 2008: 49-56

[7] Ahlgren P, Jarneving B, Rousseau R. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient[J]. Journal of the American Society for Information Science and Technology, 2003, 54(6): 550-560

[8] Deng A, Zhu Y, Shi B L. A collaborative filtering recommendation algorithm based on item rating prediction[J]. Journal of Software, 2003, 14(9): 1621-1628

[9] Leskovec J, Mcauley J J. Learning to discover social circles in ego networks[J]. Advances in Neural Information Processing Systems, 2012: 539-547

[10] Robert W P L. Chinese string searching using the KMP algorithm[C]// Proceedings of the 16th Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1996(2): 1111-1114

[11] White T. Hadoop: The definitive guide[M]. O'Reilly, 2009

[12] Lin W Q, Lu F S, Ding Z Y, et al. Parallel computing hierachical community approach based on weighted-graph[J]. Journal of Software, 2012, 23(6): 1517-1530(in Chinese)
林旺群，卢凤顺，丁兆云，等. 基于加权图的层次化社区并行计算方法[J]. 软件学报，2012，23(6)：1517-1530

[13] He L, Wu L D, Cai Yi-zhao. Survey of clustering algorithm in data mining[J]. Application Research of Computer, 2007, 24(1): 10-13(in Chinese)
贺玲，吴玲达，蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究，2007，24(1)：10-13

[14] Suo H G, Wang Y W. An improved k-means algorithm for document clustering[J]. Journal of Shandong University (Natural Science), 2008, 43(1): 60-64(in Chinese)
索红光，王玉伟. 一种用于文本聚类的改进 k-means 算法[J]. 山东大学学报(理学版)，2008，43(1)：60-64

[15] Mahdavi M, Abolhassani H. Harmony K-means algorithm for document clustering[J]. Data Mining and Knowledge Discovery, 2009, 18(3): 370-391

[16] Gupta H, Srivastava R. K-means based document clustering with automatic "k" selection and cluster refinement[J]. International Journal of Computer Science and Mobile Applications, 2014, 2(5): 7-13

[17] Mavroeidis D, Marchiori E. Feature selection for k-means clustering stability: theoretical analysis and an algorithm[J]. Data Mining and Knowledge Discovery, 2014, 28(4): 918-960

[18] Gaudani H, Lakhani K, Chhatrala R. Survey of document clustering[J]. International Journal of Computer Science and Mobile Computing, 2014, 3(5): 871-874

[19] Xie J Y, Wang Y E. K-means algorithm based on minimum deviation initialized clustering centers[J]. Computer Engineering, 2014, 40(8): 205-211, 223(in Chinese)
谢娟英，王艳娥. 最小方差优化初始聚类中心的 K-means 算法[J]. 计算机工程，2014，40(8)：205-211，223

[20] Bennett K, Robertson J, Milton P M, et al. MATLABapplications for the practical engineer[M]. InTech, 2014