软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(4):798−809 [doi: 10.3724/SP.J.1001.2013.04267]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向领域的学术文献检索框架*

邱江涛[1], 唐常杰[2], 李 庆[1]

[1](西南财经大学 经济信息工程学院,四川 成都 610074)

[2](四川大学 计算机学院,四川 成都 610065)

通讯作者: 邱江涛, E-mail: Jiangtaoqiu@gmail.com

摘 要: 在学术文献检索中,如果检索系统根据用户提交的查询返回相关领域的文献,并将文献按重要程度进行排序,可以帮助用户快速了解相关学术领域.提出一个面向领域的学术文献检索框架,结合引用网络分析和内容分析来发现并排序相关领域重要文献.该框架设计了一个评分函数进行检索,包含两个方面:(1) 论文在所查询领域的重要性;(2) 论文与该领域的相关性.首先研究了一个"社区核"发现算法,从引用网络上发现和查询领域相关的一个文献子集,并对论文计算重要性评分.设计了一种有监督非负矩阵分解算法,该算法使用确定的领域相关文献为先验知识对其他论文进行分类并给出一个评分,以确定论文和查询学术领域的相关性.在真实数据集和合成数据集上的实验,证实了方法的有效性.

关键词: 非负矩阵分解;随机游走;文献检索;引用网络;链接分析

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 邱江涛,唐常杰,李庆.面向领域的学术文献检索框架.软件学报,2013,24(4):798−809. http://www.jos.org.cn/1000-9825/4267.htm

英文引用格式: Qiu JT, Tang CJ, Li Q. Framework for domain-oriented academic literatures retrieval. Ruanjian Xuebao/Journal of Software, 2013,24(4):798−809 (in Chinese). http://www.jos.org.cn/1000-9825/4267.htm

Framework for Domain-Oriented Academic Literatures Retrieval

QIU Jiang-Tao[1], TANG Chang-Jie[2], LI Qing[1]

[1](School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu 610074, China)

[2](School of Computer Science, Sichuan University, Chengdu 610065, China)

Corresponding author: QIU Jiang-Tao, E-mail: Jiangtaoqiu@gmail.com

Abstract: A literature retrieval system, which returns user papers domain-related with queries and ranks papers by importance, can help users quickly learn one academic domain. This paper develops a framework for the domain-oriented literature retrieval, which combines links and contents analysis to search and rank important papers in one academic domain. This framework designs a score function that evaluates both importance of the paper and its relevance to the domain. The study first proposes a community-core discovery algorithm, which is capable of finding a collection of papers domain-related with query from citation network and calculates an importance score for each paper. To assign other papers a domain-related score, a supervised non-negative matrix factorization method, using identified domain-related paper as prior knowledge, is also developed. The experiments conducted on synthetic and real datasets demonstrate the feasibility and applicability of this framework.

Key words: non-negative matrix factorization; random walk; literature retrieval; citation network; link analysis

新进入某个学术领域的研究者通常希望有这样一个检索系统:当提交关于某个领域的查询时,系统能够自

* 基金项目: 国家自然科学基金(61170133); 国家教育部人文社科青年基金(09YJCZH101); 中央高校基本科研业务费(JBK120214); 西南财经大学 211 工程青年教师成长项目(211QN10061)

收稿时间: 2012-01-05; 修改时间: 2012-03-19; 定稿时间: 2012-05-29

动地返回关于该领域的重要文献,并将文献按照重要程度进行排序.传统的文献检索系统,例如中国知网、IEEE Xplore Digital Library 通常是根据用户提交的查询进行关键词匹配检索.它们不能理解查询所指向的学术领域,返回的结果充斥大量非该领域的论文,而且没有对论文的重要性进行排序.Google Scholar 可以对检索结果排序,但没有对论文所属领域进行区分.本文将研究一个面向领域的学术文献检索框架 DOLER(domain-related literature retrieval).

在 DOLER 中,用户提交查询将获得查询所描述的相关研究领域中重要论文的排序.不同于传统信息检索任务根据用户提交查询返回与查询相似度或相关性更高的文档,本文中,我们提出一个 DOLER 评分作为检索的评分函数.DOLER 评分包含两方面内容:(1) 论文在所查询领域的重要性;(2) 论文与该领域的相关性.本文要解决的第 1 个问题是,如何确定查询的领域以及计算领域的重要性评分.学术论文包含参考文献,相互引用的论文通常属于同一个研究领域.将论文间的引用关系绘制成图,可以建立一个引用网络.借用社区发现的思想,本文提出一种基于图上的随机游走的“社区核”发现方法.该方法可以确定一个与查询所描述的领域相关的文献集合,而且对文档进行重要性评分.

本文研究的第 2 个问题是计算论文与查询领域的相关性.引用网络上其他与查询领域相关的文献可能未包含在社区核发现算法确定的一个相关文献集合中.我们使用已确定的文献来判断其他论文与查询领域的相关性.因此,这是一个补集很大的半监督二分类问题.在我们的面向领域的文献检索任务中,不但要进行文本分类,还要对论文进行相关性评分.本文研究了一种基于有监督非负矩阵分解来进行半监督文本分类并进行领域相关性评分的方法.

本文的主要贡献是,开发了一个结合内容分析和引用网络分析进行面向领域学术文献检索的框架.

1 相关工作

1.1 社区发现

社区发现是社会网络分析中的传统研究领域.社区发现的研究主要包括图划分方法[1−3]、基于 Centrality(本文译为中心度)度量的方法[4−6]以及其他方法[7−9]等.MinCut[1]是最简单的图划分算法,该算法寻找移除最少的边将连通图划分成两个子图,此时的子图即为从网络中获得的社区.MinCut 在一些情况下会将图划分得不够均衡,其中一个子图会很小.一些改进算法,如 Normalized Cut[2],Min-Max Cut[3]等,研究解决这一问题.

中心度 Centrality 是对一个节点在网络中重要性的度量,它包括 Degree Centrality,Betweenness Centrality,Closeness Centrality 和 Eigenvector Centrality 等多种度量方法.基于这些网络节点的中心度度量,发展了很多社区发现算法.Newman 的研究[4,5]基于边的 betweenness 度量和评估函数 Modularity 来发现社区.Leicht 等人[6]开发了一种基于特征向量中心度(eigenvector centrality)的算法来分析社会网络中的社区结构.

上述的社区发现算法属于硬划分,即一个节点必须属于一个社区,即使该节点本不属于任何社区.本文研究的“社区核”发现算法将发现社区中更紧密的团,以确定属于该社区的节点.

1.2 非负矩阵分解

非负矩阵分解(non-negative matrix factorization,以下简称 NMF)是矩阵分解研究方式中的一种.不同于其他矩阵降阶方法,如主成分分析、概率潜语义索引等其他方法,NMF 的约束条件包括待分解的数据非负,且分解后得到的低阶数据也非负.NMF 最初由 Lee 和 Seung[10]提出,他们的分解方法在每一次迭代过程中通过乘法更新规则计算矩阵分解结果.Lin[11]提出一种边界约束优化技巧来进行非负矩阵分解,称为投影梯度法.该方法比 Lee 和 Seung 的方法收敛速度更快、运行更稳定.非负矩阵分解已在文本挖掘、图像处理、生物信息学等许多领域得到了成功应用.本文研究一个有监督的 NMF,以实现一种可以评分的二分类算法.

有监督的矩阵分解是将先验知识引入矩阵分解的过程来影响最后的矩阵分解结果.Zhu 等人[12]提出了一种有监督矩阵分解方法,通过建立监督矩阵来计算损失函数,用损失函数影响矩阵分解的过程.对于有监督的非负矩阵分解的研究还不多见.Chen 等人[13]提出一种半监督非负矩阵分解方法 SS-NMF,他们对数据对象给出约

束条件,然后建立奖励和惩罚矩阵作为约束条件,参与到矩阵分解的计算中.本文借助文献[12]的思想,通过引入监督矩阵实现一个有监督的 NMF.

1.3 结合链接分析和内容分析的知识发现

本文的研究采用了结合链接分析和内容分析的方法进行“面向领域的学术文献检索”.已有一些工作[14−17]将链接分析应用于知识发现.它们或者应用 PageRank 算法来评估文档集中的论文重要性[14,17],或者通过分析合作者网络评估作者的影响力[16],或者使用文献的引用分析来评价学者的声誉和受欢迎程度[15].

一些工作结合网络分析和内容分析从文档集合进行文本挖掘.Bolelli[18]将从学术论文集合构建的引用网络和聚类算法结合来发现不同话题的文档簇.Jo 等人[19]将描述了话题的词项分布和从引用网络获得的链接分布相结合,从而在有链接的文档集合上检测话题.Guo 等人[20]提出一个贝努力过程话题(Bernoulli process topic)模型,在引用网络上进行知识发现,该模型为文档集合在文档层和引用层两个层次上建模,然后发现文档集合中的潜在话题.

Yin 等人[21]提出的检索框架结合了文本内容分析和链接分析,他们的研究发现,采用网络节点的“度”作为链接分析更适合生物医学的文献检索.在基于内容的信息检索模型中引入度的权重计算,可以提高生物医学文献检索的性能.Yin 的研究与本文相似,但他们的研究面向生物医学文献检索.

2 面向领域的学术文献检索框架

我们首先通过布尔“或查询”操作检索文献,建立初步检索集合,在初步检索集合上构建一个引用网络.引用网络上,属于相同领域的论文因为相互之间的引用关系会呈现聚类特性.我们由此提出一种“社区核”发现方法,从引用网络上获得一个与查询领域相关的文献集合,并为每篇文档分配一个重要性评分.实践中我们发现,仅使用社区核发现方法获得领域相关文献,检索性能不如预期的好.我们进一步提出一种有监督矩阵分解方法,以社区核为先验知识,对初步检索集合中的每篇论文进行半监督二分类,并分配一个领域相关性评分.结合重要性评分和领域相关性评分,我们最终为初始检索集合中的每篇文档分配一个最终检索评分.图 1 展示了这个框架的结构.

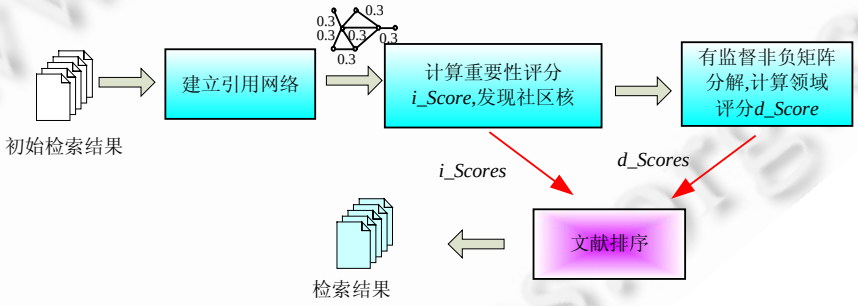


Fig.1 A framework for domain-oriented literature retrieval

图 1 面向领域的学术文献检索框架

2.1 学术引用网络上的社区核发现

社区发现方法从社会网络中发现的社区是联系紧密的成员的集合.因为有引用关系的文献在研究领域上存在相似性,将社区发现方法应用在引用网络上发现的社区则是有一定领域相似性的文献集合.特别地,位于社区中心的一个文献子集之间的领域相关性应该非常大.因此,我们考虑获得社区中的位于社区中心的一个联系更紧密的成员(文献)集合,称为“社区核”.定义 1 中给出了社区核的形式化定义.

基于初步检索结果构建引用网络时,我们计算引用网络上有边的两个节点(论文)之间的内容相似性,然后,以内容相似度作为边的权重,就可以初步把内容分析和引用网络分析结合起来,此时建立的学术引用网络是一

个加权无向图.

为了发现社区核,需要对社区中的每个成员计算一个重要性评分,分值最高的成员即社区的中心.我们在一个加权无向图上考察每个节点的重要程度.其中,连接的边越多且边权重越大的节点,重要性也越高.我们改进了 PageRank 算法,在加权无向图上计算每个节点的重要性分值.在多次迭代计算中,根据上次迭代的分值,计算本次迭代的分值.其计算公式如下所示:

$$R_{k+1}[i]=\frac{1-d}{N}+d\sum_{j\in M(i)}G_{ij}\frac{R_k[j]}{G_j} \tag{1}$$

$R_{k+1}[i]$表示在第 k+1 次迭代中,节点 i 的重要性分值.计算时,将考察所有连接到节点 i 的节点 $M(i)$.N 为节点的总数,d 是阻尼因子调节迭代过程的收敛性,G_{ij} 是连接节点 i 和节点 j 的边权重,G_j 是与节点 j 相连的所有边的权重之和.当$\|R_{k+1}-R_k\|_1$ 小于某个阈值时,迭代过程终止.

图上的随机游走[22]有聚类特性,我们利用图上的随机游走来发现社区核.寻找从网络的某个节点开始正向 t 步随机游走最可能到达的结束节点,通过随机游走重新绘制网络节点之间的关系.从网络上信息流动的角度考虑,一个节点更可能与一个比自己重要性高的节点进行连接.因此,如果结束节点的重要性分值比起始节点的分值要高,则我们在一张新的图(称为成员关联图)上,从起始节点到结束节点绘制一条边.当考察网络上从每个节点进行正向 t 步随机游走的结束节点时,绘制完成的成员关联图上的联通子图就是发现的社区.

图 2 是一个从引用网络中发现社区的例子,成员关联图描述了从引用网络重新获取的节点间的关系.图中包含两个连通子图,每一个子图即为一个社区.

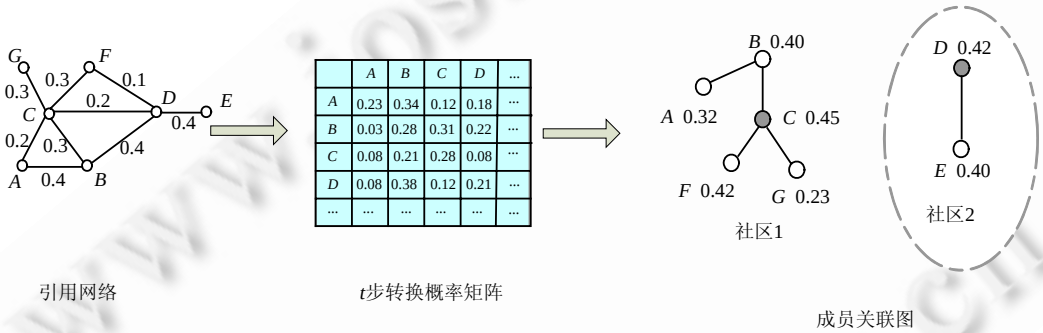


Fig.2 Discovering community and community-core

图 2 社区和社区核的发现

下面的算法 1 对绘制成员关联图的方法进行了具体描述.

算法 1. 绘制成员关联图.

输入:引用网络 G;

输出:成员关联图 C.

步骤:

1. Loop
2. For each node $p_i\in G$
3. $$R_{k+1}[i]=\frac{1-d}{N}+d\sum_{j\in M(i)}G_{ij}\frac{R_k[j]}{G_j}$$
4. End
5. $e\leftarrow\|R_k\|_1-\|R_{k+1}\|_1$
6. $R_{k+1}\leftarrow R_{k+1}+e\times S$
7. $\delta\leftarrow\|R_{k+1}-R_k\|_1$
8. $R_{k+1}\leftarrow R_k$

9. while ($\varepsilon<\delta$)
10. C←[*null*,…,*null*]
11. A←由引用网络 G 产生 1 步随机转换矩阵
12. $Z_{jj}=\sum_i [A^t]_{ij}$
13. M^t←$A^t\times Z^{-1}$
14. For each member i in R
15. k←argmax$_j M^t[i][j]$
16. if $R[k]>R[i]$
17. $C[i]=k$
18. End
19. Return C

算法 1 的时间复杂度是 $O(N\times(iter+1))$,N 是网络的节点数量,$iter$ 是改进 PageRank 算法的迭代次数.

算法 1 中,向量 R 的分值代表成员的重要性,分值越大的成员所代表的论文在其所在领域的重要性越高.由引用网络产生一个 1 步随机游走转换矩阵,矩阵的元素表示网络中两个节点的转换概率.A 是一个行随机矩阵,即 $\sum_j A_{ij}=1$.$A[j,k]$是从节点 j 到 k 的 1 步转换概率:

$$P_{t+1|t}(k\mid j)=\begin{cases}(1-s)G_{jk}\Big/\sum_i G_{ji}, & \forall k\neq j\\ s, & k=j\end{cases},$$

其中,s 表示自转换概率,G_{ji} 是节点 j,i 之间的边的权重.M^t 是 t 步概率转换矩阵,$M^t=A(A\ldots(A))$.为了对 t 步概率转换矩阵规范化,需乘上对角阵 Z 的逆矩阵 Z^{-1}.基于 t 步概率转换矩阵 M^t,我们寻找从节点 i 开始正向随机游走 t 步后最有可能到达的节点 j.如果节点 j 的重要性分值大于 i 的分值,则设 $C[i]=j$.当所有操作完成时,如果 $C[i]$值为 *null*,则表示未能在网络中发现一个比节点 i 的分值更高的从节点 i 开始随机游走的结束节点.

最后,从向量 C 中我们可以绘制一张如图 2 所示的成员关联图,其中每个连通子图就是一个社区.

定义 1(社区核). 设Φ是使用算法 1 从一个引用网络中获得的一个联通子图,即一个社区,则该联通子图中的最高分节点以及有边与这个最高分节点相连的节点集合称为Φ的社区核.

在图 2 中,节点集合{*B*,*C*,*F*,*G*}是社区 1 的社区核.

2.2 有监督非负矩阵分解

采用算法 1,从引用网络有可能发现多个社区核.我们将最高得分成员所在的社区核作为与查询领域相关的文献集合.如果将这个社区核中的论文看作是贴有正例标签的样本,而判断引用网络中其他论文是否与社区核属于同一个学术领域,就是一个半监督二分类问题.因为分类时还要对论文评分,评价论文在该学术领域的重要程度(或属于该领域的程度),因此我们提出一种有监督非负矩阵分解算法来完成这项任务.

非负矩阵分解就是给定目标函数:

$$J=\|V-WH\|_F^2 \tag{2}$$

求解 $\min\limits_{W,H}\|V-WH\|_F^2$ 的问题.其中,V 是 $n\times m$ 待分解矩阵,W 是分解后的 $n\times k$ 矩阵,H 是分解后的 $k\times m$ 矩阵,满足 $W_{ij}>0$,$H_{ij}>0$.当 V 是文档-词项(doc-term)矩阵时,分解的 k 通常代表文档的话题、主题或类别.因此,分解后的 W 矩阵是文档-类别(doc-class)矩阵,H 矩阵是类别-词项(class-term)矩阵.

我们为初始检索文献集合建立文档-词项(doc-term)矩阵.为了实施有监督非负矩阵分解,首先给公式(2)添加损失函数,该损失函数需要一个监督矩阵 Y.Y 也是一个 $n\times k$ 矩阵.Y_{ij} 的取值有 3 种:1 表示第 i 个文档属于第 j 类,–1 表示第 i 个文档不属于第 j 类,0 则表示不知道该文档的类别.W_i 是矩阵 W 的向量.W_i 中的每个元素除以 W_i 中最大元素值,将向量 W_i 的值规范化到[0,1]范围.我们给出损失函数 $L=\lambda\sum 1-Y_{ij}g(W_{ij})$, 函数 $g(W_{ij})$将矩阵 W 的值转换成与监督矩阵表达一致的–1 或 1 值.损失函数将判断最终的 W 分解结果是否与监督矩阵一致.若一致,

则损失函数为 0;差别越大,损失函数的值越大.g 函数是一个分段函数 $g(x)=\begin{cases}1, & x>=1\\ -1, & \text{otherwise}\end{cases}$.但分段函数不能求导,不能参与到矩阵分解的计算中.因此,我们给出一个连续函数模拟该分段函数 $g(x)=1-2/(1+\exp(100x-95))$.

为了让矩阵分解的过程保持健壮性,在公式(2)中增加两个正则化(regularization) $\frac{\alpha}{2}\|W\|_F^2$ 和 $\frac{\beta}{2}\|H\|_F^2$.因此,可得监督非负矩阵分解的目标函数,见公式(3):

$$J=\|V-WH\|_F^2+\lambda\sum 1-Y_{ij}g(W_{ij})+\frac{\alpha}{2}\|W\|_F^2+\frac{\beta}{2}\|H\|_F^2 \tag{3}$$

有监督非负矩阵分解的任务就是求解 $\min_{W,H}(J)$.当按照梯度法来求解上述公式时,可以得到如下两个更新公式:

$$W_{ij}\leftarrow W_{ij}+\eta_W\frac{\partial J}{\partial W},$$

$$H_{ij}\leftarrow H_{ij}+\eta_H\frac{\partial J}{\partial H}.$$

对于 $\frac{\partial J}{\partial W}=\frac{\partial\|V-WH\|_F^2}{\partial W}+\frac{\partial\lambda\sum 1-Y_{ij}g(W_{ij})}{\partial W}+\frac{\alpha}{2}\times\frac{\partial\|W\|_F^2}{\partial W}$,因为,

- $\frac{\partial\|V-WH\|_F^2}{\partial W}=WHH^T-VH^T$;
- $\frac{\partial\lambda\sum 1-Y_{ij}g(W_{ij})}{\partial W}=G_W$;
- G_W 是一个 $n\times k$ 矩阵,它的第(i,j)个元素是$-\lambda Y_{ij}g'(W_{ij})$;
- $\frac{\alpha}{2}\times\frac{\partial\|W\|_F^2}{\partial W}=\alpha W$.

因此可得 $\frac{\partial J}{\partial W}=WHH^T-VH^T+G_w+\alpha W$.

对于 $\frac{\partial J}{\partial H}=\frac{\partial\|V-WH\|_F^2}{\partial H}+\frac{\beta}{2}\times\frac{\partial\|H\|_F^2}{\partial H}$, 因为 $\frac{\partial\|V-WH\|_F^2}{\partial H}=W^TWH-W^TV$, $\frac{\beta}{2}\times\frac{\partial\|H\|_F^2}{\partial H}=\beta H$, 因此可以得到 W_{ij} 和 H_{ij} 的更新公式:

$$W_{ij}\leftarrow W_{ij}+\eta_W(WHH^T-VH^T+G_w+\alpha W)_{ij},$$

$$H_{ij}\leftarrow H_{ij}+\eta_H(W^TWH-W^TV+\beta H)_{ij}.$$

根据文献[10]的方法,我们令 $\eta_W=-\frac{W_{ij}}{(WHH^T)_{ij}}$ 和 $\eta_H=-\frac{H_{ij}}{(W^TWH)_{ij}}$,则最终的有监督非负矩阵分解更新公式如下所示:

$$W_{ij}\leftarrow\frac{W_{ij}}{(WHH^T)_{ij}}(VH^T+G_w-\alpha W)_{ij},$$

$$H_{ij}\leftarrow\frac{H_{ij}}{(W^TWH)_{ij}}(W^TV-\beta H)_{ij}.$$

将矩阵 V 分解后得到 W 矩阵,本文中,W 矩阵是 $n\times 2$ 矩阵,矩阵中一条行向量代表一个文本,行向量 W_i 的元素 W_{ij} 的分值就是文档 i 属于类 j 的分值.文档 i 属于最高分值的元素所对应的类.下面的算法 2 对基于 SNMF 的二分类算法进行了描述.

算法 2. 基于 SNMF 的二分类.

输入:$n\times m$ 矩阵 V,$n\times 2$ 监督矩阵 Y;

输出:分类结果.

步骤:

1. 随机产生 $n\times 2$ 的矩阵 W 和 $2\times n$ 的矩阵 H
2. Loop
3. W 规范化处理
4. 计算矩阵 $Gw, Gw_{ij} \leftarrow -\lambda Y_{ij} g'(W_{ij})$
5. 更新矩阵 W 和 $H, W_{ij} \leftarrow \frac{W_{ij}}{(WHH^T)_{ij}}(VH^T + G_w - \alpha W)_{ij}, H_{ij} \leftarrow \frac{H_{ij}}{(W^TWH)_{ij}}(W^TV - \beta H)_{ij}$
6. $\delta \rightarrow \| V - WH \|_F^2$
7. while ($\varepsilon < \delta$)
8. Return W

算法 2 的时间复杂度是 $O(iter)$,$iter$ 是算法循环直到收敛的迭代次数.

3 实 验

本节将通过实验评估面向领域的学术文献检索框架 DOLER 的性能.实验在 Pentium Dual-Core 1.8G CPU,内存 3G 的计算机上进行.采用 Java 和 Matlab 来实现算法.

我们采用合成网络来评估算法 1 中的参数对算法性能的影响.为了比较所提出的有监督 NMF 方法和无监督 NMF 方法,我们使用 UCI 数据集中的 Breast Cancer Wisconsin Original(简写为 WOBC)、Breast Cancer Wisconsin Diagnostic(简写为 WDBC)、Ecoli 数据集的 *cp* and *pp* 类和 Abalone 数据集的 6 类、7 类(abalone1)、11 类、12 类(abalone2)构成测试集,在测试集上做二分类实验.

SEWM 2010 信息检索测评竞赛(http://www.cwirf.org/2010WebTrack/lt/)提供了 SIGIR,KDD,CIKM 等 11 个国际会议截止到 2009 年的部分论文集,共有 10 840 篇论文.SEWM 2010 竞赛的任务之一就是根据给出的查询检索查询相关领域中最重要的论文并排序.我们使用该论文集,竞赛提供 9 个查询和答案集(以下简称 S10C 数据集)作为评估 DOLER 性能的数据集.

3.1 实验1

在本节,我们将评估算法 1 中的两个参数随机游走步数 t 和自转换概率 s 对算法性能的影响.因为算法 1 可以发现社区,由此我们采用在合成网络上发现社区的方式来评估.实验采用文献[5]中提出的方法产生合成数据集;采用 Modularity 度量[5]作为发现社区质量的评价指标.随机产生合成网络时需要设置参数:节点数 v,社区数 c,图中每个节点平均连接边的数目 z,每个节点平均连接到外部社区的边数 $zout$.

实验 1 分别在节点数为 32,64,128,256 的合成网络上进行实验.我们先讨论节点数为 128 的合成网络,参数设置如下:v=128,c=4,z=16,$zout$=4.表示每个网络共有 128 个节点,产生 4 个社区,平均每个社区有 32 个节点,每个节点平均有 4 条边链接到外部社区.我们随机产生 100 个网络.随机游走步数 t 和自转换概率 s 被设置为

$$t=\{10,20,\ldots,100\}, s=\{0.8,0.85,0.87,0.9,0.92\}.$$

对于每组参数$\langle t,s\rangle$,在 100 个网络上应用算法 1 来发现社区,计算社区发现结果的评价 Modularity 值.

产生节点数为 32,64 和 256 的合成网络参数分别如下:

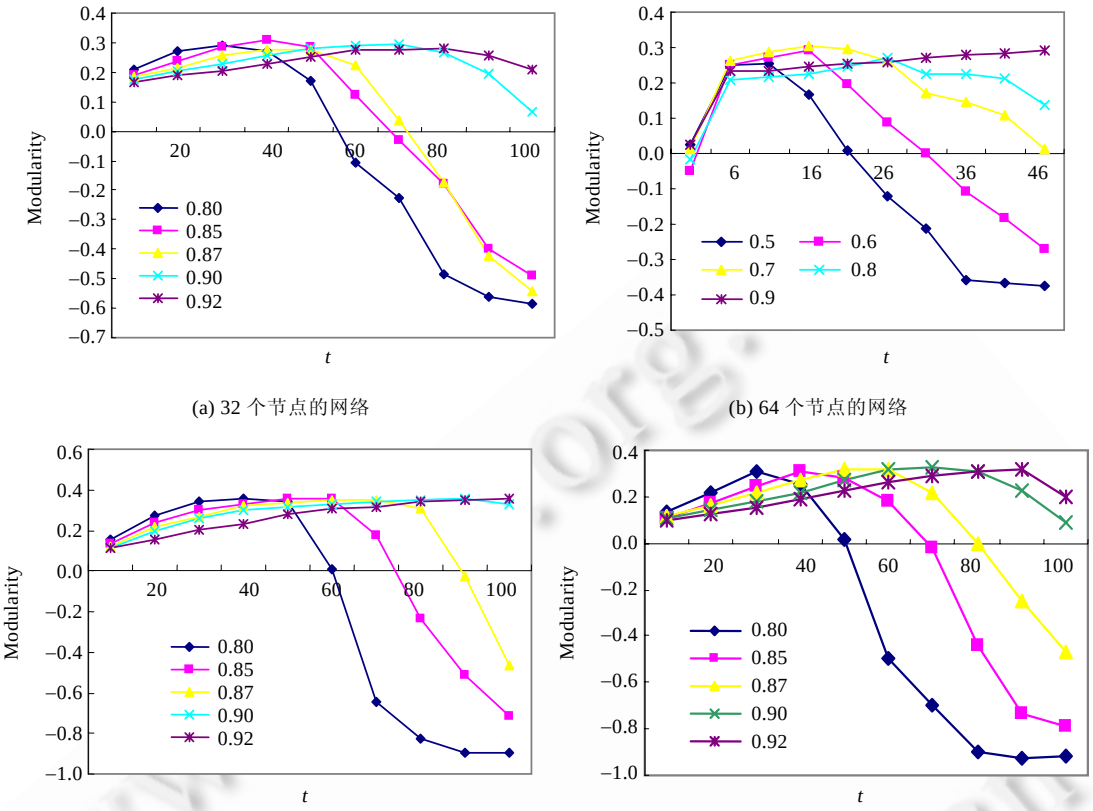
$$\{v=32,c=4,z=4,zout=1\},\{v=64,c=4,z=8,zout=2\},\{v=256,c=4,z=64,zout=16\}.$$

实验方法如前所述.实验结果如图 3(a)、图 3(b)和图 3(d)所示.

由图 3 所示实验结果可以观察到:在每个 s 值,当随机游走步数 t 开始增加时,Modularity 值会到达一个峰值;而后,再增加随机游走步数,Modularity 值开始下降.每条随机行走曲线到达的峰值几乎相等.例如,在图 3(c)节点数为 128 的网络,对于参数 s={0.8,0.85,0.87,0.9,0.92},算法 1 可以在对应的随机游走步数 t={30,40,50,70,90}时到达峰值,每条曲线的峰值约为 0.35.我们还可以观察到,对于不同规模的网络,在相同自转换概率情况下,越大规模的网络需要更大的行走步数到达峰值.例如,对于 0.9 的自转换概率,规模为 32,64,128,256 节点数的网络在

行走步数分别为 26,70,80,90 时达到峰值.

此观察结论将作为实践经验在本文中使用.其理论证明超出了本文讨论范围,我们将在以后的工作中予以研究.



(a) 32 个节点的网络 (b) 64 个节点的网络

(c) 128 个节点的网络 (d) 256 个节点的网络

Fig.3 Evaluating parameters of Random Walk steps and self-transition probability

图 3 随机游走步数和自转换概率对社区发现的影响

3.2 实验2

已有一些研究将非负矩阵分解应用于分类.本实验中将对比传统非负矩阵分解方法(称为 NMF)和本文提出的有监督非负矩阵分解方法(称为 SNMF)在二分类上的性能.分类的性能用 Accuracy 作为评价指标.

$$Accuracy=\frac{\text{正确分类的样本数目}}{\text{所有样本数目}}.$$

我们取每个数据集中 10%的数据形成 SNMF 的监督矩阵.

在非负矩阵分解的每次迭代中计算 $\delta=\|V-WH\|_F^2$,如果在第 k 次迭代中得到的δ_k和在第 k+1 次迭代中得到的δ_{k+1} 的差值小于预设阈值$\varepsilon\geqslant\delta_k-\delta_{k+1}$,则停止迭代.对于一个数据集,每次进行非负矩阵分解时并不能保证分解得到的 $\delta=\|V-WH\|_F^2$ 都是最小值,我们取δ最小的 10 次分解结果来获得数据集的分类结果,然后计算 10 次分类的 Accuracy 平均值.图 4 展示了在 5 个数据集上,使用 NMF 和 SNMF 进行分类的平均 Accuracy 结果.

从图中我们可以观察到:在其中 4 个数据集上,SNMF 都比 NMF 表现出了更好的分类性能;但在 abalone2 数据集上,SNMF 比 NMF 的分类精确度略低.可以看到,此时,NMF 和 SNMF 的分类精确度都略高于 50%,几乎都没有对数据集进行有效分类.由此推断,使用非负矩阵分解方法进行分类,在一些数据集上并不能取得良好的

效果.对这些数据集特性的分析超出了本文的讨论范围.总之,实验证实,我们提出的有监督非负矩阵分解方法可以有效地影响非负矩阵分解的过程,使得分解结果按照用户预设的方向发展.

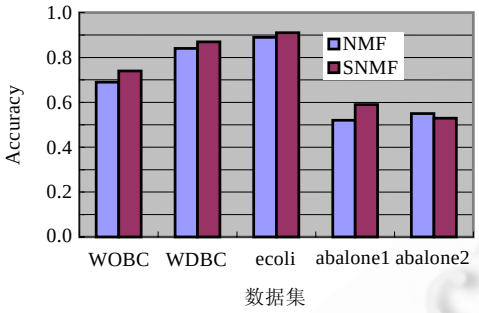


Fig.4 Classifications on UCI dataset

图 4 在 UCI 数据集上的分类实验结果

3.3 实验3

本节将使用 S10C 数据集评估 DOLER 的性能.采用 MAP(mean average precision)作为性能的评价指标.参数随机游走步数和自转换概率⟨*t*,*s*⟩的取值为 0.85 和 40.

文献[21]提出了一个通用概率检索框架(本文称其为 GPF),该框架集成了链接信息和基于内容的概率检索模型.本文研究的"面向领域的学术文献检索"与该框架完成的任务相近,因此,本节将对两个框架进行实验比较分析.GPF 的排序函数为

$$O(R\mid q,d,L_d)\propto \sum_i \log\frac{P(x_i\mid R,q)}{P(x_i\mid NR,q)}+\log\frac{P(L_d\mid R,q)}{P(L_d\mid NR,q)}.$$

R 表示相关,NR 表示不相关;q 表示查询;d 表示文档;x 是一个词项; $L_d=\sum_i^N L_{d_i}$ 指示文档 d 的链接信息.当文档 d 和文档 i 有边链接时, $L_{d_i}=1.\sum_i \log\frac{P(x_i\mid R,q)}{P(x_i\mid NR,q)}$ 表示一篇文档 d 基于内容的排序评分,由 BM25 计算得到. $\log\frac{P(L_d\mid R,q)}{P(L_d\mid NR,q)}=D_d\times\log\frac{N+0.5}{n+0.5}$ 为文档 d 的链接权重,其中,D_d 是文档 d 作为引用网络节点的度,N 是文档集合中文档的数量,n 是 $L_{d_i}=1$ 的文档数量.因此,GPF 的排序函数 GPF 可以写作

$$GPF_d=BM25(d)+D_d\times\log\frac{N+0.5}{n+0.5}.$$

BM25 模型[23]是一种应用广泛的信息检索模型,它通过对查询和文档的相关性的计算来进行文献检索.针对短查询的 BM25 模型为

$$RSV_d=\sum_{t\in q}\log\left[\frac{N}{df_t}\right]\cdot\frac{(k_1+1)tf_{td}}{k_1[(1-b)+b\times(L_d/L_{ave})]+tf_{td}}.$$

RSV_d 是为一篇文档 d 计算的评分;$\log(N/df_t)$是词项 t 的逆文档频率;tf_{td} 是词项 t 在文档 d 中的权重;L_d 和 L_{ave} 分别是文档 d 的长度及整个文档集中文档的平均长度;K_1 是一个取正值的调优参数,用于对文档中的词项频率进行缩放控制;另一个调优参数 b 决定文档长度的缩放程度.在本节,我们按照文献[24]中给出的经验值设置参数:k_1=1.2,b=0.75.

我们使用 BM25 作为基于内容的检索的评测基准;使用 PageRank 算法作为基于引用网络分析的检索方法的评测基准.Apach Lucene 是一个开源的文本搜索引擎库.Joaquin 开发了一个基于 Lucene 的 BM25 程序(http://nlp.uned.es/~jperezi/Lucene-BM25/).实验中,我们使用该程序实施 BM25 的检索模型.所有 BM25 评分和 PageRank 评分将被规范化到[0,1]之间.本文的 DOLER 框架将联合算法 1 计算的重要性评分 *i_Score* 和 SNMF 为每篇文档计算的相关性评分 *d_Score* 作为文献的最后评分 DOLER.

$$DOLER_d=(1-a)\times i\_Score(d)+a\times d\_Score(d) \quad (4)$$

重要性评分 *i_Score* 和相关性评分 *d_Score* 被规范化到[0,1]之间.我们首先评估参数 *a* 对 DOLER 性能的影响.参数 *a* 的取值从 0~1,步长为 0.1.对应每个 *a* 的取值,我们应用 DOLER 到 S10C 数据集.然后计算所有查询的平均 MAP 值.图 5 显示了实验结果.从图中可以观察到,当 *a* 取值为 0.9 时.DOLER 可以取得最佳性能.

DOLER 框架结合了引用网络分析(*i_Score*)和内容分析(*d_Score*).从图 5 中可以观察到,基于内容的检索比单独的基于引用网络分析的检索性能要好.两种检索模型的 MAP 值分别是 0.49 和 0.37;DOLER 框架计算的 DOLER 评分在 *a*=0.9 时达到了最高分值 0.51,比 *d_Score* 略有提高.我们认为,这是因为采用 SNMF 计算 *d_Score* 得分时,已经结合了内容分析和引用网络分析所致.

算法 1 使用改进的 PageRank 算法在以文档相似度作为权重的引用网络上计算论文的重要性评分 *i_Score*.图 5 中,DOLER 和 PageRank+BM25 在参数 *a*=0 和 *b*=0 时的 MAP 值分别是 0.37 和 0.18.这意味着在文献检索上,本文算法 1 中提出的重要性分值计算比基于 PageRank 的算法有更好的性能.

我们也考察一个 PageRank+BM25 的组合模型,该模型的评分 *p_Score* 按公式(5)计算.

$$p\_Score_d=(1-b)\times \mathrm{PageRank}(d)+b\times \mathrm{BM25}(d) \quad (5)$$

图 5 显示,该组合模型在 *b*=0.8 时获得最佳性能.比单独的 BM25 模型可提高 4%的 MAP 值.这意味着,即使简单地将引用网络分析和内容分析相结合,也都能够提高检索系统的性能.

我们进一步比较 5 个检索模型:BM25,PageRank,BM25+PageRank,GPF 和 DOLER 的性能,参数 *a*=0.9,*b*=0.8.应用这 5 个模型在 S10C 数据集上,基于每个查询计算评价查准率 AP(average precision).图 6 展示了实验结果,表 1 列出了这 4 个模型的 MAP 值.从图 6 可以观察到,在每个查询上,DOLER 都表现出比 BM25 更好的性能.这意味着,在学术文献检索中引入引用分析可以提高检索的性能.

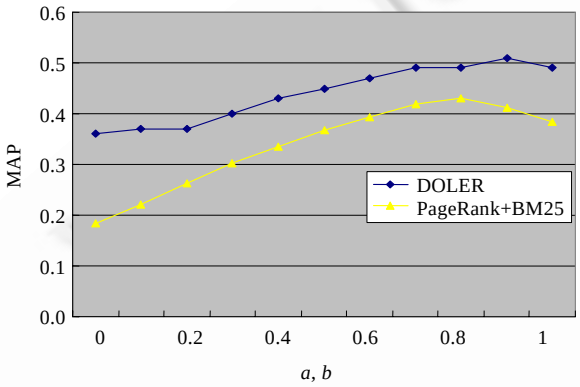


Fig.5 Experiments on academic literature retrieval

图 5 学术文献检索实验结果

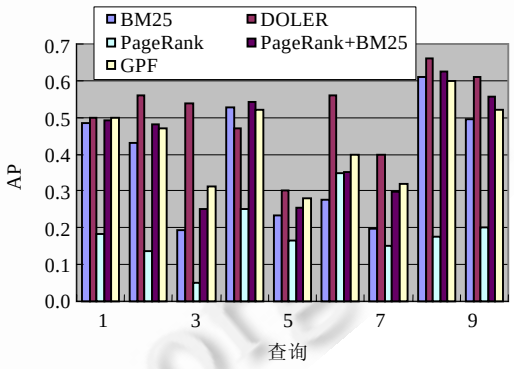


Fig.6 Comparison of retrieval models

图 6 检索模型性能比较

Table 1 MAP of five retrieval models

表 1 5 个模型的 MAP 值

	PageRank	BM25	PageRank+BM25	GPF	DOLER
MAP	0.184	0.388	0.429	0.436	0.511

GPF 模型与 PageRank+BM25 模型相比没有显著的性能优势.文献[21]描述,在生物医学文献检索方面,该模型使用引用网络节点的度作为链接权重,比 PageRank 和 Hits 等链接分析算法计算链接权重具有显著的优势.这也意味着,该模型更适用于生物医学文献检索.

虽然 DOLER 和 GPF,PageRank+BM25 模型都结合了内容分析和引用网络分析,但 DOLER 比 GPF 和 PageRank+BM25 表现出了更好的性能.这是因为,DOLER 是面向领域的文献检索模型,而 PageRank+BM25 只是简单的内容分析和引用网络分析的组合,GPF 更适合面向生物医学领域的文献检索.因此我们可以得出结论,对

于按照主题进行学术文献检索并将检索结果按重要性排序的任务,我们的方法是有效的.

4 结束语

根据用户提交的查询获取相关学术领域的文献并将文献按照重要性排序,是信息检索领域一个有趣的研究问题.针对该问题,本文提出了一个结合引用网络分析和内容分析的检索框架 DOLER.不同于传统信息检索任务,根据用户提交查询返回与查询相似度或相关性更高的文档,DOLER 设计了一个评分函数进行检索,包含两方面内容:(1) 论文在所查询领域的重要性;(2) 论文与该领域的相关性.该框架提出一种社区核发现方法发现与查询领域相关的论文集合,并计算论文重要性评分 *i_Score*.为了进一步计算论文的领域相关性评分,提出一个有监督非负矩阵分解方法.以社区核所确定的相关文献作为监督矩阵,将初步检索结果形成的 doc-term 矩阵进有监督非负矩阵分解.分解结果将论文进行分类并分配一个领域相关性的评分 *d_Score*.面向领域的文献检索框架 DOLER 使用 *i_Score* 和 *d_Score* 两个评分来计算文献的最终评分.

在 UCI 数据集上的实验结果表明,我们的有监督非负矩阵分解方法与无监督的非负矩阵分解方法相比,可以提高分类的性能.在真实数据集上的学术文献检索实验显示,我们的方法比基准方法在 MAP 值上有较大的提高,从而证实我们的方法在面向领域的学术文献检索上是有效的.

References:

[1] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(8): 888−905. [doi: 10.1109/34.868688]

[2] Hagen L, Kahng AB. New spectral methods for ratio cut partitioning and clustering. IEEE Trans. on Computed Aided Design, 1992, 11(9):1074−1085. [doi: 10.1109/43.159993]

[3] Ding C, He X, Zha H, Gu M, Simon H. A min-max cut algorithm for graph partitioning and data clustering. In: Cercone N, ed. Proc. of the 2001 IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2001. 107−114. [doi: 10.1109/ICDM.2001.989507]

[4] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6):66−72. [doi: 10.1103/PhysRevE.69.066133]

[5] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):1−15. [doi: 10.1103/PhysRevE.69.026113]

[6] Leicht EA, Clarkson G, Shedden K, Newman MEJ. Large-Scale structure of time evolving citation networks. The European Physical Journal B—Condensed Matter and Complex Systems, 2007,59(1):75−83. [doi: 10.1140/epjb/e2007-00271-7]

[7] Shen HW, Cheng XQ, Chen HQ, Liu Y. Information bottleneck based community detection in network. Chinese Journal of Computers, 2008,31(4):677−686 (in Chinese with English abstract).

[8] Yang N, Lin SX, Gao Q, Meng XF. Discovering signature of potential Web communities from clusters of MCL. Chinese Journal of Computers, 2007,30(7):1086−1093 (in Chinese with English abstract).

[9] Gan WY, He N, Li DY, Wang JM. Community discovery method in networks based on topological potential. Ruanjian Xuebao/Journal of Software, 2009,20(8):2241−2254 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3318.htm [doi: 10.3724/SP.J.1001.2009.03318]

[10] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature, 1999,401(6755):788−791. [doi: 10.1038/44565]

[11] Lin CJ. Projected gradient methods for non-negative matrix factorization. Neural Computation, 2007,19(10):2756−2779. [doi: 10.1162/neco.2007.19.10.2756]

[12] Zhu SH, Yu K, Chi Y, Gong YH. Combining content and link for classification using matrix factorization. In: Kraaij W, ed. Proc. of the 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2007. 487−494. [doi: 10.1145/1277741.1277825]

[13] Chen YH, Rege M, Dong M, Hua J. Non-Negative matrix factorization for semi-supervised data clustering. Knowledge and Information Systems, 2008,17(3):355−379. [doi: 10.1007/s10115-008-0134-6]

[14] Chen P, Xie H, Maslov S, Redner S. Finding scientific gems with Google's PageRank algorithm. Journal of Infometrics, 2007,1(1): 8−15. [doi: 10.1016/j.joi.2006.06.001]

[15] Ding Y, Cronin B. Popular and/or prestigious? Measures of scholarly esteem. Information Processing & Management, 2011,47(1): 80−96. [doi: 10.1016/j.ipm.2010.01.002]

[16] Yan EJ, Ding Y. Discovering author impact: A PageRank perspective. Information Processing & Management, 2011,47(1): 125−134. [doi: 10.1016/j.ipm.2010.05.002]

[17] Ma N, Guan JC, Zhao Y. Bringing PageRank to the citation analysis. Information Processing and Management, 2008,44(2): 800−810. [doi: 10.1016/j.ipm.2007.06.006]

[18] Bolelli L, Ertekin S, Giles CL. Clustering scientific literature using sparse citation graph analysis. Lecture Notes in Computer Science, 2006,4213:30−41. [doi: 10.1007/11871637_8]

[19] Lagoze YJC, Giles CL. Detecting research topics via the correlation between graphs and texts. In: Berkhin P, ed. Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2007. 370−379. [doi: 10.1145/1281192.1281234]

[20] Guo Z, Zhang ZF, Zhu SH, Chi Y, Gong YH. Knowledge discovery from citation networks. In: Wu XD, ed. Proc. of the 2009 IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2009. 800−805. [doi: 10.1109/ICDM.2009.137]

[21] Yin XS, Huang JXJ, Li ZJ. Mining and modeling linkage information from citation context for improving biomedical literature retrieval. Information Processing & Management, 2011,47(1):53−67. [doi: 10.1016/j.ipm.2010.03.010]

[22] Craswell N, Szummer M. Random walks on the click graph. In: Kraaij W, ed. Proc. of the 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2007. 239−246. [doi: 10.1145/1277741.1277784]

[23] Jones KS, Walker KS, Robertson SE. A probabilistic model of information retrieval: Development and comparative experiments. Information Processing & Management, 2000,36(6):779−840. [doi: 10.1016/S0306-4573(00)00015-7]

[24] Manning CD, Raghavan P, Schutze H, Wrote; Wang B, Trans. Introduction to Information Retrieval. Beijing: Post & Telecom Press, 2010. 160−161 (in Chinese).

附中文参考文献:

[7] 沈华伟,程学旗,陈海强,刘悦.基于信息瓶颈的社区发现.计算机学报,2008,31(4):677−686.

[8] 杨楠,林松祥,高强,孟小峰.一种从马尔可夫聚类簇发现潜在 Web 社区特征的方法.计算机学报,2007,30(7):1086−1093.

[9] 淦文燕,赫南,李德毅,王建民.一种基于拓扑势的网络社区发现方法.软件学报,2009,20(8):2241−2254. http://www.jos.org.cn/1000-9825/3318.htm [doi: 10.3724/SP.J.1001.2009.03318]

[24] Manning CD,Raghavan P,Schutze H,著;王斌,译.信息检索导论.北京:人民邮电出版社,2010.160−161.

邱江涛(1972－),男,四川攀枝花人,博士,副教授,CCF 会员,主要研究领域为数据挖掘,社会计算.
E-mail: Jiangtaoqiu@gmail.com

李庆(1976－),男,博士,教授,博士生导师,主要研究领域为信息检索,数据挖掘.
E-mail: kooliqing@gmail.com

唐常杰(1946－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,数据挖掘.
E-mail: cjtang@scu.edu.cn