ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 6,Nov. 2015,pp. 1246－1252
DOI:10.16337/j.1004-9037.2015.06.013

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于特征映射的微博用户标签兴趣聚类方法

秦　雨[1,2]　余正涛[1,2]　王炎冰[1,2]　石林宾[1,2]　潘华山[1,2]

(1.昆明理工大学信息工程与自动化学院,昆明,650500; 2.昆明理工大学智能信息处理重点实验室,昆明,650500)

摘　要: 针对现有的用户兴趣聚类方法没有考虑用户标签之间存在的语义相关性问题,提出了一种基于特征映射的微博用户标签兴趣聚类方法。首先,获取待分析用户及其所关注用户的用户标签,选取出现频数高于设定阈值的标签构建模糊矩阵的特征维;然后,考虑标签之间的语义相关性,利用特征映射的思想将用户标签根据其与特征维标签之间的语义相似度映射到每个特征维下,计算每个特征维所对应的特征值;最后,利用模糊聚类得到了不同阈值下的用户兴趣聚类结果。实验结果表明,本文提出的基于特征映射的微博用户标签兴趣聚类方法有效地改善了用户兴趣聚类效果。

关键词: 微博;特征映射;模糊聚类;语义相似度

中图分类号: TP391　　**文献标志码:** A

Micro-blog User Label Interest Clustering Method Based on Feature Mapping

Qin Yu[1,2], Yu Zhengtao[1,2], Wang Yanbin[1,2], Shi Linbin[1,2], Pan Huashan[1,2]

(1. Institute of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China; 2. Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, 650500, China)

Abstract: Since many methods for cluster user interest does not consider the semantic similarity of the user labels, a micro-blog user label interest clustering method is introduced based on feature mapping. Firstly, the user labels of the target users and their focus users are obtained, then the labels with the higher frequency than the threshold value is chosen. Therefore, a feature space is created. Secondly, the user labels are mapped to the feature space by calculating the semantic similarity based on the feature mapping. Finally, the fuzzy clustering is utilized to obtain the clustering result of different threshold value. Experimental results show that the method greatly improves the clustering accuracy rate for user interest clustering.

Key words: micro-blog; feature mapping; fuzzy clustering; semantic similarity

引　　言

随着社交网络的不断发展,社交网络的参与者越来越多,面对如此庞大的用户群体,如何准确地把握用户兴趣,自动为用户找到与之兴趣相近的用户,成为许多专家和学者研究的热点问题。在用户兴趣

基金项目: 国家自然科学基金(61175068)资助项目。
收稿日期: 2014-06-05;**修订日期:** 2014-06-30

挖掘领域,国内外专家学者已经开展了大量的研究工作。Huang 等[1]提出一种通过形式概念分析技术从正例文档中建立用户兴趣模型的方法。Pazzani 等[2]通过分析用户对页面的收藏行为和添加书签的行为构建用户兴趣模型。Shen 等[3-4]综合考虑用户的查询轨迹、用户浏览过的网页信息以及用户在各网站上的点击次数构建用户兴趣模型。Zhou 等[5]利用认知情感理论,以用户浏览轨迹作为数据来源,通过计算用户之间的相关程度构建用户兴趣模型。Teevan 等[6]通过收集用户的查询和浏览历史对用户兴趣进行建模等。以上方法借助分析用户的行为历史记录、浏览记录或从网页文本角度出发挖掘用户兴趣,都取得了较好的效果。对于微博用户兴趣挖掘任务,Shu 等[7]提出了一种基于 Twitter-Rank 的微博用户兴趣模型构建方法。Liu 等[8]通过提取微博中的关键词挖掘用户兴趣。Chen[9]分别利用用户本身微博和用户的粉丝微博进行了用户兴趣发现。对于微博用户兴趣挖掘,采用微博用户标签进行用户兴趣提取是较为直观的方法,现有的方法主要是基于统计学习思想对标签信息进行分析从而获取用户兴趣。阎春霖等[10]综合考虑标签的使用频率和稀疏度,通过构造邻接矩阵挖掘用户兴趣。康海潇[11]利用微博标签表示用户兴趣,使用加权二分图算法提高用户兴趣发现效果。以上利用统计学习挖掘用户兴趣的方法没有考虑用户标签之间存在的语义相关性,学习过程中选取的表征用户兴趣的特征维由于维数的限制不可能覆盖所有的标签词,从而导致一些标签词不能有效表征到用户兴趣的特征向量上。基于以上分析,本文探讨通过结合词语相似度计算和特征映射的思想来实现用户兴趣聚类。

1　基本思想

利用统计学习挖掘用户兴趣的方法在学习过程中,选取的表征用户兴趣的特征维由于维数的限制不可能覆盖所有的标签词,从而丢失一些对用户兴趣聚类有指导意义的信息。比如:通过统计获取的特征空间中存在“旅行”特征维,而对于某用户,表征其兴趣的标签集合当中没有“旅行”标签,如果按照词频统计特征进行处理,该用户在“旅行”特征维上的特征值为 0。但在特征分析时发现,该用户的标签集合中可能存在与“旅行”语义很相近的标签,如有“旅游”标签,可以考虑计算“旅游”和“旅行”的词语相似度,通过映射的方式将“旅游”标签的信息映射到“旅行”特征维标签上,这样可以更加逼近用户的真实兴趣。

基于特征映射的微博用户标签兴趣聚类方法的主要思想是:某个待分析用户的用户兴趣可以通过用户本人的标签和其所关注用户的标签构成的标签集合进行表征,所有待分析用户的标签集合可以构成一个用户标签库,然后对该标签库中存在的大量用户标签做词频统计获取用户兴趣特征维,再结合词语相似度计算和特征映射思想确定每一特征维上的特征值,最后通过模糊聚类实现用户兴趣聚类。

2　基于特征映射的微博用户标签兴趣聚类过程

2.1　特征选取

通过新浪微博应用程序编程接口(Application programming interface,API)获取每个用户和其所关注用户的标签,由于用户标签的建立是半指导方式,用户自己填写的个性化标签存在标签随意性强的特点,对该类标签的分析处理有一定困难。因此,本文的处理方式是去除所有带有特殊符号和包含英文单词的个性化标签。通过以上处理,可以得到所有满足要求的用户标签构成的标签集合。利用实验室自主开发的新浪微博标签词频统计工具统计所有标签的出现次数,对标签出现次数从大到小排序,通过设定阈值选取排序靠前的标签作为用户向量的特征维。

2.2　基于特征映射的用户特征向量表征

为了更好地表征用户特征向量,考虑标签之间的语义相关性,引入特征映射的思想将用户标签根据

其与特征维标签之间的语义相似度映射到每个特征维上,从而计算每个特征维所对应的特征值。在特征映射过程中,针对某些长标签无法直接计算词语相似度的问题,首先利用中国科学院的 ICTCLAS 分词系统对长标签进行分词,将其表示成一个词的集合,再计算用户标签与特征维标签之间的平均语义相似度。具体做法如下:为了方便标签的统一处理,对所有标签使用统一的定义,不论长标签还是正常标签都可以统一表征为一个词集合 $l_u=\{wu_1,wu_2,\cdots,wu_m\}$,其中 m 表示该标签当中所包含的词数目,每个特征维标签也同样可以表征为一个词集合 $l_d=\{wd_1,wd_2,\cdots,wd_n\}$,其中 n 表示该特征维标签中所包含的词数目。设表征每个用户的所有标签数目为 X,其中每个标签出现的次数为 x,则对于每个标签来说,其初始特征值 f_{ul} 计算公式为

$$f_{ul}=\frac{x}{X} \tag{1}$$

用户标签与特征维标签之间的语义相似度用 $Sl(l_u,l_d)$ 表示,其中 l_u 表示用户标签,l_d 表示特征维标签,其计算公式为

$$Sl(l_u,l_d)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Sim}(wu_i,wd_j)}{m\times n} \tag{2}$$

式中:$\mathrm{Sim}(wu_i,wd_j)$ 代表包含 m 个词的用户标签 $wu_i(i=1,2,3,\cdots,m)$ 与包含 n 个词的特征维标签 $wd_j(j=1,2,3,\cdots,n)$ 之间的平均语义相似度,词语相似度计算方法参考文献[12]的基于知网的词语相似度计算方法。通过依次计算某个用户的所有标签与待确定特征值的特征维标签之间的语义相似度,选取出与该特征维标签相似度最大的用户标签,将该标签本身的特征值与该最大相似度相乘,计算结果作为该特征维的特征值,这样就完成了特征维中一维的确定。重复上述工作,即可确定出特征维中每一维的特征值,从而完成用户标签到特征维标签的特征映射,特征映射过程中每一特征维的特征值的公式为

$$T(l_d)=f_{ul}((l_u)_a)\cdot\max\{Sl((l_u)_a,l_d)\}\qquad a=1,2,3,\cdots,X \tag{3}$$

式中:$\max\{Sl((l_u)_a,l_d)\}$ 表示一个用户的 X 个用户标签分别与特征维 l_d 计算相似度之后得到的相似度最大值,$f_{ul}((l_u)_a)$ 表示当 $(l_u)_a$ 与 l_d 计算取得相似度最大值时该用户标签本身的特征值,$T(l_d)$ 表示特征维 l_d 的特征值。通过以上计算,可以为每个用户构建出表征用户兴趣的特征向量。为了验证该模型的构建效果,使用模糊聚类方法对模型效果进行验证。

2.3 基于模糊聚类的用户兴趣聚类

2.3.1 数据标准化

设论域 $U=\{x_1,x_2,\cdots,x_n\}$ 为待聚类的 n 个用户,每个用户又由一个 m 维的特征向量进行表征,即 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{im})$,$i=1,2,3,\cdots,n$,于是得到原始数据矩阵 $\boldsymbol{D}$

$$\boldsymbol{D}=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1m}\\x_{21}&x_{22}&\cdots&x_{2m}\\\vdots&\vdots&\ddots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{nm}\end{pmatrix}$$

由于不同的数据通常有不同的量纲,为了能够比较不同量纲的量,本文采用标准差规格化方法对数据进行适当变换,其计算公式为

$$x'_{ik}=\frac{x_{ik}-\bar{x}_k}{s_k}\qquad i=1,2,3,\cdots,n;k=1,2,3,\cdots,m \tag{4}$$

式中:$\bar{x}_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik}$ 为样本的均值,$s_k=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}$ 为样本的标准差。

2.3.2　模糊相似矩阵

设论域 $U=\{x_1,x_2,\cdots,x_n\}$,$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{im})$,$i=1,2,3,\cdots,n$,模糊相似矩阵中的每一个元素值反映用户 x_i 和 x_j 间的相似程度,用 $r_{ij}=R(x_i,x_j)$ 表征。本文采用指数相似系数法确定 r_{ij} 的值,其计算公式为

$$r_{ij}=\frac{1}{m}\sum_{k=1}^{m}\exp\left[-\frac{3}{4}\cdot\frac{(x_{ik}-x_{jk})^2}{s_k^2}\right] \tag{5}$$

2.3.3　最佳聚类阈值

模糊相似矩阵得到后,就可基于模糊相似矩阵进行模糊聚类。在模糊聚类分析中,对于各个不同的阈值 $\lambda\in[0,1]$,对应着不同的聚类结果。找出最佳聚类阈值 λ,此时 λ 对应的聚类结果就是最佳聚类结果,本文采用 F 统计量确定最佳聚类阈值 λ。

设 $U=\{x_1,x_2,\cdots,x_n\}$ 为待聚类的 n 个用户,$\boldsymbol{x}_j=(x_{j1},x_{j2},\cdots,x_{jm})$ 为第 j 个用户,其中 x_{jk} $(k=1,2,\cdots,m)$ 为描述用户 x_j 的第 k 个特征的数据。设 r 为对应于 λ 值的类数,n_i 为第 i 类用户的个数,记 $\overline{x}_{ik}=\frac{1}{n_i}\sum_{j=1}^{n_i}x_{jk}$ $(k=1,2,\cdots,m)$ 为第 i 类用户的第 k 个特征的平均值。记 $\overline{x}_k=\frac{1}{n}\sum_{j=1}^{n}x_{jk}$ $(k=1,2,\cdots,m)$ 为全体样本第 k 个特征的平均值。引入 F 统计量,计算公式为

$$F=\frac{\sum_{i=1}^{r}n_i\sum_{k=1}^{m}(\overline{x}_{ik}-\overline{x}_k)^2(r-1)^{-1}}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{m}(x_{ik}-\overline{x}_{jk})^2(n-r)^{-1}}\sim F(r-1,n-r) \tag{6}$$

它服从自由度为 $r-1,n-r$ 的 F 分布,其分子表征类与类间的距离,分母表征类内元素间的距离。因此 F 值越大,说明类与类之间的距离越大,聚类效果就越好。如果 $F>F_\alpha(r-1,n-r)$ $(\alpha=0.05)$,则根据数理统计方差分析理论可知类与类之间差异显著,说明聚类比较合理;再在满足 $F>F_\alpha(r-1,n-r)$ 的所有情形中,取差值 $F-F_\alpha$ 最大者的 F 所对应的 λ 作为最佳 λ 值,其所对应的聚类结果即为最佳聚类结果。

3　实验与结果分析

3.1　实验数据集

为了验证本文提出的基于特征映射的微博用户标签兴趣聚类方法的有效性,在实验数据集的准备方面,利用新浪微博 API 在旅游、环保、科技、自然语言处理等 15 个领域随机收集了 3 000 位用户,为每位用户获取了其所关注的 50 位用户。对于每个用户,将用户本人及其所关注用户的用户标签构成标签集合,去除掉其中带有特殊符号和包含英文单词的标签,以备为每个用户建立用户兴趣特征向量。将所有用户的标签集合组合成用户标签库,以备统计词频确定特征空间。

3.2　不同聚类算法下的用户兴趣聚类结果对比实验

为了验证本文选取的模糊聚类算法在用户兴趣聚类任务上会有更好的效果,本实验对模糊聚类和 K-means 算法的用户兴趣聚类结果进行了对比。首先,两种不同算法在用户特征向量的构建方面,都考虑标签之间的语义相关性,引入特征映射构建用户特征向量。其次,对于模糊聚类,将这些特征向量组成原始数据矩阵,利用模糊聚类算法得到用户兴趣聚类结果。对于 K-means 方法,依次选取 2,4,6,8 作为聚类数目,选取关注用户数较多的 K 位用户作为初始聚类中心,通过 K-means 算法得到聚类结果。选取平均聚类准确度 p 作为评价标准,其计算公式为

$$p=\frac{1}{n}\sum_{i=1}^{n}p_i \tag{7}$$

式中：n 为聚类数，p_i 为各类的准确度，即类中具有相同兴趣的最大用户数与类中用户总数之比，平均聚类准确度越高，代表聚类效果越好。模糊聚类算法和 K-means 算法的实验结果对比如表 1 所示。

表 1 不同聚类算法下用户兴趣聚类结果对比

Table 1 Comparison between different clustering method of user interest

聚类数	模糊聚类	K-means
2	0.35	0.25
4	0.58	0.49
6	0.77	0.67
8	0.65	0.54

根据模糊聚类的思想，不再将每个用户的兴趣以硬划分的形式划分到某个类别当中，使用模糊聚类实现用户兴趣聚类任务更能反映用户兴趣多类属的模糊特性。从表 1 的实验结果中不难看出，在同样引入特征映射构建用户特征向量的情况下，模糊聚类方法相比于 K-means 算法在用户兴趣聚类任务中的平均聚类准确度更优。

3.3 用户兴趣聚类在是否引入特征映射条件下的对比实验

为了验证本文提出的特征映射思想对用户兴趣聚类效果的提升，该实验分别采用两种方式构造用户的特征向量：(1)考虑用户标签之间的语义相关性，利用特征映射构建用户特征向量；(2)不考虑标签语义相关性而仅仅通过硬匹配构建用户特征向量。分别将两种方式下构造的用户特征向量组成原始数据矩阵，并利用模糊聚类算法对用户兴趣进行聚类，评价标准也选用平均聚类准确度，实验结果如图 1 所示。通过分析图 1 中实验结果可以发现，考虑用户标签之间的语义相关性并引入特征映射的思想构建用户特征向量，相比使用硬匹配构建用户特征向量在用户兴趣聚类的平均聚类准确度上有了较大改善。实验结果验证了本文提出的基于特征映射的微博用户标签兴趣聚类方法的有效性。

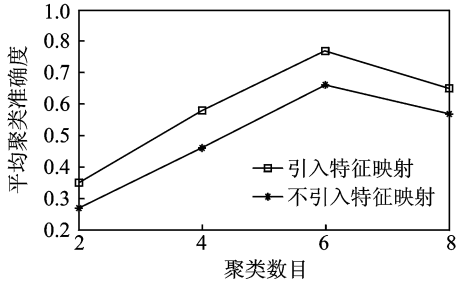

图 1 引入特征映射和不引入特征映射的平均聚类准确度对比实验

Fig. 1 Comparison of average clustering accuracy between considering feature mapping and without considering feature mapping

3.4 基于模糊聚类方法的用户兴趣聚类在不同阈值下的聚类结果分析

由于 3 000 位用户的聚类结果不便在本文中全部展示，因此本实验选取其中 15 位用户的聚类结果进行分析，这 15 位用户的用户兴趣如表 2 所示。使用本文提出的基于特征映射的用户标签兴趣聚类方法对以上 15 位用户进行兴趣聚类，在不同阈值下的聚类结果如图 2 所示。

表 2　用户及用户兴趣

Table 2　User and user interest

用户	用户兴趣	用户	用户兴趣	用户	用户兴趣
1	自然语言处理	6	护理	11	科技
2	旅游	7	保健	12	数据挖掘
3	摄影	8	环保	13	养生
4	机器学习	9	环境	14	电子商务
5	美食	10	数码	15	互联网

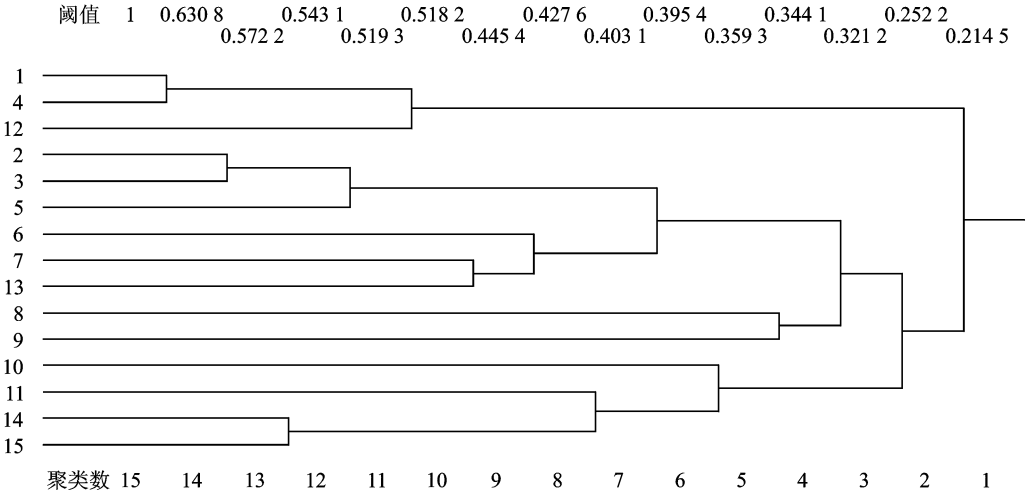

图 2　不同阈值下的用户兴趣聚类结果

Fig. 2　Clustering result of user interest with different threshold

通过观察图 2 所反映的不同阈值下的用户兴趣聚类结果,并结合表 2 中展示的 15 位用户各自的用户兴趣,可以比较直观地看到本文提出的用户兴趣聚类方法在用户兴趣聚类任务上取得了较好效果。比如当阈值取 0.630 8 时,兴趣为自然语言处理的用户 1 和兴趣为机器学习的用户 4 自动被聚为一类;再比如当阈值为 0.518 2 时,除了兴趣为自然语言处理的用户 1 和兴趣为机器学习的用户 4 聚为一类之外,兴趣为数据挖掘的用户 12 也被加入到该类当中,此外,兴趣为旅游的用户 2、兴趣为摄影的用户 3 和兴趣为美食的用户 5 被自动聚为一类,兴趣为电子商务的用户 14 和兴趣为互联网的用户 15 自动聚为一类。

4　结束语

本文针对微博用户标签之间存在一定的语义相关性问题提出了基于特征映射的微博用户标签兴趣聚类方法。考虑用户标签的语义相关性并引入特征映射的思想能够有效地提高用户兴趣聚类效果,同时验证了本文选取的模糊聚类方法相比于 K-means 算法在用户兴趣聚类任务中更具优越性。

参考文献:

[1]　Huang He, Huang Hai, Wang Rujing. FCA-Based web user profile ming for topics of interest[C]// Proceedings of the 2007 IEEE International Conference on Integration Technology. Shenzhen, China: IEEE, 2007: 20-24.

[2]　Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting web sites[J]. Machine Learning27, 1997, 27(3): 313-331.

[3] Tan Bin, Shen Xuehua, Zhai Chengxiang. Mining long-term search history to improve search accuracy[C]// Conference on Knowledge Discovery in Data. Philadelphia, PA, USA:[s. n.], 2006: 718-719.

[4] Shen X, Tan B, Zhai C. Context-sensitive information retrieval using implicit feedback[C]//Proceedings of the 28th Annual International ACM SIGIR Conference. Salvador, Brazil: ACM, 2005: 41-45.

[5] Zhou Xiaoming, Conati C. Inferring user goals from personality and behavior in a causal model of user affect[C]// Proceedings of the 8th International Conference on Intelligent User Interfaces. Miami, Florida, USA:[s. n.], 2003: 211-214.

[6] Teevan J, Dumais S T, Horvitz E. Personalizing search via automated analysis of interests and activities[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil: ACM, 2005: 449-451.

[7] Shu Wengjian, Lim E P, Jiang Jing, et al. Twitterrank: Finding topic-sensitive influential twitterers[C]// Proceedings of the 3th ACM International Conference on Web Search and Data Mining. New York City, NY, USA: ACM, 2010: 261-270.

[8] Liu Z, Chen X, Sun M. Mining the interests of Chinese microbloggers via keyword extraction[J]. Frontiers of Computer Science in China, 2012, 1(6): 76-87.

[9] Chen J, Nairn R, Nelson L, et al. Short and tweet: Experiments on recommending content from information streams[C]// Proceedings of the 28th International Conference on Human Factors in Computing Systems. New York, USA: ACM, 2010: 1185-1194.

[10] 阎春霖,张延园. 基于用户标签的社区发现方法研究[J]. 科学技术与工程, 2011, 11(6): 1237-1240.
Yan Chunlin, Zhang Yanyuan. Research of community dicovery algorithm based on user tags[J]. Science Technology and Engineering, 2011, 11(6): 1237-1240.

[11] 康海潇. 基于标签的微博用户兴趣发现算法研究及应用[D]. 杭州:浙江大学, 2013.
Kang Haixiao. Algorithm research of tag-based user interest discovery in Weibo and application[D]. Hangzhou: Zhejiang University, 2013.

[12] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学, 2002, 7(2): 59-76.
Liu Qun, Li Sujian. Word similarity computing based on How-net[J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76.

作者简介:

秦雨(1989-),男,硕士研究生,研究方向:信息检索、数据挖掘, E-mail: iamno1_2009@163.com。

余正涛(1970-),男,博士,教授,研究方向:自然语言处理、信息检索和信息抽取。

王炎冰(1988-),男,硕士研究生,研究方向:社交网络分析、信息检索。

石林宾(1989-),男,硕士研究生,研究方向:信息检索和数据挖掘。

潘华山(1989-),男,硕士研究生,研究方向:自然语言处理和数据挖掘。