标签共现的标签聚类算法研究

王娅丹,李 鹏,金 瑜,刘 宇
WANG Yadan, LI Peng, JIN Yu, LIU Yu

1.武汉科技大学 计算机科学与技术学院,武汉 430065
2.智能信息处理与实时工业系统湖北省重点实验室,武汉 430065
1.College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
2.Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430065, China

WANG Yadan, LI Peng, JIN Yu, et al. Research on tags co-occurrence for tags clustering algorithm. Computer Engineering and Applications, 2015, 51(2):146-150.

Abstract: In the social network, tag clustering analysis can deal with problems such as tag redundancy and semantic fuzziness and so on. In order to improve the effectiveness of clustering, it proposes to integrate label co-occurrence information and derive the feature vector of label, extracts the feature vector to calculate the similarity. The traditional clustering algorithm uses the geometric distance to calculate the distance to the object and the center of the object, now uses the Pearson correlation coefficient to calculate. The tag clustering algorithm that combines with *K*-means clustering algorithm to cluster label is proposed, and then analyzes the complexity of the algorithm. Finally, doing relevant comparative experiments for different clustering algorithms, the experimental results show that the proposed clustering algorithm enhances the clustering performance than other clustering algorithms, and verify the availability and effectiveness of the proposed clustering algorithm.

Key words: tag clustering; tag co-occurrence; *K*-means; Pearson correlation coefficient; feature vector

摘 要: 在社会网络中,标签聚类研究可以解决标签冗余和语义模糊等问题。为了提高聚类有效性,提出综合标签共现信息确定标签特征向量,通过特征向量的提取计算相似度,将传统聚类算法中用几何距离计算对象与中心对象的距离改为用皮尔森相关系数计算,提出结合*K*-means聚类算法对标签进行聚类的标签共现聚类算法,并分析了算法的复杂度。最后对不同聚类算法进行了相关对比实验,实验结果表明该聚类算法效果要好于其他的聚类算法,从而验证了该聚类算法的有效性和可行性。

关键词: 标签聚类; 标签共现; *K*-means; 皮尔森系数; 特征向量
文献标志码: A **中图分类号:** TP301.6 doi:10.3778/j.issn.1002-8331.1404-0359

1 引言

标签是用户对信息的主观理解,是联系客观信息和主观认识的中介。在社会网络中信息通过相同的标签联系在一起,用户也通过使用标签与其他资源及用户联系在一起,这样人与人之间就可以通过标签进行交友。标签作为在线社会化网络的一部分,已得到了广泛的研究,Flickr、del.icio.us、豆瓣网和Youtobe等网站都采用了标签的协同标注及聚类研究,但目前针对标签之间相关联系的研究比较少。现阶段对标签系统进行优化的研究主要集中于标签云[1-2],标签的有序化组织[2-3]。标签之间关联度的研究有助于对信息进行分类检索与浏览,同时也可以挖掘出用户之间的相似性,从而可以对用户进

基金项目: 国家自然科学基金(No.61303117);湖北省重点实验室开放基金资助项目(No.znss2013B012);湖北省教育厅科研基金(No.B2014085,No.B20101104);武汉科技大学大学生科技创新基金研究项目(No.12ZRC061)。

作者简介: 王娅丹(1992—),女,本科在读;李鹏(1981—),男,通讯作者,讲师,主要研究方向为分布式网络、移动计算;金瑜(1973—),女,副教授,主要研究方向为网络计算、信任模型;刘宇(1980—),男,讲师,主要研究方向为分布式计算。E-mail:lipeng@wust.edu.cn

收稿日期: 2014-04-23 **修回日期:** 2014-05-27 **文章编号:** 1002-8331(2015)02-0146-05
CNKI**网络优先出版:** 2014-07-11, http://www.cnki.net/kcms/doi/10.3778/j.issn.1002-8331.1404-0359.html

行个性化推荐。标签的聚类就可以形成一个个的社区网络[4]，随着标签的不断增多网络也会随之扩大。本文对标签之间的共现信息进行提取，然后用聚类算法[5]对标签进行聚类。文章从聚类有效性进行比较可以发现不同的分类最后聚类效果有很大的差别，由此可见选择有效的聚类方法对于标签的聚类是很必要的。

2　相关工作

在标签分类方法中，大众分类法（Folksonomy）[6]是由美国信息架构专家 Thomas Vander Wal 和 Gene Smith 于 2004 年首先提出，由“Folks”和“Taxonomy”组合而来，含义是“由大众的一致意见而产生的基于用户的分析体系”，中文翻译为“通俗分类”，“大众分类”，“自由分类”等。

Folksonomy 使得传统的分类法摆脱了固化的现象，并且跟大众的认知程度密切地结合起来，同时这种分类方法也为群体用户和信息之间建立了一个联系的桥梁。然而正是因为用户参与的广泛性，标注的随意性，使得大众标注过于自由，个性化。因此会导致一系列的问题，标签的意义可能混淆，系统的推荐很不合理，用户标签时存在错误等。针对这些问题，研究者开始对标签集之间的关联进行分析。Begelman 等人[7]提出采用聚类技术对大量标签进行自动聚类的方法来改善自由分类法的检索和浏览。还有 Heymann 等人尝试构建标签的层次分类，根据标签的相似度确定相关标签，然后对标签进行聚类[8]。

现有的研究成果表明，对标签进行合理的聚类有助于实现标签的有序化组织。本文在以上研究的基础上，基于对以往的标签聚类算法进行改进，综合标签共现的信息计算出每个标签的特征向量，然后利用 *K*-means 算法[9]对标签进行聚类。传统 *K*-means 中用几何距离计算对象与中心对象之间的相似度，这里改为利用皮尔森相关系数[10]去计算。通过该方法，解决了标签描述资源准确度低，组织混乱，存在语义模糊等问题。

3　标签共现分析

共现分析[11]的理论基础之一是心理学的邻近联系法则：曾经在一起感受过的对象往往在想象中也联系在一起，以至于想起他们中的某一个的时候，其他的对象也会以曾经同时出现的顺序被想起，即只有存在语义关联的词汇才能被作者在相邻的位置记录下来。

由于标记过程范围广泛，经常出现用不同的标签标记相同资源的现象。两个标签之间的相似性可通过多种形式进行评价，最简单的方式就是计算共现的次数。若标签 t_1 和 t_2 共同标注了某一资源，则 t_1 和 t_2 共现。其基本思想就是：如果两个标签在资源中同时出现的次数越多，则它们的相关度就越高。可以用如下的数学公式来定义两个标签之间的相似度[12]：

$$\begin{cases} Sim(t_i, t_j)=1, i=j \\ Sim(t_i, t_j)=\dfrac{\left|A_i \cap B_j\right|}{\left|A_i \cup B_j\right|}, i\neq j \end{cases} \tag{1}$$

其中 A_i 和 B_j 是两个标签描述的资源集。得到任意两个标签之间的相似度之后就可以构建相似度矩阵。假设有 n 个标签则可以构建一个 $n\times n$ 型的矩阵，其中的每个元素 t_{ij} 就是利用公式(1)计算出的第 i 个标签和第 j 个标签的相似度。然后根据这些相似度进行聚类。

以往的聚类大多都是根据标签与资源之间的关联来计算相关矩阵与参数然后进行聚类，本文是根据标签与标签之间的共现关联来进行计算聚类。相关定义见本文第 4 章。

4　算法描述

本文在对标签共现分析的基础上，综合提取标签共现信息来为每个标签构造一个特征向量，利用 *K*-means 对其进行聚类，由于传统的 *K*-means 大多都是用几何距离去计算对象与中心对象相似度，本文用皮尔森相关系数计算。

4.1　特征向量提取

定义 1（标注矩阵）该矩阵 $\boldsymbol{U}_{m\times n}$ 是 $m\times n$ 型矩阵，m 为资源个数，n 为标签个数，矩阵中的元素 u_{ij} 表示标签 t_i 标注资源 j 的频度。

在标注矩阵中，频度代表标签 t_i 标注资源 j 的次数。根据定义 1，得到标注矩阵的物理含义表示为 m 个标签和 n 个资源之间的关联。

定义 2（共同标注矩阵）该矩阵 $\boldsymbol{C}_{n\times n}$ 是 $n\times n$ 型矩阵，n 为标签个数，矩阵中的元素 c_{ij} 表示标签 t_i 和标签 t_j 共现频度，即

$$c_{ij}=\frac{W(t_i, t_j)}{\sum_{k=1}^{n} W(t_i, t_j)} \tag{2}$$

其中 $W(t_i, t_j)$ 表示标签 t_i 和标签 t_j 共同出现的次数，当 $i=j$ 时，$W(t_i, t_j)$ 为标签 t_i 标注过的资源数。在一定程度上，这个度量越大说明标签 t_i 和标签 t_j 共同出现的几率越高，即标签 t_i 与标签 t_j 之间的关系就越密切。

定义 3（标签重要度矩阵）该矩阵 $\boldsymbol{A}_{n\times n}$ 是 $n\times n$ 型矩阵，n 为标签个数，矩阵中的元素 a_{ij} 表示标签 t_i 在整个 m 个资源内的重要度，即

$$a_{ij}=c_{ij}\times \lg\left(\frac{n}{1+\Gamma(t_i)}\right) \tag{3}$$

其中，$\Gamma(t_i)$ 表示在 m 个资源中，与标签 t_i 共同出现过的

标签的个数，c_{ij} 表示标签 t_i 和标签 t_j 同出现的频度，可由公式(2)得到。在公式中，分母加1防止分母为0的情况。这个度量的物理含义代表在 m 个资源内标签出现的高频率，以及该标签在整个资源集合中的低共现频率，可以产生出高权重的 a_{ij}，该值越大说明标签 t_i 在整个资源集合中越重要。

通过公式(3)的计算得到的标签重要度矩阵中，每个行向量即代表该标签的特征向量。

定义4(相似度矩阵) 该矩阵 $\boldsymbol{S}_{n\times n}$ 是 $n\times n$ 型矩阵，n 为标签个数，矩阵中的元素 s_{ij} 表示，即

$$s_{ij}=\frac{n\cdot\sum \boldsymbol{A}_i\cdot\boldsymbol{A}_j-\sum \boldsymbol{A}_i\cdot\sum \boldsymbol{A}_j}{\sqrt{n\cdot\sum \boldsymbol{A}_i^2-\left(\sum \boldsymbol{A}_i\right)^2}\cdot\sqrt{n\cdot\sum \boldsymbol{A}_j^2-\left(\sum \boldsymbol{A}_j\right)^2}} \quad (4)$$

其中，$\boldsymbol{A}_i$ 表示标签重要度矩阵中的每个行向量，即对应标签的特征向量 $\boldsymbol{A}_i(a_{i1},a_{i2},\cdots,a_{in})$，通过计算式(4)后得到两个向量之间的相似度。该公式反映了两个变量线性相关程度的统计量。

4.2 标签聚类算法

K-means聚类算法[13]用欧氏距离作为相似性度量和距离计算，计算各数据点到其类别中心的距离平方和。本文提出的标签共现的标签聚类算法(简称Tag co-occurrence算法)，首先根据公式(2)，公式(3)计算出标签特征向量，然后对*K*-means的相似性和距离度量进行了改进，用公式(4)来进行两个向量相似度计算，就可以对标签集合进行聚类，得到最终聚类结果，具体见算法1。

算法1 Tag co-occurrence算法

输入：聚类的类别数目 k；标签个数 m；标签集合 $T=\{t_1,t_2,\cdots,t_m\}$；资源集合 R；标签标注资源的关系集合 A

输出：m 个标签的聚类结果(每个类别有哪些标签)

1：FOR 所有标签 t_i，资源 j DO

2：　　计算标注矩阵的频度 u_{ij}

3：　　根据公式(2)，计算共现频度 c_{ij}；

4：　　根据公式(3)，计算重要度 a_{ij}；

5：　　得到标签 t_i 的特征向量 $\boldsymbol{A}_i(a_{i1},a_{i2},\cdots,a_{in})$；

6：END FOR

7：选择 k 个对象作为初始聚类中心；

8：$J=\sum_{i=1}^{k}\sum_{j=1}^{n_j}d(\boldsymbol{A}_j,Z_i)$；(准则函数，$\boldsymbol{A}_j$ 为相应聚类中的标签特征向量，Z_i 为相应聚类的聚类中心)

9：While(准则函数 J 取值不再发生变化)

10：　根据公式(4)，计算每个标签与这 k 个聚类中心的相似度 s_{ij}

11：　选择相似度最大的 s 将其归为相应的聚类中心

12：　重新计算每个聚类的均值(新的聚类中心)；

13：END WHILE

14：RETURN

该算法是对*K*-means进行了一些改进，传统的*K*-means是用欧氏距离作为对象的相似度度量，聚类度量衡量的是空间各点间的绝对距离，跟各个点所在的位置坐标(即个体特征维度的数值)直接相关，距离值越小个体间相似度越大。该算法用的是皮尔森相关系数公式作为对象相似度的度量，分别对两个对象基于自身总体标准化后计算空间向量的余弦夹角。度量值越小，说明差异越大。这两种相似度的计算方法，分别适用于不同的数据分析模型：欧氏距离能够体现个体数值特征的绝对差异，所以更多地用于需要从维度的数值大小中体现差异的分析；而皮尔森相关系数更多地是从方向上区分差异，而对绝对的数值不敏感，更多地用于使用用户对内容评分来区分用户兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题。

接着分析算法的时间复杂度和空间复杂度。首先，在算法的1~5行，计算标签的频度，共现频度以及重要度，时间复杂度为 $O(m\times n)$，m 为资源个数，n 为标签个数。其次，在算法7~11行中，类似*K*-means算法，时间复杂度为 $O(n\times t\times k)$，t 为迭代次数，n 为维数，k 为聚类的数目。由于 t 和 k 均小于 n，因此，整个算法的时间复杂度为 $O(n^3)$。

对于空间复杂度，算法在运行的过程中维护4个集合，一个是频度集合，其大小为 $O(m\times n)$，一个是共线频度集合，其大小为 $O(n^2)$，一个是重要度集合，其大小为 $O(n^2)$，最后一个是相似度集合，大小是 $O(n^2)$，因此，算法的空间复杂度为 $O(n^2)$。

5 实验结果及分析

5.1 实验相关背景

实验所用的数据是从社会化书签网站del.icio.us上获取的。从del.icio.us上收集100个资源集，从中提取20个热门标签作为待聚类的标签。为了能够对实验结果有一个更加客观的评价，对三种聚类算法进行对比：仅仅考虑标签共现次数的聚类(Frequency of co-occurrence)，传统的*K*-means聚类算法，以及本文提出的标签共现的标签聚类算法(Tag co-occurrence)。

(1)Frequency of co-occurrence算法是仅仅考虑了共现次数，即随机选取 K 个标签为聚类中心，标签与哪个中心标签共现次数多就把它归为其中。

(2)传统的*K*-means算法在仅仅考虑共现次数的基础上先确定标签特征向量即矩阵 $\boldsymbol{X}_{n\times n}$，元素 x_{ij} 表示标签 t_i 和标签 t_j 共同出现的次数，$\boldsymbol{X}_{n\times n}$ 矩阵中每个行向量表示对应标签的特征向量，然后用欧式距离确定相似度*K*-means进行聚类。

(3)本文提出Tag co-occurrence算法综合了共现信

息,根据公式(2),公式(3)确定特征向量,然后利用公式(4)计算向量相似度再进行聚类。

实验中一方面从有效性指标(精确度和召回率)比较这三种聚类算法,另一方面用相似性分析(标签资源的关联和标签共线信息的关联)比较标签与其标注主题相似性的稳定度。

5.2　评价指标

5.2.1　聚类有效性指标

在评价一个聚类算法的聚类效果时,Purity来评价聚类有效性,只需计算正确聚类的标签数占总标签数的比例。这是一种极为简单的评价方法[14]:

$$Purity(W,T)=\frac{1}{N}\sum_{k}\max_{j}\left|w_k\cap t_j\right| \tag{5}$$

其中 $W=\{w_1,w_2,\cdots,w_k\}$ 是聚类的集合,w_k 表示第 k 个聚类的集合。$T=\{t_1,t_2,\cdots,t_j\}$ 是标签集合,t_j 表示第 j 类标签。N 表示标签总数。

第二种方法是计算精确度与召回率,精确度与召回率是常用的指标[15]。下面介绍一下要求出精确度与召回率需要确定的几个参数:TP(True Positives),FP(False Positives),TN(True Negatives),FN(False Negatives)。

(1)TP:聚类算法将一对标签分在了同一类别中,并且在先验类别中它们也在相同的类别中。

(2)FP:聚类算法将一对标签分在了同一类别中,但在先验类别中它们属于不同的类别。

(3)TN:聚类算法将一对标签分在了不同类别中,并且在先验类别中它们也属于不同类别。

(4)FN:聚类算法将一对标签分在了不同类别中,但在先验类别中它们属于相同的类别。

精确率为:

$$Precision=\frac{TP}{TP+FP} \tag{6}$$

召回率为:

$$Recall=\frac{TP}{TP+FN} \tag{7}$$

5.2.2　相似性计算指标

首先引入两个概念用来计算标签与其标注主题的相似性。

质心 O_t 用来表示标签 t 标注主题的质心。

$$o_t=\frac{1}{u(t)}\sum_{u_i\subset u(t)}^{n}u_i \tag{8}$$

在用标签资源的关联去进行计算时,$\boldsymbol{u}_t$ 表示定义1中标注矩阵的列向量,$U(t)$ 表示标签 t 标注的资源个数。在用标签与标签共现信息关联去计算时,$\boldsymbol{u}_t$ 表示定义3中标签重要度矩阵的列向量,$U(t)$ 表示与标签 t 有关联的标签个数。

$$Tcs=\frac{1}{u(t)}\sum_{\boldsymbol{u}_t\subset u(t)}\cos(\boldsymbol{u}_t,o_t) \tag{9}$$

Tcs 表示标签与主题的相似性,同样在用标签资源的关联去进行计算时,$\boldsymbol{u}_t$ 表示定义1中标注矩阵的列向量,$U(t)$ 表示标签 t 标注的资源个数。在用标签与标签共现信息关联去计算时,$\boldsymbol{u}_t$ 表示定义3中标签重要度矩阵的列向量,$U(t)$ 表示与标签 t 有关联的标签个数。

5.3　结果分析

首先是从聚类结果的有效性,分别计算三种聚类的Purity,精确度,召回率,将其在不同标签个数的情况下进行对比。接着是将从两个方面计算出的标签与标注主题相似度进行分析,一方面是从标签资源的关联中去计算相似度,另一方面是从标签共现信息的关联中去计算相似度。将这两种方式计算出来的相似度在不同标签个数的情况下进行对比。

5.3.1　聚类有效性分析

为了能够更加直观,精确地描述实验结果,下面对实验结果进行定量分析。

采取对这三种聚类算法分别在5,10,15,20个标签的情况下聚类结果的Purity及精确度和召回率的比较来进行分析。

先对Purity进行比较三种聚类的Purity随标签个数增加的变化趋势,具体如图1所示。

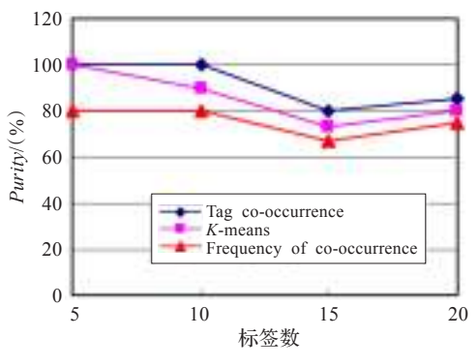

图1　聚类的Purity

接着对精确度进行比较,3种聚类的精确度随标签个数增加的变化趋势如图2所示。

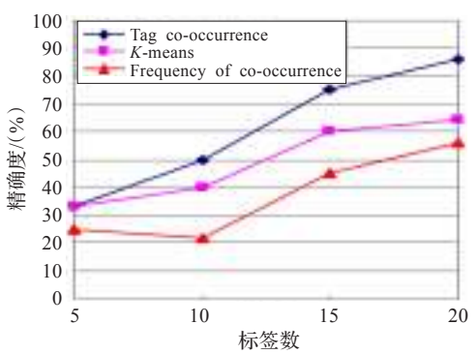

图2　聚类的精确度

由图2可以直观地看出随着标签个数的增多,本文的标签共现聚类算法的精确度高于其他的聚类算法精

确度,仅仅考虑标签共现次数的聚类算法精确度最低。

最后对召回率进行比较,3种聚类的召回率随标签个数增加的变化趋势如图3所示。

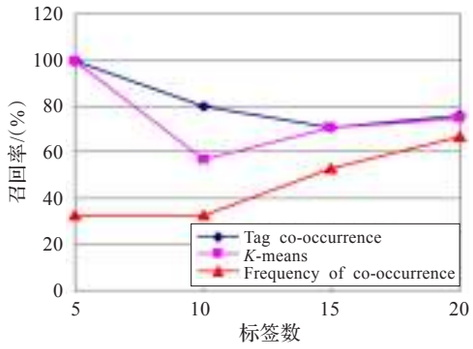

图3　聚类的召回率

由图3可以直观地看出标签共现聚类算法的召回率最高。

从Purity,精确度和召回率的比较来看,本文提出的标签共现的聚类算法聚类效果更好。这是因为本文综合了标签共现的信息,提取出了每个标签的特征向量,利用*K*-means聚类,这种方法要明显优于仅仅考虑标签共现次数的标签聚类算法。

5.3.2　相似性计算分析

根据公式(8),公式(9)计算两种方式下,标签与标注主题的相似性。图4是5个标签时分别用两种方法计算的相似度,由图可以直观地看出根据综合标签共现信息计算出的标签与它标注主题的相似度相对比较平稳。

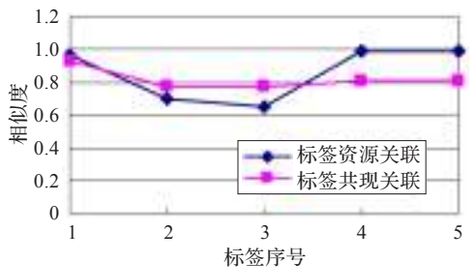

图4　5个标签相似度

图5是10个标签时分别用两种方法计算的相似度。

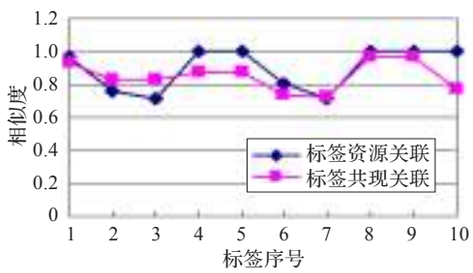

图5　10个标签相似度

图6是15个标签时分别用两种方法计算的相似度。图7是20个标签时分别用两种方法计算的相似度。由上面可以看出无论是5个,10个,15个还是20个

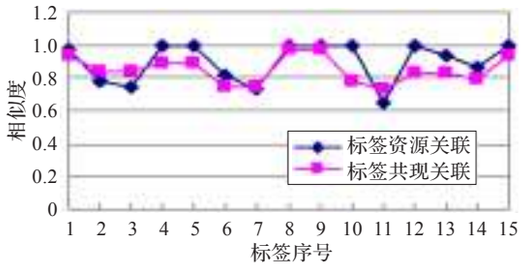

图6　15个标签相似度

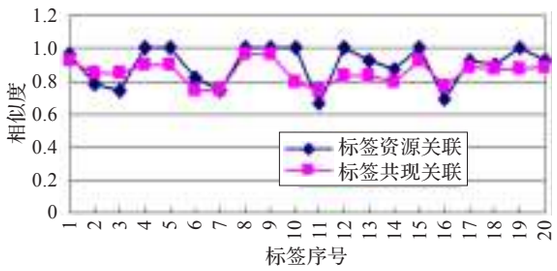

图7　20个标签相似度

标签,本文提出的标签共现聚类算法计算出的标签与标注主题的相似性更加稳定,更能很好表达它所标注的一类主题。

从以上两个方面的分析都可以看出本文用标签共现信息提取标签的特征向量对于标签聚类,标签表达主题都有更好的优势。

6　结束语

本文首先分析了在仅仅考虑标签共现次数的情况下,得到的分类结果不准确的问题。为了解决该问题,本文综合标签共现信息,将其进行利用提取出标签的特征向量,利用*K*-means算法进行聚类,并将*K*-means中计算相似度的公式由几何距离改为皮尔森相关系数。通过实验,验证了本文提出的聚类方法更加有效。

参考文献:

[1] Golder S A,Huberman B A.Usage patterns of collaborative tagging systems[J].Journal of Information Science,2006,32(2):198-208.

[2] Owen K,Daniel L.TagCloud drawing:algorithms for cloud visualization[C]//Proceedings of Tagging and Metadata for Social Information Organization(WWW2007),2007.

[3] Golder S A,Huberman B A.Usage patterns of collaborative tagging systems[J].Journal of Information Science,2006,32(2):198-208.

[4] Lin Y R,Chi Y,Zhu S,et al.Analyzing communities and their evolutions in dynamic social network[J].ACM Transactions on Knowledge Discovery from Data(TKDD),2009,3(2):1-31.

[5] 孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,2008,19(1):48-61.

(下转208页)