文章编号:1671-9352(2015)09-0036-06 DOI:10.6040/j.issn.1671-9352.3.2014.287

基于有指导LDA用户兴趣模型的微博主题挖掘

王立人[1,2],余正涛[1,2],王炎冰[1,2],高盛祥[1,2],李贤慧[1,2]
(1. 昆明理工大学信息工程与自动化学院,云南 昆明 650500;
2. 昆明理工大学智能信息处理重点实验室,云南 昆明 650500)

摘要:用户发布的微博内容能够体现用户兴趣,微博中用户的转发、评论、回复、他人评论等微博行为对用户兴趣具有很强的指导作用。为了有效利用用户微博行为,提出了一种基于有指导LDA(latent dirichlet allocation)的微博内容用户兴趣建模方法。首先通过分析对微博的转发、评论、回复、他人评论这4个因素对用户微博兴趣主题的影响,定义了4种约束关系;然后基于用户微博内容,将4种约束关系融合到LDA模型中构建有指导的LDA微博主题生成模型,最后得到用户的微博主题分布,从而获得用户兴趣模型。实验结果表明,相比LDA模型,该方法的准确率有很大提高,引入4种信息对微博用户兴趣发现有非常重要的指导作用。

关键词:微博内容;兴趣挖掘;微博行为;有指导LDA
中图分类号:TP393 **文献标志码**:A

Micro-blogging topic mining based on supervised LDA user interest model

WANG Li-ren[1,2], YU Zheng-tao[1,2], WANG Yan-bing[1,2], GAO Sheng-xiang[1,2], LI Xian-hui[1,2]
(1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, China; 2. Intelligent Information Processing Key Laboratory, Kunming University of Science and Technology, Kunming 650500, Yunnan, China)

Abstract: The content of users Micro-blogging can reflect users' interests. Forwarding, commenting, replying and other behavior about Micro-blogging have a strong guiding role to discovering users' interests. In order to using Micro-blogging behavior effectively, we proposed users' interest modeling method based on supervised-LDA Micro-blogging contents. First of all, through analyzing the impact elements, including forwarding, commenting, replying, and other behavior, four constraint relations were defined. Second, based on the contents of Micro-blogging, the four constraint relations were put into the LDA model and the supervised-LDA Micro-blogging theme generation model were constructed. And then the distribution of the users' theme and the users' interests' model were obtained. The experimental results show that compared with the LDA method, this model has high accuracy, and the four introduced guiding information have a significant role in discovering Micro-blogging users' interests.

Key words: Micro-blogging content; interest in mining; Micro-blogging behavior; supervised LDA

0 引言

随着社交网络的不断发展,微博的社交地位也越来越突出。目前主流的微博平台有Twitter、新浪等。基于微博内容,挖掘用户兴趣,进而发现具有相同兴趣爱好的用户社区,对于用户聚合和资源整合具有非常

收稿日期:2015-03-03; **网络出版时间**:2015-07-22 15:38
网络出版地址:http://www.cnki.net/kcms/detail/37.1389.N.20150722.1538.010.html
基金项目:国家自然科学基金资助项目(61175068)
作者简介:王立人(1990-),男,硕士研究生,研究方向为自然语言处理. E-mail: wlr901112@163.com

大的帮助。

当前在用户兴趣挖掘研究方面的方法主要有以下几类:一是基于用户行为和文档内容的用户兴趣挖掘。如 Curious Browser 系统将用户的主动评价信息与用户通过网页浏览的时间长短等所反映的兴趣相结合来预测用户所需要的信息[1];UCAIR 系统为用户搜索建立个性化模型,通过记录用户行为找出用户兴趣[2];文献[3]提出利用给出的示例文档,通过分析文档特征、文档类别以及段落关系来完成对用户的个性化推荐。二是基于微博内容的用户兴趣挖掘。如文献[4]提出了一种基于 Twitter-Rank 的微博用户兴趣模型构建方法,该方法将每个用户的所有微博整合成一个文档进行主题分析,提取用户兴趣;文献[5]利用 LDA(latent dirichlet allocation)模型挖掘隐藏在文本内的不同主题与词之间的关系得到文本主题分布,并对文本进行聚类;文献[6]提出的基于 LDA 的文本分类方法避免了文本表示方法的高维稀疏特征空间的问题。三是融合内容及微博交互关系的用户兴趣挖掘。如文献[7]提出了一种 MB-LDA 的微博主题挖掘方法,该方法结合微博内容进行主题挖掘,并发现用户兴趣。目前,随着微博的普及,各种微博行为成为用户兴趣发现的主要分析来源,如转发、回复这些微博行为越来越频繁,从而对用户本身的兴趣更具有指导意义,但上述模型还未出现基于微博行为的兴趣发现方法。

本文在传统 LDA 的基础上,提出了一种基于有指导 LDA 的微博生成模型,综合考虑了转发、评论、回复和他人评论 4 个因素对微博主题分布的影响,并将这 4 个影响因素引入微博生成模型,对每一条用户微博进行分析,得出用户的微博主题分布,进而对用户所有的微博进行主题分布统计,得到用户的兴趣主题分布。

1　模型建立

1.1　基于有指导 LDA 的微博主题发现

LDA 模型[8]包含词、主题、文档三个层次,一篇文档当中包含若干个主题,每一个主题通过若干个词表征。在 LDA 模型中,只对文档 - 主题的概率分布 θ 加入了狄利克雷先验,而对主题 - 词概率分布没有进行任何先验假设,整个模型的求解方法利用了 PLSA 的 EM 推导方法。文献[9]对 LDA 模型进行了改进,对主题 - 词概率分布 φ 也加上了狄利克雷先验,求解方法则是利用了多项式分布与狄利克雷分布的共轭特性。一般来说学习估计主要采用近似推理算法,常用的方法主要有变分最大期望算法[10]、吉布斯抽样算法[7]和期望传播算法[11]。其中,吉布斯采样算法由于简单易懂和运行速度快等优点,常被用来求解 LDA 模型,而吉布斯采样算法的关键就是构造目标概率分布函数[12]。

本文通过对大量微博数据进行统计分析发现微博之间存在转发、评论、回复和他人评论 4 个关系,这 4 种关系可以被视为指导信息,辅助挖掘用户微博当中的主题信息。因此,我们在传统 LDA 的基础之上,将这 4 种因素作为指导信息,进行统一建模,构成融合上述 4 种微博指导信息的微博 LDA 主题模型。图 1 为该模型的贝叶斯网络图,其中,θ_d、θ_c、θ_{rt}和 θ_{re}分别代表原微博的主题分布、被评论微博的主题分布、被转发微博的主题分布和被回复评论微博的主题分布,α、α_c、α_{rt}、α_{re}分别代表 θ_d、θ_c、θ_{rt}和 θ_{re}的超参数,χ 代表被评论微博的所有相关微博以参数为 γ 的狄利克雷分布所抽样出的相关微博的影响分布,r 代表在 χ 分布上所抽取出的影响微博,z 代表词的主题,w 代表微博当中的词,N 表示词的数量,M 表示微博的篇数,φ 表示主题下的词分布,β 是 φ 的超参数,T 是主题总数,π 代表微博类型。

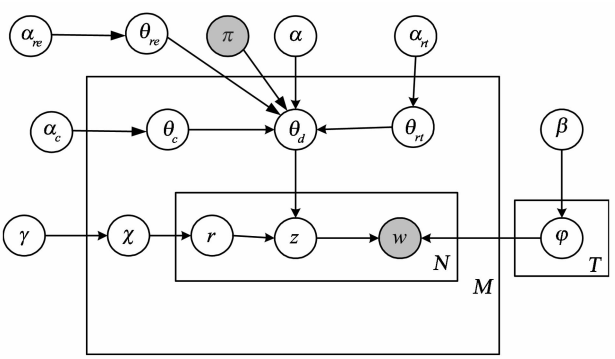

图 1　有指导 LDA 贝叶斯网络图

Fig. 1　Guide-LDA Bayesian network diagram

首先,本文模型从参数为 β 的狄利克雷分布中抽取主题与单词之间的关系 φ,在生成一条微博时,首先根据数据来源和数据形式判断微博类型,如果该微博为一条原创微博,且没有任何评论,则 π 取 0,表示该微博为一条独立原创微博,此时需要从参数为 α 的狄利克雷分布中抽样出该微博与各个主题之间的关系 θ_d;如果该微博来自新浪 API 所获取的用户评论微博,则 π 取 1,表示该微博为一条评论微博,此时需要从参数

为 α_c 的狄利克雷分布中抽样出该微博所评论的微博与各个主题之间的关系 θ_c,并把它赋值给微博 d 与各个主题之间的关系 θ_d;如果该微博为一条转发微博,则 π 取 2,表示该微博为一条转发微博,此时需要从参数为 α_{rt} 的狄利克雷分布中抽样出该微博所评论的微博与各个主题之间的关系 θ_{rt},并把它赋值给微博 d 与各个主题之间的关系 θ_d;如果该微博为一条回复他人评论的微博,则 π 取 3,此时该微博的主题分布受到被回复的评论微博以及该评论微博所评论的原创微博两方面的影响,但影响权重不同,引入影响参数 μ,代表被回复的评论微博的影响权重,则该评论微博所评论的原创微博的影响权重为 $1-\mu$,此时计算综合被回复的评论微博以及该评论微博所评论的原创微博两种主题分布,得到混合主题分布,并把计算得到的值赋值给微博 d 与各个主题之间的关系 θ_d。综上,根据微博类型,其主题分布的计算公式如下:

$$P(\theta|\alpha,\mu)=\begin{cases}\theta_d,\ \pi=0,\ \alpha=\alpha;\\ \theta_c,\ \pi=1,\ \alpha=\alpha_c;\\ \theta_{rt},\ \pi=2,\ \alpha=\alpha_{rt};\\ \mu\theta_{re}+(1-\mu)\theta_c,\ \pi=3,\alpha=\{\alpha_{re},\ \alpha_c\}。\end{cases}\tag{1}$$

对于有他人参与评论的原创微博,考虑到他人的评论对于该微博本身的主题分布有一定的影响,我们把该原创微博及其评论微博放在一起作为一个数据集 S_d,在 S_d 上增加一个 χ_d 的分布,该分布由 γ 作为参数的狄利克雷分布抽样得来。对于微博文本中的每一个单词,首先从 χ_d 分布中抽取出一个影响微博 r,该微博是隶属于 S_d,然后以参数 γ_d 的狄利克雷分布中抽样出该影响微博与各个主题之间的关系 θ_γ,并在该关系上抽取出主题 z,然后在 φ 上抽取出一个单词来填充到微博当中的对应位置。在整个模型中,θ 的概率分布如公式(2)所示:

$$P(\theta|\alpha,\mu,\gamma_d)=xP(\theta|\alpha,\mu)+(1-x)P(\theta_\gamma|\gamma_d),\tag{2}$$

其中,x 是布尔值,当微博为转发微博、评论微博、无他人评论原创微博或者回复评论微博时,取 1,否则取 0,表示该微博为原创微博。

1.2 模型推导与用户兴趣主题挖掘

通过前面的分析得出微博与词、主题的联合概率分布 $P(r,z,w|\varphi,\theta,\chi)$,可表示为

$$P(r,z,w|\varphi,\theta,\chi)=\prod_{d\in D}\frac{\Delta(M_d^r+\gamma_d)}{\Delta(\gamma_d)}\cdot\prod_{d\in D}\frac{\Delta(N_d^z+\alpha)}{\Delta(\alpha)}\cdot\prod_{z\in T}\frac{\Delta(N_z^w+\beta)}{\Delta(\beta)},\tag{3}$$

其中,M_d^r 代表微博 d 当中的单词的影响微博的计数向量,N_d^z 代表被微博 d 所影响的可观察的主题的计数向量,N_z^w 为主题 z 当中的单词的计数向量。

对公式(3)进行分解,按照下面的方法进行迭代吉布斯采样:

$$P(z_i=z',r_i=r'|z_{-i},r_{-i},w)=\frac{P(z,r,w)}{P(z_{-i},r_{-i},w_{-i})}\cdot\frac{1}{P(w_i|z_{-i},r_{-i},w_{-i})},\tag{4}$$

得到后验分布公式:

$$P(z_i=z',r_i=r'|z_{-i},r_{-i},w)\propto\frac{P(z,r,w)}{P(z_{-i},r_{-i},w_{-i})}=\frac{N_{r'z'}^{-i}+\alpha}{N_{r'}^{-i}+T\alpha}\cdot\frac{M_{dr'}^{-i}+\gamma_d(r')}{\sum_{r\in S_d}(M_{dr'}^{-i}+\gamma_d(r))}\cdot\frac{N_{z'w_i}^{-i}+\beta}{N_{z'}^{-i}+V\beta}。\tag{5}$$

其中,$N_{r'z'}^{-i}$ 代表影响微博 r' 与主题 z' 的共现次数,$N_{r'}^{-i}$ 代表影响微博 r' 与所有主题的共现次数,$N_{z'w_i}^{-i}$ 代表单词 w_i 与主题 z' 的共现次数,$N_{z'}^{-i}$ 代表主题 z' 与所有单词的共现次数。对应于不同的微博类型,α 和 β 会对应不同的参数值。

对公式(3)进行反复迭代,并对所有的主题进行抽样,最终达到抽样结果稳定。由于抽单词和抽主题都满足多项式分布,θ_d、θ_{rt}、θ_c、θ_{re}、φ_z 以及 χ 的结果分别如下:

$$\theta_d=\frac{N_{dz}+\alpha}{N_d+T\alpha};\ \theta_c=\frac{N_{cz}+\alpha_c}{N_c+T\alpha_c};\ \theta_{rt}=\frac{N_{rtz}+\alpha_{rt}}{N_{rt}+T\alpha_{rt}};$$

$$\theta_{re}=\frac{N_{rez}+\alpha_{re}}{N_{re}+T\alpha_{re}};\ \varphi_z=\frac{N_{zw}+\beta}{N_z+V\beta};\ \chi_d(r)=\frac{M_{dr}+\gamma_d(r)}{\sum_{r\in S_d}(M_{dr}+\gamma_d(r))}。$$

通过吉布斯采样求解,得到微博在主题上的概率分布和主题在词上的概率分布。

对于每一个用户来说,通过加和每个主题下的概率,然后除以用户的微博数,就可以得到该用户在每个

主题下的概率分布,进而得到用户在各个主题下的概率分布和用户 - 主题特征向量,对每个主题来说,用户关于该主题的概率计算公式如下:

$$P_u(z_i) = \frac{\sum_{i=1}^{N} P(z_i)}{N}。 \tag{6}$$

以主题作为特征维,以用户在该主题下的概率作为特征值,利用向量空间模型对用户进行建模,可以得到基于微博内容的用户兴趣模型。

2　实验与结果分析

2.1　实验环境

在实验室的 PC 上运行由实验室开发的基于新浪 API 的微博内容爬取工具,运行环境:CPU 为 Intel G620 2.6 GHz,物理内存为 4 G,硬盘为 320 G,操作系统为 Windows XP Professional。微博生成模型的程序运行环境为:CPU 为 Intel G620 2.6 GHz,物理内存为 4 G,硬盘为 320 G,操作系统为 Windows XP Professional,开发工具为 My Eclipse 8.5。最后采用 Matlab 7.0 对生成数据进行分析。

2.2　实验数据集

人工选取了新浪微博中具有较大影响力的微博达人和微博名人等共 400 人,爬取了他们的所有微博数据共计 12 845 条。出于对网站安全以及服务器的负载能力的考虑,新浪对 API 的调用有诸多限制,并且设置了登录验证机制,想要爬取用户相关数据,必须通过 oauth 认证,并且每个 API 的调用次数有一定的限制。为了解决这个问题,在实际爬取数据的时候,采取的是每天分时间进行数据爬取,这样就避免了新浪 API 的调用限制所带来的数据爬取受限问题。按照微博类型将数据分别存储在 3 个数据表当中,其中第一个数据表为用户微博的原创部分,包括原创微博、转发微博的原创部分、评论微博以及回复微博,属性列包括用户 ID、用户名、用户微博 ID、用户微博内容。第二个数据表为用户微博所转发及评论的微博列表,属性列包括原微博 ID,转发、评论或回复微博 ID,转发、评论或回复微博内容,转发、评论或回复微博用户 ID。第三个数据表存储的是原创微博的对应评论微博,属性列包括微博 ID、评论微博 ID、评论微博内容和评论微博用户 ID。

2.3　实验数据预处理

(1)@ 符号。该符号后面跟某个用户的昵称,代表该条微博为指向型微博,我们选择将 @ 后直到下一个符号出现之前的内容删除。

(2) // 符号。该符号主要出现在转发微博当中,用于连接多个转发用户的转发时所附加的微博,我们在保存数据的时候,还要将转发的部分单独保存。

(3)# 符号。表示某一条微博围绕的话题,其结构为“# 话题内容 #”,内容一般为超链接的形式。针对此形式,我们选择将两个 # 号之间所包含的内容去掉。

(4) 文本中包含的超链接。有些用户发布微博时可能只有一两个字,针对这样的微博形式,利用网络爬虫,保存该 url 所对应的网页内容,对微博进行辅助分析。

通过以上处理得到了初步处理后的微博文本,然后利用中科院的 ICTCLAS 汉语分词系统进行分词处理,最后生成预料数据。

2.4　实验及结果分析

为了测试本文当中提出的有指导 LDA 微博生成模型的主题挖掘效果,分别进行两个对比实验,一是在引入 4 种指导信息的情况下,与传统 LDA 模型进行对比实验;二是对比不同指导信息对最后主题挖掘效果的影响。对于实验结果的评测,本文采用当前比较权威的 Perplexity 指标进行度量。该指标表示预测数据时的不确定度,值越小,复杂度越小,性能越好。Perplexity 的计算公式如下:

$$\text{Perplexity}(W) = \exp\left\{ -\frac{\sum_m \ln p(w_m)}{\sum_m N_m} \right\}, \tag{7}$$

其中,W 为测试集,w_m 为测试集当中可以观测到的词,N_m 为词的数量。

在相同的参数设置下,对Perplexity进行计算结果如表1所示。图2是利用Matlab对该数据的作图结果。从图2可以看出,通过引入4种指导信息对微博进行主题挖掘,有效地提高了模型的性能和推广性。

表1 两种模型的Perplexity对比数据

Table 1 The Perplexity contrast data of two models

迭代次数	LDA	有指导LDA
50	7 123.6	6 498.7
100	6 365.4	6 136.5
150	6 187.5	6 002.6
200	6 022.3	5 679.7
250	5 765.6	5 308.7
300	5 415.5	5 008.2

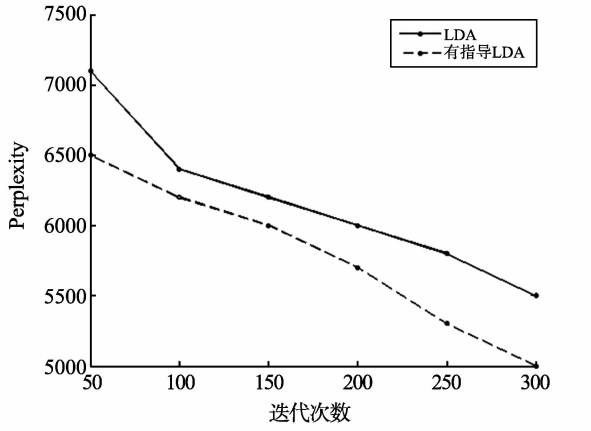

图2 两种模型的Perplexity对比

Fig. 2 Compared to the Perplexity of two models

为了验证不同的指导信息对微博主题挖掘效果的影响,本文将引入了4种指导信息的LDA模型、分别单独引入一种指导信息的LDA模型以及传统LDA模型对于微博主题提取的效果进行对比实验,实验结果如表2所示。

表2 不同指导信息的Perplexity对比数据

Table 2 The Perplexity contrast data by different guidance

迭代次数	传统LDA	转发关系	回复关系	评论关系	他人评论	4种信息综合
50	7 123.6	6 743.2	6 532.7	6 653.6	6 842.5	6 498.7
100	6 365.4	6 237.4	6 154.5	6 187.4	6 337.4	6 136.5
150	6 187.5	6 023.7	6 005.6	6 013.5	6 102.6	6 002.6
200	6 022.3	5 894.8	5 682.7	5 732.3	5 932.7	5 679.7
250	5 765.6	5 342.8	5 312.7	5 315.6	5 543.8	5 308.7
300	5 415.5	5 213.7	5 034.2	5 134.5	5 332.1	5 008.2

上述实验结果表明,在迭代次数相同时,引入4种不同的微博关系的Perplexity值普遍降低,均对模型的性能有了一定的提升,并且通过分析发现回复关系的Perplexity值下降幅度最大,从而表明了回复关系对用户兴趣聚类起到了更好的支撑作用。

3 结语

本文提出了一种基于有指导LDA的微博主题发现方法。实验表明4种微博关系对于用户兴趣主题挖掘具有一定的支撑作用,并且回复关系较其它3种关系有更好的效果。接下来的工作是要在分析用户兴趣主题的基础上,考虑针对主题的情感分布对用户兴趣分布的影响,进一步优化用户兴趣发现的效果。

参考文献:

[1] CLAYPOOL M, LE P, WASEDA M, et al. Implicit interest indicators[C]//Proceedings of the 6th International Conference. New York:ACM, 2001:30-40.

[2] SHEN Xuehua, TAN Bin, ZHAI Chengxiang. Implicit user modeling for personalized search[C]//Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management. New York:ACM, 2005, 10(5):5-6.

[3] 林鸿飞, 杨元生. 用户兴趣模型的表示和更新机制[J]. 计算机研究与发展, 2002, 39(7):843-847.
LIN Hongfei, YANG Yuansheng. The representation and update mechanism for user profile[J]. Journal of Computer Research and Development, 2002, 39(7):843-847.

[4] WENG Jianshu, 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:ACM, 2010:261-270.

[5] 董婧灵,李芳,何婷婷,等. 基于LDA模型的文本聚类研究[C]//中国计算语言学研究前沿进展,北京:清华大学出版社, 2011:455-461.
DONG Jingling, LI Fang, HE Tingting, et al. Document clustering method based on LDA model[C]//Advances of Computa-

tional Linguistics in China. Beijing:Tsinghua University Press, 2011:455-461.
[6] YAO Quanzhu, SONG Zhili, PENG Cheng. Research on text categorization based on LDA[J]. Computer Engineering and Applications, 2011, 47(13):150-153.
[7] 张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展, 2011, 48(10):1795-1802.
ZHANG Chenyi, SUN Jianling, DING Yiqun. Topic mining for Micro-blog based on MB-LDA model[J]. Journal of Computer Research and Development, 2011, 48(10):1795-1802.
[8] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.
[9] GRIFFITHS T, STEYVERS M. Finding scientific topics[C]//Proceedings of the National Academy of Sciences of the United States America. [S. l.]: [s. n.], 2004, 101:5228-5235.
[10] ROSEN-ZVI M, GRIFITHS T, STEYVERS M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Virginia: AUAI Press, 2004:487-494.
[11] HOFMANN T. Probabilistic latent semantic indexing[C]//Proceedings of the 22th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99). Berkeley: ACM, 1999:50-57.
[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 Suqun. Word similarity computing based on How-net[J]. International Journal of Computational Linguistics & Chinese Language Processing, 2002, 7(2):59-76.

(编辑:许力琴)

(上接第 35 页)

[6] ZHANG Y C, BLATTNER M, YU Y K. Heat conduction process on community networks as a recommendation model[J]. Phys Rev Lett, 2007, 99(15):4301-4305.
[7] LIU J G, ZHOU T, GUO Q. Information filtering via biased heat conduction[J]. Phys Rev E, 2011, 84(3): 7101-7105.
[8] QIU T, WANG T T, ZHANG Z K, et al. Heterogeneity involved network-based algorithm leads to accurate and personalized recommendations[J]. Physics and Society, 2013, arXiv:1305.7438vl.
[9] ZHOU T, KUSCSIK Z, LIU J G, et al. Solving the apparent diversity accuracy dilemma of recommender systems[C]//Proceedings of the National Academy of Sciences of the United States of America. Washington: Natl Acad Sciences, 2010, 107:4511-4515.
[10] SCOTT A G, BERNARDO A H. Usage patterns of collaborative tagging systems[J]. Journal of Information Science, 2006, 32(2):198-208.
[11] ZHANG Z K, ZHOU T, ZHANG Y C. Tag-aware recommender systems: a state-of-the-art survey[J]. Journal of Computer Science and Technology, 2011, 26(5): 767-777.
[12] ZHANG Z K, LIU C, ZHANG Y C, et al. Solving the cold-start problem in recommender systems with social tags[J]. Europhysics Letters, 2010, 92(2): 8002-8010.
[13] MICHAEL J P, DANIEL B. Content-based recommendation systems[J]. Lecture Notes in Computer Science, 2007, 4321:325-341.
[14] CANTADOR I, BELLOGÍN A, VALLET D. Content-based recommendation in social tagging systems[C]//Proceedings of RecSys'10. New York: ACM, 2010:237-240.
[15] JIANG Shengyi, SONG Xiaoyu, WANG Hui, et al. A clustering-based method for unsupervised intrusion detections[J]. Pattern Recognition Letters, 2006, 27(7): 802-810.
[16] BURKE R. Hybrid Recommender systems: survey and experiments[J]. User Model User-Adap Interact, 2007, 12(4):331-370.
[17] JOACHIMS T. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization[C]//Proceedings of the 14th International Conference on Machine Learning. New York: ACM, 1997:143-151.

(编辑:许力琴)