DOI:10.16451/j.cnki.issn1003-6059.2014.08.001

考虑用户标注状态的标签推荐方法*

于 洪 邓明瑶 胡 峰

(重庆邮电大学 计算智能重庆市重点实验室 重庆 400065)

摘 要 为进一步提升标签推荐的质量,提出一种考虑用户当前标注状态的标签推荐方法. 首先根据统计分析方法发现社会标签系统中用户使用的标签总数随时间有一定的变化规律,因此提出当前用户标注状态可能属于下列 3 种情况之一: 成长态、成熟态和休眠态,并给出相关定义. 然后根据 3 种用户标注状态的不同特点,提出不同策略下计算标签的概率分布,为用户推荐最可能使用的标签. 对比实验表明文中方法能提供更准确的推荐结果.

关键词 社会标签, 标签推荐, 标注状态, 概率分布

中图法分类号 TP 391

Tag Recommendation Method Considering Users Tagging Status

YU Hong, DENG Ming-Yao, HU Feng

(*Chongqing Key Laboratory of Computational Intelligence*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065)

ABSTRACT

To improve the quality of tag recommendation, a tag recommendation method considering users current tagging status is proposed. Firstly, the statistical analysis shows the total number of tags used by a user is changed with time in a social tagging system. Then, three tagging statuses are defined, i. e. the growing status, the mature status and the dormant status, and a user current tagging status is one of the above statuses. Finally, according to the characteristics of the current tagging status, different strategies are developed to compute the tag probability distribution to recommend tags to users. Results of comparative experiments show that the proposed method has better accuracy of tag recommendation.

Key Words Social Tagging, Tag Recommendation, Tagging Status, Probability Distribution

* 国家自然科学基金项目(No. 61379114,61272060)、重庆市自然科学基金项目(No. cstc2013jcyjA40063)资助
收稿日期: 2013 - 06 - 26; 修回日期: 2013 - 11 - 20
作者简介 于洪(通讯作者),女,1972 年生,博士,教授,主要研究方向为粗糙集、智能信息处理、Web 智能、数据挖掘等. E-mail: yuhong@ cqupt. edu. cn. 邓明瑶,女,1987 年生,硕士研究生,主要研究方向为智能推荐、数据挖掘. 胡峰,男,1978 年生,博士,教授,主要研究方向为人工智能、软件工程、数据挖掘.

1 引 言

社会标签(Social Tag)是一种大众索引方法,广泛应用于各种收藏、检索、共享网站. 如 CiteULike(http://www.citeulike.org)是由 Springer 提供的一个免费协助用户存储、管理和分享学术论文的网站. 用户使用标签检索、组织及共享信息,这在一定程度上可缓解用户面临的信息超载问题. 因此标签推荐(Tag Recommendation)受到学术界及互联网企业的广泛关注.

近年来,学者们提出各种标签推荐算法,主要有基于协同过滤的标签推荐算法[1-3]、基于图模型的标签推荐算法[4-7]、基于内容的标签推荐算法[8-10]和混合的标签推荐算法[11-13]等.

基于协同过滤的标签推荐算法在数据转换中存在信息丢失问题;基于图模型的标签推荐算法具有内存空间代价和时间代价敏感的特点,且不具有可拓展性;基于内容的标签推荐算法利用资源的内容、标签的语义信息,采用机器学习方法进行标签推荐,不仅计算复杂而且只能推荐那些易提取特征的资源. 如文献[4]根据标签计数推荐目标用户和目标资源的最流行标签,未考虑个性化标签推荐;文献[5]虽然将 FolkRank 表示为个性化网页排名的线性组合,但是它在空间和时间上都消耗较大;文献[6]仅在目标用户和目标资源的标签空间内计算标签的权重排名.

已有的标签推荐算法大都未考虑用户的标注行为随时间变化的情况,及不同用户在不同时间段内标注行为的不同. 然而本文通过统计分析发现,社会标签系统中用户使用过的标签总数随时间变化具有一定的变化规律. 因此本文提出用户标注状态的概念,在某个时间,用户标注状态处于如下 3 种状态之一:成长态、成熟态和休眠态. 根据用户标注状态,本文采用统计语言模型(Statistical Language Model)计算标签在用户标签空间和资源标签空间的概率分布,提出不同的标签推荐策略,为用户提供个性化的标签推荐. 对比实验表明本文的推荐策略具有较好的推荐结果.

2 基本概念介绍

2.1 社会化标签系统

Folksonomy(http://vanderwal.net/folksonomy.html)是 Wal 在 2004 年提出的术语,表示社会标签系统的基本数据结构. Folksonomy 又称“大众分类法”,由社会性书签服务中最具特色的自定义标签(Tag)功能衍生而来.

在社会标签系统中,一个 Folksonomy 是一个四元组,即 $\mathcal{F}=(U,R,Tag,Y)$,其中

$$U=\{u_1,\cdots,u_k,\cdots,u_K\}$$

表示用户集合,

$$R=\{r_1,\cdots,r_l,\cdots,r_L\}$$

表示资源集合,

$$Tag=\{tag_1,\cdots,tag_m,\cdots,tag_M\}$$

表示标签集合,K、L 和 M 分别表示用户数目、资源数目和标签数目;Y 表示用户、资源和标签间的一个三元关系,即

$$Y=\{(u,r,tag)\mid u\in U,r\in R,tag\in Tag\}\subseteq U\times R\times Tag.$$

用户 u_k 使用过的标签记为标签集合 Tag_{u_k},资源 r_l 的标注标签记为标签集合 Tag_{r_l}.

用户、资源和标签之间的三元关系 Y 可转换为 3 个二元关系:用户和标签的二元关系 $U\times Tag$,用户和资源的二元关系 $U\times R$,资源和标签的二元关系 $R\times Tag$. 用户-标签矩阵 $\boldsymbol{UTag}_{K\times M}=[w_{u_k tag_m}]$ 表示二元关系 $U\times Tag$,其中,$w_{u_k tag_m}$ 表示用户 u_k 的标签 tag_m 的权重,即用户 u_k 使用标签 tag_m 标注的资源数目. 资源-标签矩阵 $\boldsymbol{RTag}_{L\times M}=[w_{r_l tag_m}]$ 表示二元关系 $R\times Tag$,其中,$w_{r_l tag_m}$ 表示资源 r_l 的标签 tag_m 的权重,即使用标签 tag_m 标注资源 r_l 的用户数目.

标签推荐是根据标签推荐算法得到推荐标签集合

$$\hat{Tag}(u_q,r_q)\subseteq Tag,$$

其中,$u_q\in U$ 为目标用户,$r_q\in R$ 为目标资源. 按照某种规则排序标签集合,从中选择前 n 个标签作为标签集合 $\hat{Tag}^n(u_q,r_q)$ 推荐给用户 u_q.

标签匹配 a 是用户 u_k 为资源 r_l 添加标签的标注行为,即

$$a=(u_k,r_l,Tag(u_k,r_l)),$$

集合 $Tag(u_k,r_l)$ 表示用户 u_k 标注资源 r_l 时所使用的标签的集合,即

$$Tag(u_k,r_l)=\{tag\in Tag\mid(u_k,r_l,tag)\in Y\}.$$

一个社会标签系统中所有的标签匹配组成标签匹配集合 A.

2.2 统计语言模型

统计语言模型广泛应用于自然语言处理领域,如语音识别、机器翻译和信息检索等. 它是一个概率分布模型,主要描述自然语言的统计和结构方面的

内在规律. 所有合法的字符串的集合称为一个语言，一个语言模型就是这个语言里的字符串的概率分布模型.

在信息检索领域，统计语言模型的最基本思想是把一个查询 q 和文档 d 的相关性解释为从文档中产生查询的概率模型[14]，即

$$p_{LM}(q \mid d) = \prod_{w \in q} p(w \mid d),$$

其中，w 是查询中的一个词. $p(w \mid d)$ 是从文档 d 中产生查询词 w 的概率，

$$p(w \mid d) = \frac{N_d}{N_d + \lambda} \frac{tf(w,d)}{N_d} + \left(1 - \frac{N_d}{N_d + \lambda}\right) \frac{tf(w,D)}{N_D}, \quad (1)$$

N_d 为文档 d 以词为单位计算的长度; $tf(w,d)$ 是文档 d 中词 w 的词频; $tf(w,D)$ 是所有文档集中词 w 出现的词频; N_D 是所有文档集中词的总数; λ 是一个狄雷克雷特平滑因子，它的取值根据文档集中的平均文档长度来设定，即 $\lambda = N_d/N_D$.

3　基于标注历史的用户标注状态分析

3.1　历史标签与新标签定义

在一段时间 T 内观察用户拥有的标签总数的变化，设刚开始观察的时刻是 T_0，取等间隔的时间段作为观察点(后文实验中以月为单位时间)，那么下一个时刻记为 T_1. 假设当前时刻是 T_t，

$$TAG^{u_k,T_t} = \{tag_1^{u_k,T_t}, \cdots, tag_n^{u_k,T_t}, \cdots, tag_N^{u_k,T_t}\}$$

表示用户 u_k 在一个单位时间间隔(即 $T_{t-1} - T_t$) 内使用的标签集合，其中 N 表示用户 u_k 在 $T_{t-1} - T_t$ 时间内使用的标签个数. 对于 $\forall i \in [1,N]$, $j \in [1,N]$，如果 $i \neq j$，则

$$tag_i^{u_k,T_t} \neq tag_j^{u_k,T_t}.$$

$f_{u_k}(T_t)$ 表示在 $T_{t-1} - T_t$ 时间内用户 u_k 使用的标签数量，即

$$f_{u_k}(T_t) = |TAG^{u_k,T_t}| = N.$$

$g_{u_k}(T_t)$ 表示在 $T_0 - T_t$ 时间内用户 u_k 已使用过的所有的标签数量，即

$$g_{u_k}(T_t) = \left| \bigcup_{\tau=0}^{t} TAG^{u_k,T_\tau} \right|.$$

用户 u_k 在时刻 T_{t-1} 之前已使用过的标签称为用户 u_k 在时刻 T_t 的历史标签，其数量

$$g_{u_k}(T_{t-1}) = \left| \bigcup_{\tau=0}^{t-1} TAG^{u_k,T_\tau} \right|.$$

用户 u_k 在时刻 T_{t-1} 之前未使用过但在最近时间间隔 $T_{t-1} - T_t$ 内使用的标签称为用户 u_k 的新标签，其数量

$$g_{u_k}(T_t) - g_{u_k}(T_{t-1}) = \left| \bigcup_{\tau=0}^{t} TAG^{u_k,T_\tau} \setminus \bigcup_{\tau=0}^{t-1} TAG^{u_k,T_\tau} \right|.$$

显然在 $T_{t-1} - T_t$ 时间内，用户要么有标注行为，要么没有标注行为，即有如下 2 种情况.

情况 1　用户 u_k 在 $T_{t-1} - T_t$ 时间内没有标注行为，即 $f_{u_k}(T_t) = 0$.

情况 2　用户 u_k 在 $T_{t-1} - T_t$ 时间内有标注行为，即 $f_{u_k}(T_t) \neq 0$，此时用户使用历史标签或新标签.

针对情况 2，又分为如下 2 类.

1) 用户在 $T_{t-1} - T_t$ 时间内使用新标签，即

$$g_{u_k}(T_t) - g_{u_k}(T_{t-1}) > 0.$$

这种情况下，需考虑 2 个方面.

(1) 用户 u_k 在 $T_{t-1} - T_t$ 时间内既使用新标签又使用历史标签，即

$$0 < \frac{g_{u_k}(T_t) - g_{u_k}(T_{t-1})}{f_{u_k}(T_t)} < 1.$$

(2) 用户 u_k 在 $T_{t-1} - T_t$ 时间内仅使用新标签，没有使用历史标签，即

$$\frac{g_{u_k}(T_t) - g_{u_k}(T_{t-1})}{f_{u_k}(T_t)} = 1.$$

2) 用户 u_k 在 $T_{t-1} - T_t$ 时间内没有使用新标签，只使用历史标签，即

$$g_{u_k}(T_t) - g_{u_k}(T_{t-1}) = 0.$$

3.2　社会标签系统数据集统计分析

本文基于学术文献共享网站 CiteULike 数据集、电影推荐系统 MovieLens2 数据集、书签分享网站 Delicious2 数据集和音乐网站 Last. fm 数据集分别统计分析在一段时间内用户使用标签的情况. 限于篇幅，在此仅以 CiteULike 数据集为例介绍统计分析实验及其结果.

实验统计在 2004 年 11 月 ~ 2008 年 09 月(共 47 个月) 期间，CiteULike 数据集中 2 925 个用户的标签总数随时间的变化情况(时间以月为单位)，分别统计每个月内每个用户使用过的不同标签总数.

在一段时间内，如果用户拥有的标签总数(即这段时间内用户标注资源所使用的不同标签的数量) 缓慢增长或急剧增长，那么这段时间内用户使用新标签，同时用户也可能使用历史标签. 如果用户拥有的标签总数保持不变，那么这段时间内用户只使用历史标签，或用户这段时间内没有标注行为.

分析统计结果发现用户的标注状态有 3 种情况: 1) 在一段时间内用户的标签总数快速增加; 2)

用户的标签总数缓慢增加;3) 用户在一段时间内没有标注行为. 用户的标注状态是第 1 种情况时,他为收藏或共享的不同资源添加很多的新标签;用户的标注状态是第 2 种情况时,用户可能使用历史标签标注资源,这时新标签的使用会相对减少;用户的标注状态是第 3 种情况时,用户在这段时间内没有为资源添加标签.

作为示例,图 1 给出用户($ID = 1845$) 的标签总数随着时间变化的曲线. 图中,$totalTagNum$ 表示随着月份增长用户不同标签总数的变化,即 $g_{u_k}(T_t)$ 的变化曲线;$perMonthTagNum$ 表示每个月内用户的不同标签总数的变化,即 $f_{u_k}(T_t)$ 的变化曲线. 可见从第 1 个月至第 4 个月,用户($ID = 1845$) 没有标注行为,这期间用户的标签总数为 0,说明在这段时间内该用户的标注状态处于第 3 种情况;从第 5 个月至第 19 个月,用户标签总数快速增长,且这期间每个月内的标签总数都不为 0,这说明在这段时间内该用户的标注状态处于第 1 种情况;从第 20 个月至第 47 个月,用户标签总数增长缓慢,且这期间每个月内用户的标签总数大多数不为 0,说明在这段时间内该用户的标注状态处于第 2 种情况.

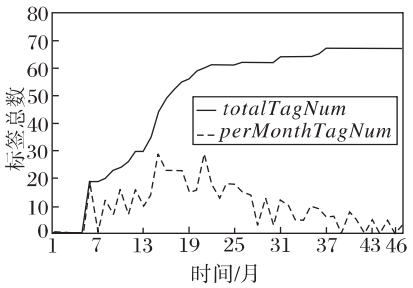

图 1 用户($ID=1845$) 的标签总数随时间变化的曲线

Fig. 1 Curves of the total number of different tags changing with time for user $ID=1845$

3.3 用户标注状态

通过上述统计分析发现在社会标签系统中用户的标注状态具有 3.2 节描述的 3 种情况,因此本文将上述用户的标注状态的 3 种情况分别定义为成长态、成熟态和休眠态,并在本节给出相应定义.

在时间段 T 内,用户拥有的标签总数是增大的,且这段时间内标签总数的平均增长率大于等于阈值 α,说明这段时间内用户的标注行为较活跃,且在社会标签系统中添加很多新标签,称用户此时的用户标注状态处于成长态.

定义 1 成长态 若用户 u_k 在时间段 $[T_{t-\Delta t}, T_t]$ 内,

$$\frac{g_{u_k}(T_{t-1}) - g_{u_k}(T_{t-\Delta t})}{\Delta t} \geqslant \alpha,$$

则用户 u_k 在时刻 T_t 的用户标注状态处于成长态.

在时间段 T 内,用户拥有的标签总数是增大的,且这段时间内标签总数的平均增长率小于阈值 α,说明这段时间内用户在社会标签系统中添加少量新标签,使用很多的历史标签,称用户此时的用户标注状态处于成熟态.

定义 2 成熟态 若用户 u_k 在时间段 $[T_{t-\Delta t}, T_t]$ 内,$\exists T_{t'} \in [T_{t-\Delta t}, T_t]$,使得 $f_{u_k}(T_{t'}) \neq 0$ 且

$$0 \leqslant \frac{g_{u_k}(T_{t-1}) - g_{u_k}(T_{t-\Delta t})}{\Delta t} < \alpha,$$

则用户 u_k 在时刻 T_t 的用户标注状态处于成熟态.

在时间段 T 内用户没有标注行为,用户拥有的标签总数是不变的,称用户此时的用户标注状态处于休眠态.

定义 3 休眠态 若用户 u_k 在时间段 $[T_{t-\Delta t}, T_t]$ 内,$\forall T_{t'} \in [T_{t-\Delta t}, T_t]$,使得 $f_{u_k}(T_{t'}) = 0$,则用户 u_k 在时刻 T_t 的用户标注状态处于休眠态.

3.4 用户标注状态确定算法

考察在时刻 T_t 之前一个时间段 Δt 内,用户 u_k 的标注历史,即用户拥有的标签总数的变化,根据定义 1 ~ 定义 3 可判断用户 u_k 在某时刻 T_t 的标注状态. T_0 表示用户 u_k 刚开始使用社会标签系统时,如果用户 u_k 使用社会标签系统的时间小于 Δt,且其近期有标注行为,则认为其处于成长态. 由图 1 可见,每个人都有自己的个性,一个用户在社会标签系统中所处的各种标注状态并不是等时的. 为简化计算,下面给出的用户标注状态确定算法只考察当前时刻之前的一个固定时间段的历史标签变化情况.

算法 用户标注状态确定算法

输入 当前时刻值 T_t,阈值 α,Δt

输出 用户 u_k 在时刻 T_t 的用户标注状态

step 1 如果 $T_t - T_0 < \Delta t$,对 $\forall T_{t'} \in [T_0, T_t]$,均有 $f_{u_k}(T_{t'}) = 0$,则用户 u_k 在时刻 T_t 的用户标注状态处于休眠态;若 $\exists T_{t'} \in [T_0, T_t]$,使得 $f_{u_k}(T_{t'}) \neq 0$,则用户 u_k 在时刻 T_t 的用户标注状态处于成长态;否则执行 step 2.

step 2 如果 $\forall T_{t'} \in [T_{t-\Delta t}, T_t]$,使得 $f_{u_k}(T_{t'}) = 0$,则用户 u_k 在时刻 T_t 的用户标注状态处于休眠态;否则执行 step 3.

step 3 计算

$$differT = \frac{g_{u_k}(T_{t-1}) - g_{u_k}(T_{t-\Delta t})}{\Delta t}.$$

step 4 如果 $differT \geqslant \alpha$,则用户 u_k 在时刻 T_t 的

用户标注状态处于成长态; 如果 $0 \leqslant differT < \alpha$,则用户 u_k 在时刻 T_t 的用户标注状态是成熟态.

4　考虑用户标注状态的标签推荐方法

首先根据用户的标注历史判断此刻用户标注状态; 然后针对这 3 种用户标注状态的不同特点,分别采取不同的标签概率统计方法作为标签推荐策略,为用户推荐其最可能使用的标签. 本文将这种标签推荐方法命名为 TR-CUTS(Tag Recommendation Considering User Tagging Status) ,并将用户标注状态处于成长态、成熟态和休眠态时采用的标签推荐策略分别命名为 TR-GU(Tag Recommendation for Growing Users) 、TR-MU(Tag Recommendation for Mature Users) 和 TR-DU (Tag Recommendation for Dormant Users) . 下面详细描述各种标签推荐策略.

4.1　用户标注状态处于成长态

用户 u_q 标注资源 r_q 时,若此刻用户标注状态是成长态,则在此刻之前的一段时间内,用户 u_q 标注的资源数量不断增加,用户 u_q 拥有的标签总数不断增大. 本文利用用户 u_q 及用户 u_q 所属组内的用户标签和资源 r_q 及与资源 r_q 相似的资源标签,采用语言模型方法计算标签的概率分布以进行标签推荐. 这样既保证为用户 u_q 推荐的标签的个性化,也提高为用户 u_q 推荐的标签的多样性.

step 1　利用资源-标签矩阵 $\boldsymbol{RTag}_{L\times M}$,采用余弦相似性方法计算资源 $r_l(r_l \in R\backslash\{r_q\})$ 与资源 r_q 的相似性,并选择相似性最高的前 S 个资源作为与资源 r_q 相似的资源的集合.

资源-标签矩阵 $\boldsymbol{RTag}_{L\times M}$ 的每行用向量 $\boldsymbol{r}$ 表示. $\boldsymbol{r}_i$、$\boldsymbol{r}_q$ 分别表示资源 r_i、r_q 的标签集合里每个标签的权重信息,即

$$\boldsymbol{r}_i = (w_{r_i tag_1}, w_{r_i tag_2}, \cdots, w_{r_i tag_m}, \ldots, w_{r_i tag_M}),$$

$$\boldsymbol{r}_q = (w_{r_q tag_1}, w_{r_q tag_2}, \cdots, w_{r_q tag_m}, \ldots, w_{r_q tag_M}),$$

则 r_i 与 r_q 的相似性计算如下:

$$sim(\boldsymbol{r}_i, \boldsymbol{r}_q) = \frac{\boldsymbol{r}_i \cdot \boldsymbol{r}_q}{\|\boldsymbol{r}_i\| \ \|\boldsymbol{r}_q\|}.$$

step 2　在资源 r_q 及与资源 r_q 相似的资源标签空间,根据 LM 得到资源 r_q 被标签 tag_m 标注的概率 $p(tag_m \mid r_q)$:

$$p(tag_m \mid r_q) = \frac{N_{Tag_{r_q}}}{N_{Tag_{r_q}} + \lambda_{r_q}} \frac{TF(tag_m, Tag_{r_q})}{N_{Tag_{r_q}}} + \left(1 - \frac{N_{Tag_{r_q}}}{N_{Tag_{r_q}} + \lambda_{r_q}}\right) \frac{TF(tag_m, TagS)}{N_{TagS}},$$

其中,$TF(tag_m, Tag_{r_q})$ 表示资源 r_q 的标签 tag_m 的权重,

$$TF(tag_m, Tag_{r_q}) = w_{r_q tag_m};$$

$N_{Tag_{r_q}}$ 表示资源 r_q 的标签权重之和; $TagS$ 表示资源 r_q 及与资源 r_q 相似的资源标签集合,且对于 $\forall tag \in TagS$,资源的标签权重

$$w'_{r_k tag} = w_{r_k tag} \cdot sim(r_q, r_k);$$

N_{TagS} 表示标签集合 $TagS$ 内的标签权重之和; $TF(tag_m, TagS)$ 表示资源 r_q 及与资源 r_q 相似的资源标签 tag_m 的权重之和; λ_{r_q} 是一个狄雷克雷特平滑因子,$\lambda_{r_q} = N_{Tag_{r_q}}/N_{TagS}$.

step 3　类似地,在用户 u_q 及用户 u_q 所属组的组内用户的标签空间,根据 LM,利用用户-标签矩阵 $\boldsymbol{UTag}_{K\times M}$,得到用户 u_q 使用标签 tag_m 的概率 $p(tag_m \mid u_q)$:

$$p(tag_m \mid u_q) = \frac{N_{Tag_{u_q}}}{N_{Tag_{u_q}} + \lambda_u} \frac{TF(tag_m, Tag_{u_q})}{N_{Tag_{u_q}}} + \left(1 - \frac{N_{Tag_{u_q}}}{N_{Tag_{u_q}} + \lambda_u}\right) \frac{TF(tag_m, Tag_{Uq})}{N_{Tag_{Uq}}},$$

其中,$TF(tag_m, Tag_{u_q})$ 表示用户 u_q 的标签 tag_m 的权重,

$$TF(tag_m, Tag_{u_q}) = w_{u_q tag_m};$$

U_q 表示用户 u_q 及用户 u_q 所属组的组内用户所组成的用户集合; Tag_{U_q} 表示集合 U_q 内用户标签集合; $N_{Tag_{u_q}}$ 表示用户 u_q 的标签权重之和; $N_{Tag_{U_q}}$ 表示集合 U_q 内所有用户的标签权重之和; $TF(tag_m, Tag_{U_q})$ 表示 U_q 内用户标签 tag_m 权重之和; λ_u 是一个狄雷克雷特平滑因子,$\lambda_u = N_{Tag_{u_q}}/N_{Tag_{U_q}}$.

step 4　用户 u_q 使用标签 tag_m 标注资源 r_q 的可能性 $p(tag_m \mid u_q, r_q)$:

$$p(tag_m \mid u_q, r_q) = (1-\beta)\, p(tag_m \mid u_q) + \beta p(tag_m \mid r_q),$$

其中 $\beta \in [0,1]$.

step 5　根据 $p(tag_m \mid u_q, r_q)$ 对标签排序,为用户 u_q 推荐 Top-n 标签列表:

$$\overset{\wedge}{Tag}^n(u_q, r_q) = \arg \max^n_{tag_m \in Tag} (p(tag_m \mid u_q, r_q)).$$

4.2　用户标注状态处于成熟态

用户 u_q 标注资源 r_q 时,若此刻用户标注状态处于成熟态,则在此刻之前的一段时间内用户 u_q 的标注行为趋于稳定,用户 u_q 标注的资源达到一定数

量,则该用户的标签数量缓慢增加.当用户 u_q 标注资源 r_q 时,利用用户 u_q 的标签和资源 r_q 的标签,采用简单语言模型方法计算标签概率分布进行标签推荐. 这样在保证标签推荐准确性的基础上,也降低计算的复杂度.

step 1 对于 $\forall tag_m \in Tag_{u_q}$,根据标签计数计算用户 u_q 使用标签 tag_m 的概率 $p_u(tag_m \mid u_q)$:

$$p_u(tag_m \mid u_q) = \frac{w_{u_q tag_m}}{N_{Tag_{u_q}}},$$

其中,$N_{Tag_{u_q}}$ 表示用户 u_q 的标签权重之和,

$$N_{Tag_{u_q}} = \sum_{tag \in Tag_{u_q}} w_{u_q tag}.$$

step 2 对于 $\forall tag_m \in Tag_{r_q}$,根据标签计数计算资源 r_q 被标签 tag_m 标注的概率 $p_r(tag_m \mid r_q)$:

$$p_r(tag_m \mid r_q) = \frac{w_{r_q tag_m}}{N_{Tag_{r_q}}},$$

其中,$N_{Tag_{r_q}}$ 表示资源 r_q 的标签权重之和,

$$N_{Tag_{r_q}} = \sum_{tag \in Tag_{r_q}} w_{r_q tag}.$$

step 3 用户 u_q 使用标签 tag_m 标注资源 r_q 的可能性 $p(tag_m \mid u_q, r_q)$:

$$p(tag_m \mid u_q, r_q) = (1-\gamma)\, p_u(tag_m \mid u_q) + \gamma p_r(tag_m \mid r_q),$$

其中,$tag_m \in (Tag_{r_q} \cup Tag_{u_q})$,$\gamma \in [0,1]$.

step 4 根据 $p(tag_m \mid u_q, r_q)$ 排序标签,为用户 u_q 推荐 $Top\text{-}n$ 标签列表:

$$\overset{\wedge}{Tag}^n(u_q, r_q) = \arg \max^n_{tag_m \in Tag} (p(tag_m \mid u_q, r_q)).$$

4.3 用户标注状态处于休眠态

用户 u_q 标注资源 r_q 时,若此刻用户标注状态处于休眠态,则在此刻之前的一段时间内用户 u_q 没有标注行为. 用户 u_q 开始标注资源 r_q 时,利用资源 r_q 及与资源 r_q 相似的资源标签集合,采用语言模型方法计算标签的概率分布以进行标签推荐.

step 1 与 4.1 节相同,计算得到资源 r_q 被标签 tag_m 标注的概率 $p(tag_m \mid r_q)$.

step 2 根据 $p(tag_m \mid r_q)$ 排序标签,为用户 u_q 推荐 $Top\text{-}n$ 标签列表:

$$\overset{\wedge}{Tag}^n(u_q, r_q) = \arg \max^n_{tag_m \in Tag} (p(tag_m \mid r_q)).$$

5 实验及结果分析

为验证本文的用户标注状态的有效性,本节进行 TR-CUTS 与 TR-GU、TR-MU 和 TR-DU 的对比实验. 同时为验证 TR-CUTS 的有效性,实验采用 FolkRank [5]、LocalRank [6] 和最流行标签算法(Most Popular Tags ρ-Mix) [4] 作为对比方法,并利用 Jäschke 等 [15] 提出的 LeavePostOut 方法评估各算法的标签推荐效果. 本文在 CiteULike 数据集(http://www.citeulike.org/faq/data.adp) 和 Last.fm 数据集(http://www.grouplens.org/node/462) 上进行实验. 由于在 Last.fm 数据集中没有用户所属组信息,本文把用户及其好友组成一个组.

5.1 数据集预处理

由于数据的稀疏性和“长尾”现象,利用 p-core of level k [16] 处理数据集,使每个用户、资源和标签至少在数据集中出现 k 次. CiteULike 数据集和 Last.fm 数据集预处理后的统计结果如表 1 所示.

表 1 数据集统计结果

Table 1 Statistical results of datasets

	citeULike(k=30)	Last.fm(k=10)
用户数	1700	966
资源数	32208	3870
标签数	6012	1204
标签匹配数	89076	48578
用户资源标签三元关系数	1507781	133945
数据时间区间	2004-11-04~2012-10-16	2005-08-01~2011-01-01
用户平均标注行为数	52	50
标签匹配中的平均标签数	17	3
用户标签使用数平均值	113	20
资源被标注的标签数平均值	38	16

对于预处理过的 CiteULike 数据集,实验选择其在 2004-11-04 ~ 2012-10-01 期间的数据集作为训练集,选择其在 2012-10-01 ~ 2012-10-16 期间的数据集作为测试集.

5.2 评价方法

本文采用信息检索领域中的标准度量方法评估标签推荐的效果 [5],即计算 Top-n 推荐时的召回率(Recall)、精确率(Precision) 和 F-measure: $R@n$、$P@n$ 和 $F@n$,其中 n 表示推荐标签集合中标签的个数.

设测试集中的用户得到的用户子集记为 $\widetilde{U}$,有 $\widetilde{U} \subseteq U$. 对于 $\widetilde{U}$ 中的每个用户 u_q 的每个标签匹配

$$a = (u_q, r_q, Tag(u_q, r_q)),$$

利用训练集为 a 得到一个推荐标签集合 $\hat{Tag}^n(u_q, r_q)$. 实验对比测试集中每个标签匹配的标签集合 $Tag(u_q, r_q)$ 与相应的推荐标签集合 $\hat{Tag}^n(u_q, r_q)$，计算 $R@n$、$P@n$ 和 $F@n$.

$R@n$ 表示推荐标签集合中正确推荐的标签在该标签匹配的标签集合中所占的比例:

$$R@n = \frac{|Tag(u_q, r_q) \cap \hat{Tag}^n(u_q, r_q)|}{|Tag(u_q, r_q)|}.$$

$P@n$ 表示推荐标签集合中正确推荐的标签在推荐标签集合中所占的比例:

$$P@n = \frac{|Tag(u_q, r_q) \cap \hat{Tag}^n(u_q, r_q)|}{|\hat{Tag}^n(u_q, r_q)|}.$$

由于 n 的大小影响标签推荐的 $P@n$ 和 $R@n$，n 越大会有更大的 $R@n$ 和更小的 $P@n$，因此实验中采用 $F@n$ 表示 $R@n$ 和 $P@n$ 之间的权重调和平均:

$$F@n = \frac{2R@n \cdot P@n}{R@n + P@n}.$$

实验中 $n = 1, 2, \cdots, 5$，计算测试集中每个标签匹配的 $R@n$、$P@n$ 和 $F@n$. 为获得标签推荐算法在整个测试集上的推荐效果，本文实验结果记录相应平均值.

5.3　实验结果

实验分为 3 部分: 第 1 部分给出确定测试集中用户的标注状态; 第 2 部分获得本文标签推荐策略中涉及的阈值及最流行标签算法中的参数值 ρ; 第 3 部分是 TR-CUTS 与 TR-GU、TR-MU 和 TR-DU 的对比结果，及 TR-CUTS 与 FolkRank、LocalRank 和最流行标签算法的对比结果. 限于篇幅，本文以 CiteULike 数据集为例，给出详细的实验过程及对比结果，在 Last. fm 数据集上的实验只给出对比实验结果.

5.3.1　确定用户标注状态

在 CiteULike 数据集上，实验以月为时间间隔. 对于 $\forall u_k \in \tilde{U}$，首先利用 Microsoft SQL Server 2008 统计 2004 - 11 - 04 ~ 2012 - 10 - 01 期间(共 86 个月) 每个月用户 u_k 的标签总数 $f_{u_k}(T_t)$ 和 $g_{u_k}(T_t)$. 然后根据 3.4 节中的判断用户标注状态算法得到用户 u_k 在每个月份的用户标注状态.

图 2 以 CiteULike 数据集上用户($ID = 103$) 为例给出从第 1 个月至第 86 个月之间不同标签总数变化曲线. *totalTagNum* 表示随月份增长用户($ID = 103$) 的不同标签总数的变化，即 $g_{u_k}(T_t)$ 的变化曲线，*perMonthTagNum* 表示每个月内用户($ID = 103$) 的不同标签总数的变化，即 $f_{u_k}(T_t)$ 的变化曲线. 表 2 给出用户($ID = 103$) 某些月通过用户标注状态确定算法计算得到的该用户所处的用户标注状态. 图 2 中的结果与表 2 中标签总数的变化趋势相符合.

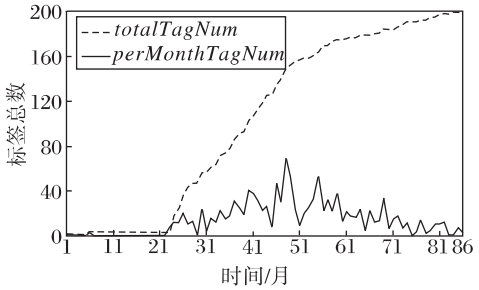

图 2　用户($ID = 103$) 的标签总数随时间变化的曲线

Fig. 2　Curves of the total number of different tags changing with time for user $ID = 103$

表 2　用户($ID = 103$) 在部分月份的用户标注状态

Table 2　Tagging status of user $ID = 103$ in several months

时间/月	1	3	6	16	26	36	46	56	66	76	86
标注状态	休眠态	休眠态	成长态	成长态	成长态	成长态	成长态	成熟态	成熟态	成熟态	成熟态

在 CiteULike 数据集上，实验得到阈值 $\Delta t = 4$，$\alpha = 3$. 在 Last. fm 数据集上，采用同样方法，实验得到阈值 $\Delta t = 2$，$\alpha = 2$.

5.3.2　参数确定

根据 3.4 节用户标注状态确定算法得到用户子集 $\tilde{U}$ 中每个用户的用户标注状态. 按照用户标注状态，把 CiteULike 数据集的测试集分为 3 个部分: 成长态测试集、成熟态测试集和休眠态测试集. 然后分别对这 3 个测试集上的每个用户 - 资源对 (u_q, r_q) 采用相应的标签推荐策略进行标签推荐，对比相应的阈值取不同值时标签推荐的效果，得到 $F@n$ 值较优时相应的阈值取值.

同时利用文献 [8] 中最流行标签算法，在 ρ 取不同值时，对测试集上每个用户 - 资源对 (u, r) 进行标签推荐，从而得到标签推荐的效果 $F@n$ 最优时的 ρ 值.

5.3.2.1　在成长态测试集上的实验

对于成长态测试集上每个用户 - 资源对 (u_q, r_q)，采用4.1节介绍的策略TR-GU进行标签推荐.

实验中资源 r_q 的近邻个数 S_1 值设置为 5,10,15,25,35,45,55,100,150,β 值设置为 0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0. 表 3 给出 S_1 与 β 的部分取值情况下的标签推荐效果.

表 3 成长态测试集上 TR-GU 的实验结果

Table 3 Experimental results of TR-GU on growing status subset

	$S_1=25$ $\beta=0.5$	$S_1=45$ $\beta=0.6$	$S_1=25$ $\beta=0.7$	$S_1=25$ $\beta=0.8$	$S_1=25$ $\beta=0.9$
$F@1$	0.2921	**0.2956**	0.2921	0.2921	0.2921
$F@2$	0.4202	**0.3898**	0.4202	0.4202	0.4202
$F@3$	0.4198	**0.4277**	0.4198	0.4198	0.4198
$F@4$	0.4047	**0.4132**	0.4047	0.4047	0.4047
$F@5$	0.3675	**0.3811**	0.3550	0.3550	0.3550

由表 3 可知,资源 r_q 的近邻个数 S_1 和 β 的取值增大或减小都会影响标签推荐的效果. 当 $S_1=45$, $\beta=0.6$ 时 $F@n$ 值最优,标签推荐的整体效果最好.

5.3.2.2 在成熟态测试集上的实验

对于成熟态测试集的每个用户 - 资源对(u_q,r_q),采用 4.2 节介绍的方法 TR-MU 进行标签推荐. 实验中 γ 值设置为 0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0. 为清楚显示实验结果,表 4 给出 γ 分别取值 0.3,0.4,0.5,0.6,0.7 时的标签推荐效果.

表 4 成熟态测试集上 TP-MU 的实验结果

Table 4 Experimental results of TR-MU on mature status subset

	$\gamma=0.3$	$\gamma=0.4$	$\gamma=0.5$	$\gamma=0.6$	$\gamma=0.7$
$F@1$	0.3078	0.3894	0.4549	**0.4549**	0.4262
$F@2$	0.3994	0.4439	0.4238	**0.4367**	0.4164
$F@3$	0.3772	0.4181	0.3988	**0.3928**	0.3918
$F@4$	0.3704	0.3799	0.3819	**0.3708**	0.3714
$F@5$	0.3489	0.3445	0.3608	**0.3453**	0.3413

由表 4 可知,不同的 γ 值对标签推荐效果会有不同影响. 当 $\gamma=0.6$ 时 $F@n$ 值最优,标签推荐的整体效果最好.

5.3.2.3 在休眠态测试集上的实验

对于休眠态测试集的每个用户 - 资源对(u_q,r_q),采用 4.3 节介绍的方法 TR-DU 进行标签推荐. 实验中资源 r_q 的近邻个数 S_2 值设置为 5,10,15,25,35,45,55,100,150. 为清楚显示实验结果,表 5 给出 S_2 分别取值 5,10,15,25,35 时的标签推荐效果.

由表 5 可知,资源 r_q 的近邻个数 S_2 值的增大或缩小都会影响标签推荐的效果. 当 $S_2=10$ 时 $F@n$ 值最优,标签推荐的整体效果最好.

表 5 休眠态测试集上 TR-DU 的实验结果

Table 5 Experimental results of TR-DU on dormant status subset

	$S_2=5$	$S_2=10$	$S_2=15$	$S_2=25$	$S_2=35$
$F@1$	0.3311	**0.3311**	0.3311	0.3311	0.3311
$F@2$	0.3543	**0.3415**	0.3415	0.3300	0.3300
$F@3$	0.3201	**0.3401**	0.3316	0.3316	0.3222
$F@4$	0.3151	**0.3151**	0.3151	0.3134	0.3134
$F@5$	0.2954	**0.2954**	0.2937	0.2937	0.2867

5.3.2.4 在测试集上确定 ρ 的实验

对于测试集的每个用户 - 资源对(u_q,r_q),实验采用最常用的标签算法进行标签推荐. 根据文献[4],ρ 设置为 0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0. 为清楚显示实验结果,表 6 给出 ρ 分别取值 0.6,0.7,0.8,0.9,1.0 时的标签推荐效果.

表 6 最流行标签算法中 ρ 取不同值时的实验结果

Table 6 Experimental results of the most popular tags algorithm with different ρ

	$\rho=0.6$	$\rho=0.7$	$\rho=0.8$	$\rho=0.9$	$\rho=1.0$
$F@1$	0.2984	0.3194	0.3242	**0.3346**	0.3253
$F@2$	0.3478	0.3476	0.3540	**0.3540**	0.3294
$F@3$	0.3363	0.3460	0.3526	**0.3558**	0.3271
$F@4$	0.3364	0.3476	0.3476	**0.3476**	0.3104
$F@5$	0.3276	0.3301	0.3301	**0.3325**	0.2873

由表 6 可知,ρ 取不同值对标签推荐的效果会有不同的影响. 当 $\rho=0.9$ 时 $F@n$ 值最优,标签推荐的整体效果最好.

根据 $F@1\sim F@5$ 的值,上述实验确定在 CiteULike 数据集上对用户标注状态分别处于成长态、成熟态和休眠态的用户进行标签推荐取得较好效果时的对应阈值. 对于用户标注状态处于成长态的用户,实验选择 $S_1=45$,$\beta=0.6$. 对于用户标注状态处于成熟态的用户,实验选择 $\gamma=0.6$. 对于用户标注状态处于休眠态的用户,实验选择 $S_2=10$. 对于最流行标签算法,实验选择 $\rho=0.9$.

同样对于 Last.fm 数据集,实验确定对用户标注状态分别处于成长态、成熟态和休眠态的用户进行标签推荐取得较好效果时的对应阈值. 对于用户标注状态处于成长态的用户,实验选择 $S_1=5$,$\beta=0.9$. 对于用户标注状态处于成熟态的用户,实验选择 $\gamma=0.5$. 对于用户标注状态处于休眠态的用户,实验选择 $S_2=5$. 对于最流行标签算法,实验选择 $\rho=1.0$.

对于 FolkRank,参数设置同文献[5],即阻尼系数 $d=0.7$,迭代次数为 10,对偏好向量的对应项权

重设置为 $1+|U|$ 与 $1+|R|$.

5.3.3　用户标注状态定义的有效性验证

为相对客观地进行实验对比,下面的实验中,本文的 TR-CUTS 和最流行标签算法中的参数取值都采用 5.3.2 节得到的较好情况下的阈值.

为验证本文提出的用户标注状态定义是有效的,在 CiteULike 数据集中,对比 TR-CUTS 与 TR-GU、TR-MU 和 TR-DU 在测试集上的标签推荐效果,实验结果见表 7.

表 7　TR-CUTS、TR-GU、TR-MU 和 TR-DU 的标签推荐结果

Table 7　Experimental results of tag recommendation by TR-UTS, TR-GU, TR-MU and TR-DU

	TR-CUTS	TR-GU	TR-MU	TR-DU
F@1	**0.3972**	0.3907	0.3938	0.3074
F@2	**0.4064**	0.4000	0.4004	0.3134
F@3	**0.3854**	0.3740	0.3712	0.3055
F@4	**0.3634**	0.3567	0.3474	0.2827
F@5	**0.3370**	0.3352	0.3287	0.2681

由表 7 可得,TR-CUTS 的标签推荐效果 $F@n$ 值略高于 TR-GU 和 TR-MU,明显高于 TR-DU,其 $F@n$ 的平均值比 TR-GU、TR-MU 和 TR-DU 分别提高 1.71%、2.58%、21.78%. 这说明本文提出的标签推荐策略,即判断用户标注状态之后根据用户所处用户标注状态的不同选择不同的标签推荐策略,其整体标签推荐效果好于单个的标签推荐策略. 同时 TR-GU 的标签推荐效果略好于 TR-MU,而明显好于 TR-DU,这说明标签推荐过程中考虑同组用户的标签和相似资源的标签是有效的.

5.3.4　对比实验

为观察 TR-CUTS 的推荐效果,本文在 CiteULike 数据集和 Last. fm 数据集上进行对比实验,对比算法包括 FolkRank、LocalRank 和最流行标签算法.

在 CiteULike 数据集上的对比实验结果见表 8,其中最流行标签算法简称为 Popular 0.9-mix. 由表 8 的数据可知,除在 $P@5$ 和 $R@5$ 上,TR-CUTS 略差于 LocalRank; 其他各种情况下,TR-CUTS 均优于 FolkRank、LocalRank 和最流行标签算法. 本文提出 TR-CUTS 的 $F@n$ 的平均值分别比 FolkRank、Popular 0.9-mix 和 LocalRank 提高 18.25%、7.14%,1.45%. 特别观察推荐效果 $F@1$ 的值,TR-CUTS 的标签推荐效果明显高于 FolkRank 和 Popular 0.9-mix,分别提高 46.56%、20.29%.

在 Last. fm 数据集上的对比实验结果见表 9. 由表 9 的数据可知,TR-CUTS 的召回率 $R@n$ 均高于 FolkRank、LocalRank 和最流行标签算法. 对于精确率 $P@n$,当 $n=1,2$ 时,TR-CUTS 的 $P@n$ 略小于对比算法; 而当 n 取其他值时,TR-CUTS 优于对比算法. TR-CUTS 的 $F@n$ 的平均值分别比 FolkRank、Popular 1.0-mix 和 LocalRank 提高 1.46%、2.04% 和 5.13%. 特别观察标签推荐效果 $F@3$ 的值,TR-CUTS 比 FolkRank、Popular 1.0-mix 和 LocalRank 分别提高 3.93%、5.38% 和 6.47%.

上述对比实验结果说明 TR-CUTS 的推荐效果优于 FolkRank、Popular 1.0-mix 和 LocalRank.

表 8　CiteULike 数据集上各算法的实验结果

Table 8　Experimental results of each algorithm on CiteULike dataset

	TR-CUTS	FolkRank	Popular 0.9-mix	LocalRank
P@1	**0.6095**	0.4486	0.5047	0.5794
P@2	**0.4190**	0.3738	0.3692	0.4112
P@3	**0.3365**	0.3115	0.3084	0.3271
P@4	**0.2857**	0.2593	0.2710	0.2827
P@5	**0.2462**	0.2243	0.2411	0.2486
R@1	**0.2945**	0.1941	0.2503	0.2864
R@2	**0.3945**	0.3293	0.3400	0.3839
R@3	**0.4509**	0.4086	0.4205	0.4449
R@4	**0.4993**	0.4557	0.4847	0.4979
R@5	**0.5342**	0.4880	0.5353	0.5524
F@1	**0.3972**	0.2710	0.3346	0.3834
F@2	**0.4064**	0.3502	0.3540	0.3971
F@3	**0.3854**	0.3535	0.3558	0.3770
F@4	**0.3634**	0.3306	0.3476	0.3606
F@5	**0.3370**	0.3073	0.3325	0.3429

表 9　Last. fm 数据集上各算法的实验结果

Table 9　Experimental results of each algorithm on Last. fm dataset

	TR-CUTS	FolkRank	Popular 1.0-mix	LocalRank
P@1	**0.3990**	0.4166	0.4068	0.4088
P@2	**0.3400**	0.3532	0.3415	0.3449
P@3	**0.3141**	0.3034	0.2992	0.3067
P@4	**0.2783**	0.2722	0.2617	0.2749
P@5	**0.2501**	0.2472	0.2326	0.2500
R@1	**0.2010**	0.1959	0.1944	0.1979
R@2	**0.3184**	0.3104	0.3026	0.3106
R@3	**0.4168**	0.3866	0.3844	0.3936
R@4	**0.4759**	0.4545	0.4403	0.4618
R@5	**0.5241**	0.5077	0.4800	0.5158
F@1	**0.2674**	0.2666	0.2665	0.2630
F@2	**0.3288**	0.3268	0.3304	0.3208
F@3	**0.3583**	0.3447	0.3400	0.3365
F@4	**0.3512**	0.3446	0.3405	0.3283
F@5	**0.3386**	0.3367	0.3325	0.3133

6 结 束 语

根据在 CiteULike 和 Last. fm 等数据集上的统计分析,本文发现在社会标签系统中用户的标签总数随其在系统中所处的时间有如下规律: 在一段时间内,用户的标签总数快速增加,或用户的标签总数缓慢增加,或用户标签总数不变. 因此本文定义 3 种用户标注状态: 成长态、成熟态和休眠态,并给出其形式化的定义及计算方法. 然后根据不同用户标注状态的不同特点,分别采取不同标签概率统计方法作为其标签推荐策略,为用户推荐最可能使用的标签. 最后对比本文方法与 FolkRank、LocalRank 和最流行标签算法,实验结果表明本文的标签推荐策略在推荐准确性上更优,同时也说明本文的用户标注状态的概念是有效的.

本文在判断用户当前所处标注状态时,是通过分析用户标注时刻的前一段时间内用户的标注行为确定的. 考虑到用户兴趣特点、用户行为模式等个性化特征,建立自适应不同用户的时间段将是下一步的研究工作重点.

参 考 文 献

[1] Hamouda S, Wanas N M. PUT-Tag: Personalized User-Centric Tag Recommendation for Social Bookmarking Systems. Social Network Analysis and Mining, 2011, 1(4): 377 - 385

[2] Lu C M, Hu X H, Park J R, *et al*. Post-Based Collaborative Filtering for Personalized Tag Recommendation // Proc of the iConference. Seattle, USA, 2011: 561 - 568

[3] Liu K P, Fang B X, Zhang W Z. Exploring Social Relations for Personalized Tag Recommendation in Social Tagging Systems. IEICE Trans on Information and Systems, 2011, 94 - D(3): 542 - 551

[4] Jäschke R, Marinho L, Hotho A, *et al*. Tag Recommendations in Social Bookmarking Systems. AI Communications, 2008, 21(4): 231 - 247

[5] Kim H N, El Saddik A. Personalized PageRank Vectors for Tag Recommendations: Inside FolkRank // Proc of the 5th ACM Conference on Recommender Systems. Chicago, USA, 2011: 45 - 52

[6] Kubatz M, Gedikli F, Jannach D. LocalRank - Neighborhood-Based, Fast Computation of Tag Recommendations // Proc of the 12th International Conference on E-commerce and Web Technologies. Toulouse, France, 2011: 258 - 269

[7] Ramezani M. Improving Graph-Based Approaches for Personalized Tag Recommendation. Journal of Emerging Technologies in Web Intelligence, 2011, 3(2): 168 - 176

[8] Krestel R, Fankhauser P. Personalized Topic-Based Tag Recommendation. Neurocomputing, 2012, 76(1): 61 - 70

[9] Zhang B, Zhang Y, Gao K N, *et al*. Combining Relation and Content Analysis for Social Tagging Recommendation. Journal of Software, 2012, 23(3): 476 - 488 (in Chinese)
(张 斌,张 引,高克宁,等. 融合关系与内容分析的社会标签推荐. 软件学报, 2012, 23(3): 476 - 488)

[10] Feng W, Wang J Y. Incorporating Heterogeneous Information for Personalized Tag Recommendation in Social Tagging Systems // Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 1276 - 1284

[11] Nieminen I T. Combining Tag Recommendations Based on User History [EB/OL]. [2013 - 04 - 20]. http: // ceur-ws. org/vol_497/paper_08. pdf

[12] Zhang Y, Zhang N, Tang J. A Collaborative Filtering Tag Recommendation System Based on Graph // Proc of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Bled, Slovenia, 2009: 297 - 306

[13] Gemmell J, Schimoler T, Ramezani M, *et al*. Improving FolkRank with Item-Based Collaborative Filtering // Proc of the 3rd ACM Conference on Recommender System. New York, USA, 2009: 17 - 24

[14] Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval // Proc of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 275 - 281

[15] Jäschke R, Marinho L, Hotho A, *et al*. Tag Recommendations in Folksonomies // Proc of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases. Warsaw, Poland, 2007: 506 - 514

[16] Batagelj V, Zaveršnik M. Fast Algorithms for Determining (Generalized) Core Groups in Social Networks. Advances in Data Analysis and Classification, 2011, 5(2): 129 - 145