基于标签的个性化推荐技术研究

王金燕，刘亚军

东南大学 计算机科学与工程学院，江苏 南京 211189

摘 要：传统的标签推荐技术强调标签推荐的精确性和规范性，忽略了标签的个性化特征。针对此问题，文章认为标签推荐应符合个性化特征，提出个性化的判断标准 coverage 和 ranking，并对传统 *ρ-mixed* 算法进行改进，提出了 *ρ*-Co-*mixed* 算法。该算法在已知资源最热门标签集合的基础上，结合考虑用户个人标签库中标签间的共现概率调整标签分值，最终产生推荐集。实验结果证明，*ρ*-Co-*mixed* 算法在 coverage 和 ranking 上优于 *ρ-mixed* 算法，即可推荐出更具个性化的标签。

关键词：标签推荐；个性化；共现概率

Tag-based Personalized Recommendation Research

Jinyan Wang, Yajun Liu

Computer Science and Engineering Department, Southeast University, Nanjing, China, 211189

Email: erinwjy@gmail.com

Abstract: Conventional tag recommendation mainly focuses on tags' precision and normalization, but ignores its personalization. In order to solve this problem, it is proposed in this paper coverage and ranking as the new standards of personalized tags, and gives an improved algorithm based on algorithm. The new algorithm takes into account the tags' co-occurrence in user's tag library after getting the most popular tags of a resource, modifies the tags' scores and generates tag recommendations according to the final scores. The experiment results prove that the new algorithm performs are better in both coverage and ranking and can get more personalized tag recommendation.

Key words: Tag Recommendation; Personalized; Co-occurrence

引 言

随着 Web2.0 网站的发展，互联网上信息的爆炸式增长使个性化的信息服务显得越来越重要。Web2.0 作为一个时代的产物，其成功的背后有一个核心原则：即借助网络的力量，利用从用户的行为中得到集体智慧。社会性标签正是这种集体智慧的一个产物。标签作为一种特殊的元数据，来源于标注者对资源主观感受的概括，被用户用于描述和分类资源。当多个用户对多个资源添加标签后，标签就具有了社会性，成为社会化标签，称为 folksonomy；当一个用户对多个对象添加标签后，就形成了其个人标签库，称为 personomy。对资源进行标注是系统加于用户的一个操作，为了更好地帮助用户标注资源，便于个人自愿的浏览、回顾、组织、管理、检索资源，诸多社会标注系统在用户标注资源时提供了推荐机制。

现有标签推荐的研究成果包括： Bruno Oliveira[1]、Marek Lipczak[2]、Pierpaolo Basile[3]等研究的基于资源内容的标签推荐； Leandro Balby Marinho[4]、 Xu Yanfei 和 Zhang Liang[5]、A.T.Ji[6]等研究的基于二元矩阵的协同推荐；Zhichen Xu[7]等提出的高质量的标签要具有的标准； Andreas Hotho[8]等受到 PageRank 算法启发提出的 FolkRank 算法以及 Symeonidis 等 [9]研究的张量降维技术。

上述基于标签的推荐工作大多集中在推荐更精确、更规范、更具概括性的标签，即过分关注标签作为 folksonomy 的方面，而忽略了标签作为 personomy 的方面，使得推荐的标签不具个性化，不符合用户的标注习惯。本文提出一种基于标签的个性化推荐算法，该算法同时考虑了资源的热门标签以及用户个人标签库中标签间的共现概率，通过实验论证，该算法可以提高标签推荐的个性化水平。

1 *ρmixed* 算法

Robert Jäschke[10]提出了使标签推荐集具个性化特征的 *ρ- mixed* 算法，该算法综合考虑了资源最热门标签集和用户最常使用标签集。标签推荐公式如下：

$$T'(u,r) = \arg\max_{t\in T}^{n}\left(\rho * norm_r\left(t\right) + \left(1-\rho\right) * norm_u\left(t\right)\right) \tag{1}$$

其中 $norm_r\left(t\right)$ 为资源 r 中的热门标签规格化后的分值，$norm_u\left(t\right)$ 为用户 u 标签库中的常用标签规格化后的分值。算法复杂度：

$$O((m+n)\log(m+n)) = O\left(m\log m\right) \tag{2}$$

其中 m 表示算法规定的资源最热门标签集合的大小，n 表示算法规定的个人标签库中最常使用的标签集合大小，通常情况下 m、n 为同一数量级。*ρ-mixed* 算法的推荐效果优于协同过滤算法，且推荐速度快于 FolkRank 算法，所以更具实用性。然而 *ρ-mixed* 算法虽然考虑到用户的个性化标签库，但只是简单将资源热门标签与用户热门标签通过一定比例相加，推荐结果中可能会出现与资源内容不符的标签却推荐分值过高的情况。

2 个性化标签判断标准

用户通过一定时间的标注行为，会形成自己的标签库，每个标签对应曾标注过的资源。如果一味强调推荐标签的精确性、规范性、概括性，忽视用户的个性化标注行为，那么即使是相似资源，用户也可能用不同的标签标注，因此不利于用户对其曾标注过的资源分类、管理和回顾。针对当前标签推荐中被忽视的问题，本文提出个性化标签两个判断标准：覆盖率和排名率。

2.1 覆盖率

覆盖率是指推荐集中的标签在用户最终采用的标签集中所占的比例。推荐集中的标签在保证准确性的同时，应尽可能多地覆盖用户曾标注过的资源，便于用户回顾管理。如果标签覆盖的历史资源多，说明该标签是用户经常使用，符合用户标注习惯的。高覆盖率是指标签被用户使用的可能性更高，即推荐集中含有更多的用户最终采用的标签。用户 u 对资源 o 的一次标注行为，其覆盖率用公式（3）计算。

$$Coverage\left(u;o\right) = \frac{\left|T_{u;o} \cap T'_{u;o}\right|}{\left|T_{u;o}\right|} \tag{3}$$

平均覆盖率表示用户集 U 在资源集 R 上的覆盖率的平均情况，平均覆盖率的计算如公式（4）所示。

$$\overline{Coverage}\left(U;R\right) = \frac{1}{\left|U\right|}\sum_{u\in U}\left(\frac{1}{\left|O_{u;R}\right|}\sum_{o\in O_{u;R}} Coverage\left(u;o\right)\right) \tag{4}$$

公式（3）和公式（4）中 $T_{u;o}$ 表示用户 u 对资源 o 实际标注的标签集合，$T'_{u;o}$ 表示系统向用户 u 推荐的资源 o 的标签集合，$O_{u;R}$ 表示 R 中用户 u 标注过的资源集合。

2.2 排名率

排名率表示推荐集中的标签在用户最终采用的标签集中的排名情况。分析 delicious 网站提供的 2004-12-31~2005-12-31 的用户数据，可知共有 532894 位用户对 17262097 个资源进行了 47256954 次标注行为，平均每个用户对每个资源添加 3 个标签。所以在较大的标签集中，应将更符合用户标注习惯的标签排名靠前，则用户在浏览少量的标签后就可获得所需的标签。

令 $position(t,T'_{u;o})$ 表示标签 t 在 $T'_{u;o}$ 中的位置，令 $p_{u;o;t} = \frac{1}{position(t,T'_{u;o})}$，若 $t\in T'_{u;o}$，则 $p_{u;o;t}$ 越大，标签在推荐集中的位置越靠前，其推荐效果越好。考虑如下表的两种情况的标签推荐集。假设用户实际使用的标签不分主次，

同等重要，其中 1 表示用户实际使用的标签，0 表示用户未使用的标签。

表 1 两种推荐情况

实际使用	1	1	1			
推荐集 1	1	0	0	0	0	1
推荐集 2	0	1	1	0	0	0

推荐集 1 中 $\overline{p_{u;o}}=\frac{3.5}{6}$，推荐集 2 中 $\overline{p_{u;o}}=\frac{2.5}{6}$，根据计算结果推荐集 1 优于推荐集 2。然而推荐集 1 中的一个标签排名第一，另一个标签排名靠后，但推荐集 2 中的两个标签在推荐集中排名都相当靠前，所以推荐集 2 的整体推荐效果要优于推荐集 1。为了降低标签分布跨度过大对最终结果带来的影响，须要修正 $p_{u;o;t}$ 的计算公式。由于平均每个用户对每个资源使用 3 个标签，本文以 3 个标签为一组，且认为同组内的标签排名相同。令修正后 $p_{u;o;t}{}'=\frac{1}{\lceil position(t,T'_{u;o})/3\rceil}$，则对用户 u 对资源 o 的一次标注行为，定义排名率如公式（5）所示：

$$Ranking(u;o)=\frac{1}{|T_{u;o}|}\sum_{t\in T_{u;o}}1/\left(p_{u;o;t}{}'\right) \tag{5}$$

平均排名率表示用户集 U 在资源集 R 的排名率的平均情况，定义平均排名率如公式（6）所示：

$$\overline{Ranking}(U;R)=\frac{1}{|U|}\sum_{u\in U}\left(\frac{1}{|O_{u;R}|}\sum_{o\in O_{u;R}}Ranking(u;o)\right) \tag{6}$$

3 *ρ- mixed* 改进算法

根据上文提出的个性化标签判断标准，本文对 *ρ- mixed* 算法进行改进，提出新的推荐算法，即 *ρ*-Co-*mixed* 算法。

3.1 *ρ*-Co-*mixed* 算法的基本思想

设 $p(t;o)$ 表示在资源 o 中标签 t 被推荐的概率；$p(t\mid k;u)$ 表示对于用户 u，若已使用标签 t，同时使用标签 k 的概率；ρ 表示用户标签库中的标签在最终的分值中所占的比例；$S(t;u;o)$ 表示向用户 u 对资源 o 推荐标签 t 的合适性，其值在计算过程中会被不断调整；$T_{k;u}$ 表示在用户 u 的标签库中和标签 k 共同出现概率最高的标签集合；R_1 表示资源最热门标签，R_2 表示待推荐集，$\forall t\in R_1$，加入 R_2，并赋初值 $S(t;u;o)=p(t;o)$；

$\forall k\in R_1$，$\forall t\in T_{k;u}$，如果 $t\in R_2$，则依据公式（7）调整分值：

$$S(t;u;o)=S(t;u;o)+p(k;o)*p(t\mid k;u)*\rho \tag{7}$$

否则将标签 t 加入 R_2 中，并根据公式（8）赋予初值：

$$S(t;u;o)=p(k;o)*p(t\mid k;u)*\rho \tag{8}$$

3.2 *ρ-Co-mixed* 算法描述

算法名称：*ρ*-Co-*mixed*

算法输入：当前用户 u，当前标注的资源 o

算法输出：标签推荐集 R

步骤：

a）首先统计资源 o 上已有的最热门标签，获得待推荐集 R_1，统计 R_1 中各标签的出现次数及所有标签的出现次

数总和，求出各标签的推荐概率，新建标签集 R_2 并初始化；

b）依次分析推荐集 R_1 中的每个标签，从用户 u 的个人标签库中找出对应的共现概率最高的标签集合，并根据公式（7）（8）调整待推荐集 R_2；

c）对 R_2 中的所有标签根据分值从高到低排序，推荐 Top-K 的标签得到最终推荐集 R；

d）算法结束。

ρ-Co-*mixed* 算法的时间复杂度如公式（9）所示：

$$O\left(mc+\left(mc\right)\log(mc)\right)=O\left(\left(mc\right)\log(mc)\right) \tag{9}$$

其中 m 表示算法规定的资源最热门标签集合的大小，c 表示 R_1 中的每个标签在用户个人标签库中找寻的共现概率最高的标签个数。

与传统 ρ-*mixed* 算法相比，ρ-Co-*mixed* 算法不再是将资源 o 上的最热门标签集合 R_1 以及用户 U 最常使用的标签集合这两部分通过系数 ρ 简单相加，而是在 R_1 的基础上，逐一考虑 R_1 中的标签在用户个人标签库中与其他标签间的共现概率。因此可以在有限的推荐集中包含更多符合资源内容和用户标注习惯的标签，并提高相应标签的排名，在保证准确率的同时提高标签推荐集的个性化水平。

4 实验结果与分析

实验采用的数据源为 delicious 网站提供的 2004-12-31~2005-12-31 的用户数据。因为本文是为了提供更符合用户标注行为的个性化标签，所以需要用户标注过一定数量的资源，标签库相对稳定，实验中选取已标注资源总数前 500 的用户，并随机选取 id 号小于 50000 的资源 1000 个，作为分析的内容。本实验中取每个标签的共现标签集大小为 5，推荐集大小为 10。实验环境如表 1 所示。

表 2 实验环境参数

操作系统	Windows Server 2003
数据库系统	SQLServer2008
硬件配置	Intel(R) Xeon(R) CPU E7330 2.4GHZ+4GMemory
开发平台	MyEclipse 6.0
编程语言	Java

4.1 实验结果分析

取 id 为 88 的用户对 id 为 560 的资源的一条标注行为作分析。数据如表 3 所示：

表 3 不同 ρ 对结果的影响分析

实际标签	ρ = 0	ρ = 0.7	ρ = 0.9	ρ = 1.7
tools	widgets	**tools**	**tools**	**tools**
computers	yahoo	widgets	software	software
utilities	software	software	widgets	widgets
yahoo	**tools**	yahoo	yahoo	yahoo
	widget	windows	windows	**utilities**
	windows	widget	**utilities**	windows

分析上述数据： 当 ρ =0.7 时，可发现 tools 已经从原来的排行 4 的升到排行 1，这与用户的实际标注情况吻合；当 ρ = 0.9 时，推荐集合中第一次出现了 utilities 标签，用户实际标注中也确实使用了此标签；当 ρ = 1.7 时，utilities

标签的排名又升高一位，从原先的 6 升至为 5。

分析用户的标注历史，其中使用过 46 次 yahoo，同时使用 yahoo 和 tools 有 10 次，使用过 software 311 次，同时使用 software 和 tools 有 146 次，所以 $p(tools \mid yahoo;u) = 0.217$， $p(tools \mid software;u) = 0.441$，从而 tools 的分值提高，排名靠前。同理分析标签 utilities， $p(utilities \mid yahoo;u) = 0.043$， $p(utilities \mid software;u) = 0.284$，$p(utilities \mid tools;u) = 0.174$ 所以虽然标签 utilities 在最流行标签集中未出现，但由于用户的历史标注行为中 utilities 与其他标签的共现概率高，所以也得到了推荐的机会。

4.2 实验结果比较

图 1 显示了使用 ρ-Co-*mixed* 算法推荐标签的效果。依据公式（4）计算当 ρ 取不同值时平均覆盖率的情况，如图 a 所示；依据公式（6）计算当 ρ 取不同值时平均排行率的情况，如图 b 所示。

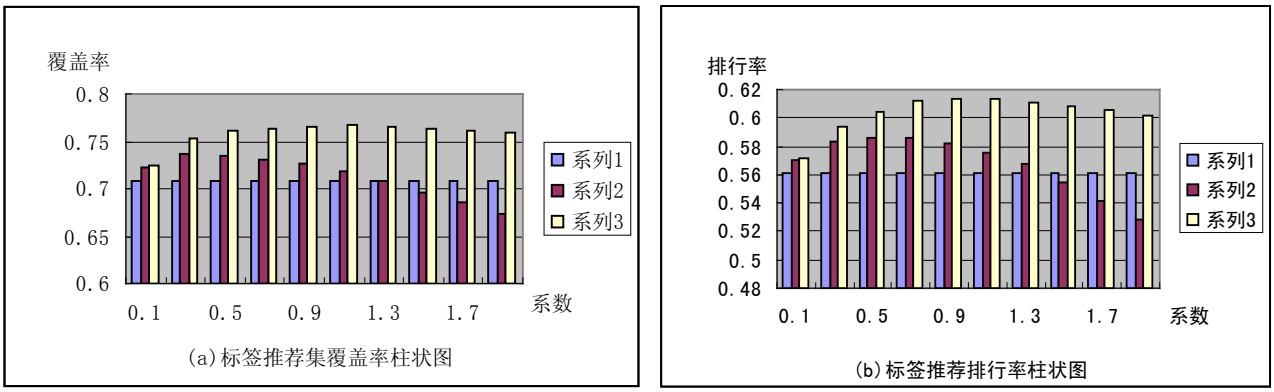

图 1 标签推荐效果图

其中系列 1 表示推荐最热门标签，系列 2 表示 ρ-*mixed* 算法，系列 3 表示 $\rho - Co - mixed$ 算法。由图 1，图 2 可知，使用 ρ-Co-*mixed* 算法后，标签推荐集的覆盖率和排名率都优于 ρ-*mixed* 算法，且在 ρ = 1.1 时平均覆盖率和平均排行率都达到最高值。可见在资源热门标签的基础上，同时考虑用户的个人标签库确实能在有限的推荐集中推荐出更多符合用户标注习惯的标签，同时能将更符合用户标注习惯的标签排名靠前。

4.3 推荐速度

公式（2）和公式（9）分别表示 ρ-*mixed* 算法和 ρ-Co-*mixed* 算法的算法复杂度，通常取 c 为较小的整数，本实验中 c=5，因此可认为两者算法复杂度相同。经实验得 $\rho - mixed$ 算法的平均耗时 64.522ms，ρ-Co-*mixed* 算法平均耗时 160.184ms。虽然 ρ-Co-*mixed* 算法比 ρ-*mixed* 算法耗时更多，但仍与其处于同一数量级上，用户无法分辨出此种数量级上的差异，所以仍有较高的实用性。

5 结语

本文通过分析当前标签推荐存在的问题，提出了基于标签的个性化推荐方法和个性化推荐的判断标准，并在资源最热门标签的基础上，利用用户标签库中的标签共现性，对传统 ρ-*mixed* 算法进行改进，提出了 ρ-Co-*mixed* 算法。实验证明，改进后的算法在覆盖率和排名率上表现更优，即标签推荐的个性化特点更突出，对于本文所选定的数据集和实验环境来说，当系数 ρ = 1.1 时推荐效果达到最佳，推荐时间为毫秒数量级。

由于标签本身具有随意性、冗余性、歧义性等缺点，对于新资源和新用户的推荐可能会出现冷启动的情况，所以如何推荐出真正高质量的标签仍有待进一步研究。

参考文献

[1] Bruno Oliveira, Pavel Calado, and H. Sofia Pinto. Automatic Tag Suggestion Based on Resource Contents. Lecture Notes in Computer

Science, 2008, Volume 5268/2008:255-264.

[2] Marek Lipczak. Tag Recommendation for Folksonomies Oriented towards Individual Users. In Proceedings of ECML PKDD Discovery Challenge (RSDC08). 84-95.

[3] Pierpaolo Basile, Domenico Gendarmi, Filippo Lanubile etc. Recommending Smart Tags in a Social Bookmarking System. In: Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007): 22–29.

[4] Leandro Balby Marinho, Lars Schmidt–Thieme. Collaborative Tag Recommendations. Studies in Classification, Data Analysis, and Knowledge Organization, 2008, Data Analysis, Machine Learning and Applications, VIII: 533-540.

[5] Yanfei Xu, Liang Zhang. Personalized Information Service Based on Social Bookmarking.Lecture Notes in Computer Science, 2005, Volume 3815, Digital Libraries: Implementing Strategies and Sharing Experiences: 475-476.

[6] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, Robin Burke. Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering [2009-5-1]. Proceedings of the 2008 ACM conference on Recommender systems: 259-266.

[7] Z. Xu, Y. Fu, J. Mao, and D. Su. Torwards the semantic web: Collaborative tag suggestions. In Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference, (WWW 2006), 22. Edinburgh, Scotland (2006).

[8] Andreas Hotho, Robert Jaschke, Christoph Schmitz, Gerd Stumme FolkRank: A Ranking Algorithm for Folksonomies. In Proceedings of LWA. 2006:111-114.

[9] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. Tag recommendations based on tensor dimensionality reduction. In RecSys ’08: Proceedings of the 2008 ACM conference on Recommender systems: 43-50.

[10] Robert Jäschke Leandro Marinho Andreas Hotho etc. Tag recommendation in social bookmarking systems. AI Communications, Volume 21 Issue 4, December 2008, Pages 231-247.

【作者简介】

王金燕（1987-），女，汉，硕士研究生，研究方向：数据库应用及技术，
2005-2009 南京邮电大学计算机科学与技术，2009~至今 东南大学计算机应用方向；
Email：erinwjy@gmail.com。

刘亚军 （1953-），女，汉，教授，研究方向：数据库应用及技术；
Email：yjliu@seu.edu.cn。