基于标签和 PageRank 的重要微博用户推荐算法

王嵘冰 安维凯 冯 勇 徐红艳
(辽宁大学信息学院 沈阳 110036)

摘 要 海量的微博信息使新进用户很难获取到其感兴趣的内容,重要微博用户推荐为新用户提供了一条有效获取信息的途径。目前,由于用户间的关系没有被充分考虑及缺乏对用户个性化标签的处理,导致重要微博用户推荐的准确率不高。为此,提出了一种基于标签和 PageRank 的重要微博用户推荐算法。该算法首先对个性化标签进行分词、去噪、设置权重等处理,并将其作为用户兴趣的代表;然后根据 PageRank 计算模型来分析用户间的关系,结合标签相似度计算向新用户推荐与其兴趣相似的重要微博用户。实验表明,该算法由于融入了对微博用户关系和用户个性化标签的重要性分析,因此与基于标签和协同过滤的个性化推荐算法相比具有更高的重要微博用户推荐准确率。

关键词 个性化推荐,PageRank,标签,微博
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.047

Important Micro-blog User Recommendation Algorithm Based on Label and PageRank

WANG Rong-bing AN Wei-kai FENG Yong XU Hong-yan
(School of Information, Liaoning University, Shenyang 110036, China)

Abstract Massive micro-blog information makes it difficult for new users to obtain the content they are interested in. Important micro-blog user recommendation provides an effective way for new users to access information. At present, inadequate consideration of the relationship between users and the lack of user personalized label processing make the recommendation accuracy of important micro-blog user be not high. Therefore, an important micro-blog user recommendation algorithm based on label and PageRank was proposed. Firstly, the personalized label is processed by word segmentation, de-noising and setting weight, and the processed result is used as the representative of user interest. Secondly, the relationship between users is analyzed by PageRank calculation model. Finally, important micro-blog users are recommended to new users with similar interests by label similarity calculation. The experiment shows that the proposed algorithm improves the recommendation accuracy of important micro-blog users compared with the recommendation algorithm based on label and collaborative filtering, because the analysis of the importance of micro-blog user relationship and user's personalized label is integrated into this algorithm.

Keywords Personalized recommendation, PageRank, Label, Micro-blog

1 引言

随着互联网技术进入 Web 2.0 时代,用户更愿借助微博、播客等社交媒体展现自我,网络信息量不断增加,导致"信息迷航"问题愈发严重[1]。尤其以微博为主,用户可以根据自己的兴趣爱好来关注好友,但由于微博用户的剧增,添加哪些用户为自己的好友变成了一个困难的问题。面向 Web 2.0 应用现状,个性化信息推荐技术已经成为提高信息利用率和信息服务水平的一种有效工具[2]。

近年来,在个性化信息推荐领域,有学者将标签应用于推荐算法中以提高推荐质量[3]。其中,邢千里等人通过使用标签来预测用户间的关注关系,但是未考虑标签的稀疏性,导致推荐质量较低[4]。李瑞敏等人在个性化音乐推荐方面使用重启型随机游走算法和二部图节点结构相似性来分析用户、标签、项目两两之间的联系,但没有考虑到用户与用户之间的关系[5]。蔡强等人通过用户对资源的偏好相似度与资源标签的相似度来进行个性化推荐,但只考虑了单用户的资源推荐,没有考虑到用户推荐,即没有考虑用户之间的关系[6]。上述算法虽然在一定程度上提升了推荐的准确度,但由于忽视了用户间的关系、兴趣领域中权威用户的重要性和标签稀疏等因素,因此其个性化服务水平还有待提高。

微博是 Web 2.0 的代表技术之一,为人们提供了广泛的社交服务。在微博的某兴趣领域,权威用户拥有相对全面而准确的信息,对新用户来说具有重要意义。为了向新进微博

到稿日期:2017-05-01 返修日期:2017-08-01 本文受辽宁省博士科研启动基金(201601099),辽宁省档案科技项目(L-2016-8-7)资助。
王嵘冰(1979—),男,博士,讲师,主要研究方向为云计算、大数据、个性化推荐;安维凯(1991—),男,硕士生,主要研究方向为个性化推荐;冯 勇(1973—),男,博士,教授,CCF 会员,主要研究方向为数据挖掘,E-mail:fengyong@lnu.edu.cn(通信作者);徐红艳(1972—),女,硕士,副教授,主要研究方向为个性化推荐、Deep Web。

用户提供更好的个性化信息服务，本文提出了一种基于标签和 PageRank[7] 的重要微博用户推荐算法。该算法首先采用 PageRank 思想来确定用户之间的关系以及用户在关系网络中的重要性，再通过设置标签权重和计算用户标签间的相似度进行重要微博用户的推荐，最后通过实验与基于标签和协同过滤的推荐算法[6]进行了比较分析。本文所提算法融入了用户间关系的分析，并通过采用标签分词和设置权重解决了标签稀疏的问题，从而有效提高了重要微博用户推荐的准确率。

2　基于标签和 PageRank 的重要微博用户推荐算法的框架

本文利用 UCI 官网中的 microblogPCU 数据集[8]来获取用户关系信息和用户带有的标签信息。对于新加入的微博用户来说，其既希望关注具有相同偏好的用户，也希望关注该领域中的权威用户。为满足新用户的社交需求，本文提出了一种基于标签和 PageRank 的重要微博用户推荐算法，所提算法的主要步骤包括数据预处理、PageRank 重要用户的发现、用户标签相似度的计算和重要用户的推荐。其中 PageRank 重要用户的发现和用户标签相似度的计算是核心环节，算法流程如图 1 所示。

图 1　算法流程图

Fig. 1　Flowchart of the proposed algorithm

(1)数据预处理。分析用户关注的信息，构建以用户 ID 为节点、以用户之间的关系为边的有向网络——用户关系网。

(2)基于 PageRank 发现重要用户。基于数据预处理环节中构建的用户关系网，使用改进的 PageRank 算法来计算每位用户的重要度 PR_i。

(3)用户标签相似度的计算。分析数据集中的标签信息，对标签进行分词，使用 TF-IDF 对标签设置权重，计算新用户与用户关系网中每个用户的标签相似度 S_i。

(4)重要用户的推荐。根据步骤(2)和步骤(3)的计算结果，通过调节参数 α 来分配 PR_i 和 S_i 的权重。计算公式如式(1)所示：

$$Sim_i=\alpha * S_i+(1-\alpha) * PR_i \tag{1}$$

其中，i 表示用户 ID，Sim_i 为新用户与用户 i 的综合相似度。由于步骤(2)中得到的是各个兴趣领域的重要用户，其中有一些用户与新用户的兴趣领域不符，需要对这些用户结合标签相似度加以筛选，最终根据综合相似度 Sim_i 的大小进行 Top-N 推荐。

3　核心环节的描述

3.1　基于 PageRank 发现重要用户

本文以微博用户为节点，以用户之间的关注关系为有向边来构建用户关系网[9]，而重要用户的发现是指发现各微博领域中的关键、核心用户，即受关注度高的用户。本文提出的用户重要度传播的思想是：在微博领域中被多个用户关注的用户是重要的，而被重要用户关注的用户也是重要的。此想法与 PageRank 思想相似，因此引入 PageRank 算法来发现重要用户。

最初，Google 使用 PageRank 算法来计算页面相关性和重要性，通过页面的链入链出来计算页面等级[10]，而这种模型与微博用户关注模型很相似，因此本文引入了 PageRank 算法，并借此获取微博用户的重要度，向新用户推荐领域中重要度高的用户，具体公式如式(2)所示：

$$PR(u_i)=\frac{1-q}{N}+q\left(\frac{PR(u_1)}{C(u_1)}+\frac{PR(u_2)}{C(u_2)}+\cdots+\frac{PR(u_n)}{C(u_n)}\right) \tag{2}$$

其中，$PR(u_i)$是用户 u_i 的重要度；q 是介于(0,1)区间的阻尼系数，一般取值 0.85；$u_1,u_2,\cdots,u_n$ 为关注用户 u_i 的微博用户；$C(u_j)(1<j<n)$表示用户 u_j 所关注的用户数目。

具体计算流程描述如下。

输入：所有用户关注列表 D，阻尼系数 q，用户个数 n

输出：用户 u_i 的 PR 值

步骤：

1. for k=1 to n do
2. D_{list}←find_user_relation(D)；//扫描 D，得到用户的关注关系列表 D_{list}
3. M_{init}←init_matrix(D_{list})；//根据关注关系列表来初始化概率转移矩阵
4. end for
5. 根据用户个数来初始化单元矩阵 unit_Matrix
6. M←M_{init} * q+(1−q) * unit_Matrix/n；//根据概率转移矩阵和阻尼系数构建矩阵 M
7. M←M * X；//不停迭代计算，最终收敛，得到特征向量 X
8. 根据特征向量 X 来获取每个用户的 PR 值
9. PR(u_i)←Calculation(D_{List_ui})；//D_{List_ui}为关注用户 u_i 的微博用户列表，代入式(2)计算用户 u_i 的 PR 值

3.2　用户标签相似度的计算

用户之间最常用的相似度计算方法有余弦相似度计算和皮尔逊相似度计算，这两种相似度计算方法都是基于“用户-项目”评分矩阵来进行的[11]。本文由于是对微博用户的标签进行相似度计算，即对“用户-标签”矩阵进行计算[12]，因此采用余弦相似度来计算用户之间的标签相似度。

3.2.1　获取用户标签并设置标签权重

本文通过 UCI 官网中的 microblogPCU 数据集来获取用户关系网中用户的自定义标签，首先采用 Lucene 全文检索工具包[12]下的 IKAnalyzer 分词器来给标签分词并构建标签库 $\overrightarrow{T_{total}}=(t_1,t_2,\cdots,t_i)$，然后采用 TF-IDF 来计算标签的权重并构建标签库权重向量$\overrightarrow{W_{total}}=(w_1,w_2,\cdots,w_i)$。但是考虑到用户自定义标签分词后会有相似标签和共现标签，相似标签之间的影响会提高标签的权重，本文首先使用爬虫来爬取微博

中热门微博分类下的微博信息，并对微博信息分词采用 TF-IDF 来计算词权重，选取值靠前的多个标签作为各分类下的相似标签库。然后提出了一种相似标签权重设置方案。例如，对于相似标签(t_1，t_2，t_3)，其对应的相似标签共现次数为(n_1，n_2，n_3)，不考虑相似标签前的权重向量为(w_1，w_2，w_3)，考虑到相似标签的影响，对于 t_1 标签的权重的具体计算公式如式(3)所示：

$$w_1' = w_1 + \frac{n_2}{n_1+n_2+n_3} * w_2 + \frac{n_3}{n_1+n_2+n_3} * w_3 \tag{3}$$

计算权重后重新构建标签库权重向量 $\overrightarrow{W'_{total}} = (w_1', w_2', \cdots, w_i')$。如果不考虑相似标签的影响，那么当用户标签共现次数较少时权重也小；但是该用户标签存在多个相似标签，根据式(3)其会提高用户标签的权重，以保证标签权重的公平性。

3.2.2 标签相似度的计算

当新用户设置自己的标签 $\overrightarrow{T_{new}}$ 后，可通过 $\overrightarrow{W_{total}}$ 得到新用户的标签权重 $\overrightarrow{W_{new}}$，以及对应标签库的向量 $\overrightarrow{T_{new}} = (\langle t_1, w_1\rangle, \langle t_2, w_2\rangle, \cdots, \langle t_i, w_i\rangle)$。结合 3.1 节中 PageRank 计算获得的用户排序，依次计算新用户与排序好的用户的标签相似度[13]，相似度计算公式如式(4)所示：

$$sim(u_{new}, u_i) = \frac{\overrightarrow{T_{new}} \cdot \overrightarrow{T_{u_i}}}{\sqrt{|\overrightarrow{T_{new}}| \times |\overrightarrow{T_{u_i}}|}} \tag{4}$$

其中，$\overrightarrow{T_{u_i}}$ 表示用户 i 的标签向量。用户标签相似度的计算流程描述如下。

输入：用户标签列表 L，新用户标签列表 L_{new}

输出：各个用户与新用户的标签相似度 sim

方法：

```
1. for i=1 to n do
2.    L_T←get_tags(L);//获取每个用户标签，并创建标签库 L_T
3. end for
4. for k=1 to n do
5.    L_FC←Fenci(L_T);//对标签库 L_T进行分词处理
6.    for j=1 to m do
7.       If(L_FC.contains(L_Tj));//判断 L_FC中是否包含 L_T标签，如果
         包含则增加标签的重复次数，否则不做处理
8.          TF_IDF(L_FC);//通过 TF-IDF 来设置标签权重
9.    end for
10.      sim←get_sim(L_new,L_T);//根据式(4)计算新用户与其他用户
         的标签相似度
11. end for
```

4 实验结果与分析

本实验数据来源于 UCI 的 microblogPCU 数据集[1)]。该数据集中包括 59191 位用户，其中 262 位用户带有自定义标签，标签总数为 1411 个，标签分类如表 1 所列。数据集还包括用户之间的 142368 条关注关系，通过用户之间的关注关系来构建用户关系网。运用本文所提算法计算新用户与关系网络中用户的综合相似度，根据综合相似度的高低进行个性化推荐，并与传统的个性化推荐算法进行比较。其中 TCF[6] 为结合标签和协同过滤的推荐算法，该算法依据标签计算用户的偏好程度和资源特征相似度，结合基于资源的协同过滤推荐实现对资源的个性化推荐；DBCF[14] 为基于动态社会行为和用户背景的协同推荐算法，该算法通过动态社会化标签和用户背景信息分别得出用户动态兴趣和用户相似度，同时得到用户的最近邻居集并进行个性化推荐。

表 1 标签分类

Table 1 Classification of lable

序号	领域	用户数	标签总数
1	时尚娱乐	165	326
2	体育	102	204
3	科技	98	197
4	教育就业	186	418
5	养生	136	266

实验设定新用户的自定义标签为：演员、电影、综艺节目、数码控、智能手机、平板电脑；然后运用本文所提算法与 TCF 算法[6] 和 DBCF 算法[14] 分别产生推荐，实验结果如表 2 所列。

表 2 新用户推荐列表

Table 2 List of new user recommendation

用户 ID	用户标签	*PR* 值	标签相似度		
			TCF 算法	DBCF 算法	本文算法
3239617560	演员，手机	0.8204	0.3433	0.3527	0.5279
3867947139	电影，综艺节目	0.5319	0.3532	0.3659	0.4246
3914146316	美食，演员	0.6137	0.1695	0.1803	0.3472
3771310213	明星，数码，健身	0.5283	0.1628	0.1785	0.3091
3649943230	节目主持人，美女	0.3569	0.1048	0.1184	0.20564

由表 2 可知，用户 3239617560 和用户 3867947139 的 *PR* 值相差较大，但标签相似度接近，由于本算法结合了 PageRank思想，*PR* 值大的用户更可能被新用户所关注，因此所提算法经综合计算后向新用户推荐 *PR* 值大的用户。对于用户 3867947139 和用户 3771310213，其 *PR* 值接近，但标签相似度值相差较大，导致本文所提算法和 TCF 算法的推荐顺序是一致的，即都是推荐标签相似度大的用户。

由于所提算法中涉及到 α 参数的确定，因此本文实验将随机抽取 90%的数据作为训练集，而剩下的 10%的数据作为测试集，实验推荐的标准度采用准确率(Precision)、召回率(Recall)、F 度量值。α 参数值对推荐算法的影响如图 2 所示。

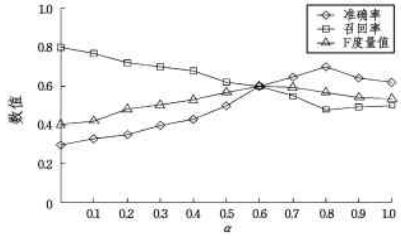

图 2 α 参数值对推荐算法的影响

Fig. 2 Effect of parameter α on the recommendation algorithm

当 α=0.6 时，F 度量值达到最大，算法的性能达到最佳

1) http://archive.ics.uci.edu/ml/datasets/microblogPCU

状态。在不同大小的数据集下，本文算法与 TCF 算法和 DBCF 算法的 F 度量值比较结果如图 3 所示。

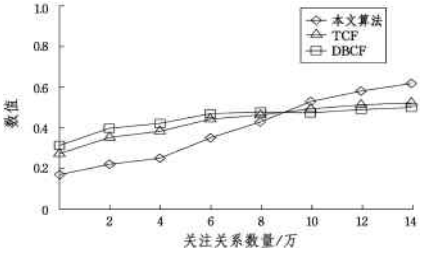

图 3　F 度量值对比

Fig. 3　Comparision of F-measure

由图 3 可以看出，当用户数量一致时，用户关注关系数量越多，算法的准确率越高；在用户关注关系数量小于 9 万条时，用户之间的关系比较稀疏，用户 *PR* 值普遍较小，结合标签相似度反而降低了推荐准确率；而在用户关注关系数量大于 9 万条时，用户 *PR* 值成为了体现用户重要度的指标，结合标签相似度可以提高推荐重要微博用户的准确率。本文算法在 F 度量值方面明显优于 TCF 算法和 DBCF 算法，这表明越复杂的关系网络越可以体现用户 *PR* 值的重要性，结合标签相似度可以更有效地提高推荐重要用户的质量。

结束语　本文通过应用 PageRank 思想来考量用户间的关系并计算用户重要度，对用户个性化标签的分词、去噪和 TF-IDF 设置权重，并结合标签余弦相似性计算，提出一种基于标签和 PageRank 的重要微博用户推荐算法。与 TCF 算法和 DBCF 算法的对比实验表明，所提算法能够有效地解决个性化标签的稀疏性问题；同时所提算法融入了对微博用户关系和用户个性化标签的重要性分析，较大程度地提升了推荐的质量。

参 考 文 献

[1] ZHANG R, JIN Z G, WANG Y. Recommendation Model of Microblog User Tags Based on Hybrid Grain[J]. Computer Science, 2016, 43(4): 192-196. (in Chinese)
张瑞，金志刚，王颖. 一种基于混合粒度的微博用户标签推荐模型[J]. 计算机科学，2016，43(4)：192-196.

[2] WEI S, ZHENG X, CHEN D, et al. A Hybrid Approach for Movie Recommendation via Tags and Ratings[J]. Electronic Commerce Research & Applications, 2016, 18(C): 83-94.

[3] YANG A T, TANG Y, WANG J B, et al. Personalized Friends Recommendation System Based on Game Theory in Social Network[J]. Computer Science, 2015, 42(9): 191-194. (in Chinese)
杨阿祧，汤庸，王江斌，等. 基于博弈的社会网络个性化好友推荐算法研究[J]. 计算机科学，2015，42(9)：191-194.

[4] XING Q L, LIU L, LIU Y Q, et al. Study on User Tags in Weibo[J]. Journal of Software, 2015, 26(7): 1626-1637. (in Chinese)
邢千里，刘列，刘奕群，等. 微博中用户标签的研究[J]. 软件学报，2015，26(7)：1626-1637.

[5] LI R M, LIN H F, YAN J. Mining Latent Semantic on User-Tag-Item for Personalized Music Recommendation[J]. Journal of Computer Research and Development, 2014, 51(10): 2270-2276. (in Chinese)
李瑞敏，林鸿飞，闫俊. 基于用户-标签-项目语义挖掘的个性化音乐推荐[J]. 计算机研究与发展，2014，51(10)：2270-2276.

[6] CAI Q, HAN D M, LI H S, et al. Personalized Resource Recommendation Based on Tags and Collaborative Filtering[J]. Computer Science, 2014, 41(1): 69-71. (in Chinese)
蔡强，韩东梅，李海生，等. 基于标签和协同过滤的个性化资源推荐[J]. 计算机科学，2014，41(1)：69-71.

[7] WANG X Y, REN G S. Improved PageRank Algorithm Based on User Behavior and PageAnalysis[J]. Computer Engineering, 2016, 42(2): 164-168. (in Chinese)
王旭阳，任国盛. 基于用户行为与页面分析的改进 PageRank 算法[J]. 计算机工程，2016，42(2)：164-168.

[8] REN X Y, SONG M N, SONG J D. Context-Aware Point-of-Interest Recommendation in Location-Based Social Networks[J]. Chinese Journal of Computers, 2017, 40(4): 824-841. (in Chinese)
任星怡，宋美娜，宋俊德. 基于位置社交网络的上下文感知的兴趣点推荐[J]. 计算机学报，2017，40(4)：824-841.

[9] LIANG T T, LI C Q, LI H S. Top-k Learning Resource Matching Recommendation Based on Content Filtering PageRank [J]. Computer Engineering, 2017, 43(2): 220-226. (in Chinese)
梁婷婷，李春青，李海生. 基于内容过滤 PageRank 的 Top-k 学习资源匹配推荐[J]. 计算机工程，2017，43(2)：220-226.

[10] OLVERA E P, GODOY D. Evaluating Term Weighting Schemes for Content-based Tag Recommendation in Social Tagging Systems[J]. IEEE Latin America Transaction, 2012, 10(4): 1973-1980.

[11] LIU J, ZHANG K, CHEN X. Personalized Recommendation Algorithm Based on Tags and Collaborative Filtering[J]. Computer & Modernization, 2016(2): 62-65. (in Chinese)
刘健，张琨，陈旋. 基于标签和协同过滤的个性化推荐算法[J]. 计算机与现代化，2016(2)：62-65.

[12] SONG Y, ZHANG L, GILES C L. Automatic Tag Recommendation Algorithms for Social Recommender Systems[J]. ACM Transactions on the Web, 2011, 5(1): 4.

[13] DU W H, RAN J W, HUANG J W, et al. Improving the Quality of Tags Using State Transition on Progressive Image Search and Recommendation System[C]// IEEE International Confe-rence on Systems, Man, and Cybernetics. IEEE, 2012: 3233-3238.

[14] JIANG S, WANG Z Q, XIU Y, et al. Collaborative Filtering Recommendation Method Based on Dynamic Social Behavior and Users' Background Information[J]. Computer Science, 2015, 42(3): 252-255. (in Chinese)
蒋胜，王忠群，修宇，等. 基于动态社会行为和用户背景的协同推荐方法[J]. 计算机科学，2015，42(3)：252-255.