基于 RBLDA 模型和交互关系的微博标签推荐算法①

余 勇, 郭躬德

(福建师范大学 数学与计算机科学学院, 福州 350007)
(网络安全与密码技术福建省重点实验室(福建师范大学), 福州 350007)

摘 要: 随着互联网技术的发展, 个性化标签推荐系统在海量信息或资源过滤中起着重要的角色. 在新浪微博平台中, 用户可以自主的给自己添加标签来表明自己的兴趣爱好. 同时, 用户也可以通过标签来搜索与自己兴趣爱好相似的用户. 针对新浪微博中大部分用户没有添加标签或添加标签数目较少的问题, 提出了一种基于 RBLDA 模型和交互关系的微博标签推荐算法, 它首先利用 RBLDA 模型来产生用户的初始标签列表, 然后再结合用户的交互关系而形成的交互图来预测用户标签的算法. 通过在新浪微博真实数据集上的实验发现, 该方案与传统的标签推荐算法相比, 取得了良好的实验效果.

关键词: 个性化标签; 标签推荐; 主题模型; 交互网络; 新浪微博

RBLDA Model and Interaction Relation Algorithm for User Tags Recommendation in Microblog

YU Yong, GUO Gong-De

(School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)
(Key Laboratory of Network Security and Cryptography in Fujian Province (Fujian Normal University), Fuzhou 35007, China)

Abstract: With the development of internet technology, the personalized tag recommendation system plays an important role in information or resources filtering. In Sina microblog website, an user can freely tag himself to indicate his interests. Meanwhile, users can also search other users who have the similar interests through tags. For the issue that there are no tags or few tags for the most users in Sina microblog website, an algorithm based on RBLDA model and users' interaction graph for tags recommendation is proposed in this paper. The algorithm utilizes the RBLDA model to produce the intial list of tags, and combines with users' interaction graph generated from actions of interaction between users to predict the final tags. The experimental results carried on some real data sets show that the proposed method performs better than traditional tag recommendation algorithms in comparison.

Key words: personalized tag; tag recommendation; topic model; interaction network; Sina microblog

随着 web2.0 技术的兴起与发展, 社交网络迅速发展和壮大, 它给互联网用户的生活带来了巨大的影响. 微博作为一种新兴的具有代表性的社交平台, 允许人们通过虚拟网络来获取海量的、实时的信息, 因而吸引着大量的用户群体. 然而, 互联网的飞速发展也带来了数据量的急剧暴增, 如何快速从海量数据中搜寻所需要的信息或资源, 已成为互联网用户所面临的一个难题. 其中, 个性化服务显得尤为突出. 推荐功能是个性化服务的一个重要途径, 很多学者都对其进行了广泛研究, 而且这一技术手段也已成功应用于各大商业网站. 众所周知, 人们对其关注的信息或资源进行打标签, 将会大大提高信息或资源的推荐效率与准确率[1].

标签一词, 根据维基百科的定义, 是一种无层次化结构, 用来描述信息的关键词, 可以用来描述物品的语义. 正是由于社会化标签的广泛应用, 很多网站取得了很大的成功. 作为标签推荐系统里的开山鼻祖之称的 Delicious, 允许用户给互联网上的每个网页打标签, 从而通过标签来重新组合整个互联网; CiteULike 允许用户提交或者收藏自己感兴趣的论文

① 收稿时间:2014-12-01;收到修改稿时间:2015-01-12

并且给论文打上标签; Last.fm 通过分析用户的听歌行为预测用户对音乐的兴趣, 从而给用户推荐个性化的音乐[2]. 国内很多网站也应用了标签推荐系统, 如豆瓣网支持用户对图书和电影等进行标注和评分, 借此获得图书和电影的内容信息和语义, 并用这种信息改善推荐效果. 这些社会化标签系统允许用户自行对信息或资源加以标注, 给人们对信息过滤带来了极大的帮助. 与其他常规标签不同, 微博用户的个性化标签是用户给自己添加的标注, 是对自身的一种描述方式, 用以体现用户的个性化特征, 同时也给微博中的好友推荐和其他信息推荐提供了更加丰富的内容.

但是在新浪微博平台中, 大部分用户都没有给自己添加标签, 这使得用户的个性化特征不是很明显, 不利于标签推荐系统的应用, 也不利于用户的兴趣挖掘. 本文研究的重点, 就是针对微博中用户添加标签不足的情形, 利用 RBLDA(Relationship Bind Latent Dirichlet Allocation)主题模型 [2], 并结合微博中用户间的好友关系和交互关系, 提出了一种微博用户的标签推荐算法, 可以让用户主动地添加与自身相关的个性化标签, 从而更好地为微博用户提供个性化服务. 通过对用户推送与自身兴趣爱好相关的标签, 不仅能够增强用户自身的个性化特征, 还有利于在微博平台中快速找到志同道合的好友, 有利于微博的数据挖掘.

1 相关工作

微博中推荐算法的应用十分广泛, 如根据用户的资料和兴趣, 可以为其推荐可能感兴趣的好友; 根据用户历史转发的微博和关注的话题, 可以为其推荐可能感兴趣的热点话题等. 如王晟[3]等提出一种基于贝叶斯个性化排序的微博推荐算法, 对用户进行个性化微博推荐. 该算法中提出一种微博对的概念来解决数据的稀疏性和不对称性, 并取得了良好的效果. 近年来, 对于标签推荐的研究, 主要可以概括为两大方法: 基于网络拓扑图的方法和基于内容的方法. FolkRank[4]是一种基于 PageRank 算法的改进算法, 通过标签在图中权重的排序, 为其推荐标签, 其核心思想是被重要的用户使用的标签标注的资源, 其自身也是重要的. FolkDiffusion[5]利用主题漂移概念提出了一种新的标签排序算法, 拓扑图中的标签传播就像物理学中的热传导现象一样, 从用户或资源相关度高的一方传播到另一方, 并最终保证被推荐标签跟目标用户的主题相关性. Guan[6]等人结合文档的相关度和用户偏好, 提出了一种多类别相关对象的排序方法, 为用户推荐标签, 取得了良好的效果. Zhou[7]等人结合物质扩算过程提出了针对用户-对象-标签三部图的物质扩散算法, 大大提高了刚刚进入系统的用户和物品的推荐准确度. 汪祥[8]等人基于微博用户之间的转发和提及关系构建用户之间的交互网络, 并依据用户之间的交互行为让标签在交互网络中以一定概率从一个用户转移到相关用户, 最终完成推送标签. 针对内容的推荐方法, 则主要集中在主题模型(Topic Model)方面. 主题模型是一种使用概率的产生式模型来挖掘文本主题的新方法. 如 Harvey[9]等人针对数据稀疏性问题, 提出了一种用户-资源-标签的主题模型 TTM(Tripartite Topic Model), 有效解决了标签数据的稀疏性问题; Li[10]等人从网络的主题和结构特性之间的关系出发, 结合 LDA(Latent Dirichlet Allocation)主题模型和 Girvan-Newman 社区发现算法, 提出了一种 TTR-LDA-Community(Tagger Tag Resource-latent Dirichlet Allocation-community)方法, 使得标签推荐准确率相比传统推荐算法有了很大的提高; 陈文涛[11]等通过对三种主体模型(TwitterLDA、AuthorLDA、UserLDA)进行比较, 发现 AuthorLDA 模型产生的主题具有较高的区分度, 而 UserLDA 和 AuthorLDA 模型则能更好地反映出用户的社交网络关系. 徐彬[2]等借助微博中用户的好友关系对用户标签及用户发表内容进行聚类,并按聚类后关键词的主题分布权重进行排序, 为用户推荐标签, 从一定程度上解决了文本过于短小造成的主题模型区分度不足的缺陷. 而张晨逸[12]等则提出了一种 MB-LDA 模型(MicroBlog-latent Dirichlet Allocation), 用于挖掘微博文本的主题分布, 此外, 该模型还能挖掘出联系人关注的主题, 以及利用挖掘的结果对微博集作个性化推荐, 如推荐相似微博和感兴趣的用户等.

总结上述提出的方法, 大多数方法都是基于用户、资源、标签之间的相似性, 或基于网络拓扑图的方法, 或基于主题模型的方法, 为用户推荐资源的标签, 而本文所研究的是为社交网络中微博用户自身推荐标签, 并综合考虑两大主要方法. 本文将主题模型和网络拓扑图两种方法相结合, 综合考虑微博用户之间的关系: 好友关系和交互关系, 针对每个用户的标签词过于短小, 利用 RBLDA 模型[2]对用户的好友标签进行主题分析, 并对标签词进行聚类, 为用户推送初

始的标签列表. 然后考虑微博用户的交互关系图, 让标签在交互图中以一定的概率转移, 最终根据标签在图中权重的排序, 选取合适的标签为其推荐.

2 微博用户标签分析

新浪微博作为国内访问量最大的微博平台之一, 允许用户给自己添加个性化标签, 但最多只能添加十个标签. 由于用户添加标签的随意性, 导致添加的标签并没有规范性. 图 1 为通过新浪微博 API 随机抓取的约 150 万个微博用户标签数量的统计情况, 其中约占 82.54%的用户都没有添加标签, 而添加标签数大于 5 的用户只约占 6.02%. 在这些用户中, 一共添加有 130887 个不同的标签, 其中音乐、电影等大众标签出现的频率较高, 而个性化标签出现的频率较低. 统计发现, 在这些标签中, 出现一次的标签约占 75.04%, 甚至 95%以上的标签出现的次数不大于 10. 只有极少数的标签使用频率较高, 这一部分正是大众标签. 图 2 所展示的是标签词汇使用的情况, 不难发现, 微博用户最常用的标签都是一些大众化的标签. 如果按照此种方法推荐, 就不能体现出微博用户个性化特征的多样性. 因而, 如何将这些个性化标签推荐给没有添加标签或只添加了少数标签的微博用户, 是我们所要研究的重点.

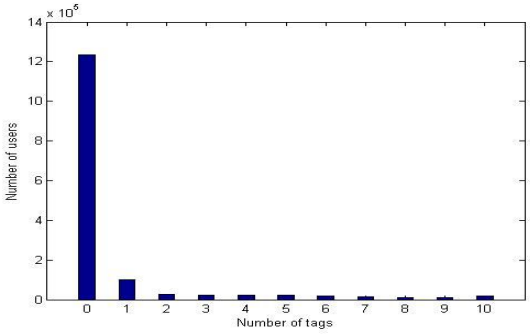

图 1 用户标签数目分布情况

新浪微博中, 每个用户都可以凭借自己的兴趣爱好, 添加关注自己感兴趣的用户. 因而, 好友添加的标签必定与用户之间存在着某种必然的联系. 正如文献[2]所提出的结论, 由于社交网络中存在的同质性, 好友之间的标签选用存在一定的主题相关性. 故利用用户间的好友关系, 给用户推荐标签是一种很好的方法.

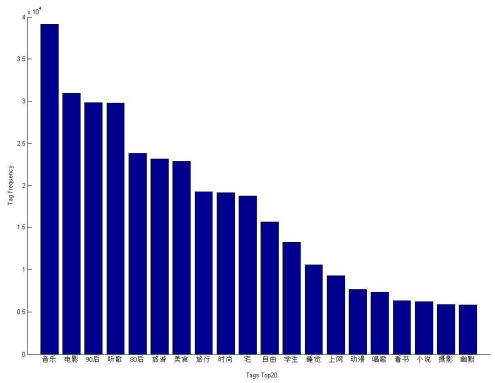

图 2 数据集中出现次数前 20 的标签

新浪微博中, 除了关注和被关注的关系, 还有一些比较重要的动态交互关系, 如转发、提及和评论. 用户可以通过对其他用户发布的微博进行转发, 也可以通过自己发布的微博@其他用户, 还可以对其他用户发布的微博进行评论. 通过对爬取的数据集中用户转发、提及和评论关系的分析发现, 大多数用户转发、提及和评论其他用户的数目较小. 其中, 只有 28.37%的用户转发、提及和评论其他用户的数目大于 5, 19.63%的用户转发、提及和评论其他用户的数目大于 10. 如果一个用户转发、提及和评论另一个用户, 可以从某种程度上说明用户之间对某一话题的共同兴趣[12]. 因而, 通过用户间的交互行为, 可以为用户推送与自身兴趣相关的标签.

3 微博用户标签推荐算法

3.1 基于好友关系约束的主题模型标签推荐

主题模型(Topic Model)是一种使用概率的产生式模型来挖掘文本主题的新方法, 被广泛应用于主题挖掘、文本检索、引文分析和社交网络分析等领域[12]. 其基本思想是每个文本可以表示成由一系列主题的混合分布, 而每个主题则由一系列的关键词依据概率分布生成. LDA 模型[13]在主题模型中最具有代表性, 是一个比较完备的主题模型[12]. 它是由 Blei 等人[13]在 2002 年提出的一种“文档-主题-单词”三层的贝叶斯概率模型, 语料库中的每一篇文档 d 与 K 个主题的一个 Dirichlet 分布 θ_d 相对应, 即文档-主题分布; 每个主题又与多个关键词 wdn 的一个多项式分布 β_k 相对应, 即主题-单词分布, 如图 3 所示[12]. 其中, α 和 η 分别对应 Dirichlet 分布 θ_d 和多项式分布的 β_k 超参数.

LDA 模型描述如下[12]: 对于 D 篇文档集中的任何一篇文档 d, 可以计算出该文档中的每个关键词所在的主题编号 z_{dn}, 每次从 K 个主题集中抽取一个主题, 通过迭代, 最终可以得出每篇文档在主题上的概率分布, 以及每个主题上相应关键词的概率分布. 在 LDA 模型中, 任何一篇文档中的关键词都是可观测的数据, 而文档中的主题是隐式变量, 根据文档中的已有数据和生成规则, 就可以求得该篇文档的主题分布结构[12].

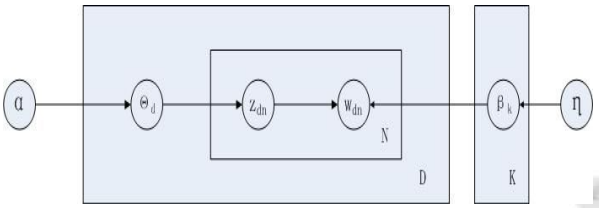

图 3　LDA 主题模型

新浪微博中, 关注是用户的一种主动行为, 不需要对方同意就能添加成功. 而被关注(粉丝)是一种被动行为. 用户通常通过添加关注获取自己感兴趣的好友, 以此来扩大自己的交际圈. 因而, 从某种程度上, 用户间的好友关系可以用来表征用户的兴趣所在. 个性化标签正是体现一个用户兴趣爱好最好的载体, 因此, 我们可以通过好友关系来帮助用户推荐标签.

徐彬等[2]通过对用户好友的标签信息进行分析发现, 由于社会网络中存在同质性, 好友之间的标签选用存在一定的主题相关性. 在实际的应用过程中, 由于传统的 LDA 模型在文本过于短小的文档集中进行主题挖掘, 将会出现同一个词在两篇短文本中的概率过小, 因而很难度量两者的相似性. 对于微博而言, 如果仅考虑某个微博用户所添加的标签, 就会出现这种状况. 为了弥补传统 LDA 的不足, 徐彬等[2]结合微博用户好友之间的标签相关的特点, 引入用户好友关系, 构建用户-标签对, 并提出一种好友关系约束的主题模型(RBLDA). 通过该模型可以计算出每个用户-标签对之间的主题相关度, 从而完成对自身用户的标签推荐. 该主题模型如图 4 所示[2], 其中, M_f 为该用户的粉丝集, M_a 为该用户所关注的好友集.

RBLDA 模型主要有两部分构成[2]: 第一部分是当前用户 u 与其粉丝好友 v_{jf} 组成的好友对(u, v_{jf}); 第二部分是当前用户 u 与其关注好友 v_{ka} 组成的好友对(u, v_{ka}). 对于每一好友对(u, v_{jf})或者(u, v_{ka}), 先推断用户 $v_{jf}(v_{ka})$的主题分布 θ_{ij} (θ_{ik}), 然后再以 θ_{ij} (θ_{ik})约束进一步推断用户 u 的主题分布. 其中, α_f 和 α_a 分别对应于粉丝好友集和关注好友集中的推断参数.

由于好友之间的标签选用存在一定的主题相关性, 又由于单一用户的标签组成的文档过于短小, 利用 RBLDA 模型[2]正好可以克服这一缺陷. 对用户好友的标签进行主题分析, 从而推断出用户的主题分布, 并对标签词进行聚类分析, 从而为用户推送标签.

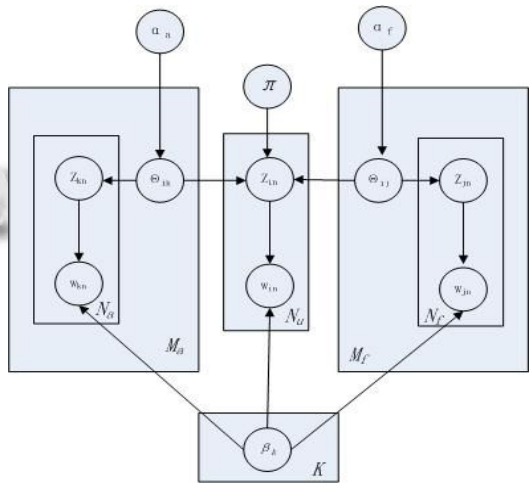

图 4　好友关系约束的 RBLDA 模型

3.2 基于交互关系的微博用户标签推荐

微博中用户的关系除了静态的好友关系, 还包括转发关系、提及关系和评论关系. 其中转发、提及和评论是用户在社交网络中通过交互行为而产生的一种交互关系, 具有一定的动态性. 如果一个用户转发、提及或评论另一个用户, 则更能反映用户之间对某一特定话题的共同兴趣[8]. 在标签推荐系统中, 除了考虑用户间的好友关系, 还更应当考虑用户间的转发、提及和评论关系. 因而, 本文在考虑了静态好友关系的基础上, 进一步考虑了用户间的转发、提及和评论关系.

微博中用户间通过交互关系可以形成一个很大的一个社交网络, 所以, 我们可以用一种加权的有向图来表示用户间交互关系的交互图 $G=(V,E,W)$[8]: 对于某一用户 u_i, T_{u_i} 表示用户 u_i 的标签集合, W_{u_i} 表示用户 u_i 的所有标签与 u_i 的相关程度的集合, w_{u_i,t_j} 表示用户 u_i 和标签 t_j 之间的相关度, 其中 $t_j \in T_{u_i}$. 初始时, 用户标签的初始权值都为 $1/|T_{u_i}|$. 在图 G 中一个顶点 v_i ($v_i \in V$), 可以表示为 $v_i=(u_i,T_{u_i},W_{u_i})$; 顶点之间的边表示用户之间的交互关系, 即转发、提及和评论关系.

如果用户 u_i 转发(或提及或评论)了用户 u_j, 那么可以认为在图 G 中有一条从用户 u_j 到用户 u_i 的有向边, 即 e_{ji} ($v_j \rightarrow v_i$). 边上的权值表示为标签从一个用户转移到另一个用户的转移概率.

用户之间的交互关系, 说明用户之间有共同感兴趣的话题, 因此可以认为, 用户 u_i 转发、提及或评论用户 u_j, 说明用户 u_i 受到了用户 u_j 的影响. 根据文献[8]得出的比较结果, 如果用户 u_i 转发、提及或评论用户 u_j, 则认为用户 u_j 影响了用户 u_i, 同时, 用户 u_i 的标签验证了用户 u_j 的标签, 即用户 u_j 和用户 u_i 之间的标签传播是一种双向传播.

在交互图 G 中, 有向边边 e_{ji} ($v_j \rightarrow v_i$)的权值反映了用户之间的影响强度, 本文采用用户间的转发、提及和评论关系的次数来计算有向边边 e_{ji} 的权值. 如果用户 u_i 转发、提及或评论用户 u_j 的次数越多, 则 u_i 对 u_j 的影响强度越大, 即用户 u_i 和 u_j 的标签有更大的概率在彼此之间进行相互传播. 利用用户间转发、提及和评论的次数, 计算用户间的标签转移概率,公式如下[8]:

$$P(u_j \rightarrow u_i) = \frac{f(u_i,u_j)}{\sum_{u_s \in Set(u_j)} f(u_s,u_j)} \quad (1)$$

其中, $f(u_i,u_j)$ 表示用户 u_i 转发、评论或提及用户 u_j 的总次数, $Set(u_j)$ 表示所有转发、评论或提及了用户 u_j 的用户集合, 且满足 $0 \le P(u_j \rightarrow u_i) \le 1$.

考虑到转发、提及和评论在微博中的作用不同, 分别给予不同的权重 ω_1、ω_2、ω_3, 则用户间的标签转移概率计算公式如下:

$$P(u_j \rightarrow u_i) = \omega_1 \frac{f_1(u_i,u_j)}{\sum_{u_s \in Set(u_j)} f(u_s,u_j)} + \omega_2 \frac{f_2(u_i,u_j)}{\sum_{u_s \in Set(u_j)} f(u_s,u_j)} + \omega_3 \frac{f_3(u_i,u_j)}{\sum_{u_s \in Set(u_j)} f(u_s,u_j)} \quad (2)$$

其中, $f_1(u_i,u_j)$, $f_2(u_i,u_j)$, $f_3(u_i,u_j)$ 分别表示 u_i 转发、提及和评论用户 u_j 的总次数, 且满足 $0 \le P(u_j \rightarrow u_i) \le 1$, $\omega_1+\omega_2+\omega_3=1$.

在整个庞大的交互网络中, 用户 u_i 会和多名用户产生交互关系, 而这些与之产生交互关系的用户会将标签以一定的转移概率传播到用户 u_i, 这样就可以为用户推荐准确的标签信息. 假设与用户 u_i 有交互关系的用户集合为 $UserSet(u_i)$, 则可以传播到用户 u_i 的标签集合 $TagSet(u_i)$表示为[8]:

$$\mathrm{TagSet}(u_i) = \bigcup_{u_s \in UserSet(u_i)} T_{u_s} \quad (3)$$

标签的权值是所有相关用户当前标签的权值乘以其转移概率后的总和, 标签 t_i ($t_i \in TagSet(u_i)$)的权值采用如下公式进行计算[8]:

$$weight(t_m) = \sum_{u_s \in TagSet(u_i)} P(u_s \rightarrow u_i) \bullet w_{u_s t_m} \quad (4)$$

在这个庞大的交互网络中, 用户间的标签通过一定的概率转移到下一个被转发、提及和评论的用户. 通过不断的迭代传播, 直到算法达到一个稳定的状态. 最后, 通过对每个标签的权值进行排序, 选取权值最大的前 k 个作为最终的推荐标签.

4 综合机制的微博用户标签推荐算法

4.1 算法基本思想

由于微博中大部分的用户都没有添加标签信息, 很大程度上不利于用户的标签推荐系统的应用. 本文综合考虑用户之间的好友关系, 以及具有交互功能的转发、提及和评论关系, 提出一种综合机制的用户标签推荐算法. 首先针对大多数用户都没有添加标签的问题, 初始时, 设定一定数量的标签作为初始标签: 在好友关系网络中, 利用 RBLDA 模型[2]对用户好友的标签词进行聚类分析, 为用户推荐一定数量的初始标签. 然后, 利用用户间具有动态交互关系的转发、提及和评论关系, 在整个交互关系网络中, 让标签以一定的转移概率传播到与之产生交互关系的用户当中, 直到整个交互网络达到稳定状态. 最后, 依据 Top-k 原则, 从推荐的标签中按照概率大小选取前 k 个标签作为用户的推荐标签.

4.2 算法基本步骤

输入 新浪微博数据集 S

输出 给定用户推荐的 k 个标签

Begin

Step1: 根据给定的数据集 S 建立用户的好友关系网络图 M, 以及用户的交互关系网络图 G;

Step2: 推送初始标签. 依据 RBLDA 模型[2]为 M 中每个用户推送初始的标签, 保证每个用户都能有一定数量的原始标签;

Step3: 计算标签的转移概率. 依据用户间的交互关系(转发、提及和评论), 在交互网络图 G 中利用标签概率转移公式(1)或(2)计算用户的每个标签的转移概率;

Step4: 计算标签权值. 利用公式(3), 计算当前所

有用户的每个标签所对应的权值;

Step5: 重复 Step3~Step4, 直到整个交互网络图 G 中每个用户的标签权值不再变化, 即达到稳定状态;

Step6: 标签选取. 针对以上步骤计算出的每个标签的最终的权值, 进行排序, 选取前 k 个标签最为用户最终的推荐标签.

End

5 实验结果与分析

5.1 实验数据集与实验环境

本文所采用的实验数据集原始数据来源于新浪微博, 该数据集共收集了 150 万名微博用户的信息, 其中包括用户的好友信息, 转发、提及和评论信息, 以及用户的标签信息等. 对所爬取的信息进行预处理, 主要是去除无效用户, 以及用户信息中没有或者只有少量粉丝或者关注好友的用户. 预处理后, 共筛选出 207359 个有效用户信息, 如表 1 所示.

表 1 新浪微博数据集基本信息

统计量	统计值
微博用户数	207359
标签数	109273
好友关系边数	1058364
交互关系边数	46082977

实验中所采用的实验环境为: Intel(R) Core(TM) i5-3470 CPU @3.20GHz 3.20GHz 处理器, 4GB 内存, Windows 7 操作系统, Eclipse 开发平台.

5.2 标签预测实验

表 2 为某微博用户在两种不同推荐算法下标签推荐结果对比. 该用户为新浪微博认证用户, 职业为摄影师. 表中两种算法都给出了推荐度排名前十的标签词, 我们可以发现, RBLDA 模型[2]推荐的结果中, 虽然推荐有摄影师、摄影、果粉这些与用户相关的标签, 但大众化标签(电影、音乐、美食、旅游等)仍占据大多数, 并且这些大众化标签在推荐结果中都处在靠前的位置; 本文提出的算法的推荐结果中, 虽然也有大众化标签, 但推荐更多的是与用户职业或兴趣相关的标签, 如: 摄影师、婚纱摄影、创作艺术等标签词, 且所推荐的标签区分度更高.

表 2 某微博用户在两种算法中的推荐结果

用户基本信息	用户 ID	awei1101
	关注好友数	395
	粉丝好友数	8536
	微博数	283
	用户原始标签	摄影师, 技术宅, 果粉
RBLDA 推荐标签	电影, 音乐, 美食, 摄影师, 旅游, 摄影, 美女, 果粉, 会记学, 快乐	
本文算法推荐标签	电影, 摄影师, 青春期, 婚纱摄影, 创作艺术, 数码控, 时尚, 果粉, 旅游, 美学	

5.3 准确度实验评估

为了进一步评估本文算法的性能, 本文提出对推荐标签的准确度进行度量. 从数据集中选取标签数大于 N 的用户, 并将其标签删去, 利用本文提出的算法给这些删去标签的用户重新打上标签, 最后将重新得到的标签与用户原始标记的标签进行比较.

为保证实验的正确性, 本文选取三组测试集, 每组测试集都包含 5000 个用户, 原始标签数 N=5. 同时为了保证测试集数据选取的随机性, 同样采用随机数的方法来判断某一符合条件的数据加入测试集中. 第一组测试集选取转发、提及或评论其他用户的总数在 50~60 之间, 且被其他用户转发、提及或评论的用户总数也在 50~60 之间的用户; 第二组测试集选取转发、提及或评论其他用户的总数在 100~150 之间, 且被其他用户转发、提及或评论的用户总数也在 100~150 之间的用户; 第三组测试集选取转发、提及或评论其他用户的总数在 200 以上, 且被其他用户转发、提及或评论的用户总数也在 200 以上的用户.

为增加算法的对比性, 本文添加了几组对比实验: 算法 1 为 RBLDA 主题模型[2]的标签推荐算法; 算法 2 为没有给用户初始标签的算法, 且采用公式(1)计算标签的转移概率, 即文献[8]中的原始算法; 算法 3 为本文提出的算法, 且采用公式(1)计算标签的转移概率; 算法 4 也为本文提出的算法, 但采用公式(2)计算标签的转移概率, 实验过程中, 参数 ω_1、ω_2、ω_3 分别取值为 0.25、0.35、0.4.

对三组测试集进行试验, 并采用前 N 条结果的准确率 $P@N$ 和前 N 条结果的召回率 $R@N$ 对算法性能进行评价, 实验过程中准确率 N 取值 1、3、5、10, 召回率 N 取值 20. 实验结果如表 3 所示.

表 3 三组测试集上不同算法的测试性能对比

测试集	算法	*P@1*	*P@3*	*P@5*	*P@10*	*R@20*
第一组测试集 50~60	算法 1	0.4011	0.2673	0.2014	0.0767	0.2868
	算法 2	0.3869	0.3517	0.2708	0.1210	0.3395
	算法 3	0.4106	0.3594	0.3047	0.1409	0.3767
	算法 4	0.4263	0.3814	0.3026	0.1536	0.3906
第二组测试集 100~150	算法 1	0.3974	0.2786	0.1981	0.0731	0.2893
	算法 2	0.3218	0.2975	0.2467	0.0936	0.3105
	算法 3	0.3940	0.3283	0.2605	0.1190	0.3281
	算法 4	0.3939	0.3553	0.2783	0.1219	0.3509
第三组测试集 200 以上	算法 1	0.3896	0.2632	0.1827	0.0695	0.2830
	算法 2	0.3018	0.2744	0.1792	0.0701	0.2798
	算法 3	0.3779	0.2912	0.1900	0.0795	0.2859
	算法 4	0.3786	0.3409	0.2159	0.0910	0.2947

从表 3 可以看出, 三组测试集中, 本文所提出的算法(算法 3 和算法 4)要比原始算法(算法 1 和算法 2)的性能好, 且随着 N 值增大准确率和召回率都有所下降. 三组测试集中的交互用户总数都是在增加, 算法 1 的各项评价指标显示的性能变化不是很大, 而其他三种算法的各项评价指标显示的性能, 并不随着测试集中交互总数的增加而呈现正相关, 反而随着交互用户总数的增加有所下降. 究其原因, 算法 1 并不依赖于测试集中用户之间的交互关系, 而针对其他算法, 由于在整个社交网络中, 用户添加的标签不足, 而且绝大多数标签很少被其他用户使用, 导致在交互网络中, 交互关系越多的用户接收到不相关的标签的几率增大. 比较算法 2 和算法 3, 可以看出, 在算法 2 的基础上用 RBLDA 模型[2]的方法给用户推荐初始的标签列表, 取得了很好的效果. 比较算法 3 和算法 4, 算法 4 针对不同的交互关系给予不同的权重来计算标签转移概率, 这与现实中用户对待不同交互关系的重要程度有关.

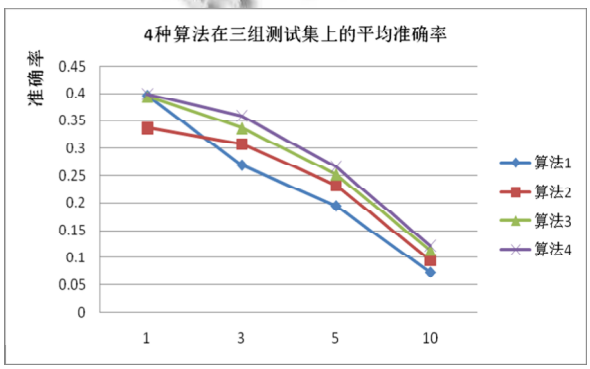

图 5 四种算法在三组测试集上的平均准确率

相对评论和提及行为而言, 转发往往是用户的一种随意的行为. 图 5 显示的是四种算法在三组测试集上平均准确率, 我们可以清晰的看出, 算法 4 的实验效果最好. 理论分析与实际状况是相一致的.

6 结语

本文针对微博中用户添加标签信息不足的问题提出了一种利用用户间好友关系的 RBLDA 模型[2]为用户推荐初始标签, 并考虑用户间的交互关系的推荐算法, 来为用户推荐感兴趣的标签. 在真实的新浪微博数据集上对算法进行了测试, 实验结果表明, 本文所提出的方法取得了良好的效果. 今后的研究工作中, 将进一步考虑用户的交互关系在交互网络中的作用, 以及交互过程中用户的个人影响力对推荐准确度的影响.

参考文献

1 Sigurbjörnsson B, Van Zwol R. Flickr tag recommendation based on collective knowledge. Proc. of the 17th International Conference on World Wide Web. ACM. 2008. 327–336.

2 徐彬,杨丹,张昱,等.面向微博用户标签推荐的关系约束主题模型.计算机科学与探索,2014,8(3):288–295.

3 王晟,王子琪,张铭.个性化微博推荐算法.计算机科学与探索,2012,6(10):895–902.

4 Jäschke R, Marinho L, Hotho A, et al. Tag recommendations in folksonomies. Knowledge Discovery in Databases: PKDD 2007. Springer Berlin Heidelberg, 2007: 506–514.

5 Liu Z, Shi C, Sun M. FolkDiffusion: A graph-based tag suggestion method for folksonomies. Information Retrieval Technology. Springer Berlin Heidelberg, 2010: 231–240.

6 Guan Z, Bu J, Mei Q, et al. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. Proc. of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. 2009. 540–547.

7 Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation. Physical Review E, 2007, 76(4): 046115.

8 汪祥,贾焰,周斌,等.基于交互关系的微博用户标签预测.计算机工程与科学,2013,35(10): 44–50.

9 Harvey M, Baillie M, Ruthven I, et al. Tripartite hidden topic models for personalised tag suggestion. Advances in Information Retrieval. Springer Berlin Heidelberg, 2010: 432–443.

10 Li DF, Ding Y, Sugimoto C, et al. Modeling topic and community structure in social tagging: the TTR-LDA Community model. Journal of the American Society for Information Science and Technology, 2011, 62(9): 1849–1866.

11 陈文涛,张小明,李舟军.构建微博用户兴趣模型的主题模型的分析.计算机科学,2013,40(4):127–130.

12 张晨逸,孙建伶,丁轶群.基于 MB-LDA 模型的微博主题挖掘.计算机研究与发展,2011,48(10):1795–1802.

13 Beli DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3: 993–102.