融合标签关联关系与用户社交关系的微博推荐方法

马慧芳,贾美惠子,张 迪,蔺想红
(西北师范大学计算机科学与工程学院,甘肃兰州 730070)

摘 要: 通过分析微博特点及现有微博推荐算法的缺陷,提出一种融合了标签间关联关系与用户间社交关系的微博推荐方法. 采用标签检索策略对未加标签和标签较少的用户进行加标,构建用户-标签矩阵,得到用户标签权重,为了解决该矩阵中稀疏的问题,通过挖掘标签间的关联关系,继而更新用户-标签矩阵. 考虑到多用户之间社交关系对挖掘用户兴趣并进行微博推荐的重要性,构建用户-用户社交关系相似度矩阵,并与更新后的用户-标签矩阵进行迭代,得到最终的用户兴趣并进行相关推荐. 实验证明了该算法针对微博信息推荐是有效的.

关键词: 微博推荐; 标签检索; 用户-标签矩阵; 用户标签权重; 标签关联关系; 用户-用户社交关系相似度矩阵

中图分类号: TP393.092 **文献标识码:** A **文章编号:** 0372-2112 (2017)01-0112-07

电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.01.016

Microblog Recommendation Based on Tag Correlation and User Social Relation

MA Hui-fang, JIA Mei-hui-zi, ZHANG Di, LIN Xiang-hong
(*College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu* 730070, *China*)

Abstract: A novel microblog recommendation method combining the tag correlation with the user social relation is proposed via analyzing microblog features and the deficiencies of existing microblog recommendation algorithm. Specifically, we establish a tag retrieval strategy to add tags for unlabeled users and users with few tags, and then build the user-tag matrix and obtain user-tag weights. In order to solve the problem of sparsity of the matrix, we investigate the correlation between the tags to update the user-tag matrix. Considering the significance of user social relation for microblog recommendation, a user-user social relation similarity matrix is constructed and a mechanism is designed to iteratively obtain user interest. Experimental results show that the algorithm is effective in microblog recommendation.

Key words: microblog recommendation; tag retrieval; user-tag matrix; user-tag weight; tag correlation; user-user social relation similarity matrix

1 引言

随着 Web2.0 技术、无线网络技术和移动通信 4G 技术的发展,微博社交网络平台应运而生. 如何在此平台中为用户提供个性化的服务,筛选出高质量的内容变得非常重要,而准确的发现用户的兴趣则是实现这种个性化服务的前提. 在此前提下,大量的用户兴趣推荐算法应运而生. 针对微博长度短,内容稀疏的特点,多数研究人员多从扩展文本特征,丰富语义的角度出发进行文本表示. 主流的方法多使用外部数据库进行语义扩充[1]. 少数研究人员对文本特征向量进行缩减,利用极少数词表示用户的兴趣[2]. 由于微博中的标签是用户专门添加来描述自身特征的,其对于描述用户兴趣具有重要意义. 在互联网其他领域的研究中,已经有一些工作开始考虑使用标签信息来表示用户兴趣[3]. 在推荐研究[4,5]方面,虽然已有人采用标签来发掘用户

收稿日期:2015-06-02;修回日期:2016-01-25;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61163039, No. 61363058);甘肃省科技厅青年基金(No. 145RJYA259);甘肃省自然科学基金(No. 1506RJZA127);中科院智能信息处理重点实验室开放课题(No. IIP2014-4)

兴趣[6,7],也已有研究人员挖掘用户的关系进行推荐[8,9],但并未考虑多标签之间存在着一定的关联关系,更是很少有人将标签间关系与用户间关系融合表征用户兴趣进行相关微博推荐.

本文提出了一种融合了标签间关联关系与用户间社交关系的微博推荐方法. 对于无标签以及标签较少的用户,通过标签检索策略获取相应标签,继而构建用户-标签矩阵,得到初始用户标签权重,考虑标签与标签间的关联关系,通过挖掘被同一用户标注的多标签的内联关系与被不同用户标注的多标签外联的关系,构建合理的多标签关联关系矩阵,对用户-标签矩阵进行更新. 另外,考虑用户与用户之间的社交信息关系,构建合理的用户间社交关系相似度矩阵,并与更新后的用户-标签矩阵进行迭代,得到最终的用户标签权重. 与忽略标签与用户间关系的微博推荐算法对比,本文提出的推荐方法能够更有效地进行微博推荐. 图 1 为文章算法程序流程图,主要由标签关联关系与用户社交关系两部分组成,该算法的输入是微博信息流,输出为微博的推荐序列.

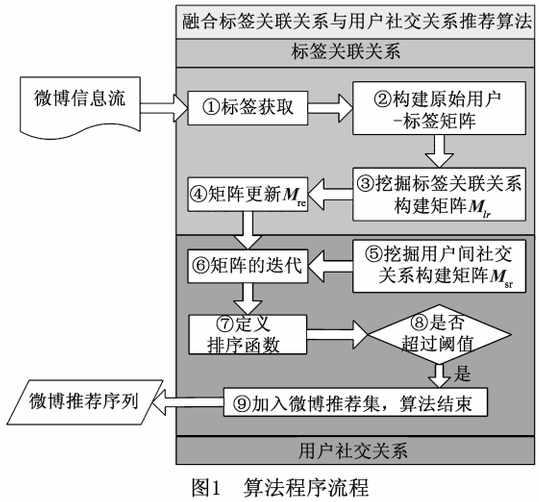


图1　算法程序流程

2　用户标签获取与多标签关联关系

2.1　用户标签的获取

对于用户的个人标签,可通过用户其本身设定以及从用户所发布的微博中检索两种方式获取,且可表示为用户的兴趣[10]. 若用户自身设定标签较多,则不需对微博标签进行获取;若用户自身设定标签较少或用户并未自身设定标签,则需对用户的标签进行获取.

2.1.1　标签检索

针对未添加标签以及标签较少的用户,设计标签检索与加权策略,获取用户的标签. 表 1 为在标签检索阶段各符号的定义.

表 1　标签检索阶段各符号的定义

符号	定义
$U=\{u_1,u_2,\cdots,u_i,\cdots,u_N\}$	需进行加标微博用户集合
N	用户的个数
$D_i=\{d_{i1},d_{i2},\cdots,d_{iM_i}\}$	用户 u_i 所发布的微博集合
$M_i,i=(1,2,\cdots,N)$	用户 u_i 所发布的微博个数
$D=\bigcup_{i=1}^{N}D_i$	所有用户发布的微博集合
$T_i=\{t_{i1},t_{i2},\cdots,t_{im_i}\}$	微博集合 D_i 中的词集
$m_i,m_i \ll M_i$	词集中词语的个数
$L_i=\{l_{i1},l_{i2},\cdots,l_{in_i}\}$	用户 u_i 的标签集合
n_i	被用户 u_i 标注的标签个数
$L=\bigcup_{i=1}^{N}L_i$	所有标签集合
n	标签集合 L 中标签的总数

在标签检索的过程中最重要的部分是如何从用户以往发布过的微博中选取具有代表性的关键词[11]作为用户的标签. 选择关键词作为相应标签应基于以下两点:一是能够较好地揭示用户所发布微博的内容;二是有局部指示性,即该词应能表示特定的主题. 针对前者,采用词频 TF(Term Frequency)或词频-逆文档频率 TF-IDF(Term Frequency-Inverse Document Frequency)加权策略[12]. 针对后者,采用清晰度 clarity 权衡所选查询关键词的局部指示性. 所选查询关键词的 clarity 得分是查询语言模型与选择语言模型间的 KL(Kullback-Leibler)距离. 前者是与一个给定的查询关键词最佳匹配的一系列微博,后者是用户 u_i 微博集合. 第 j 个查询关键词 l_{ij} 为所选择的候选词,利用 l_{ij} 作为单一查询关键词,查询检索最具有相关性的 g_i 篇微博,定义为 $Q_{l_{ij}}$. 式(1)定义查询关键词 l_{ij} 的 Clarity 得分.

$$\text{Clarity}(l_{ij}) = \sum_{l_{ij}^{'}\in L_i} P(l_{ij}^{'}\mid Q_{l_{ij}}) \log \frac{P(l_{ij}^{'}\mid Q_{l_{ij}})}{P(l_{ij}^{'}\mid D_i)} \quad (1)$$

若 l_{ij} 具有局部指示性且能够代表特定的主题,那么匹配的文档与 l_{ij} 具有相同的主题,该主题由集合中出现概率较高且数量较少的词表示.

定义用户 u_i 所发布的微博集中第 j 个词的得分如下:

$$s_j = tf_j \times \text{clarity}(l_{ij}) \quad (2)$$

选定 n_i 个最高权重的查询关键词被作为用户 u_i 的标签. 每个标签初始权重的归一化如式(3)所示:

$$\text{normalized}(s_j) = \frac{s_j}{\sum_{x=1}^{n_i} s_x} \quad (3)$$

2.1.2　用户标签矩阵

针对用户 u_i 构造一个标签权重向量 $\boldsymbol{V}_i=(w_{i1},w_{i2},\cdots,w_{in})$ 以用来存储标签的权重[13].

用户标签的初始权重如下：

$$w_{ij}=\begin{cases}1/Z_i, & \text{用户通过标签服务获得}\\ \text{normalized}(s_j), & \text{标签检索}\\ 0, & \text{其他情况}\end{cases} \tag{4}$$

若标签是通过用户所发布的微博中检索得到，则式(3)作为标签的初始权重. 否则，如若标签是通过微博系统所提供的标签服务得到的，每个标签应有相等的重要性. 假定用户 u_i 有 Z_i 个自己所标注的标签，$1/Z_i$ 为用户 u_i 的第 j 个标签 l_j 的初始权重.

基于以上的用户权重向量，构建一个 $N\times n$ 的用户-标签矩阵 $\boldsymbol{M}_{ul}$. 其中 N 为用户的数量，n 为标签的数量，矩阵中的元素 w_{ij} 为第 i 个用户 u_i 的第 j 个标签 l_j 的权重.

由于该用户-标签矩阵的列向量为所有待加标用户的标签集合，并且该集合中的标签并不可能被所有用户标注. 故该矩阵存在非常稀疏的问题，并不能很好的表示出用户的兴趣.

2.2 多标签关联关系

在 2.1 节中所构建的用户-标签矩阵中，由于矩阵有其自身的局限性，在本节中，将挖掘多标签间的关联关系，以此更新原始的用户-标签矩阵.

图 2 中 u_1 和 u_2 分别表示用户 1 与用户 2，白色的三角形表示被用户 1 与用户 2 共同标注的标签，箭头左右两边的实线框分别表示标签的两种关联关系模式.

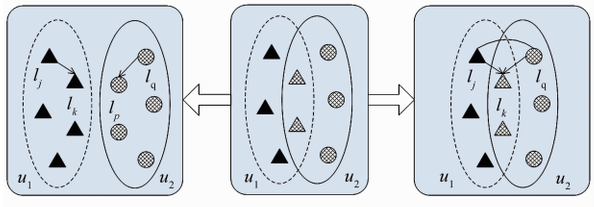


图2 标签关联关系

2.2.1 多标签内联关系

定义 1 若两个标签被同一个用户所标注，则这两个标签间存在**内联关系**. 该模型如图 2 左边箭头指向的实线框所示.

l_j 和 l_k 是被用户 u_1 标注的标签，故标签 l_j 和 l_k 在用户 u_1 中具有内联关系，根据 Jaccard 相似度公式，定义标签 l_j 和 l_k 的内联关系由式(5)所表示：

$$LIR(l_j,l_k)=\frac{1}{|H|}\times\sum_{y_i\in H}\frac{w_{ij}w_{ik}}{w_{ij}+w_{ik}-w_{ij}w_{ik}} \tag{5}$$

其中：w_{ij} 和 w_{ik} 分别表示被用户 u_i 标注的第 j 个标签 l_j 与第 k 个标签 l_k 的权重，w_{ij} 的权重可由式(4)表示.

H 表示在集合 $H=\{y_i|(w_{ij}\neq 0)\&(w_{ik}\neq 0)\}$ 中的元素个数. 若 $H=\varnothing$，则 $LIR(l_j,l_k)=0$. 由式(5)可得到标签间内联关系的语义信息. 并将其值规范化到[0,1]之间，则归一化标签 l_j 和 l_k 的内联关系公式如下：

$$N\text{-}LIR(l_j,l_k)=\begin{cases}1, & j=k\\ \dfrac{LIR(l_j,l_k)}{\sum\limits_{j=1,j\neq k}^{n}LIR(l_j,l_k)}, & j\neq k\end{cases} \tag{6}$$

其中，n 表示标签的数量，$N\text{-}LIR(l_j,l_k)$ 表示被同一用户标注的标签 l_j 和 l_k 的内联关系. 对于所有的标签 l_j 来说，当 l_j 和 l_k 不相同时，有 $\sum\limits_{j=1,j\neq k}^{n}N\text{-}LIR(l_j,l_k)=1$.

2.2.2 多标签外联关系

定义 2 存在两个用户 u_1 和 u_2，若有标签同时被用户 u_1 和 u_2 共同标注，那么分别在 u_1，u_2 中与该标签有内联关系的两个标签间会存在一定的**外联关系**. 该模型如图 2 右边箭头指向的实线框所示.

若给定了标签 l_j 和 l_k，则至少存在一个标签 l_q，使得 $N\text{-}LIR(l_j,l_q)>0$ 且 $N\text{-}LIR(l_k,l_q)>0$，那么则称标签 l_j 和 l_k 是具有外联关系的，其中标签 l_q 为标签 l_j 和 l_k 的链接标签，标签 l_j 和 l_k 通过标签 l_q 的外联关系如下：

$$LOR(l_j,l_k|l_q)=\min(N\text{-}LIR(l_j,l_q),N\text{-}LIR(l_k,l_q)) \tag{7}$$

对所有与标签 l_j 和 l_k 同时具有外联关系的链接标签求最终的外联关系，并规范化至[0,1]之间如下：

$$N\text{-}LOR(l_j,l_k)=\begin{cases}0, & j=k\\ \dfrac{\sum\limits_{\forall l_q\in E}LOR(l_j,l_k|l_q)}{|E|}, & j\neq k\end{cases} \tag{8}$$

其中，$E=\{l_q|(N\text{-}LIR(l_j,l_q)>0)\&(N\text{-}LIR(l_k,l_q)>0)\}$ 表示链接标签 l_q 的个数. 若 $E=\varnothing$，则表示两标签间没有共有的链接标签，即标签间没有外联关系. 式(8)表明若两个标签共有的链接标签越多，则这两个标签的关系就越密切，其多标签间的外联关系就越大.

至此，上述文章中充分挖掘出多标签间的全部的关联关系. 通过如下公式计算出多标签间的关联关系：

$$LR(l_j,l_k)=\begin{cases}1, & j=k\\ \alpha\times N\text{-}LOR(l_j,l_k)+(1-\alpha)\times N\text{-}LIR(l_j,l_k), & \text{otherwise}\end{cases} \tag{9}$$

其中，$\alpha(\alpha\in[0,1])$ 决定了多标签间的内联关系与外联关系所占的比例.

2.2.3 用户-标签矩阵的更新

构建 $n\times n$ 的多标签关联关系矩阵 $\boldsymbol{M}_{lr}$，且 $LR(l_j,l_k)$ 为矩阵 $\boldsymbol{M}_{lr}$ 中的元素. 由上述方法得到的多标签关联关系矩阵可对原始的用户-标签矩阵进行更新，更新后的矩阵不仅可以解决稀疏的问题，而且还具有更丰富的语义信息，可以更好地表示出用户的兴趣. 下式为包含多标签关联关系的用户-标签矩阵：

$$\boldsymbol{M}_{re}=\boldsymbol{M}_{ul}\times\boldsymbol{M}_{lr} \tag{10}$$

式(10)中的矩阵 $\boldsymbol{M}_{re}$ 为更新后的 $N\times n$ 用户-标签矩阵.

3　用户间社交关系与迭代算法描述

文章第二部分所构建的用户-标签矩阵中仅仅挖掘了用户与标签以及标签与标签的关系,而并未考虑用户间的社交关系,而用户间的社交关系对挖掘用户兴趣非常重要. 已有研究人员对微博用户的相似度进行了相关研究[14]. 受此文章启发,本节将分析用户间的社交关系.

3.1　用户间社交关系

在微博系统的社交网络中,用户通过关注信息与粉丝信息彼此相连,由于关注信息与粉丝信息是一种非常复杂社交关系结构,如若用户 u_i 对用户 u_j 所发布的微博感兴趣,那么 u_i 会有意地关注 u_j;又如用户 u_i 与 u_k 虽未彼此关注,但都同时关注了用户 u_j,通过链接用户 u_j,可以说明 u_i 与 u_k 也是具有一定的关系等. 通过挖掘社交关系可以进一步地揭示出用户的兴趣.

定义 3　用户 u_i 的**社交关系**为:

$$SR(u_i)=\{\boldsymbol{Fe}(u_i),\boldsymbol{Fr}(u_i)\}(Social\ Relation(u_i)=\{\boldsymbol{Followee}(u_i),\boldsymbol{Follower}(u_i)\})$$

$SR(u_i)$ 表示用户 u_i 的社交关系,其中包括两种属性信息,分别是关注信息与粉丝信息,首先可分别将这两种属性信息表示为两个向量:关注向量 $\boldsymbol{Fe}(u_i)$,粉丝向量 $\boldsymbol{Fr}(u_i)$,其次,将文中 N 个用户编号$\{1,2,\cdots,i,\cdots,j,\cdots,N\}$,如若用户 u_i 关注了用户 u_j,那么在关注向量 $\boldsymbol{Fe}(u_i)$ 中第 j 个分量为 1,否则为 0. 同理,若用户 u_i 被用户 u_j 所关注,那么在粉丝向量 $\boldsymbol{Fr}(u_i)$ 中第 j 个分量为 1,否则为 0.

3.1.1　用户间社交关系相似度计算

对于两个用户 u_i 和 u_j,他们之间的社交关系可分别表示为 $SR(u_i)=\{\boldsymbol{Fe}(u_i),\boldsymbol{Fr}(u_i)\}$,$SR(u_j)=\{\boldsymbol{Fe}(u_j),\boldsymbol{Fr}(u_j)\}$. 因此用户 u_i 和 u_j 的社交关系相似度计算,可转换为对社交关系的两种属性:关注信息与粉丝信息相似度计算. 文章采用余弦相似度计算方法.

(u_i,u_j) 的关注信息相似度为:

$$sim(\boldsymbol{Fe}(u_i),\boldsymbol{Fe}(u_j))=\frac{\boldsymbol{Fe}(u_i)\cdot\boldsymbol{Fe}(u_j)}{\|\boldsymbol{Fe}(u_i)\|\times\|\boldsymbol{Fe}(u_j)\|}\tag{11}$$

(u_i,u_j) 的粉丝信息相似度为:

$$sim(\boldsymbol{Fr}(u_i),\boldsymbol{Fr}(u_j))=\frac{\boldsymbol{Fr}(u_i)\cdot\boldsymbol{Fr}(u_j)}{\|\boldsymbol{Fr}(u_i)\|\times\|\boldsymbol{Fr}(u_j)\|}\tag{12}$$

用户 u_i 和 u_j 之间社交关系的相似度为:

$$sim(SR(u_i),SR(u_j))=sim(\boldsymbol{Fe}(u_i),\boldsymbol{Fe}(u_j)+sim(\boldsymbol{Fr}(u_i),\boldsymbol{Fr}(u_j))\tag{13}$$

对式(13)进行归一化处理,使结果规范化到[0,1]之间且($i\neq j$). 如下:

$$N-sim(SR(u_i),SR(u_j))=\frac{sim(SR(u_i),SR(u_j))-\min\ sim(SR(u_i),SR(u_j))}{\max\ sim(SR(u_i),SR(u_j))-\min\ sim(SR(u_i),SR(u_j))}\tag{14}$$

3.1.2　用户间社交关系相似度矩阵

通过对用户间社交关系即关注信息和粉丝信息的研究计算,可构建一个 $N\times N$ 的用户-用户社交关系相似度矩阵 $\boldsymbol{M}_{sr}$,其中 N 为用户的总数.

$$\boldsymbol{M}_{sr}=\begin{bmatrix}m_{11}&m_{12}&\cdots&m_{1N}\\m_{21}&m_{22}&\cdots&m_{2N}\\\vdots&\vdots&\vdots&\vdots\\m_{N1}&m_{N2}&\cdots&m_{NN}\end{bmatrix}\tag{15}$$

矩阵 $\boldsymbol{M}_{sr}$ 中各元素为两两用户间社交关系相似度,如式(16)所示:

$$m_{ij}=\begin{cases}N-sim(SR(u_i),SR(u_j)), & 存在社交关系\\1, & u_i=u_j\\0, & 未存在社交关系\end{cases}\tag{16}$$

3.2　迭代推荐算法描述

3.2.1　矩阵的迭代

通过以上计算可得出 $N\times N$ 的用户-用户社交关系相似度矩阵 $\boldsymbol{M}_{sr}$ 以及更新后的 $N\times n$ 用户-标签矩阵 $\boldsymbol{M}_{re}$,将矩阵 $\boldsymbol{M}_{sr}$ 与矩阵 $\boldsymbol{M}_{re}$ 相乘,可得一个 $N\times n$ 的矩阵,该矩阵的行向量为融合了标签关联关系与用户社交关系的新的用户标签权重,为得到收敛的用户标签权重,可将矩阵 $\boldsymbol{M}_{sr}$ 与矩阵 $\boldsymbol{M}_{re}$ 的相乘多次.

矩阵 $\boldsymbol{M}_x$ 为矩阵 $\boldsymbol{M}_{sr}$ 与矩阵 $\boldsymbol{M}_{re}$ 第 x 次迭代的结果,如式(17)所示[13]:

$$\boldsymbol{M}_x=\beta\boldsymbol{M}_{re}+(1-\beta)\boldsymbol{M}_{sr}\times\boldsymbol{M}_{x-1}\tag{17}$$

其中,$\boldsymbol{M}_0=\boldsymbol{M}_{re}$ 且 $\beta\in(0,1]$. 假定在经过 δ 次迭代后,算法的精确度趋于稳定,则矩阵收敛,那么矩阵 $\boldsymbol{M}_\delta$ 的行向量为更新后的标签权重向量 $\boldsymbol{V}_i'$.

至此,更新后的标签权重向量可以更好地表示出微博用户的兴趣,从而能更加准确地进行微博推荐.

3.2.2　推荐算法描述

定义 4　对于一个在线用户 u_i,若给定一篇微博 d_p,则**排序函数** $f(u_i,d_p)$ 由如下公式定义[13]:

$$f(u_i,d_p)=\underset{P}{\boldsymbol{E}}\cdot(\boldsymbol{V}_i')^{\mathrm{T}}\tag{18}$$

其中,$f(u_i,d_p)$ 表示用户 u_i 与微博 d_p 之间的相关性. 每篇微博 d_p 可被表示为 $\underset{p}{\boldsymbol{E}}=(w_{p1},w_{p2},\cdots,w_{pn})$,若微博 d_p 包含标签 l_j,则 $w_{pj}=1$,否则 $w_{pj}=0$. $\boldsymbol{V}_i'=(w_{i1}',w_{i2}',\cdots,w_{in}')$ 为更新后的用户 u_i 的标签权重向量. 该排序函数可

广泛的用作测量微博间的关联性. 预先设定阈值 γ_i,若该排序函数的值大于阈值 γ_i,微博 d_p 将推荐给用户 u_i.

算法 1 融合标签关联关系与用户社交关系的迭代微博推荐算法

输入:需进行加标的微博用户集合 $U=\{u_1,u_2,\cdots,u_i,\cdots u_N\}$ 用户所发布的所有微博信息流 $D_i=\{d_{i1},d_{i2},\cdots,d_{iM_i}\}$,$i\in\{1,2,\cdots,N\}$ 所有待推荐用户发布的微博集合 $D=\cup_{i=1}^{N}D_i$.

输出:微博推荐序列

第一步:对于待加标的用户,通过标签检索算法,获取标签;

第二步:根据所获取的标签构建原始的 $N\times n$ 用户-标签矩阵 $\boldsymbol{M}_{ul}$,得到原始的用户标签权重;

第三步:通过挖掘多标签间的内联及外联关系,构建 $n\times n$ 的多标签关联关系矩阵 $\boldsymbol{M}_{lr}$;

第四步:通过融合多标签关联关系,更新用户-标签矩阵 $\boldsymbol{M}_{re}=\boldsymbol{M}_{ul}\times\boldsymbol{M}_{lr}$

第五步:对用户间的社交关系进行分析,构建 $N\times N$ 的多用户间社交关系相似度矩阵 $\boldsymbol{M}_{sr}$;

第六步:利用迭代公式 $\boldsymbol{M}_x=\beta\boldsymbol{M}_{re}+(1-\beta)\boldsymbol{M}_{sr}\times\boldsymbol{M}_{x-1}$ 将更新后的用户-标签矩阵 $\boldsymbol{M}_{re}$ 与用户-用户社交关系相似度矩阵 $\boldsymbol{M}_{sr}$ 迭代,直至收敛;

第七步:定义排序函数 $f(u_i,d_p)=\boldsymbol{E}_p\cdot(\boldsymbol{V}_i')^{\mathrm{T}}$,并计算其相关性;

第八步:将超过阈值 γ_i 的微博 d_p 加入推荐集,并作为推荐结果.

4 实验性能与分析

为了验证本文微博推荐方法的有效性及推荐结果的准确性,首先设计实验对本文方法进行验证,其次提出相应的评价指标对实验结果进行评价.

4.1 数据描述

本文实验数据来源于新浪微博 API. 实验采集了 8524 位用户在 2014 年 3 月 21 日至 2014 年 4 月 27 日发布的大量微博. 对实验数据进行预处理,首先去除微博文本中的链接、转发标志、@ 用户名和表情符号、简繁转换等噪音信息之后,对其进行分词,去除停用词、高频词和低频词,得实验数据集合. 其中分词采用 python 开源分词系统 stammer,停用词表采用新浪提供的 1028 个停用词. 最后去除所有在少于 10 篇文档中出现过的低频词项,得到最终的实验数据集合,该集合中微博条数为 698653,词项总数为 355765,经人工处理,分为了 20 个类,在这 20 个类别组成的数据集中,共包含 585873 条训练样本及 114000 条测试样本. 如表 2 所示.

表 2 实验数据统计信息

类别	训练样本	测试样本	类别	训练样本	测试样本
体育	41540	8000	军事	18790	3000
科技	29700	6000	育儿	25100	6000
房地产	27836	6000	环保	30053	6000
股票	23000	4000	健康	27820	6000
情感	46703	8000	旅游	35200	8000
娱乐	55430	8000	医学	24991	5000
政治	18905	3000	商品	39860	8000
宗教	13680	3000	教育	25074	5000
健身	27481	6000	美食	37440	8000
艺术	21530	4000	家装	15740	3000

4.2 实验结果与相关分析

为了验证本文方法的有效性,设计了四个实验:(1)通过实验比较,验证使用 TF * Clarity 作为标签检索方法的有效性. (2)通过改变变量 α、β 的数值,得到该微博推荐算法的性能. (3)迭代算法收敛性验证. (4)不同算法性能的比较.

4.2.1 标签检索选取策略对推荐算法的影响

该阶段推荐算法的准确性由准确率 Precision 进行对比分析. 在标签检索选取查询词作为用户标签的阶段有以下 4 种策略:TF、TF * IDF、TF * Clarity、TF * IDF * Clarity. 在每个方案中,选取得分最高的{1、3、5、7、9、11}的词作为用户的标签. 该算法的准确性由图 3 所示.

从图 3 的结果中可得出:首先,用户选择越多的标签进行标注该算法的准确性会呈现递增的趋势,如从 1 个标签增加到 7 个标签时会显示出更好的推荐效果,而标签的数量超过 7 时增长趋势越来越小. 经实验验证,应对每个用户选取 9 个标签进行标注. 其次,在以上所有策略中,用户选择超过 3 个标签时选取 TF * Clarity 的策略时,推荐算法准确性要优于其余三种策略. 由于在标签检索阶段采取的 Clarity 策略是由关键词所匹配的一系列微博与该用户所有微博集合之间的 KL 散度,如若采用 TF * IDF 与 Clarity 结合的方式,当 IDF 的值很高时,关键词匹配微博数量就会降低,从而导致 TF * IDF * Clarity 效果并不理想,由实验结果也可证实. 所以,基于以上讨论,在标签检索阶段,最终选取 TF * Clarity 的策略.

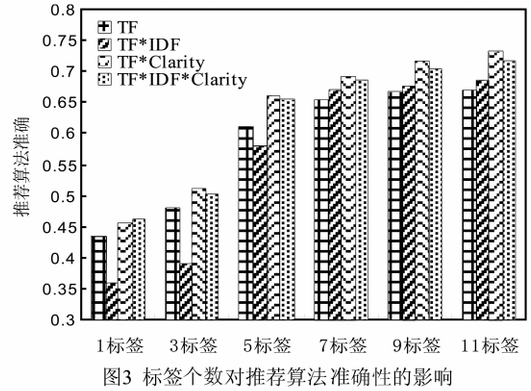


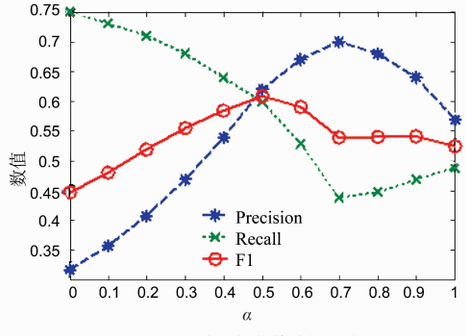
图3 标签个数对推荐算法准确性的影响

4.2.2　不同参数对推荐算法的影响

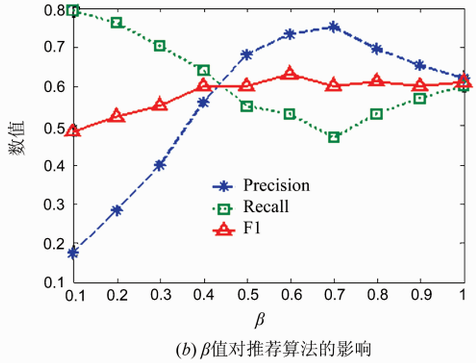
$\alpha \in [0,1]$决定了多标签间的内联关系,与外联关系所占的比例. 而$\beta \in (0,1]$决定了标签间关联关系,与用户间社交关系所占的比例. 图4(a)、(b)中可分别看出,α 与 β 值的改变会对本文的微博推荐算法的性能产生一定的影响. 当 α、β 值变化时,微博推荐算法的准确率 Precision、召回率 Recall、F1 度量值均呈现一定的变化.

如图4(a)所示,当 $\alpha = 0.5$ 时,算法的性能达到最佳状态,即标签间的内联关系与外联关系同等重要. 当 $\alpha = 1$ 时,只考虑多标签间的外联关系,该微博推荐算法的性能要优于当 $\alpha = 0$ 时,只考虑多标签间的内联关系. 由于用户所标注的标签是有限的,若仅考虑到被同一用户标注的标签间的内联关系势必不能充分挖掘出这些标签间的信息,而在用户间通过关联标签产生的标签间的外联关系中,由于关联标签较多,所以更能通过标签间的关系挖掘出用户的兴趣. 因此,仅考虑多标签间的外联关系时,微博推荐算法的性能要优于仅考虑标签的内联关系.

如图4(b)所示,当 $\beta = 0.6$ 时,算法的性能达到最佳状态. 即标签关联关系与用户的社交关系并不是同等重要. 前者对该推荐算法的贡献要比后者略大一些. 且 $\beta \neq 0$,因为当 $\beta = 0$ 时,$\boldsymbol{M}_0 = \boldsymbol{M}_{sr}$ 与文章中所设的初始值 $\boldsymbol{M}_0 = \boldsymbol{M}_{re}$ 不符,且此时只包含用户的社交关系,并不能表示出用户的兴趣. 当 $\beta = 1$ 时,$\boldsymbol{M}_0 = \boldsymbol{M}_{re}$,此时虽并未将用户的社交关系融入算法中,但可挖掘出用户的兴趣,并可进行推荐. 故 β 的取值范围为 $\beta \in (0,1]$.



(a)　α值对推荐算法的影响



(b) β值对推荐算法的影响

图4　不同参数对推荐算法的影响

4.2.3　迭代算法收敛性验证

为对迭代算法的收敛性进行验证,文章将最大迭代次数 Δ 设置为35,并计算在不同迭代次数的情况下本文算法的准确率 Precision,进而确定该算法的收敛条件.

如图5所示,当 $\delta < 7$ 时,随着迭代次数的增加,算法的精确度急剧增加,当 $\delta > 7$ 时,算法的精确度逐步趋于平稳,即算法的迭代过程趋于收敛. 因此,进行7次迭代可确保本文算法的收敛性.

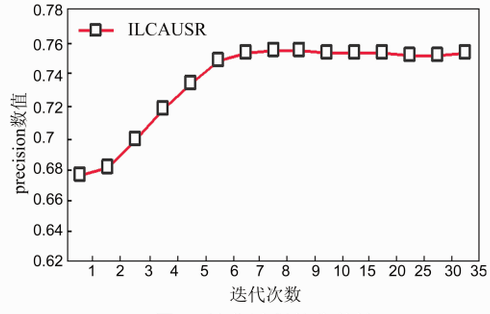


图5　迭代过程的收敛性

4.2.4　不同算法的推荐性能对比

为了验证该推荐算法的有效性,在选定参数 $\delta = 7$、$\alpha = 0.5$、$\beta = 0.6$ 时,选择朴素推荐算法(NA)(Naive Approach)[13]、标签关联关系推荐算法(LC)(Label Correlation)[10]、融合标签关系与用户关系推荐算法(未迭代)(LCAUSR)(Label Correlation And User Social Relation)与本文算法(ILCAUSR)分别从平均值 AP、准确率 Precision、召回率 Recall 以及 F1 度量值进行对比. 结果如表3所示.

表3　不同算法的推荐性能比较

	AP	Precision	Recall	F1
NA	0.45	0.531	0.63	0.58
LC	0.49	0.62	0.6	0.61
LCAUSR	0.55	0.675	0.581	0.628
ILCAUSR	0.58	0.754	0.546	0.65

由表3的推荐结果可知,与其余三种算法相比,本文的 ILCAUSR 算法在平均值、准确率以及 F1 度量值这3项指标上均取得了最好的推荐结果. ILCAUSR 算法的优越性主要体现在以下方面:

首先,采取标签检索策略,从用户以往发布过的微博中选取最具有代表性的词作为用户的标签,这些标签在初始阶段已具有代表性,并能表示用户的兴趣. 挖掘被选取的标签内联及外联关系,在很大程度上减少

了原始用户-标签矩阵的稀疏性. 不仅如此,此方法还具有更丰富的语义信息.

其次,将用户间社交关系融入微博推荐算法中,通过用户间关注信息与粉丝信息建立的社交关系网,与用户标签相结合,可较准确的表示出用户的兴趣.

最后,为得到收敛的用户标签权重,将更新后的用户-标签矩阵与用户社交关系矩阵迭代相乘多次,可更准确的揭示出用户的兴趣,并得到更好的推荐结果.

5 结束语

鉴于微博短文本的特性,本文提出了一种融合多用户间社交关系和多标签关联关系的迭代的微博推荐算法. 首先,对众多用户中未加标签以及添加少量标签的用户采用标签检索的策略对其进行加标;其次,挖掘多标签的内联关系与外联的关系;最后,分析用户间社交关系,并将其与标签关联关系相结合,进行微博推荐.

参考文献

[1] Tang J L, Wang X F, et al. Enriching short text representation in microblog for clustering [J]. Frontiers of Computer Science, 2012, 6(1): 88 – 101.

[2] Sun A. Short text classification using very few words [A]. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. Portland, Oregon, USA: ACM, 2012. 1145 – 1146.

[3] Li X, Guo L, et al. Tag-based social interest discovery [A]. Proceedings of the 17th International Conference on World Wide Web [C]. New York, USA: ACM, 2008. 675 – 684.

[4] 吴永辉,王晓龙,等. 基于主题的自适应、在线网络热点发现方法及新闻推荐系统[J]. 电子学报,2010,38(11):2620-2624.
Wu yonghui, Wang xiaolong, et al. Adaptive on-line web topic detection method for web news recommendation system [J]. Acta Electronica Sinica, 2010, 38(11): 2620 – 2624. (in Chinese)

[5] 黄震华,张波,等. 一种社交网络群组间信息推荐的有效方法[J]. 电子学报,2015,43(6):1090 – 1093.
Huang zhenhua, Zhang bo, et al. an efficient algorithm of information recommendation between groups in social networks [J]. Acta Electronica Sinica, 2015, 43(6): 1090 – 1093. (in Chinese)

[6] 张引,张斌,等. 面向自主意识的标签个性化推荐方法研究[J]. 电子学报,2012,40(12):2353 – 2359.
Zhang yin, Zhang bin, et al. Autonomy oriented personalized tag recommendation [J]. Acta Electronica Sinica, 2012, 40(12): 2353 – 2359. (in Chinese)

[7] Yamaguchi Y, Amagasa T, et al. Tag-based user topic discovery using twitter lists [A]. Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM) [C]. Kaohsiung: IEEE, 2011. 13 – 20.

[8] Abel F, Gao Q, et al. Analyzing user modeling on Twitter for personalized news recommendations [A]. Proceedings of 19th International Conference UMAP on User Modeling, Adaption and Personalization [C]. Girona, Spain: UMAP, 2011. 1 – 12.

[9] Liu Q W, Xiong Y, Huang W C. Combining user-based and item-based models for collaborative filtering using stacked regression [J]. Chinese Journal of Electronics, 2014, 23(4): 712 – 717.

[10] Ma H F, Jia M H Z, et al. A Microblog Recommendation Algorithm Based on Multi-tag Correlation [A]. Knowledge Science, Engineering and Management [C]. Springer International Publishing, 2015. 483 – 488.

[11] Liu Z Y, Chen X X, et al. Mining the interests of Chinese microbloggers via keyword extraction [J]. Frontiers of Computer Science, 2012, 6(1): 76 – 87.

[12] Wu W, Zhang B, et al. Automatic generation of personalized annotation tags for twitter users [A]. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics [C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 689 – 692.

[13] Zhou X K, Wu S, et al. Real-time recommendation for microblogs [J]. Information Sciences, 2014, 279: 301 – 325.

[14] 徐志明,李栋,等. 微博用户的相似性度量及其应用[J]. 计算机学报,2014,37(1):207 – 218.
Xu Zhiming, Li Dong, et al. Measuring similarity between microblog users and its application [J]. Chinese Journal of Computes, 2014, 37(1): 207 – 218. (in Chinese)

作者简介

马慧芳 女,1981 年 7 月出生,甘肃兰州人. 2010 年获中国科学院计算技术研究所计算机软件与理论博士学位,现为西北师范大学计算机科学与工程学院副教授、硕士生导师. 研究领域为数据挖掘与机器学习.
E-mail: mahuifang@ yeah. net

贾美惠子 女,1991 年 4 月出生,陕西宝鸡人. 现为西北师范大学计算机科学与工程学院硕士研究生. 研究方向为互联网数据挖掘与机器学习.
E-mail: jmhuizi@ yeah. net