计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20131551
52(5):1014-1021,2015

社区热点微博推荐研究

彭泽环[1] 孙 乐[1,2] 韩先培[1,2] 陈 波[1]
[1](中国科学院软件研究所基础软件国家工程研究中心 北京 100190)
[2](计算机科学国家重点实验室(中国科学院软件研究所) 北京 100190)
(pengzehuan@yahoo.cn)

Community Hot Statuses Recommendation

Peng Zehuan[1], Sun Le[1,2], Han Xianpei[1,2], and Chen Bo[1]
[1](*National Engineering Research Center of Fundamental Software*, *Institute of Software*, *Chinese Academy of Sciences*, *Beijing* 100190)
[2](*State Key Laboratory of Computer Science*(*Institute of Software*, *Chinese Academy of Sciences*), *Beijing* 100190)

Abstract Micro-blog recommendation is an effective technique to resolve the information overload problem in micro-blog systems. In this paper, we summarize and model several key factors which affect a user's interest on a specific status, including implicit features, content features (i.e., content similarity, user tags, and user's favorites), social network features, and status features. Based on the above features, we propose a community hot status recommendation algorithm—CMR (community micro-blog recommendation), which combines both explicit features and implicit features for better recommendation. Specifically, we propose a learning method to rank based framework, which learns a user's interest model of status from his preference data, including his retweets, favorites, comments, etc. Then new statuses are scored and ranked using the learned interest model. In order to measure our method's performance, we conduct a series of experiments in three community data sets (including NLP, Photography and Basketball). Experimental results show that: 1) by combining both implicit features and explicit features, our method achieves better recommendation performance than that using a single type of features; 2) compared with the MRR (micro-blog repost rank based recommendation), CMR gets better recommendation performance; 3) MRR prefers recommending hot statuss in the whole micro-blog system, in contrast CMR usually recommends community-specific statuses.

Key words micro-blog; recommendation; community; latent factor model; information overloading

摘 要 分析并总结了影响用户对特定微博兴趣的若干因素,在此基础上基于潜在因素模型提出了1个融合显式特征和潜在特征的社区热点微博推荐算法(community micro-blog recommendation, CMR),并将其用于发现微博兴趣社区热点信息.算法在3个兴趣社区上进行了实验,结果表明:1)融合2种特征信息的微博推荐效果好于使用单一特征信息的推荐;2)CMR的推荐效果好于基于转发次数的对照实验(micro-blog repost rank based recommendation, MRR);3)通过分析各个算法所推荐的微博内容,发现CMR倾向于为用户推荐兴趣社区相关微博,而MRR倾向于为用户推荐公共热点微博.

关键词 微博;推荐;社区;潜在因素模型;信息过载

中图法分类号 TP391

收稿日期:2013-11-29;**修回日期**:2014-10-10

基金项目:国家自然科学基金项目(61433015,61272324);国家"八六三"高技术研究发展计划基金项目(2015AA015405);网络文化与数字传播北京市重点实验室开放课题(ICDD201204)

微博是目前最热门的互联网应用之一,吸引了数以亿计的用户.用户通过微博系统可自由地关注感兴趣的人,同时发布、分享、评论感兴趣的信息.可以说微博已经成为与人们生活息息相关的信息平台.但是随着用户关注人数的增多,相应的问题也逐渐凸显出来,这些问题主要包括:

1) 信息过载问题.1 个关注几百个好友的用户,每天出现在首页动态栏中的最新微博消息可能达到几百到几千,用户需要花费大量的时间去"刷微博",也就是一一查看这些消息以发现其中有用的信息,研究表明[1]微博中大部分信息是用户间的对话、心情微博、日常琐事等无用的信息,查看这类信息占据了微博用户的大量时间.

2) 信息冗余展示问题.有着相同兴趣爱好的用户间常常会相互关注,形成 1 个相互交织的关注关系网.上述结构通常被称为兴趣社区(圈子),如图 1 所示:

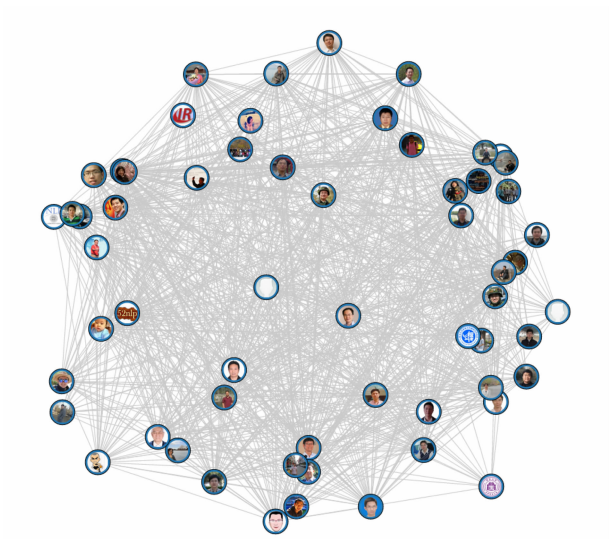

Fig. 1 A graph of micro-blog interest community.

图 1 微博兴趣社区示意图

兴趣社区中存在大量如图 2 所示的子结构,其中用户 0 关注了用户 1 和用户 2,用户 1 也关注了

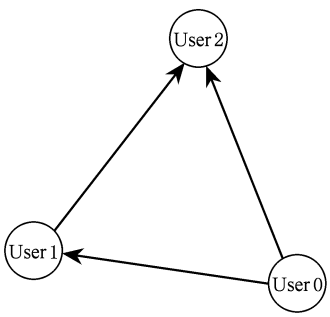

Fig. 2 A graph of triangle following relation.

图 2 微博用户三角关注关系图

用户 2.社区中存在的上述子结构导致了社交信息冗余,即如果用户 2 发布 1 条微博 A,A 将出现在用户 0 和用户 1 的动态栏中,若用户 1 对这条信息很感兴趣,转发了 A,那么信息 A 将会 2 次出现在用户 0 的动态栏中.

此外,作为微博最重要的特性之一,时效性也给用户获取信息带来时效性差的重要信息被覆盖的问题.也就是,用户首页动态栏通常按照时间顺序呈现微博,时效性差的信息因此往往会被时效性强的信息覆盖,这可能导致用户错过一些重要但时效性不强的信息.

3) 无法获得关注人之外重要信息的问题.通常具有某个兴趣的用户会对整个兴趣社区的热点微博感兴趣.虽然用户可以自由地关注兴趣社区的用户,但用户一般仅关注兴趣社区中很少的人,所以常常会错过兴趣社区中的热点微博.针对上述问题本文从微博用户兴趣社区(圈子)的角度为用户推荐感兴趣的热点微博,帮助用户发现有用的信息,以期在某种程度上解决用户社交信息过载问题.

1 相关工作

微博中可推荐的内容信息很多,例如热点话题、图片、视频等.内容推荐是传统推荐系统的主要任务之一,也是微博不可或缺的组成模块.内容推荐帮助用户发现感兴趣的用户微博、话题、热门事件、标签以及多媒体信息,过滤兴趣无关的信息,从而一定程度上解决社交过载问题.

Liu 等人[1]通过分析 Google News 登录用户的点击日志建立用户兴趣模型,并结合基于内容的推荐方法和协同推荐方法,构建一个贝叶斯用户新闻推荐系统.日志分析表明用户的新闻兴趣随着时间变化并且受到局部新闻变化趋势影响.实验表明该混合推荐系统优于已存在的协同过滤推荐系统,有效地提高了新闻推荐的质量,为 Google News 网站带来了更多访问流量.Lerman 等人[2]分析社交新闻投票网站 Digg 的新闻推荐,将 Digg 中的好友接口视为社交推荐系统,向用户推荐其好友发布的新闻和好友标记喜欢的新闻.

直观上,人们倾向于信任朋友、熟人的推荐.借助社交网络中的好友可以有效地提高内容推荐的质量.Sinha 等人[3]比较在线系统和朋友对书籍和电影的推荐效果,结果表明好友推荐质量高于在线系统的推荐,在线系统推荐的物品通常比较"新"、"未

预见”,好友推荐的物品倾向于和用户以往兴趣相关. Koren 等人[4]针对电影推荐问题,采用基于矩阵分解的协同排序方法,将用户和电影建模为同一潜在特征空间的多项式分布,通过计算用户和电影的特征向量内积衡量用户对电影的兴趣程度,该方法因其优秀的推荐性能获得 NetFlix 大赛的大奖. 本论文正是受到该方法的启发,但和 Koren 等人的方法不同,本文不是将微博表示为潜在特征空间的向量,而是将微博中的关键词表示为潜在特征空间向量,这样有效地解决了用户微博数据稀疏问题.

Chen 等人[5]分析采用 3 种因子(线索长度(thread length),主题相关度(topic relevance),连接强度(tie-strength))向 Twitter 用户推荐感兴趣的会话,当用户信息需求是社交目的时基于 Tie-Strength 因子的算法和其他算法相比推荐效果明显偏好. Yan 等人[6]提出一个基于图论的 tweet 推荐模型,模型基于 3 种网络同时对 tweet 和用户进行协同排序,3 种网络分别为用户社交网、tweet 网、用户 tweet 混合网络,基于 tweet 和用户在排序时相互加强的假设采用协同排序算法以向用户推荐感兴趣的 tweet.

除上述以外,研究人员还提出大量的社交网络推荐算法和系统. Koga 等人[7]基于 Twitter 用户的关注、发布和转发等信息建模用户潜在话题模型,话题模型采用隐含狄利克雷分配(latent Dirichlet allocation, LDA)[8]建模,推荐时计算用户话题模型的 Kullback-Leibler(KL)距离,KL 距离越短,用户间潜在联系越强推荐得分越高. Peng 等人[9]通过研究分析 3 种典型的特征,基于排序学习算法提出一个 Twitter 微博排序策略. Geyer 等人[10]比较在一个企业社交网络中的推荐“about you”相关对象的方法,证明基于社交关系的推荐效果好于基于内容的推荐. Guy 等人[11]认为在社交网络中可以为用户构建 2 个子网络:相似用户网络(similarity)、熟悉用户网络(familiarity),通过比较 Lotus 系统中 2 个子网络的推荐情况,证明基于熟悉用户网络的推荐效果要好于传统的基于相似用户网络的推荐,这也与现实世界中人们的认知习惯相符合.

2 社区热点微博推荐

2.1 用户微博评分模型

影响用户对 1 条微博感兴趣的因素很多,比如微博的内容、微博的发布者与用户是否是熟人、用户关注的群体对微博的关注程度等等. 这些因素总的来说可分为 2 类:潜在特征(主题、社交关联)和显式特征(关系特征、内容相关特征、发布者权威特征等). 为综合上述 2 类特征,本文基于文献[12]提出计算用户对微博兴趣度评分:

$$\hat{r}_{ui} = \alpha \sum_{j \in J} \boldsymbol{b}_j^u w_j^{ui} + (1-\alpha) \frac{\boldsymbol{p}_u}{|T_i|} \sum_{t \in T_i} \boldsymbol{q}_t, \quad (1)$$

其中,$\alpha \in (0,1)$是特征权重参数,用于控制兴趣度中潜在特征和显式特征所占的比例;u 表示用户;i 表示微博;J 表示〈用户,微博〉对的显式特征空间;w_j^{ui} 表示基于〈用户,微博〉对的显式特征值(将在 2.3 节详细介绍这些特征及意义);$\boldsymbol{b}_j^u$ 表示显式特征值对应的权重. 一个显式特征对不同的用户具有不同权重,例如有的用户偏向关注粉丝数较高的用户发布的微博,有的用户偏向关注转发次数较多的微博;$\boldsymbol{p}_u$ 是用户在潜在因素空间中的向量表示,可以理解为表示用户在若干个 topic 上的兴趣分布;T_i 是微博 i 中的词的集合(去掉了停用词、和微博表情符号),$|T_i|$是微博 i 中词的个数,用于对微博长度进行惩罚;$\boldsymbol{q}_t$ 是微博 i 中的词语 t 在潜在因素空间的向量表示,可以理解为该词在用户 topic 空间的概率分布. $\sum_{j \in J} \boldsymbol{b}_j^u w_j^{ui}$ 部分建模显式特征对微博评分的影响,$\frac{\boldsymbol{p}_u}{|T_i|} \sum_{t \in T_i} \boldsymbol{q}_t$ 建模潜在特征对微博评分的影响.

2.2 潜在特征

本文采用 $\frac{\boldsymbol{p}_u}{|T_i|} \sum_{t \in T_i} \boldsymbol{q}_t$ 计算词语与用户在潜在空间中的关联,其主要思想是将用户是否对一条微博感兴趣转化为用户是否对微博中的特定词语感兴趣[12]. 这样做的好处有 2 个:1)可以较直接地将用户兴趣与词关联;2)用户的名字也常常出现在微博中,如果 t 是出现在微博中的名字,则词语 t 的潜在向量可以直接表示用户的兴趣. 例如用户“@白硕”发布的微博中常常出现用户“@马少平 THU”名字,词“@马少平 THU”的潜在向量就可以用来描述该用户的表兴趣.

2.3 显式特征

影响用户对一条微博兴趣的因素很多,在文献[12]的基础上,本文总结了微博中 5 类影响用户对微博的兴趣的显式特征,它们包括内容特征、社交关系特征、微博特征、微博发布者特征和微博转发者特征. 下面详细介绍这些特征,并分析每类特征的意义和作用.

1) 内容特征

该特征描述微博 s 的内容和用户 u 兴趣的关联度. 用户以往发布或者转发的微博在一定的意义上表征用户的兴趣,如果一条微博和这些以往发布的

微博内容相似,则可以认为用户对这条微博有兴趣.基于内容的特征包括以下3种:

① 微博相似度

这个特征衡量微博 s 与用户 u 以往发布的微博的相似度.本文计算微博 s 与每一条用户历史微博的相似度,然后将这些相似度相加计算该特征值:

$$relevance(s,u)=\sum_{us\in statuses(u)} statusesrelevance(s,us), \tag{2}$$

其中,$statuses(u)$表示用户以往发布的微博(包括转发的),$statusesrelevance(s,us)$表示微博 s 和微博 us 的相似度,可以采用信息检索领域经典的词频-逆文档频率(term frequency-inverse document frequency,TFIDF)文档权重向量的余弦相似度计算.

② 用户标签

用户标签是用户在完善个人资料时指定的一组描述用户兴趣爱好的关键字,是表征用户兴趣爱好的一种有效手段.如果一条微博中出现了标签中的关键字,则可以认为用户对这条微博有兴趣,且出现的关键字越多用户对这条微博越感兴趣.本文用该特征表示微博中包含用户标签中关键词的个数.

③ 用户收藏

微博中用户看到一条很感兴趣的微博,可以收藏之,便于日后重复查看.通常可以认为用户收藏一条微博表明用户对这条微博十分感兴趣,其感兴趣程度高于转发、评论的微博.

2) 社交关系特征

社交关系特征中的关系指用户 u 与微博发布者 p 的社交关系.关系特征通过分析用户与发布者在社交网络中的关系来衡量用户对微博兴趣度,本文使用如下3个社交关系特征.

① 共同关注得分

这个特征描述用户与发布者的关注的好友集合的相似度,通常二者关注的好友集合重合度越高,二者兴趣关联越大.对于用户 u 和发布者 p,二者的共同关注得分可以采用Jaccard相似度计算:

$$cofollowee_score=\frac{|followee(u)\cap followee(p)|}{|followee(u)\cup followee(p)|}, \tag{3}$$

其中,$followee(u)$表示用户关注的好友集合,$followee(p)$表示微博发布者关注的好友集合.

② ufollowee_pfollower

这个特征描述了用户 u 关注的好友中关注发布者 p 的人数,直观上如果 u 的好友中关注发布者 p 的人越多,则意味着二者的兴趣关联越大.

③ 双向关注关系

表示用户与发布者是否为双向关注关系,即用户关注了发布者,发布者也关注了用户,这是一种较强的社交关系,双向关注的用户与发布者通常是熟人、一个领域的人或者具有共同的兴趣爱好等.

3) 微博特征

微博本身的一些属性,也会决定用户对微博的兴趣程度.用户通常都偏好一些高质量、信息量大的微博,下面的这些特征可以较好地描述微博包含的信息量.

① Hash标签数

Hash标签指发布者在发布微博时用2个#号包含突出显示的一个文本片段.Hash标签通常都含有较大的信息量,可以作为对这个微博的一个总结.微博中的Hash标签数越多,信息量越大,用户兴趣度越高,该特征描述微博中Hash标签的个数.

② 是否有url

这个特征表示微博中是否含有url链接,发布者在编辑微博时常常附上链接url,作为对微博内容的补充或者详情.包含url的微博,信息量通常较大.

③ 转发次数

转发次数描述了这个微博系统中用户对这条微博的热度的投票,转发次数越多,这个微博越可能是热点微博,用户对这个微博感兴趣的可能性也越大.

④ 评论次数

这个特征一般和转发次数特征有着较强的关联,表示用户对这个热点消息的关注度,转发的次数较多的微博一般被评论次数也比较多.

⑤ 首句是否以“【】”开始

这个特征描述微博是否包含用“【】”标记的主题句,包含则该特征为1,不包含该特征值为0.微博主题句类似新闻信息的标题,是整条微博有着主要内容的概括.包含“【】”首句的微博消息一般都类似新闻消息,因此用户对这个样的微博感兴趣的可能性也较大.

4) 微博发布者特征

微博发布者的特征包括以下3种:

① 发布者的粉丝数

该特征记录关注发布者的人数,基于此排名生成微博人气排行榜,该特征是对用户权威度最直观的反映.

② 关注者数

该特征表明用户关注的好友数是用户活跃度的

一个指标.越活跃的用户，发布的信息传播的范围也越广，活跃的用户更能引起关注.

③ 发布微博数

该特征表明用户发布的微博总数，也可以作为用户活跃度的一个指标.

5）微博转发者特征

和微博发布者特征类似，信息转发者的权威度和活跃度也会影响用户对一条微博的感兴趣程度，例如一条被李开复转发的微博很可能会吸引用户的关注.本文使用转发者粉丝数、好友数和发布微博数3个特征衡量转发者的权威度和活跃度.

2.4 评分模型求解

训练数据并没有显式地表示用户对微博的兴趣度评分，但是系统可以获知用户对某条微博的兴趣偏好，例如对一条用户转发、评论或者收藏的微博和一条没有任何动作的微博，上述数据表明用户对前者更有兴趣.这种“对偏好”机制类似排序学习中的pairwise机制，用户 u 对这样的微博对 $\langle k,h\rangle$（对前者更有兴趣）的偏好程度可以采用式(4)建模.

$$P\left(pref(k) > pref(h) \mid u = \frac{1}{1+\mathrm{e}^{-(\hat{r}_{uk}-\hat{r}_{uh})}}\right), \tag{4}$$

其中，$pref(k)$ 是用户对一条微博的兴趣程度.系统可以获得用户偏好对集合 $D=\{\langle u,k,h\rangle \mid k\in R(u), h\notin R(u)\}$，$R(u)$ 表示用户转发、评论和收藏的微博集合.为了在 D 上学习出评分模型，模型要解决的问题是最大化 D 上所有对的偏好度之和的对数似然，其计算如下：

$$\ell = \ln \sum_{\langle u,k,h\rangle \in D} P(pref(k) > pref(h) \mid u). \tag{5}$$

这个问题可以进一步转化为最小化式(6)的最优化问题，其中 λ_1，λ_2 调节拟合数据的程度，$\lambda_2 \| \boldsymbol{b}^u \|^2$ 和 $\lambda_1(\| \boldsymbol{p}_u \|^2 + \sum_{t\in T_h} \| \boldsymbol{q}_t \|^2 + \sum_{t\in T_k} \| \boldsymbol{q}_t \|^2)$ 分别是 L_2 正规化惩罚因子.

$$\ell = \sum_{\langle u,k,h\rangle \in D} \ln(1+\mathrm{e}^{-(\hat{r}_{uk}-\hat{r}_{uh})}) + \lambda_1(\| \boldsymbol{p}_u \|^2 + \sum_{t\in T_h} \| \boldsymbol{q}_t \|^2 + \sum_{t\in T_k} \| \boldsymbol{q}_t \|^2) + \lambda_2(\| \boldsymbol{b}^u \|^2). \tag{6}$$

本文采用随机梯度下降算法求解式(6)的最优化问题，每个变量的梯度分别如式(7)～(10)所示，其中，$e_{ukh}=1-P(pref(k)>pref(h)\mid u)$.

$$\frac{\partial \ell}{\partial p_{un}} = -e_{ukh}\left(\frac{\sum_{t\in T_k} q_{tn}}{|T_k|} - \frac{\sum_{t\in T_h} q_{tn}}{|T_h|}\right) + \lambda_1 p_{un}, \tag{7}$$

$$\frac{\partial \ell}{\partial q_{tn}} = -e_{ukh}\frac{p_{un}}{|T_k|} + \lambda_1 q_{tn},\ t\in T_k, \tag{8}$$

$$\frac{\partial \ell}{\partial q_{tn}} = e_{ukh}\frac{p_{un}}{|T_h|} + \lambda_1 q_{tn},\ t\in T_h, \tag{9}$$

$$\frac{\partial \ell}{\partial b_j^u} = -e_{ukh}(w_j^{uk} - w_j^{uh}) + \lambda_2 b_j^u, \tag{10}$$

注意这里 q_{tn} 有2个梯度，一个是 $t\in T_k$，另外一个是 $t\in T_h$，一条微博既可能出现在偏好对的左边，也可能出现在偏好对的右边.上述梯度对应变量的更新公式分别为式(11)～(14)所示：

$$p_{un} \leftarrow p_{un} - \gamma\frac{\partial \ell}{\partial p_{un}}, \tag{11}$$

$$q_{tn} \leftarrow q_{tn} - \gamma\frac{\partial \ell}{\partial q_{tn}},\ t\in T_k, \tag{12}$$

$$q_{tn} \leftarrow q_{tn} - \gamma\frac{\partial \ell}{\partial q_{tn}},\ t\in T_h, \tag{13}$$

$$b_j^u \leftarrow b_j^u - \gamma\frac{\partial \ell}{\partial b_j^u}. \tag{14}$$

在训练数据集 D 上采用上述算法学习后，系统就可以获取每个用户和微博中词语在潜在因素空间的向量以及每个用户的显式特征的权重，然后根据式(1)就可以估计出用户对每条微博的兴趣度得分.

2.5 社区热点微博发现

根据式(1)，系统可以预测用户对某条微博的兴趣度评分.微博社区热点微博推荐过程如下：

1）抽取一个社区所有用户当日发布的所有微博 I；

2）计算社区中用户 $u(u\in U)$ 对每条微博 $i(i\in I)$ 的评分 $\hat{r}_{ui}$；

3）计算社区中每条微博 $i(i\in I)$ 的平均评分 $\hat{r}_i = \sum_{u\in U}\frac{\hat{r}_{ui}}{|U|}$；

4）根据每条微博的平均评分的高低对所有微博 I 重排序，将前若干条推荐给用户.

3 实 验

为进行试验，作者使用3个微博用户兴趣社区进行实验，分别是：自然语言处理社区(357人)，简称为自然语言处理(natural language processing，NLP)；摄影爱好者社区(200人)，简称Photography；篮球社区(280人)，简称Basketball.社区用户选取方式如下：1)手动选取具有特定社区标签的用户，作为相应社区的种子用户；2)手动过滤种子用户的关注者和被关注者，只保留真正的社区用户.每个兴趣社区内部的用户相互关注密集.本文使用新浪微博

开放 API 抽取每个兴趣社区用户当天发布的最新微博以及用户过去的一段时间发布和转发的所有微博,此外本文还抽取了兴趣社区的关注关系以及用户标签.

3.1 数据集

为了进行实验,本文连续 5 天抽取 3 个兴趣社区用户当日发布的微博集合(测试数据)以及用户在过去的一段时间发布的所有微博集合(训练数据).每个社区每天的微博数统计如表 1～3 所示,其中表 1 是每个社区每天抽取的训练用微博条数;表 2 是根据每个社区训练数据生成的用于训练的偏好对个数;表 3 是每个社区每天发布的新微博数,即测试用数据,也就是要推荐的候选微博集合.

Table 1 The Number of Micro-blog in Training Set

表 1 训练集微博数目统计量

Day	NLP	Photography	Basketball
The 1st Day	18 623	10 088	14 281
The 2nd Day	18 677	10 126	14 395
The 3rd Day	18 782	10 198	14 623
The 4th Day	18 788	10 267	14 713
The 5th Day	18 795	10 314	15 001

Table 2 The Number of Preference-pair in Training Set

表 2 训练集偏好对统计量

Day	NLP	Photography	Basketball
The 1st Day	13 199	5 531	9 177
The 2nd Day	13 206	5 702	9 514
The 3rd Day	13 250	5 617	9 240
The 4th Day	13 266	5 868	9 301
The 5th Day	13 302	5 690	9 416

Table 3 The Number of Micro-blog in Testing Set

表 3 测试集微博数目统计量

Day	NLP	Photography	Basketball
The 1st Day	340	244	345
The 2nd Day	359	256	423
The 3rd Day	362	264	458
The 4th Day	316	320	526
The 5th Day	354	370	614

3.2 对比实验及评价方式

本文提出的社区热点微博推荐算法(CMR)每天在每个社区中检测出 N 条热点微博,记为集合 P_{CMR}.为了验证 CMR 算法的有效性,本文在表 1～3 的数据上进行了一系列实验,并与 3 种对比系统比较实验结果,这 3 个对比系统分别是:

1) 对比系统 1.仅使用显式特征推荐,记为 ECMR (explicit community micro-blog recommendation),返回圈子 N 条热点微博,记为 P_{ECMR};

2) 对比系统 2.仅使用潜在特征推荐,记为 LCMR(latent community micro-blog recommendation),返回圈子 N 条热点微博,记为 P_{LCMR};

3) 对比系统 3.根据当日微博的转发(评论)次数的高低对所有微博排序,记为 MRR,返回其中前 N 条热门微博,记为 P_{MRR}.

本文合并微博集合 P_{CMR},P_{ECMR},P_{LCMR} 和 P_{MRR},假设为 M 条,然后将这 M 条微博发给圈子中的 K 个用户,让用户手工标记出感兴趣的微博.假设各个用户标记的微博集合为 $M_1, M_2, \cdots, M_k$,本文选取至少被 2/3 用户标记为感兴趣的微博作为整个社区用户感兴趣微博,记为集合 M_{interest}.

为了评价各个系统的推荐效果好坏,本文采用信息检索领域的检索准确度评价指标 $P@N$,每个系统推荐性能的 $P@N$ 指标计算方法如式(15)所示:

$$P@N = \frac{|M_{\mathrm{interest}} \cap P_{\mathrm{system}}|}{5N}, \tag{15}$$

其中,M_{interest}表示当天社区微博中用户认为有兴趣的微博,P_{system}是相应的推荐系统推荐的 N 条热点微博,系统连续 5 d(对应于索引 d)的推荐性能采用平均准确度 AP(average accuracy)指标计算,如式(16)所示:

$$AP = \sum_{d=1}^{5} \frac{|M_{\mathrm{interest}}^{d} \cap P_{\mathrm{system}}^{d}|}{5N}. \tag{16}$$

3.3 实验结果及分析

本文选择 N 值为 5,10,15 和 20 这 4 种情况,系统使用 5 个用户对结果进行人工评价($K=5$),各系统在不同社区不同 N 值上的推荐性能结果如图 3～5 所示.

由图 3～5 的实验结果我们可以得出以下结论:

1) 融合潜在特征和显式特征的社区热点微博推荐算法 CMR 可以较有效地为用户推荐热点微博信息;

2) 推荐系统 ECMR 和 LCMR 的推荐性能均不如 CMR 的推荐效果,验证了融合 2 种信息特征联合推荐的效果要好于仅仅采用一种信息特征的推荐效果;

3) CMR 的推荐性能好于基于转发(评论)次数的推荐系统 MRR.

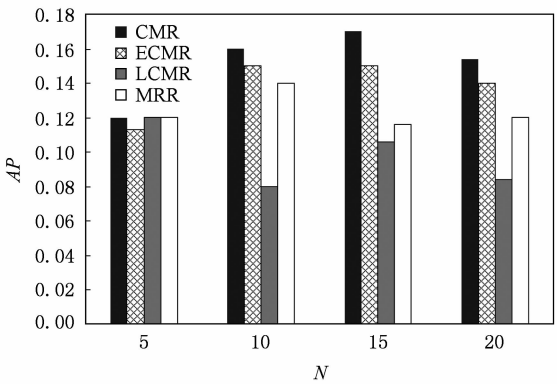

Fig. 3 Recommendation performance of 4 systems in NLP community.

图 3 各推荐系统在 NLP 圈子中推荐性能

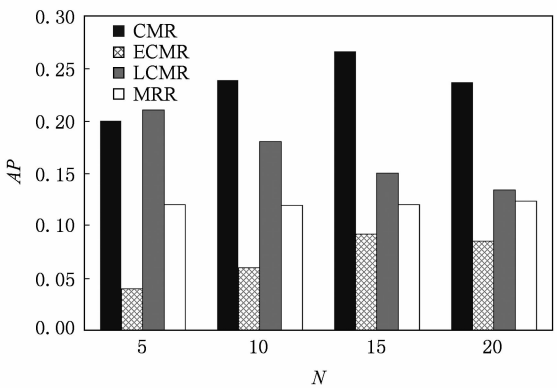

Fig. 4 Recommendation performance of 4 systems in photograph community.

图 4 各推荐系统在摄影圈子中推荐性能

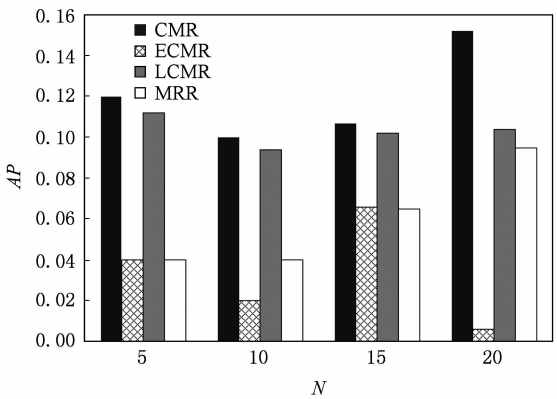

Fig. 5 Recommendation performance of 4 systems in basketball community.

图 5 各推荐系统在篮球圈子中推荐性能

为了进一步解释上述实验结果，我们分析了 CMR 算法和 MRR 算法返回的热点微博内容，将返回的热点微博分为公共热点微博、社区相关热点微博和其他 3 类. 公共热点微博通常是社会热点新闻实事相关等大部分用户感兴趣的微博；社区相关热点微博是兴趣社区内部用户较为感兴趣的微博信息，社区外的用户通常没有兴趣. CMR 算法和 MRR 算法返回的微博中各类微博的比例如图 6 所示：

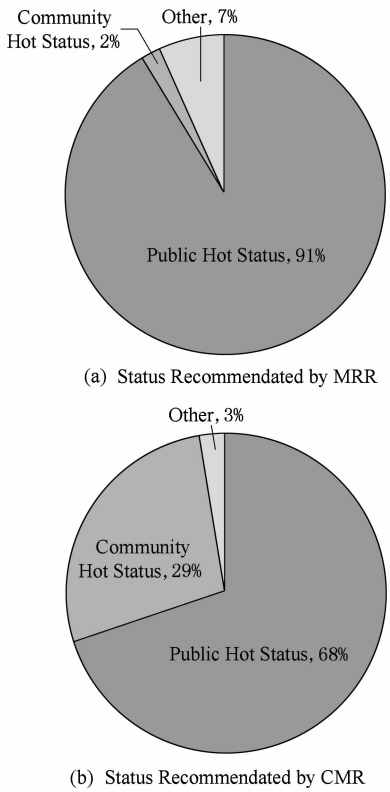

Fig. 6 The proportion of recommended status by CMR and MRR systems.

图 6 CMR 和 MRR 返回的各类微博的比例

根据图 6 可以得出 3 个结论：

1) 公共热点微博是用户普遍感兴趣的微博信息，公共热点微博占 MRR 系统和 CMR 系统推荐热点微博的大部分；

2) 推荐系统 MRR 倾向于推荐公共热点微博；

3) 推荐系统 CMR 可以同时推荐公共热点微博和社区相关的热点微博，社区相关的热点微博通常对用户是更“有用”的信息，CMR 算法更真实地反映某个社区中用户当日关注的热门微博.

4 结论及下一步工作

本文提出了社区热点微博推荐算法 CMR. 该算法具有如下 3 方面的特点：1) 基于显式特征预测用户对微博的显式评分；2) 利用用户和微博的潜在语义特征，采用潜在因素模型将用户和微博词语建模为潜在特征空间的向量，计算潜在因素的评分；3) 融

合 2 个评分预测用户对微博的兴趣评分. 基于这个评分算法可以发现一个兴趣社区的当天热点微博并进行推荐.

因微博数据获取的限制,实验中的 3 个圈子都是预设定的微博用户社交图,这些圈子中的用户是整个微博中该兴趣社区用户集合的子集,如何自动化地从整个微博系统中发现这些兴趣社区是将来的研究问题之一;此外,如何进行特征选择构建高效的微博信息推荐框架也是下一步的研究方向之一.

参 考 文 献

[1] Liu Jiahui, Dolan P, Pedersen E R. Personalized news recommendation based on click behavior [C] //Proc of the 15th Int Conf on Intelligent User Interfaces. New York: ACM, 2010: 31-44

[2] Lerman K. Social networks and social information filtering on digg [J]. Journal IEEE Internet Computing Archive, 2007, 11(6): 16-28

[3] Sinha R R, Swearingen K. Comparing recommendations made by online systems and friends [C] //Proc of the 2nd DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries. Dublin: Dublin City University, 2001

[4] Koren Y. The bellkor solution to the netflix grand prize [EB/OL]. 2009 [2013-11-29]. http://www. netflixprize. com/assets/GrandPrize2009_BPC_BellKor. pdf

[5] Chen Jilin, Nairn R, Chi E. Speak little and well: Recommending conversations in online social streams [C] // Proc of the 2011 SIGCHI Conf on Human Factors in Computing Systems. New York: ACM, 2011: 217-226

[6] Yan Rui, Lapata M, Li Xiaoming. Tweet recommendation with graph co-ranking [C] //Proc of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2012: 516-525

[7] Koga H, Taniguchi T. Developing a user recommendation engine on twitter using estimated latent topics [C] //Proc of the 14th Int Conf on Human-Computer Interaction: Design and Development Approaches. New York: ACM, 2011: 461-470

[8] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3(4/5): 993-1022

[9] Peng Zehuan, Sun Le, Han Xianpei. Micro-blog user recommendation using learning to rank [J]. Journal of Chinese Information Processing, 2013, 27(4): 96-102 (in Chinese)
(彭泽环, 孙乐, 韩先培. 基于排序学习的微博用户推荐[J]. 中文信息学报, 2013, 27(4): 96-102)

[10] Geyer W, Dugan C, Millen D R, et al. Recommending topics for self-descriptions in online user profiles [C] //Proc of the 2008 ACM Conf on Recommender Systems. New York: ACM, 2008: 59-66

[11] Guy I, Zwerdling N, Carmel D, et al. Personalized recommendation of social software items based on social relations [C] //Proc of the 3rd ACM Conf on Recommender Systems. New York: ACM, 2009: 53-60

[12] Chen Kailong, Chen Tianqi, Zheng Guoqing, et al. Collaborative personalized tweet recommendation [C] //Proc of the 35th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2012: 661-670

Peng Zehuan, born in 1987. Graduate student at the Institute of Software, Chinese Academy of Sciences from 2013. His main research interests include information storage and information retrieval.

Sun Le, born in 1971. Received his PhD degree from Nanjing University of Science and Technology in 1998. Currently professor and PhD supervisor in the Institute of Software, Chinese Academy of Sciences. His main research interests include information retrieval and natural language processing.

Han Xianpei, born in 1984. Received his PhD degree from Institute of Automation, Chinese Academy of Sciences in 2010. Currently associate professor in the Institute of Software, Chinese Academy of Sciences. His main research interests include information extraction, knowledge base population and natural language processing.

Chen Bo, born in 1989. Graduate student in the Institute of Software, Chinese Academy of Sciences. His main research interests include Semantic Parsing and Information Retrieval.