软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(7):1626−1637 [doi: 10.13328/j.cnki.jos.004655]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

微博中用户标签的研究*

邢千里, 刘 列, 刘奕群, 张 敏, 马少平

(清华大学 计算机科学与技术系,北京 100084)

通讯作者: 邢千里, E-mail: xingqianli@gmail.com

摘 要: 微博环境中用户可以为自己添加标签,用户所添加的标签往往被视为是对自身特点和兴趣的重要描述信息.标签中所包含的信息可能有助于建立精确的用户描述,因此在个性化推荐、专家检索、影响力分析等应用中有潜在的应用价值.首先,在大规模数据上分析和研究了微博中用户添加标签的行为及标签内容分布的特点;之后,通过主题模型对用户的微博内容进行分析,实验结果表明:用户的标签越相似,微博内容也越相似,反之亦然;随后,分析了用户关注关系与微博和标签内容之间的联系,实验结果显示,有关注关系的用户之间微博和标签的内容越相似;基于这个发现,分别使用标签内容和微博内容对真实微博数据中的用户关注关系进行预测,结果表明:基于标签的预测方法其效果明显优于基于微博内容的预测方法,显示出用户标签在描述用户兴趣方面的价值.

关键词: 微博;用户标签;主题模型;关注关系预测

中图法分类号: TP391

中文引用格式: 邢千里,刘列,刘奕群,张敏,马少平.微博中用户标签的研究.软件学报,2015,26(7):1626−1637. http://www.jos.org.cn/1000-9825/4655.htm

英文引用格式: Xing QL, Liu L, Liu YQ, Zhang M, Ma SP. Study on user tags in Weibo. Ruan Jian Xue Bao/Journal of Software, 2015,26(7):1626−1637 (in Chinese). http://www.jos.org.cn/1000-9825/4655.htm

Study on User Tags in Weibo

XING Qian-Li, LIU Lie, LIU Yi-Qun, ZHANG Min, MA Shao-Ping

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Weibo allows users to add text tags in their profiles, which are descriptive to one's personality and interests. The tag information can be very useful to user profiling in applications such as personalized recommendation, expert finding and social influence measuring. This paper first studies the characteristics of users' tagging behavior and content of the tags based on large-scale data. By adopting topic model on users' Weibo posts, it finds that the more tags two users have in common, the more similar their Weibo posts are and vice versa. It also finds that the users with connections to each other have more similar tags and Weibo posts. Based on this observation, this study uses tags and Weibo posts to predict user connections separately on real-world data. The experimental results show that the tag-based approach is significantly better than the approach based on Weibo posts, thus validating the effectiveness of user tags in describing user interests.

Key words: Weibo; user tag; topic model; connection prediction

微博中的领域专家寻找和针对微博用户的个性化推荐是当前社会计算的研究热点[1−5],挖掘用户的兴趣并建立有效的用户描述文档(user profile)是其中的关键步骤之一,用户描述文档的质量对于个性化推荐和专家检索的最终效果有着直接影响.已有工作通常使用用户之间的链接关系[1,2]、用户所发布的文本内容[4,5],以及其他个人描述信息[3]来建立用户描述文档.基于链接关系的方法试图利用用户关注关系之间所隐含的用户同质性来进行相似用户和内容的挖掘;而基于内容的方法则试图从用户曾经发布的微博等文本中发掘出用户所感兴

* 基金项目: 国家高技术研究发展计划(863)(2011AA01A205); 国家自然科学基金(60903107, 61073071)
收稿时间: 2013-08-01; 修改时间: 2013-10-31, 2014-01-10; 定稿时间: 2014-05-21

趣的主题,从而进行个性化推荐.然而,由于微博平台所固有的特点,用户所发布的微博往往长度短、内容杂,既包含用户感兴趣主题的相关内容,也有与之无关的感情抒发或是聊天内容,导致基于微博内容的方法常常受到噪声的困扰,很难非常准确地提取出用户的兴趣所在.

在国外知名微博站点 Twitter.com 中,用户可以对其所关注的对象添加分组描述信息(称为 List 功能),并且分组名称和描述信息对所有用户公开.Ghosh 等人[3]巧妙地利用了这个功能来对用户建立描述文档,他们收集其他用户对一个用户的分组描述信息,然后使用出现最多的一部分描述词作为对这个用户的描述.当需要查找某个特定领域的用户时,就可以根据这些描述信息进行检索.由于描述信息来自许多其他用户,因此出现频率较高的描述具有较高的可信度,往往能够在寻找领域专家方面得到不错的效果.这种方法存在的主要问题是:(1) 只有关注量较多的用户会获得足够的分组描述信息,而其他一些用户则完全没有或者只有很少的分组描述,导致只能为有限的一部分用户建立起描述文档;(2) 由于分组描述来自于其他用户,针对一个用户的描述信息往往反映出该用户在他人眼中的属性,例如"家人"、"朋友"、"明星"这样的描述,更多的用户从个人组织社交网络的角度对其他用户的描述,而非其感兴趣话题.

新浪微博[6]是中国目前用户规模最大的微博平台之一,它虽然也提供了为其他用户分组和描述的功能,但是这部分信息并不像 Twitter 那样是公开的,因此,研究者无法获得用户为其他用户添加的分组描述信息.但除此之外,新浪微博还提供了一个允许用户对自己添加标签的功能,此功能允许用户用最多 10 个关键词对自己进行描述.新浪微博对用户标签的定义是:"添加描述自己职业、兴趣爱好等方面的词语,让更多人的找到你,让你找到更多同类".因此,用户对自己所添加的标签将是对自身专家领域和兴趣的直接描述,这比 Twitter 的 List 功能包含了更多的信息量,而这些信息对建立更为准确而全面的用户描述文档十分有用.

新浪微博的个人标签中包含了非常有价值的用户描述信息,但目前针对微博用户标签进行的研究相对较少.本文将对新浪微博中用户添加标签的行为及其内容特点进行研究,并且分析标签内容与用户微博内容和用户关注关系之间的联系,最终我们将通过关注关系预测任务来验证用户标签在实际应用中的价值.本文的主要贡献有以下几点:

1) 全面分析了微博用户添加标签的行为特点,验证了标签数与用户活跃度之间的联系,发现了不同标签位置上的总标签种类数目的变化规律;
2) 研究了用户的标签内容与微博内容之间的联系,实验结果表明:标签越相似的用户,其微博内容也越相似,反之亦然.这从侧面反映出短小的标签能够在一定程度上反映用户的微博内容;
3) 研究了用户的关注关系与标签和微博内容的联系,实验结果表明:存在关注关系的用户标签与微博内容比不存在关注关系的用户之间更相似,并且使用标签或微博内容的相似度进行关注关系预测的效果远远好于随机预测的效果.最后我们指出,使用标签内容进行预测的效果远好于使用微博内容进行预测的效果,说明了标签在描述用户兴趣方面的价值.

本文首先介绍相关工作.第 2 节对新浪微博中用户添加个人标签的行为特点进行研究.第 3 节研究标签的内容分布及其与微博内容之间的联系.第 4 节研究标签和微博与用户关注关系之间的联系,并且分别使用标签和微博进行用户关注关系的预测.最后,给出总结与未来工作展望.

1 相关工作

目前,对微博环境下用户添加标签的行为进行研究的相关工作数量较少.已有工作中,陈渊等人[7]提出了一种结合标签扩散和微博内容关键词提取的标签推荐方法,他们指出:在好友个数不足的情况下,从微博内容简单根据词频提取关键词作为标签推荐即可得到较好的效果.但是他们在文献[7]中通过选取的个别用户实例只在直观上说明标签推荐效果的好坏,并没有进行定量分析.在 Liang 等人[8]的工作中,用户标签被用来发现微博中能够鉴别流言的领域专家.在这个工作中,标签被看作是微博用户对其自身专长领域的描述.给出一条流言,可以计算出一个用户通过各个标签与该流言产生关联的概率,进而获得与流言最相关的用户.他们使用了一个公开的流言数据集,通过人工方法标注出与每条流言相关的专家用户.实验结果表明,基于标签的方法效果好于基

于微博内容的语言模型方法.这个工作是用户标签在微博中的一个具体应用,实验结果表明了标签信息在实际应用中的价值,但是他们并没有对标签内容的分布、标签内容与微博内容之间的关联进行分析.

国外知名微博客网站 Twitter 并没有为用户提供给自己添加标签的功能,因此也没有与之直接相关的研究工作.然而,Ghosh 等人[3]利用 Twitter 中的好友分组信息为用户建立描述信息,从而实现领域专家的查找工作.受此工作的启发,我们认为新浪微博中用户的个人标签既包含对专业领域的描述,也包含对兴趣爱好的描述,因此在专家检索和个性化推荐方面有更大的利用价值.

以上提到的工作虽然涉及微博中的用户标签,但都没有对用户添加标签的行为、标签内容的特点和标签与其他用户信息(如微博内容、关注关系)之间的联系进行充分的研究,而这些将是本文所关注的主要问题.

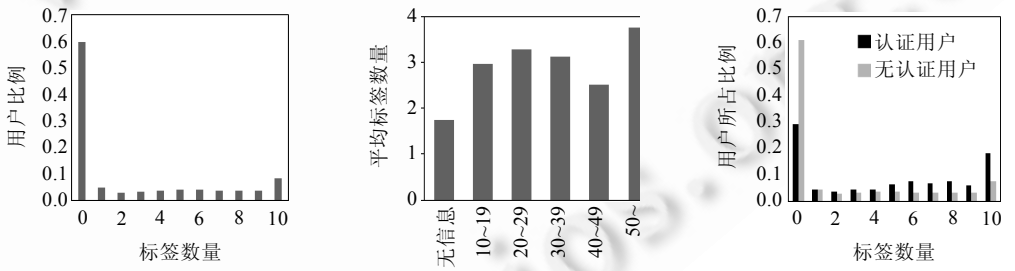
2 用户标签行为分析

2.1 微博数据集

我们首先从新浪微博中选取了一部分人气较高的账号作为种子,随后使用链接扩散方法抓取了 2 631 313 个用户的个人信息和关注关系数据.用户个人信息中包括了性别、所在地、出生日期、个人描述、微博官方认证信息、个人标签、工作信息、毕业院校和博客(传统博客)地址.其中,性别和所在地作为注册时必填,每一条用户的数据中都包含这两项;其他的信息则为用户选填,因此并非每个用户的个人信息中都会包含以上列出的所有项目.在我们的数据中,有 52.6%的女性用户和 47.4%的男性用户.在地域分布上,广东、北京、上海是用户分布最多的 3 个城市,分别占总用户数的 20.9%,10.7%和 8.5%.

2.2 用户添加标签行为的分析

本节对用户添加标签的行为进行统计分析.首先,我们统计了数据集中所有用户的标签数量分布,结果如图 1 所示.40.6%的用户至少添加了一个标签,而 59.4%的用户没有为自己添加任何标签.没有添加标签的用户可能并不知道新浪微博提供了添加标签的功能,或是知道这个功能但并没有使用.根据图 1(a)所示的分布,在有标签的用户中,我们发现只有 1 个标签的用户和添加满 10 个标签的用户数量最多,而中间的用户数量相对较少.这个现象可以解释为在为自己添加标签的用户中存在着两种心理:一种是为了体验添加标签这项功能,所以只象征性地添加了 1 个标签;另一种则是非常乐意为自己添加尽可能多的标签,从而添加了系统规定的上限个数的标签.



(a) 用户的标签数量分布 (b) 不同年龄段用户的标签数目 (c) 认证/非认证用户的标签数分布

Fig.1

图 1

进一步细分用户群体我们发现,男性用户和女性用户在为自己添加的标签数量上几乎没有差别.而在图 1(b)中,不同年龄段的用户则表现出了一些差异:50 岁及以上的用户和 20 岁~29 岁的用户添加的平均标签数量最多,而年龄信息为空(none)的用户平均标签数最少.其中的原因我们推测:可能是 50 岁以上的用户有更多的时间投入到微博使用中,而 20 岁~29 岁的年轻人是微博最活跃的用户群,因此,这些用户的信息完善程度更高.而没有年龄信息的用户相应的其他各项信息填写也不完善,因而导致标签数量少.此外,图 1(c)的结果表明:经过新浪

微博官方认证的用户(加 V 用户),明显比没有认证信息的用户倾向于添加更多的标签.

根据上述结果我们想到,用户的标签数量可能在某种程度上能够反映出用户的活跃程度.我们进一步对拥有不同数量标签的用户在其他用户属性上进行了分析,图 2 显示出拥有不同数量标签的用户在所发微博数量、关注人数和关注者人数上的表现情况.从图中可以看到:标签越多的用户,其发布的微博平均数量也越多,关注的其他用户越多;关注者的数量走势虽然有一些波动,但整体也呈现上升趋势.这个现象表明,用户添加标签的数量与其微博活跃程度呈现出正相关关系.因此我们认为,用户为自己添加标签的行为可以作为衡量用户微博活跃程度和影响力的一个因素考虑在内.

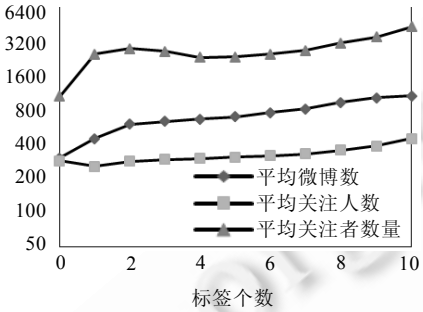


Fig.2 Involvement of users with different number of tags

图 2 不同标签数量的用户的微博活跃程度

为了进一步验证用户标签数量和其他用户行为之间的关联性,我们使用第 2.2 节提到的除标签以外的其他用户信息作为特征,训练了一个二值分类器来预测用户是否会为自己添加标签.我们从数据集中随机抽取了 25 000 个用户的数据作训练,使用 C4.5 决策树,在 10 交叉验证情况下得到预测准确率为 73.6%.由于随机预测的准确率为 60%,该结果说明,其他用户信息对于预测用户添加标签的行为是有帮助的,这个结果也验证了标签与其他用户信息之间的关联关系.但是应该注意到:在实际应用中,我们所用到的用户特征信息并不一定是先于用户标签而被添加的,因此在真实环境下并不一定能通过这些特征来预测用户的标签行为.

3 用户标签的内容分析

3.1 标签词语分布

上一节中我们分析了用户添加标签的行为特点,本节中,我们对用户添加的标签内容进行分析.为了解用户一般使用哪些词语对自己进行描述,我们首先在实验数据集上统计了标签词语的频率分布.如图 3 所示:在横轴和纵轴都使用了对数坐标的情况下,标签词语的频率分布明显呈现出幂率分布[9]的形态,即,在对数坐标上近似线性分布.这意味着有大量的标签只出现过很少的次数,而只有很少一部分标签会频繁出现.

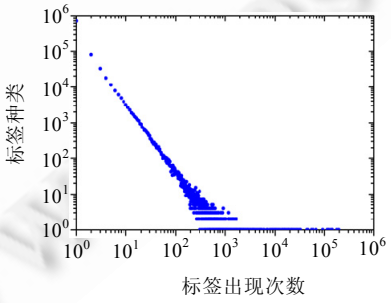


Fig.3 Number of distinct tags on different frequencies

图 3 不同频率上的标签个数分布

表 1 展示了我们的数据集中出现最频繁的 10 个标签以及它们在总的标签出现次数中所占的比例.可以看到:少数的高频标签占据了相当多的总出现次数,如“音乐”这个标签占了所有标签出现次数的 3.06%.表中所示的频率最高的 10 个标签加起来总共占了 21.2%的标签出现次数.

Table 1 Top ten popular tags

表 1 最热门的前 10 条标签

序号	标签	出现次数	所占比例(%)
1	音乐	195 542	3.06
2	电影	179 982	2.81
3	80 后	146 621	2.29
4	美食	136 534	2.13
5	旅游	134 527	2.10
6	时尚	122 104	1.91
7	90 后	120 425	1.88
8	听歌	116 286	1.82
9	旅行	113 369	1.77
10	宅	92 763	1.45
总计		1 358 153	21.20

从表 1 中可以观察到:这些热门标签的内容多是大众性的兴趣爱好的描述,如“音乐”、“电影”、“美食”等;或者是对一些常见人群的描述,如“80 后”、“90 后”、“宅”.这些标签之所以被频繁使用,一是因为这其中的一些标签在用户添加标签的页面作为系统推荐选项出现,因此有更大的概率被用户看到和选中,而不用手动输入;二是此类标签对于新浪微博用户具有普适性,即,很多微博用户都会发现这样的标签在某种程度上符合对自己的描述.例如,在我们的数据集中,在有出生日期的用户中,有 46.5%的用户出生于 1980 年~1989 年之间,有 44.6%的用户出生于 1990 年~1999 之间.“80 后”、“90 后”两个标签非常符合对这些用户的描述,因此成为高频标签.

新浪微博最多允许用户为自己添加 10 个标签,而且用户所添加的标签是有顺序的.我们发现:在不同的位置上,标签的分布情况有所不同.首先,我们统计了不同位置上不重复的标签个数.为保证每个位置上总的标签个数一致,使得结果可比,我们将所有的用户数据分成了 10 份,编号为 1~10,第 i 份数据中只包括恰好有 i 个标签的用户.之后,我们对每一份数据分别进行统计.这样的统计方法保证了在每一份数据中,在每一个位置上总的标签出现次数是相同的.统计结果显示:对于所有的 10 份数据,随着标签位置越来越靠后,在该位置的不重复标签的个数(即标签的种类)呈现递减趋势.在越靠前的位置上,不重复标签数越多.图 4 展示了对有 10 个标签的用户所统计出的各个位置上的不重复标签数,随着位置的靠后,可以观察到一个非常明显的线性下降趋势.

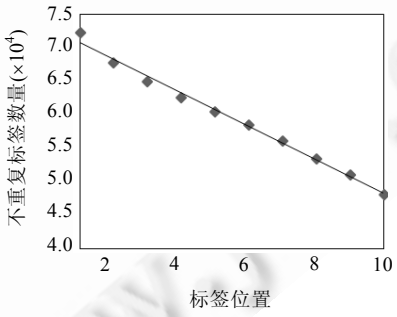


Fig.4 Number of distinct tags at different positions

图 4 不同位置上的不重复标签数

一个位置上的不重复标签的数量一定程度上能够反映出该位置上标签分布的多样性,但是并不包含每个标签的出现频率信息,因此不能排除噪音干扰的可能.为了更好地衡量每个位置上标签分布的多样性,我们为每个位置计算标签分布的熵,其计算方法如下所示:

$$H=-\sum_{i=1}^{n}p_i\log_2 p_i,$$

其中,p_i是第 i 个不重复标签在当前位置上的出现概率.熵 H 能够反映分布的混乱程度,H 的数值越大,表示分布的混乱程度越高.图 5 展示了对于有不同数量标签的用户,在各个位置上的标签分布的熵.其中,颜色越深的块表示熵越低,即,分布的混乱程度越小.可以看到:在图中每一行,颜色都是从左到右逐渐加深.说明对于每一组用户数据,标签位置越靠后,标签分布的混乱程度越低.这个变化规律与之前的不重复标签数的变化规律是一致的.

我们进一步统计了热门标签(出现最频繁的标签)在不同位置上所占的比例.图 6 展示了表 1 中前 3、前 5、前 10 的标签在各个标签位置上所占的比例.可以看出:在越靠后的位置,热门标签所占比例越高.图 4 和图 5 的结果表明:在越靠后的位置,用户越倾向于添加常见的热门标签,因而导致不重复标签数和标签分布熵都小于靠前的位置.造成这一现象的原因可以是:当用户在添加标签时,首先想到的是最具个性化的标签,这类标签最能反映出自己与他人的区别,因此最先被用户添加,正是由于个性化标签的多样性,导致不重复标签数量多并且分布混乱;而在越靠后的位置,用户能够想到的个性化标签越来越少,这时就更有可能添加一些与自己比较相关的大众标签.按照上述现象及解释,我们认为:位置靠前的标签比位置靠后的标签更能描述用户的个性特征,因此可能在个性化推荐中具有更大的利用价值.在后续实验中,我们将根据这一发现对不同位置上的标签赋以不同的权重.

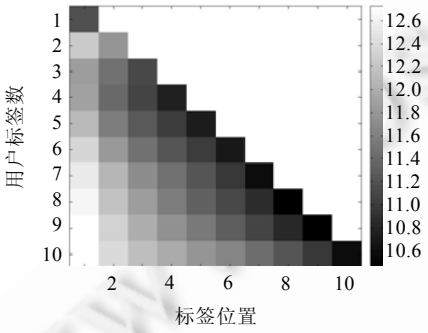


Fig.5 Entropy of tag distribution at different positions

图 5 不同位置上的标签分布熵(深色表示低熵值)

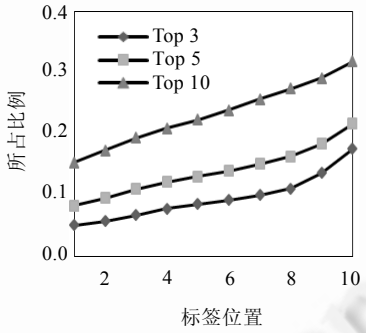


Fig.6 Percentage of top tags at different positions

图 6 不同位置上热门标签所占比例

3.2 用户标签与微博内容的联系

3.2.1 从微博内容中提取主题

用户所发布的微博内容能够在一定程度上反映出用户的兴趣所在,因此常被用来提取用户所关注的主题[4].标签作为一种更短更直接的用户描述,是否与用户的微博内容表达了相同的主题?在这一节中,我们研究用户标签与微博内容之间的关系,探索能否使用标签帮助或替代微博内容进行用户关注主题的提取.

主题模型(topic model)[10,11]常被用来发掘文档-主题-词语之间的潜在生成关系.在主题模型中,一个文档被看作是由词语及其出现频率组成的向量(词袋模型),输入一个文档-词语的矩阵,主题模型会估计出两类参数:一类参数是文档在主题上的概率分布,另一类参数是主题在词语上的概率分布.

如果将一个用户发布的所有微博内容看作该用户的描述文档,我们可以使用主题模型估计出该用户在不同主题上的概率分布,主题数量一般人为设定,数量远远少于文档中的词语数,因此,用户在主题上的概率分布可以看作是其在低维空间上的表示.在主题模型训练完成后,一个用户 u 可以用如下的向量进行表示:

$$\overrightarrow{V_u}=(p_u^1,p_u^2,...,p_u^T),$$

$$\text{s.t. }\sum_{i=1}^{T}p_u^i=1,$$

其中,p_u^i 是 u 发布的微博内容产生主题 i 的概率,T 是所有主题的数目.由于主题模型的计算代价较大(在

TwitterRank[4]一文中,作者只在 1 000 个 Twitter 用户上使用了主题模型),为了将运行算法的时间控制在可接受的范围内,我们从数据中抽取出一部分用户的数据展开实验.从第 2.1 节里所描述的数据集中,我们提取出标签中含有“互联网”字符串的用户,共 5 901 个.将用户做这样的限定是为了使得选出的用户在标签和微博内容上有一定的相似性,并且在关注关系上不至于过于稀疏,以保证后面实验中的相似度计算和关注关系预测不会得到过多为 0 的结果.我们抓取了这些用户在 2012 年 11 月之前最新发布的至多 1 000 条微博内容,共得到 3 861 174 条微博数据.对于每一个用户,我们首先对其微博进行了必要的噪音过滤,如去除文本长度过短(如内容只包括“赞”、“呵呵”等信息量较小的微博).随后,我们将其过滤后的所有微博内容合并在一起,使用中文分词工具 ICTCLAS2013[12],对合并后的文本进行分词并使用中文停用词表过滤停用词后,得到用户的描述文本(向量形式).每个用户文本平均包含 4 987 个词.

GibbsLDA[13]是一个常用的主题模型工具包,它使用 Gibbs 采样方法进行参数估计.我们使用它在上面得到的用户描述文档上训练主题模型.在设置训练模型的参数时,我们将迭代次数设置为 1 000,并将主题个数设置为 20 和 100,分别训练出两个模型,其他参数则使用了 GibbsLDA 的默认值.我们在配有 40G 内存,8 核 2.3GHz CPU 和 Linux 环境的服务器下运行 GibbsLDA,在我们的数据集上完成一次训练耗时为 7~8 小时.

3.2.2 计算用户微博内容的相似度

在主题模型得到用户在主题上的概率分布后,我们使用第 3.2.1 节中的用户描述向量计算两个用户微博关注主题之间的距离,距离越大,说明相似度越低.我们使用了两种计算向量之间距离的方法.

- 方法 1(*DIS*1)计算两个向量在各个维度上的差值的绝对值之和,表示为

$$DIS1(\overrightarrow{V_u},\overrightarrow{V_v})=\sum_{i=1}^{T}|p_u^i-p_v^i|.$$

它衡量了用户在每个主题上的差别之和,其形式简单且易于理解.Weng 等人[4]也使用了类似的方法来度量两个用户在某个主题上的相似度;

- 方法 2(*DIS*2)中,我们使用 KL divergence[14]来度量两个向量之间的距离.KL divergence 在语言模型中常被用来衡量两个分布的近似程度,值越大,说明越不相似,计算方法如下:

$$DIS2(\overrightarrow{V_u},\overrightarrow{V_v})=\sum_{i=1}^{T}p_u^i\log\frac{p_u^i}{p_v^i}.$$

3.2.3 标签与微博内容的关系

由于标签和微博内容都能够在一定程度上描述用户的兴趣,本节中,我们研究用户的标签内容与微博内容之间的关系.假设标签和微博都能够表示用户所关注的主题,那么标签相似的用户在微博内容上应该具有一定的相似度,反之亦然.下面我们通过实验来验证这个假设.

为保证有充足的标签数据来计算相似度,我们从第 3.2.1 节中提到的 5 901 个用户中提取出所有含有 10 个标签的用户,并让他们两两组合生成用户对,共产生不重复的用户对 2 160 081 个.由于标签数量少、内容短,对于一个用户对$\langle u,v\rangle$,我们用 u 和 v 之间的共有标签个数 k 来衡量他们在标签内容上的相似度,k 越大,说明标签相似度越高.同时,我们使用第 3.2.1 节中训练得到的用户主题分布和第 3.2.2 节中的距离度量方法来衡量用户微博内容的相似度.图 7 中展示了共有标签数为 k 的用户对微博之间的平均距离.由于共有标签在 7 个及以上的用户对个数非常少,为避免取平均值时引入的噪音,图中我们只展示了共有标签小于等于 6 的结果.

从图 7 中可以看出:在两种计算微博距离的指标下,对于 T=20 和 T=100 两种情况,用户微博的平均距离基本上随着用户共有标签数量的增多而减小.这说明,当用户的标签越相似时,他们所发的微博内容也越相似.这个结果验证了“标签相似的用户微博内容也倾向于相似”的假设,暗示了标签文本虽然短小,但通过标签也能找到有共同兴趣的用户.

图 8 显示了在两种距离度量下,用户之间的标签相似度(用共同标签数衡量)随微博内容距离变化的趋势.图中结果显示:用户间的微博内容距离越远,则标签的相似度越低,并且这个趋势在 *DIS*2 下表现得尤为明显.图 7 与图 8 中的结果均表明:用户的微博与标签之间存在着较强的正相关关系,即,标签越相似的用户,微博内容越相

似,反之亦然.这印证了微博中“标签是用户对自身特点和兴趣的描述”的设定.此外,实验结果从侧面表明:在使用主题模型产生的“用户-主题”向量计算用户之间的相似度(或距离)时,*DIS*2 比 *DIS*1 更为有效,*DIS*2 结果中,横轴量与纵轴量之间的相关系数的绝对值均超过 0.75(对 *T*=20 和 *T*=100).因此,在计算微博内容距离时,我们主要考虑使用 *DIS*2 进行度量.

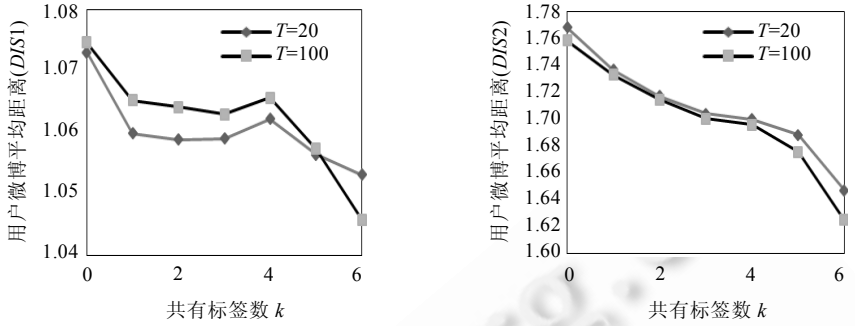


Fig.7 Weibo content distance against number of common tags

图 7 用户的微博距离随标签相似度的变化

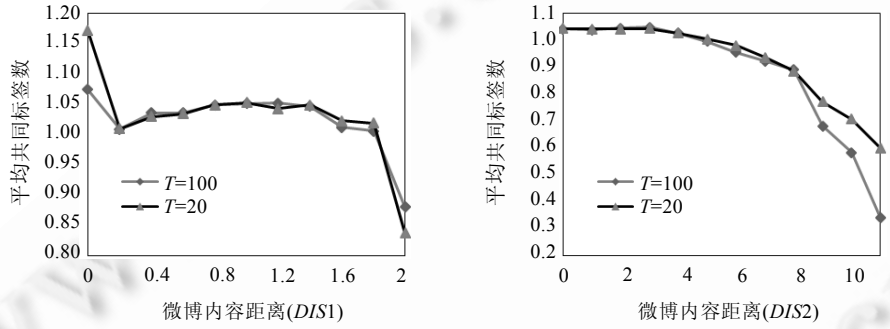


Fig.8 Average number of common tags against Weibo content distance

图 8 用户的标签相似度随微博距离的变化

需要注意的是:上述结果是在有 10 个标签的用户数据上得到的,在缺少标签的情况下,计算出的标签相似度数值的可靠性会有所减弱.因此,在挖掘用户兴趣时,使用标签信息虽然有效,但在用户标签数据量不足的情况下(图 1 中显示微博中 50%以上的用户没有为自己添加标签),标签并不能完全取代用户的微博内容.

4 使用标签预测用户间的关注关系

好友推荐对社交网站有非常重要的意义,几乎所有的社交网站在用户注册之初就会竭力帮助用户寻找或推荐好友,其目的在于增加用户黏性和社交网络本身的稠密程度.新浪微博让用户添加个人标签的初衷,也是为了能够帮助在兴趣相同的人之间建立起更多的关注关系.因此在这一节中,我们将分析用户标签与关注关系之间的联系,并且将使用标签对用户的关注关系进行预测.

4.1 与用户关注关系的联系

通常我们认为:在微博中,用户与其所关注的对象在一定程度上有着共同的兴趣.根据这个假设,存在关注关系的用户之间在微博内容和标签内容上应该有一定的相似性.现在,我们通过实验验证这个假设.在我们的数据集中有微博内容的有 5 901 个用户,其中,共存在 5 138 条关注关系(每个单向关注计算为 1 条),包括 2 994 条单向关注关系,1 072 对相互关注.我们提取出这些关注关系所对应的用户对,计算用户对中用户之间的标签相

似度和微博内容距离.作为对比,随机提取 5 138 个没有关注关系的用户对作为对比组.表 2 中列出了各种相似度计算方法下的相似度(距离)计算结果.

Table 2 Tag similarity and Weibo content similarity between user pairs

表 2 用户对的标签相似度和微博内容距离

标签相似度	相互关注	单向关注	无关注关系
TagSim	**0.072**	**0.069**	0.050
TagInter	**0.999**	**0.973**	0.701
WeightedTagSim	**1.556**	**1.522**	0.696
微博内容距离	相互关注	单向关注	无关注关系
*DIS*1(*T*=20)	**1.006**	**1.111**	1.138
*DIS*2(*T*=20)	**1.541**	**2.024**	2.048
*DIS*1(*T*=100)	**1.017**	**1.127**	1.136
*DIS*2(*T*=100)	**1.549**	**2.033**	2.058

其中,*TagInter* 是指公共标签数量.由于用户对中的两个用户可能含有不同数量的标签,我们同时使用 *TagSim* 计算两个标签列表之间归一化之后的相似度.根据第 3.1 节中的发现,即,位置越靠前的标签能表现出用户的个性,我们对不同位置上的标签赋以不同的权重后,得到带位置加权的用户标签相似度计算方法 *WeightedTagSim*,这 3 个相似度计算方法分别用下面的式子表示:

$$TagInter(S_u,S_v)=|S_u\cap S_v|,$$

$$TagSim(S_u,S_v)=|S_u\cap S_v|/|S_u\cup S_v|,$$

$$WeightedTagSim(S_u,S_v)=\sum_{t\in S_u\cap S_v}\frac{rank_{S_u}(t)}{1+rank_{S_u}(t)}+\frac{rank_{S_v}(t)}{1+rank_{S_v}(t)},$$

其中,*u* 和 *v* 表示用户,*S* 表示用户的标签列表,$rank_S(t)$表示标签 *t* 在列表 *S* 中所处的位置.*TagInter*,*TagSim*,*WeightedTagSim* 数值越大,说明两个用户的标签越相似.表 2 中的结果显示:有关注关系的用户之间的标签相似度在上述 3 个指标上均明显大于没有关注关系的用户对,即,标签更相似,并且相互关注的用户之间的标签相似度略大于只有单向关注关系的用户对.在微博内容的相似度方面,我们仍然使用了第 3.2.2 节中用到的 *DIS*1 和 *DIS*2 两种距离度量方法,*DIS*1 和 *DIS*2 值越小,代表用户微博内容越相似.表 2 中的结果表明:在主题数量 *T*=20 和 *T*=100 两种情况下,有关注关系的用户之间的微博内容距离都小于没有关注关系的用户之间的微博内容距离;并且相互关注的用户之间的微博内容距离明显小于单向关注用户之间的微博内容距离.这一结果进一步验证了我们的假设,表明有关注关系的用户之间在标签和微博内容上都更为相似.此外,从表 2 中还可以观察到:用户间的标签相似度对“有没有关注关系”比较敏感,而用户间的微博内容相似度则对“是否相互关注”比较敏感.在下一节中,我们将分别使用标签和微博对用户的单向关注关系和双向关注关系分别进行预测.

4.2 预测用户关注关系

基于上一节中得到的“有关注关系的用户在标签和微博内容上更为相似”这个结果,现在我们反过来假设标签或者微博内容越相似的用户之间越有可能存在关注关系.这一节中,我们将使用标签和微博的内容对用户的关注关系进行预测.

预测用户关注关系的任务描述为:假设我们的数据中已有的用户之间的关注关系是未知的,在已知所有用户的标签和微博内容的情况下,对任一用户 *u* 可能关注的对象进行预测.在得到预测结果后,用实际数据中存在的关注关系作为答案进行评价.

对于给定的用户 *u*,我们计算 *u* 与其他所有用户之间的相似度,并选出相似度最高的一部分用户,作为 *u* 可能关注的对象的预测结果.在这个框架下,使用不同的相似度计算方法可以得到不同的预测结果.实验中,我们对比了以下预测方法:

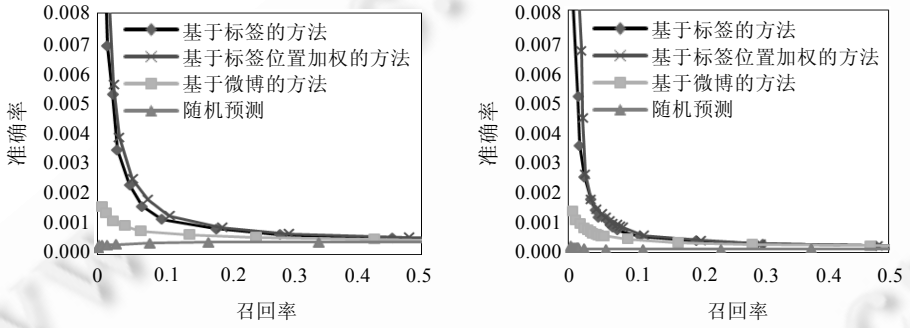
1) 基于标签的方法:根据用户标签内容进行预测,使用 *TagSim* 计算相似度;
2) 基于标签位置加权的方法:根据用户标签内容进行预测,使用 *WeightedTagSim* 计算相似度;
3) 基于微博的方法:根据从用户微博中提取的用户在主题上的分布进行预测,使用 *DIS*2 计算相似度(主

题数量 T=20);

4) 随机预测:挑选随机用户作为预测结果.

在基于标签的方法中,由于 *TagSim* 和 *TagInter* 都是没有加入标签位置权重的相似度指标,并且最终实验结果类似,因此这里只展示了 *TagSim* 的结果(方法 1).同样地,在基于微博内容的方法中,使用 *DIS*1 和 *DIS*2 在 T=20 和 T=100 下得到的结果也基本类似.因此,这里只展示使用 *DIS*2 在 T=20 时的结果(方法 3).

在 5 901 个用户组成的集合中,共有 2 204 个用户关注了至少一个这个集合之中的其他用户.对这 2 204 个用户中的每一个,我们在其他 5 900 个用户中挑选与其最相似的前 N 个用户作为关注关系的预测结果.如果将真实数据中实际存在的关注关系(单向或双向)看作正确的预测,即正例,则一种有效的预测方法应当能够将尽可能多的正例排在靠前的位置.一般地,当 N 较小时,预测准确率较高,召回率较低;当 N 变大时,准确率下降,召回率上升.随着 N 的变化,我们可以得到准确率-召回率曲线(precision-recall curve),这个曲线能够反映出一种预测方法的好坏,曲线的形态越靠右上角,说明预测效果越好.对于上面提到的 4 种预测方法,实验中分别得到了它们的准确率-召回率曲线,结果如图 9 所示.我们分别将“是否有单向关注关系”和“是否有相互关注关系”分别作为判断是否正确预测的标准,得到图 9(a)、图 9(b)两个结果.从图中可以看到:在预测单向和双向关注关系上,基于标签的方法和基于微博的方法预测效果都远远好于随机预测的结果,这验证了“标签和微博内容越相似的用户之间越可能存在关注关系”这个假设.并且,基于标签的预测方法效果明显好于基于微博的方法,而其中基于标签位置加权的方法效果最好,这印证了我们在第 3.1 节中的发现对基于标签的预测方法是有帮助的.



(a) 预测单向关注关系 (b) 预测相互关注关系

Fig.9 Precision-Recall curve of user connection prediction

图 9 关注关系预测的准确率-召回率曲线

在我们的数据集上,随机地对单向关注关系进行预测的理论准确率为 5138/(2204×5900)=0.000395,而基于标签的带位置加权的方法在 N=1 时预测准确率可以达到 0.018,是随机方法准确率的 45.6 倍,是基于微博的方法准确率的 4.96 倍;并且,其准确率在 N 较小时(在实际用户推荐的应用中,N 一般也不会取很大的数值),始终明显高于其他两种方法.在相互关注关系的预测中,当 N=1 时,基于标签的带位置加权方法的准确率是随机方法的 73.3 倍,是基于微博的方法准确率的 4.7 倍.这些结果说明了基于标签的方法在预测用户关注关系中的有效性.考虑到基于标签的方法只使用了简单的相似度计算方法就使其效果明显好于基于微博内容的方法,我们认为,标签内容本身对于描述用户兴趣、发掘潜在的关注关系十分有效.此外我们观察到:在预测相互关注关系时,基于标签的方法相对基于微博的方法的优势有所减小,这印证了表 2 中的结果,即:微博内容相似度对“是否互相关注”更为敏感,而标签内容对“是否有关注关系”更为敏感.但是,使用用户微博内容在预测用户关注关系上的整体效果仍然远远差于使用用户标签.

上述实验结果表明,用户标签在预测用户关注关系中的价值大于用户的微博内容.我们注意到,图中准确率的绝对数值整体偏低.这是由数据集中用户关注关系的稀疏性和预测任务本身的难度所决定的,加之我们使用已有的关注关系作为评价标准,这样更会使得准确率数值看上去偏低,但这并不影响我们对不同的方法进行

比较.

5 总结与未来工作

本文首先通过大规模真实微博数据研究了新浪微博用户添加标签的行为特点和标签内容的分布规律.通过实验我们发现:用户添加标签的行为与其在微博中的活跃程度和影响力存在正相关的关系,标签越多的用户微博活跃度越高,反之亦然.在标签内容的分布上,少量的热门标签占据了很大的出现比例,不同位置的标签内容分布也存在差异,在总标签数量相同的情况下,越靠前的位置上标签分布越杂乱.我们分析认为:在用户为自己添加标签的过程中,倾向于在靠前的位置添加个性化的标签,而在靠后的位置添加常见的热门标签.实验结果支持了我们的假设.

通过将标签内容与用户的微博内容进行对比,我们发现:当用户之间拥有越多的公共标签时,他们的微博内容也越相似,反之亦然.这表明:标签虽然内容很短、数量少,但是能够在一定程度上反映出用户感兴趣的微博话题.实验中我们还发现,用户标签和微博内容与关注关系之间存在很强的关联性.实验结果表明:有关注关系的用户之间的标签相似度和微博相似度大于没有关注关系的用户;反之,我们推测标签和微博内容越相似的用户,其存在关注关系的可能性也越大.根据这个假设,我们提出了基于用户标签和微博预测关注关系(单向和双向)的方法,结果表明:基于标签的方法和基于微博内容的方法都能比随机方法给出更好的预测结果,并且基于标签的方法其效果明显好于基于微博的方法.这一系列结果验证了标签在描述用户兴趣方面的价值和有效性.在未来的工作中,我们将进一步研究如何在个性化推荐、专家检索、影响力度量等应用场景中有效使用微博用户的标签信息.

References:

[1] Pal A, Counts S. Identifying topical authorities in microblogs. In: Proc. of the 4th ACM Int'l Conf. on Web Search And Data Mining (WSDM). New York, 2011. 45−54. [doi: 10.1145/1935826.1935843]

[2] Cha M, Haddadi H, Benevenuto F, Gummadi KP. Measuring user influence in twitter: The million follower fallacy. In: Proc. of the Int'l AAAI Conf. on Weblogs & Social Media. Washington, 2010. 10−17.

[3] Ghosh S, Sharma N, Benevenuto F, Ganguly N, Gummadi KP. Cognos: Crowdsourcing search for topic experts in microblogs. In: Proc. of the 35th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR). New York, 2012. 575−590. [doi: 10.1145/2348283.2348361]

[4] Weng JS, Lim EP, Jiang J, He Q. TwitterRank: Finding topic-sensitive influential Twitterers. In: Proc. of the 3rd ACM Int'l Conf. on Web Search and Data Mining (WSDM).New York, 2010. 261−270. [doi: 10.1145/1718487.1718520]

[5] Si XC. Content-Based recommendation and analysis of social tags [Ph.D. Thesis]. Beijing: Tsinghua University, 2010 (in Chinese with English abstract).

[6] Sina Weibo. http://www.weibo.com

[7] Chen Y, Lin L, Sun CJ, Liu BQ. A tag recommendation method for microblog users. Intellegent Computer and Applications, 2011,1(3):21−26 (in Chinese with English abstract).

[8] Liang C, Liu ZY, Sun MS. Expert finding for microblog misinformation identification. In: Proc. of the 24th ACL Int'l Conf. on Computational Linguistics. Mumbai, 2012. 703−712.

[9] Simon HA. On a class of skew distribution functions. Biometrika, 1995,42(3/4):425−440. [doi: 10.1093/biomet/42.3-4.425]

[10] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003,3:993−1022.

[11] Hofmann T. Probabilistic latent semantic indexing. In: Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR). ACM Press, 1999. 50−57. [doi: 10.1145/312624.312649]

[12] Zhang HP. ICTCLAS. 2012. http://ictclas.nlpir.org/

[13] Phan XH, Nguyen CT. GibbsLDA++: A C/C++ implementation of latent dirichlet allocation (LDA). 2007. http://gibbslda.sourceforge.net/

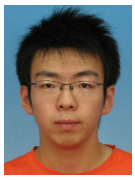
[14] Kullback S, Leibler RA. On information and sufficiency. Annals of Mathematical Statistics, 1951,22(1):79−86. [doi: 10.1214/aoms/1177729694]

附中文参考文献:

[5] 司宪策.基于内容的社会标签推荐与分析研究[博士学位论文].北京:清华大学,2010.

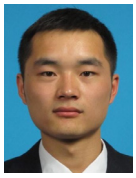
[6] 新浪微博. http://www.weibo.com

[7] 陈渊,林磊,孙承杰,刘秉权.一种面向微博用户的标签推荐方法.智能计算机与应用,2011,1(3):21−26.



邢千里(1987－),男,陕西岐山人,博士生,主要研究领域为信息检索.

张敏(1977－),女,博士,副教授,CCF 高级会员,主要研究领域为机器学习,信息检索.



刘列(1991－),男,硕士生,主要研究领域为信息检索.

马少平(1961－),男,博士,教授,博士生导师,主要研究领域为知识工程,信息检索,汉字识别与后处理,中文古籍数字化.

刘奕群(1981－),男,博士,副教授,CCF 高级会员,主要研究领域为信息检索.