基于发文内容的微博用户兴趣挖掘方法研究*

熊才伟[1,2]，曹亚男[1]

(1. 中国科学院信息工程研究所 国家重点工程实验室，北京 100093；2. 中国科学院大学 计算机与控制学院，北京 100093)

摘 要：针对微博用户兴趣属性缺失问题，提出一种基于发文内容分析的微博用户兴趣挖掘方法。利用基于短语的主题模型和自动构建的用户兴趣知识库，能够有效地从发文内容中挖掘出高质量的用户兴趣短语并标志其类别，从而实现对微博用户的兴趣挖掘。在 SMP CUP 2016 数据集上的实验结果表明，主题短语模型在困惑度和短语质量上取得的效果均优于传统的主题模型，用户兴趣挖掘的准确率和召回率最高可达到 78% 和 82%。

关键词：微博；发文内容；兴趣挖掘；主题短语模型；知识库

中图分类号：TP301.6 **文献标志码**：A **文章编号**：1001-3695(2018)06-1619-05

doi:10.3969/j.issn.1001-3695.2018.06.004

Research of microblog user interest mining based on microblog posts

Xiong Caiwei[1,2], Cao Yanan[1]

(1. *National Key Engineering Laboratory, Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093, *China*; 2. *School of Computer & Control Engineering, University of Chinese Academy of Sciences, Beijing* 100093, *China*)

Abstract: To abstract missing interests of microblog users, this paper proposed an data mining approach based on posting message analysis. Using the phrase-LDA and the user interest knowledge base constructed automatically, it could extract high-quality candidate interest phrases from posting messages and implement the interest classification. The experimental results on SMP CUP 2016 dataset show that the phrase-LDA can achieve better results than traditional topic model on perplexity and phrase quality. The accuracy rate and the recall rate of user interest mining can reach 78% and 82% at best respectively.

Key words: microblog; microblog posts; interests mining; phrase-LDA; knowledge base

0 引言

微博是基于社交关系来进行信息传播的媒体平台。作为重要的社交网站，微博引发了众多的关注和研究。随着微博平台的蓬勃发展，微博用户规模的不断增大，微博用户的属性、关系和行为分析也逐渐成为学术界和工业界研究的热点。其中，微博用户的兴趣爱好能够反映用户的倾向性，同时与用户性别、年龄、职业等属性有着紧密的关联性，对于实现更精准的用户群组划分和个性化推荐具有重要意义。目前，微博用户注册的兴趣标签缺失率达到 70% 以上[1]，只依靠用户的注册信息不足以描述用户的兴趣情况。已有研究[2~4]表明，发文内容通常隐含着丰富的兴趣信息，是挖掘微博用户兴趣的重要数据源。

在基于微博用户的发文内容来挖掘用户的兴趣信息方面，国内外学者开展了大量的研究工作。Mihalcea 等人[5]利用基于图的 TextRank 方法从微博文本中挖掘用户兴趣关键词，准确率和召回率分别达到 31.2% 和 43.1%。Vu 等人[2]利用语言规则，并结合 TFIDF 和 TextRank 算法来挖掘用户兴趣，兴趣挖掘的准确率能达到 54.5%。Tao 等人[3]则利用时间序列进行微博用户的兴趣挖掘，可将用户微博分类的准确率提高至 67%。这些方法在挖掘微博用户的兴趣信息方面取得了一定的效果，但由于没有利用文档内和文档间的统计特征，同时没有考虑兴趣词的歧义性问题，所以在兴趣挖掘的准确率和召回率上仍无法满足实际应用的需要。近年来，主题模型由于能够利用文本中潜在的主题结构，适用于处理稀疏性高的短文本特性，而被广泛地应用于微博用户的兴趣挖掘工作中。Zhao 等人[6]利用 LDA 模型来分析挖掘用户的兴趣，证明 LDA 模型能够有效地挖掘文本中潜在的兴趣主题信息。张晨逸等人[7]提出 MB-LDA 模型，能够从用户微博内容中挖掘出更高质量的关键词以表示用户的兴趣。主题模型能够获取一系列可能性最高的词来描述一个主题，从而挖掘出用户的潜在兴趣。然而用词来表示兴趣具有很强的二义性；同时，主题模型虽然能够获取文本中的潜在主题，但却无法获得主题的语义标签，从而无法对微博用户的兴趣进行明确标志。

针对以上两个问题，本文通过挖掘用户发文内容中的主题短语来识别用户兴趣。该方法分为以下两个阶段：

a)利用基于短语的主题模型挖掘用户的候选兴趣短语。相对于词，短语具有更加明确的语义信息，能够更加直观和准确地表示一个特定的主题。表 1 是在 SMP CUP 2016 数据集上，分别采用基于词和基于短语的主题模型得到某个主题下的词和短语的集合。可以明显地看出短语的歧义性较小。在该部分首先利用频繁项挖掘算法和短语结合度算法来从微博用

收稿日期：2017-01-24；**修回日期**：2017-03-14 **基金项目**：国家自然科学基金青年基金资助项目(61403369)；国家科技部重大专项资助项目(2016YFB0801300)

作者简介：熊才伟(1991-)，男，河南信阳人，硕士研究生，主要研究方向为数据挖掘、自然语言处理(wxiong126@163.com)；曹亚男(1986-)，女，副研究员，博士，主要研究方向为自然语言处理.

户的发文内容中挖掘出候选兴趣短语,然后利用主题短语模型来得到微博用户的兴趣短语集合。

表 1 相同主题下词和短语的集合

词	短语	词	短语
3D	3D 打印	性能	强劲性能
路由	360 安全路由	板砖	卫星
安全	锤子科技	相机	数码相机
手机	滴滴红包	百度	手机百度
小米	小米手机	科技	暴风科技

b)利用微博用户兴趣类别知识库对兴趣短语的类别进行标志,从而实现对微博用户的兴趣分类。其主要工作包括微博用户兴趣体系的构建、兴趣类别知识库的构建、兴趣短语类别的自动标志三部分。本文通过分析微博平台的用户兴趣分布情况,构建了一个二级的微博用户兴趣体系,并根据该体系,基于开放数据源自动构建微博用户的兴趣类别知识库;利用兴趣短语的主题分布,结合微博用户兴趣类别知识库,实现兴趣短语类别的自动标志。

本文提出的方法具有以下几个优势:a)利用统计学特性,提出一种无监督的兴趣短语挖掘方法,能够快速地从微博用户发文内容中提取出候选兴趣短语;b)利用"bag-of-phrases"代替"bag-of-words"来表示文档集合,能够获得高质量的兴趣短语集合,同时降低了主题模型的复杂度;c)构建微博用户兴趣类别知识库,通过引入知识库,实现了微博用户兴趣的细粒度划分和明确的语义类别识别。

1 国内外研究综述

现有文献已经对基于文本分析的微博用户兴趣挖掘展开了诸多研究。Salton 等人[8]利用 TFIDF 方法,根据词语出现的频率来从微博用户发文内容中提取出候选词,并根据频率对候选词进行排序,挑选出其中的 Top-M 个词作为关键词来表示用户兴趣。Mihalcea 等人[5]则尝试用 TextRank 方法来建立一个基于词的图,并在图上运用 PageRank 技术[9]来进行候选关键词的排序,以挖掘出用户兴趣关键词,能够获得 31.2% 的准确率和 43.1% 的召回率。Banerjee 等人[10]使用内容指示词(用户兴趣所属的类别)和动作指示词(兴趣类别相关的动作)的二元组集合来描述用户的兴趣,可有效挖掘出微博用户的实时兴趣。Tao 等人[3]则考虑到用户微博的时间分布规律,利用时间序列对用户微博进行分类,将用户微博分类的准确率提高至 67%,并在此基础上挖掘微博用户的兴趣。这些方法利用了文本信息中词的统计特性或语义信息,在挖掘微博用户的兴趣信息方面取得了一定的效果,但却无法利用文档内和文档间的统计特征,也无法解决兴趣词的歧义性问题。

主题模型在这方面则体现出了较好的效果。Zhang 等人[11]利用 LDA 扩展文本特征空间,然后使用频率统计的方法来挖掘出热点话题,使得热点话题的排名更加靠前。Ramage 等人[12]使用聚合信息训练 LDA 模型,实验结果显示该模型更有利于作者—特点话题的建模。Zhao 等人[6]提出了 Twitter-LDA 来对非热点话题词汇进行过滤,并与传统媒体中的热点话题分布进行比较,发现微博中有很大部分话题是关于用户日常生活的,更能体现出用户个人的兴趣爱好信息。张晨逸等人[7]在 LDA 模型的基础上提出了 MB-LDA 模型,能够从用户微博内容中挖掘出更高质量的关键词来表示用户的兴趣。以上的研究均表明,主题模型通过利用文本中词与主题间的分布以及主题与文档间的分布,能够有效地从微博这类稀疏性高的短文本中进行兴趣挖掘。但是已有工作仅对兴趣词进行主题划分,并没有对主题的语义和用户兴趣类别进行明确标志。

本文针对现有研究的不足,利用主题短语模型,从微博用户发文内容中挖掘出更高质量的兴趣短语,并结合微博用户兴趣知识库来识别微博用户兴趣的类别。

2 候选兴趣短语挖掘

本章展示了一种能够从给定的已分词的文档集中获取高质量的候选兴趣短语的方法。该方法基于一个直观的假设,即高质量的兴趣短语是由一个或多个频繁且连续的词所组成的。该方法分为两个主要阶段:a)频繁短语挖掘,即从文本中挖掘出所有满足最小支持度的短语作为初始的候选兴趣短语集合;b)短语过滤,即利用一种短语结合度算法对初始的候选兴趣短语进行过滤,得到最终的候选兴趣短语集合。

首先,对问题进行如下描述:给定包含 D 个文档的语料库,第 d 个文档由 N_d 个词组成,每个词由 $w_{d,i}(1\leqslant i\leqslant N_d)$ 表示,令 $N=\sum_{d=1}^{D}N_d$。同时,本文将该语料库中所有不重复的词进行排序,构成一个词典 V,并且 $w_{d,i}=v_k, v_k\in V$,即在第 d 个文档中的第 i 个元素是词典 V 中的第 k 个词。

定义 1 一个短语由一个或多个连续的词组成,短语用 P 表示,$P=\{w_{d,i},\cdots,w_{d,i+n}\}, n\geqslant 0$。

2.1 频繁短语挖掘

频繁短语挖掘的任务是从文档集中挖掘出满足最小支持度的所有短语。基于 Apriori 算法,本文利用以下两条性质来进行频繁短语的挖掘:

a)向下闭合引理。如果短语 P 不是频繁项,则任何包含 P 的短语也不是频繁项。

b)数据的反单调性。如果一个文档中不包含长度为 n 的频繁短语,则该文档中不包含长度大于 n 的频繁短语。

本文利用这两条性质可以有效过滤稀疏的短语,并且可以在不搜索过大候选短语空间的前提下更早地终止算法,使之具备较好的时间效率。本文利用一种长度增长的滑动窗从语料库中获取候选短语,并统计其出现次数。在第 k 轮迭代中,对于每个仍保留的文档,如果长度为 $k-1$ 的短语不满足最小支持度,则迭代结束,该文档就会被移除出下一轮计算。该条件也是本文算法的终止标准。

2.2 短语过滤

短语过滤的任务是从候选兴趣短语集合中挑选出高质量的候选兴趣短语。本节利用一种短语结合度算法来判断一个候选兴趣短语是否应当保留,从而实现短语过滤功能。

该算法是在 bag-of-phrases 的假设上推导而来。为了从统计上解释短语的出现频率,可以考虑一种虚假设,即文档集是由一系列独立的伯努利实验产生的。在这种假设下,在文档集中特定位置出现的短语是伯努利随机变量的结果,并且短语的出现频率可以用二项分布来进行解释。在文档集中,短语的总数目 L 可以设置为相当大,因此这个伯努利分布可以近似为正态分布,则随机变量 $f(P)$(短语 P 在文档集中的出现次数)的虚假设分布为

$$h_0(f(P))=N(Lp(P),Lp(P)(1-p(P)))\approx N(Lp(P),Lp(P)) \tag{1}$$

其中:$p(P)$是短语 P 的伯努利实验成功的概率。一个短语在文档集中的出现概率可以估计为 $p(P)=\frac{f(P)}{L}$。考虑一个更长的由短语 P_1 和 P_2 组成的短语,在本文的虚假设下,两者相互独立,组合成一个新的短语的平均频率为

$$\mu_0(f(P_1 \oplus P_2)) = Lp(P_1)p(P_2) \tag{2}$$

同时,由于整体方差满足最小支持度的样本数量是未知的,所以可以用样本方差来估计整体方差,即 $\sigma^2_{P_1 \oplus P_2} = f(P_1 \oplus P_2)$,$f(P_1 \oplus P_2)$是样本短语的出现次数。

本文利用一个显著性分数来计算两个短语是否应当组合成一个新短语的概率。该显著性分数表达式为

$$\mathrm{sig}(P_1 \oplus P_2) \approx \frac{f(P_1 \oplus P_2) - \mu_0(P_1 \oplus P_2)}{\sqrt{f(P_1 \oplus P_2)}} \tag{3}$$

该显著性分数计算了组合短语的实际出现频率在虚假设下偏离预期频率的标准差,高分数意味着两个短语的相关性非常高且应该被合并在一起。

利用该显著性分数,可以对文档集中的频繁短语进行合并操作。针对文档集中的每一句话,本文采用一种自底向上的合并方法。在每一次的迭代中都会合并显著性分数最高且满足阈值的一个短语对。如果所有短语均被合并在一起或者剩下的所有两两短语间的显著性分数均不满足阈值,则迭代终止。合并只发生在同一句话中,使得短语的合并是符合语义规则的,从而确保合并后的短语质量。短语合并算法具体如下所示。

算法 1　短语合并算法

```
H←—MaxHeap( )
Place all contiguous token pairs into H with their significance score key
while H. size( ) >1 do
  Best←—H. getMax( )
  if Best. Sig≥α then
    New←—Merge(Best)
    Remove Best from H
     Update significance for New with its left instance and right phrase instance
  else
    break
    end
  end
```

短语过滤正是在短语合并的过程中同时进行的。通过短语合并,对所有由多个短语组成的频繁短语进行显著性判断,并只保留显著性满足阈值的短语,以此来实现短语过滤功能。

3　基于主题模型的兴趣短语聚类

通过候选兴趣短语挖掘,已经将文档集划分成了短语集合,这些短语由一个或多个出现频繁、连续且非偶然性出现的词所组成。下面在 LDA 模型基础上,提出主题短语模型,用于进行微博用户的兴趣短语聚类。

LDA 模型假设一个文档是一系列主题的混合,每一个主题都被定义为词表中词的一个多项分布,一般的生成过程如下:

a)$\varphi_k \sim \mathrm{dir}(\beta)$,$k=1,2,\cdots,K$

b)对于第 d 篇文档,$d=1,2,\cdots,D$

(a)$\theta_d \sim \mathrm{dir}(\alpha)$。

(b)对于第 d 篇文档中的第 i 个元素,$i=1,2,\cdots,N_d$。

$z_{d,i} \sim \mathrm{multi}(\theta_d)$

$w_{d,i} \sim \mathrm{multi}(\varphi_{z_{d,i}})$

其中:K 表示的是主题个数;D 表示的是文档个数;N_d 表示的是文档 d 中元素的个数;φ_k 表示的是词在主题下的多项分布;θ_d 表示的是主题在文档中的多项分布;α 和 β 分别是 θ_d 和 φ_k 的狄里克莱分布的超参数;$z_{d,g,j}$表示的是第 d 篇文档中第 g 个短语中第 j 个词的潜在主题;$w_{d,g,j}$表示的是第 d 篇文档中第 g 个短语中的第 j 个词。

LDA 的联合分布可以写为(为简单起见,本文省略了超参数 α 和 β):

$$P_{\mathrm{LDA}}(Z,W,\Phi,\Theta) = \prod_{d,i} p(z_{d,i}|\theta_d)p(w_{d,i}|z_{d,i},\Phi)\prod_d p(\theta_d)\prod_k p(\Phi_k) \tag{4}$$

因为多项式分布与狄里克莱分布之间具有共轭性,可以很容易地计算$\{\Theta,\Phi\}$的积分,即

$$P_{\mathrm{LDA}}(Z,W) = \int P_{\mathrm{LDA}}(Z,W,\Phi,\Theta)\mathrm{d}(\Theta)\mathrm{d}(\Phi) \tag{5}$$

接下来将进行主题模型的构造。在上文中将文档集表示为短语集合,遵循这样一种设定,在同一个短语中的词很有可能共享一个主题,用一个潜在方程 $f(C_{d,g})$ 来表示,其中 $C_{d,g}$ 表示短语。由此可定义所有随机变量之间的联合分布为

$$P_{\mathrm{LDA}}(Z,W,\Phi,\Theta) = \frac{1}{C}P_{\mathrm{LDA}}(Z,W,\Phi,\Theta)\prod_{d,g} f(C_{d,g}) \tag{6}$$

其中:C 是归一化后的常量,使得公式左边是一个合法的概率分布。由式(5)可以得到该分布的简易形式:

$$P_{\mathrm{LDA}}(Z,W) = \frac{1}{C}P_{\mathrm{LDA}}(Z,W)\prod_{d,g} f(C_{d,g}) \tag{7}$$

在此,选择一个特殊的势函数来表示 $f(C_{d,g})$。

$$f(C_{d,g}) = \begin{cases} 1 & \text{if } z_{d,g,1} = z_{d,g,2} = \cdots = z_{d,g,w_{d,g}} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

该势函数约束同一个短语中的词共享一个潜在主题。接下来本文采用一个紧缩的吉布斯采样方法,从 $C_{d,g}$ 的后验 $p(C_{d,g}|W,Z_{\setminus C_{d,g}})$ 中抽取一个对照组,并利用 $C_{d,g}=k$ 来表示 $C_{d,g}$ 中的所有变量均取值为 k 的情况,最终可得到表达式:

$$p(Z,W|\alpha,\beta) = \frac{1}{C}P_{\mathrm{LDA}}(Z,W|\alpha,\beta) \tag{9}$$

这表明可采用典型的 LDA 方法中超参数 α 和 β 的调优方法。

4　基于兴趣知识库的用户兴趣标志

本文的任务是对挖掘出来的主题短语进行进一步的语义上的识别,实现微博用户的兴趣类别的自动标志。这需要外部知识库的支撑。为了更加有效地完成这个目标,首先构建微博用户兴趣体系,并根据该体系构建微博用户兴趣类别知识库,从而结合主题短语挖掘结果,实现微博用户的兴趣类别识别。

4.1　微博用户兴趣体系

为了更加有效地构建微博用户兴趣类别知识库,首先需要构建一个正交的、较为完备的微博用户兴趣体系。在大量调研的基础上,本文构建了一个二级分类体系,尽可能涵盖微博用户的主要兴趣类别。一级兴趣类别和部分二级兴趣类别如表 2 所示。

表 2　微博用户兴趣体系

一级类别	二级类别	一级类别	二级类别
影视	爱情、喜剧、科幻、动画、悬疑…	财经	理财、经济管理、金融…
音乐	民谣、电子、爵士、说唱、摇滚…	科技	计算机、电子工程、汽车、机械…
体育	足球、篮球、网球、羽毛球、乒乓球…	健康美食	饮食、烟草、医学、医疗…
游戏	手机游戏、网络游戏、单机游戏…	娱乐	明星、综艺、旅游、摄影…
读书	小说、散文、哲学、传记、管理…	社会	生活、教育、法律、房地产…
政治	公共管理、社会学、政治学…	购物	服饰、礼品、家居装饰、美容护肤…

兴趣体系的构建能够帮助明确微博用户的兴趣范围,从而更加有效地构建微博用户的兴趣类别知识库。

4.2 兴趣知识库的自动构建

为了实现用户兴趣类别的精准识别,需要根据微博用户的兴趣体系,构建一个较为完备的微博用户兴趣类别知识库。为了更加丰富知识库,本文结合两种方法来进行知识库的构建。

对于专有名词类的兴趣类别关键词,如音乐名、电影名等,本文利用爬虫程序在特定网站(如豆瓣、搜狗词库等)上爬取相关词条,作为知识库中的兴趣类别关键词。部分能够表征特定兴趣类别的关键词如 ace 球、拉杆等,在网站上往往难以以一个特定词条的形式出现,不能直接爬取。对于这类关键词,本文采用了 TextRank 方法对特定网站的内容进行分析,并选择排名靠前的候选词作为兴趣类别关键词。通过这两方面的工作,能够有效地构建一个较为完备的微博用户兴趣知识库。知识库中的部分类别关键词以及相对应的目标网站如表 3 所示。

表 3 知识库部分类别关键词及目标网站示例

兴趣类别	关键词	目标网站
音乐	我的中国心、青花瓷	豆瓣音乐、搜狗词条
电影	喜剧之王、湄公河行动	豆瓣电影、搜狗词条
读书	孔乙己、生死疲劳	豆瓣读书
政治	特朗普、奥巴马	百度百科、人民网
体育	姚明、梅西、ace 球	虎扑、腾讯体育
…	…	…

4.3 用户兴趣短语类别识别

利用微博用户兴趣类别知识库,可为聚类后的微博用户兴趣短语赋予类别标签。本文结合短语在主题下的分布情况和短语在兴趣类别下的分布情况,对用户兴趣短语的类别进行标志。

根据主题短语模型,可以得到某个短语 P 在某个主题 z 下的概率分布 $p(P|z)$。由微博用户兴趣类别知识库可以得到某个短语 P 在兴趣类别 $i(i=0,1,\cdots,k)$ 下的概率分布 $p(P|i)$。本文对 $p(P|i)$ 作出如下设定:

a)若知识库包含短语 P,$p(P|i)=\begin{cases}1 & \text{if } P\in i\\ 0 & \text{otherwise}\end{cases}$。

b)若知识库不包含短语 P,$p(P|i)=1$。

对于某个主题 z,识别其兴趣类别 i 的步骤如下:

a)基于短语在主题下的分布,依据短语分布概率 $p(P|z)$ 的大小对该主题下的所有短语进行排序,并挑选出前 M 个短语作为判别该主题的兴趣类别的标准短语。

b)基于挑选出的 M 个标准短语,利用微博用户兴趣类别知识库,统计该主题在各个兴趣类别上的概率分布情况 $p(z|i)$,$p(z|i)=\dfrac{\sum_{P\in z}p(P|i)}{\sum_{P\in z}\sum_{i=0}^{k}p(P|i)}$。

c)对该主题在各个兴趣类别下的分布概率进行排序,选择概率最大的兴趣类别作为该主题的类别标志。

通过上述步骤,能够有效结合兴趣短语的主题分布以及微博用户兴趣类别知识库,实现用户兴趣短语类别的自动识别。

5 实验及结果分析

5.1 实验数据

本文采用 SMP CUP 2016 发布的数据集作为实验数据,该数据集是新浪微博的真实数据集,包含约 4.6 万个用户,超过 30 101 194 条微博内容。

5.2 实验设计及结果

为了详细说明本文方法的有效性,将分别从兴趣短语挖掘和兴趣短语自动标志两方面的实验进行详细的说明。

5.2.1 兴趣短语挖掘实验

针对主题短语模型,本文采用困惑度这一指标对比标准的 LDA 模型来衡量该方法的有效性。困惑度是衡量主题模型效果的重要指标,其值越小,表明模型效果越好。本文将 α 和 β 的初始值分别设为 0.1 和 0.01。实验结果如图 1 所示。

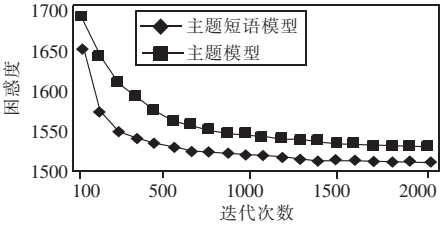

图 1 困惑度指标对比

实验结果表明,在不同的迭代次数下,主题短语模型在困惑度的表现上始终优于标准的 LDA 模型,这表明改进的主题短语模型在主题的聚类效果上表现得更为优异。

同时,为了能够更加直观地观察主题短语算法的有效性,本文针对基于词和基于短语的主题模型进行了实验验证,并使用词和短语在主题下的分布频率作为词和短语的排序标准,列举出了其中五个主题下的排名靠前的部分词语。实验结果如表 4 所示。由表 4 可以明显看出,相比于基于词的主题模型,本文的主题短语模型所得的主题词与在同一个主题下的关联性明显更强,语义也更加明确,表明本文算法所得到的短语具有较好的聚合效果。

表 4 基于词和基于短语的主题模型实验结果对比

	主题 1	主题 2	主题 3	主题 4	主题 5
词	存款	冯导	新浪	汶川	3D
	淘宝	门票	直博	屈原	路由
	京东	龙女	梅西	大赛	安全
	商城	春晚	跑友	人生	手机
	背板	投票	竞彩	生活	小米
	保暖	半月	投给	造谣	性能
	加绒	名单	男孩	签证	板砖
	手套	琅琊	揭晓	信号	相机
	下载	黄粱	跻身	黄金	百度
	资料	注册	足球快报	科技	
短语	预存款	半月传	新浪足球直播	延参法师	3D 打印
	淘宝	琅琊榜	吨位	汶川地震	360 安全路由
	保暖衣	小龙女	德安格罗	屈原	锤子科技
	郎布鲁斯	冯导	梅西	爆照	滴滴红包
	达芙妮	黄粱伊梦	穆里奇	阳光男孩	小米手机
	女款	综艺门票	补篮	生活记录	强劲性能
	肤水	春晚	盛宴	居家必备	卫星
	清润	明星	跑友	黄金屋	数码相机
	京东商城	名单	投给	剪刀手大赛	手机百度
	加绒手套	陈赫	竞彩	人生	暴风科技

此外,为了验证本文算法的时间效率,针对不同规模的数据集进行了时间效率的测试实验。候选兴趣短语挖掘方法和主题短语模型方法在运行时间上的表现如图 2 所示。

由图 2 可以看出,随着文档集规模的增加,短语构造方法和主题模型方法在运行时间上呈现出近似线性的趋势(log 级别),这表明算法具有较高的时间效率。

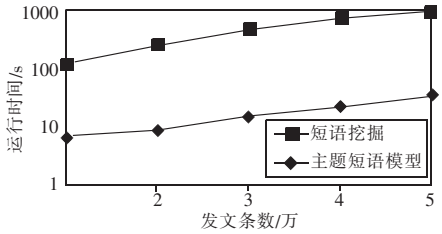

图 2　运行时间实验结果

5.2.2　兴趣短语自动标志实验

在兴趣短语的自动标志实验中,本文采取了人工判别的方式来确保实验结果的准确性。设定 M 值为 100,对主题短语模型所得到的各个主题的类别进行识别,通过统计各个主题下短语的类别识别结果,可以得到用户兴趣类别识别的准确率和召回率。其实验结果如表 5 所示。

表 5　用户兴趣类别识别实验结果

兴趣类别	准确率	召回率	兴趣类别	准确率	召回率
影视	0.753 1	0.801 2	财经	0.235 8	0.306 5
音乐	0.781 9	0.821 5	科技	0.305 6	0.385 7
体育	0.658 2	0.751 4	健康美食	0.568 9	0.601 8
游戏	0.701 4	0.732 5	娱乐	0.706 9	0.810 5
读书	0.633 5	0.523 8	社会	0.204 7	0.358 2
政治	0.303 8	0.352 3	购物	0.687 4	0.805 9

由表 5 结果可以看出,本文方法在用户兴趣类别识别上的准确率和召回率最高可达到 78% 和 82%,在影视、音乐、游戏、娱乐和购物上都具有较好的表现,这说明本文方法能够有效识别出用户兴趣的类别。另外,本文方法在财经、科技和社会等兴趣类别方面表现不佳,主要是由于微博用户兴趣类别知识库构建尚不完善,将在笔者的下一步工作中进行改进。

6　结束语

本文提出了一种基于发文内容的微博用户兴趣挖掘方法。该方法利用主题短语模型从用户发文内容中提取出高质量的兴趣短语,并通过构建微博用户兴趣类别知识库来实现兴趣短语的自动标志。通过实验验证,证明了本文方法在微博用户兴趣挖掘的准确率和召回率上具有良好的表现,能够实现微博用户兴趣的有效挖掘。在下一步的工作中,鉴于微博用户兴趣类别知识库对于用户兴趣类别精准识别的重要性,笔者考虑更加丰富和完善微博用户兴趣类别知识库来进一步提高用户兴趣挖掘的准确率和召回率。

参考文献:

[1] 丁宇新,肖骁,吴美晶,等. 基于半监督学习的社交网络用户属性预测[J]. 通信学报,2014,35(8):15-22.

[2] Vu T, Perez V. Interest mining from user Tweets[C]//Proc of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2013: 1869-1872.

[3] Tao Yang, Lee D, Su Yan. Steeler NATION, 12th man, and boo birds: classifying Twitter user interests using time series[C]//Proc of IEEE/ACM International Conference on Advances in Social Networks and Mining. New York: ACM Press, 2013: 684-691.

[4] He Li, Jia Yan, Han Weihong, *et al*. Mining user interest in microblogs with a user-topic model[J]. China Communications, 2014, 11(8): 131-144.

[5] Mihalcea R, Tarau P. Textrank: bringing order into texts[EB/OL]. (2011-01-31). https://digital.library.unt.edu/ark:/67531/metadc30962/.

[6] Zhao W X, Jiang Jing, Weng Jianshu, *et al*. Comparing Twitter and traditional media using topic models[C]//Advances in Information Retrieval. Berlin: Springer, 2011: 338-349.

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

[8] Salton G, Buckley C. Term-weight approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24(5): 513-523.

[9] Page L, Brin S, Motwani R, *et al*. The PageRank citation ranking: bringing order to the Web[R]. Palo Alto. CA: Stanford Infolab, 1999: 1-17.

[10] Banerjee N, Chakraborty D, Dasgupta K, *et al*. User interests in social media sites: an exploration with micro-blogs[C]//Proc of the 18th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2009: 1823-1826.

[11] Zhang Silong, Luo Junyong, Liu Yan, *et al*. Hotspots detection on microblog[C]//Proc of the 4th International Conference on Multimedia Information Networking and Security. Washington DC: IEEE Press, 2012: 922-925.

[12] Ramage D, Hall D, Nallapati R, *et al*. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora[C]//Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2009: 248-256.

[13] Hu Xia, Sun Nan, Zhang Chao, *et al*. Exploiting internal and external semantics for the clustering of short texts using world knowledge[C]//Proc of the 18th ACM Conference on Information and Knowledge management. New York: ACM Press, 2009: 919-928.

[14] Abel F, Gao Qi, Houben G J, *et al*. Semantic enrichment of twitter posts for user profile construction on the social Web[C]//Proc of the 8th Extended Semantic Web Conference on the Semantic Web: Research and Applications. Berlin: Springer-Verlag, 2011: 375-389.

[15] Musat C C, Velcin J, Trausan-Matu S, *et al*. Improving topic evaluation using conceptual knowledge[C]//Proc of the 22nd International Joint Conference on Artifical Intelligence. San Francisco: AAAI Press, 2011: 1866-1871.

[16] 王广新. 基于微博的用户兴趣分析与个性化信息推荐[D]. 上海:上海交通大学,2013.

[17] 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[J]. 计算机科学,2013,40(4):45-53.

[18] Welch M J, Schonfeld U, He Dan, *et al*. Topical semantics of twitter links[C]//Proc of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2011: 327-336.

[19] Ma Yunfei, Zeng Yi, Ren Xu, *et al*. User interests modeling based on multi-source personal information fusion and semantic reasoning[C]//Lecture Notes in Computer Science, vol 6890. Berlin: Springer, 2011: 195-205.

[20] Du Yajun, Hai Yufeng. Semantic ranking of Web pages based on formal concept analysis[J]. Journal of Systems and Software, 2013, 86(1): 187-197.

[21] Ramage D, Dumais S, Liebling D. Characterizing microblogs with topic models[C]//Proc of the 4th International Conference on Weblogs & Social Media. Palo Alto, CA: AAAI Press, 2010: 130-137.

[22] Hong Liangjie, Davison B D. Empirical study of topic modeling in Twitter[C]//Procs of the 1st Workshop on Social Media Analytics. New York: ACM Press, 2012: 80-88.

[23] Weng Jianshu, Lim E P, Jiang Jing, *et al*. TwitterRank: finding topic sensitive influential twitterers[C]//Proc of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010: 261-270.