

基于 LDA 模型和多层聚类的微博话题检测

刘红兵¹, 李文坤², 张仰森²

(1. 太原科技大学 电子信息学院, 山西 太原 030024;

2. 北京信息科技大学 智能信息处理研究所, 北京 100192)

摘要:随着微博这一新兴社交媒体的广泛应用,以微博为背景的相关研究不断涌现,其中基于微博的话题检测是当前研究的热点之一。结合微博文本的相关特点,文中提出了一种基于 LDA 模型和多层聚类的微博话题检测方法。首先,通过 LDA 模型对微博数据建模并提取特征;其次,利用改进的 Single-Pass 聚类 and 层次聚类对微博数据进行聚类,从而发现热点话题。通过在大规模微博数据上进行话题检测实验,通过 LDA 建模比通过 TF-IDF 进行特征选择和权重计算效果好;改进的 Single-Pass 聚类能够处理第一遍 Single-Pass 聚类未处理的微博,提高了初步聚类的精度,并且为下一步层次聚类减少了时间;多层聚类的聚类效果在准确率、召回率和 F 值三方面均比单一聚类算法的聚类效果好。显然,文中的话题检测方法是可行的,也是有效的。

关键词:LDA 模型;话题检测;改进的 Single-Pass 聚类;层次聚类

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2016)06-0025-06

doi:10.3969/j.issn.1673-629X.2016.06.006

Microblog Topic Detection Based on LDA Model and Multi-level Clustering

LIU Hong-bing¹, LI Wen-kun², ZHANG Yang-sen²

(1. College of Electronic Information, Taiyuan University of Science and Technology,
Taiyuan 030024, China;

2. Institute of Intelligence Information Processing, Beijing University of Information Science and Technology,
Beijing 100192, China)

Abstract: With the wide application of microblog, emerging social media, relevant research is being emerged on microblog. The topic detection based on microblog is one of the hotspots in current research. In combination with the relevant characteristics of microblog, a microblog topic detection based on LDA model and hierarchical clustering is proposed. First, LDA model is applied for modeling and feature extraction to microblog data. Then, the improved Single-Pass clustering and hierarchical clustering is used on microblog data clustering and the hot topic is found. Experiment on large-scale corpus shows that it is more effective through the LDA model than by TF-IDF for feature selection and weight calculation; the improved Single-Pass clustering can deal with the untreated microblog by the first Single-Pass clustering, which can improve the accuracy of the initial clustering and reduce the time of hierarchical clustering; it is more effective through the hierarchical clustering than the single clustering in accuracy, recall and F -value. Clearly, it is feasible and effective by the LDA model and multi-level clustering to detect the microblog topic.

Key words: LDA model; topic detection; improved Single-Pass clustering; hierarchical clustering

0 引言

随着互联网技术的发展及其广泛的应用,包括微博、社交网站、即时通讯等在内的一些新兴社交媒体正在从根本上改变着人们的生活。据中国互联网信息中

心(CNNIC)发布的《第 34 次中国互联网络发展状况统计报告》^[1]显示:截至 2014 年 6 月底,我国网民规模达 6.32 亿,较 2013 年底增加了 1 442 万人。然而,微博网民规模为 2.75 亿,占有网民的 43.6%。微博已

收稿日期:2014-11-14

修回日期:2015-04-08

网络出版时间:2016-05-25

基金项目:国家自然科学基金资助项目(61370139);北京市属高等学校创新团队建设与教师职业发展计划项目(IDHT20130519);北京市教委专项基金(PXM2013_014224_000042, PXM2014_014224_000067)

作者简介:刘红兵(1968-),男,副教授,研究方向为智能计算机控制。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160525.1700.006.html>

经成为人们在线交流和传播信息的主要平台,已经成为社会舆情传播的重要载体,一些重要的热点事件或商业信息都首先通过微博进行报道。微博上的热点话题一般来源于突发事件的报道、具有重要新闻价值的信息或者引起讨论、共鸣甚至争论的用户交流,很大程度上反映着当前社会的舆论方向。对这些话题进行实时检测可以帮助用户快速了解目前的热点话题、热门事件,也能够帮助政府及时了解社会动态、知道民众的想法。随着微博的进一步发展和日益普及,开展微博平台上的话题检测技术研究迫在眉睫。

1 研究现状

近年来,有关微博的研究受到了学术界和企业界的广泛关注,针对微博的研究也越来越多。同时,微博话题检测也有了相应的进展。

Peng 等^[2]总结了热门话题的特征,提出了一种基于用户喜好的热门话题检测方法。Ramage 等^[3]分析了 Twitter 数据的特征,利用 Labeled LDA 模型进行特征提取,并实现了 Twitter 排序和推荐功能。Du 等^[4]通过 PangRank 算法抽取出关键用户,然后结合语义信息提取突发特征,进而发现微博中的突发事件。孙励^[5]采用 LDA 模型发现微博热点话题,并用主题代表话题。此方法虽然能够解决微博数据稀疏问题,但是话题检测性能有待提高。邱洋^[6]分析了微博的特点,在计算相似度时融入了语义和时间参数,然后采用 Single-Pass 算法进行话题检测。路荣等^[7]利用隐语义分析解决微博短文本数据稀疏问题,然后选取每个时间窗内最有可能是谈论新闻话题的微博,最后采用 K-means 和层次聚类进行微博热点话题检测。孙胜平^[8]采用 SP&HA 混合聚类发现微博中的话题,并通过实验验证了该方法的有效性。马雯雯等^[9]首先采用隐语义分析(LSA)对微博数据建模,然后利用层次聚类的 CURE 算法确定 K-means 的初始类,最后通过 K-means 算法发现微博话题。蒋洪梅^[10]对微博的舆论影响特点进行了具体论述,并对如何更好地利用微博进行舆论引导作了尝试性的探讨。彭泽映等^[11]通过观察和分析发现基于微博的大规模短文本所具有的“长尾分布”的特性,提出了一种基于不完全聚类思想用以对这类数据进行聚类分析,一定程度上解决了传统聚类算法难以对大规模短文本进行有效处理的问题。马彬等^[12]提出了一种基于线索树双层聚类的微博话题检测方法。首先建立微博线索树,然后在线索树内部进行局部聚类,最后进行全局聚类发现微博话题。史剑虹等^[13]通过隐主题分析挖掘微博中的隐含主题信息,然后采用聚类算法和频繁项集挖掘技术进行微博话题检测并提取话题关键词集。

在前人研究的基础上,文中提出了一种新的基于 LDA 模型和多层聚类的微博话题检测方法。通过 LDA 模型挖掘微博文本中潜在的主题信息,解决微博数据的数据稀疏问题,同时采用融合改进的 Single-Pass 聚类算法和层次聚类算法进行微博话题检测。实验结果表明,该方法能够从大规模微博语料中准确地检测出当前的热点话题。

2 关键技术

2.1 LDA 模型

LDA 模型^[14]首先由 Blei 等于 2003 年提出,是现今最流行的一种文档主题生成模型。LDA 模型适于对文本进行“隐性语义分析”,可以用来识别大规模文档集或语料库中潜藏的主题信息,目的是将文档集或语料库中的每篇文档的主题按照概率分布的形式给出。而且它也是一种无监督的学习算法,不需要任何关于文档的背景知识和已标注的训练语料。

LDA 模型也是一个三层贝叶斯概率模型,包含词、主题和文档三层结构。其中,文档到主题服从 Dirichlet 分布,主题到词服从多项式分布。它采用产生式全概率模型对文档进行建模,对于给定的一个文档集, LDA 将每一篇文档用若干主题的概率表示,将每个主题用所有的词的概率表示。每篇文档的主题都服从特定的分布,主题之间也相互独立,并且被所有文档共享。LDA 模型生成文档的过程如图 1 所示。

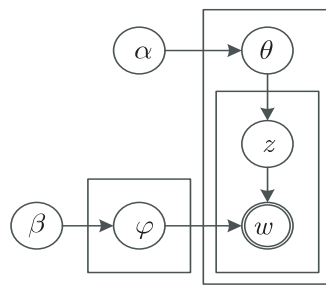


图 1 LDA 模型

图中, θ, φ, z 都是隐藏变量, w 是可见变量, 方框中的内容表示循环执行。 α 是每篇文档下主题的多项式分布的 Dirichlet 先验参数, β 是每个主题下词的多项式分布的 Dirichlet 先验参数, θ 表示该文档的主题分布, φ 表示该主题的词分布, z 表示每篇文档分配在每个词上的主题, w 表示每篇文档的词向量。概率生成模型的计算公式如式(1)所示。

$$P(\theta, z | w, \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^k p(z_n | \theta) p(w_n | z_n, \beta) \quad (1)$$

LDA 模型中隐藏参数的估计也称为 LDA 的 Inference 问题,通常采用 EM 算法和吉布斯采样(Gibbs Sampling)进行学习估计。Gibbs 采样是由 Thomas L.

Griffith 等人提出的,它是 MCMC 的一个二维实现方法,比较适合大规模数据的处理,是目前最流行的参数估计算法。这个算法的运行方式是每次选取概率向量的一个维度,给定其他维度的变量值 Sample 当前维度的值。不断迭代,直到收敛输出待估计的参数。文中也采用 Gibbs 采样对 LDA 模型的参数进行估计。LDA 模型对文档集建模的最终结果如下:

- (1) z 文件,它的每一行表示原始文档集中的一个文档。它把所有的词用该词所对应的隐主题替换,然后用这些隐主题表示文档。
- (2) ϕ 文件,即文档-主题矩阵 $M * K$ 。 M 表示文档集中的文档数, K 表示主题数。
- (3) θ 文件,即主题-词矩阵 $K * V$ 。 K 表示主题数, V 表示文档集中词的个数。
- (4) t words 文件,它将所有的主题用概率最高的那些特定的词表示,显示每个主题的具体内容。

传统 LDA 模型中, V 指文档集中所有不相同的词的个数,但是,对于话题检测来说并不是所有的词都有实际的语义。正如副词、介词、连词、助词、叹词和拟声词等,这些词都依附于实词,没有具体的语义,对话题检测没有作用,而且影响系统的性能。文中采用 LDA 建模时,对传统 LDA 模型中 V 的选择进行改进,只保留名词、动词、形容词。这样做不仅能提高 LDA 模型的性能,而且能降低建模时间。

2.2 多层聚类

2.2.1 Single-Pass 聚类

Single-Pass 聚类^[8]是单遍聚类,属于增量式聚类算法中的一种。Single-Pass 聚类算法的基本思想是:按照文档输入的顺序依次处理每个文档,把第一个文档认为是第一个话题,后续输入的每个文档都与之前创建的话题进行相似度计算,并找出与该文档相似度最大的那个话题,如果相似度大于阈值,那么将该文档归入此话题并更新话题簇,否则用该文档创建一个新话题,一直循环此过程,直到所有文档处理完毕,算法结束。

Single-Pass 算法的优点是算法逻辑简单,执行效率较高,而且该算法对输入文档的顺序敏感,比较适合微博话题检测。

文中在传统 Single-Pass 聚类的基础上进行改进,得到一个适合于微博话题检测的聚类算法,具体内容详见第三节。

2.2.2 凝聚式层次聚类

层次聚类也是一种常用的聚类算法,分为分裂式层次聚类和凝聚式层次聚类。分裂式层次聚类是自顶向下的层次聚类,凝聚式层次聚类是自底向上的层次聚类。凝聚式层次聚类非常适合话题检测,其运用到

话题检测的思想是:把每一个文档当作初始的类簇,然后计算各个类簇之间的相似度并找出最大相似度和相应的类簇,如果该值大于预定的阈值,那么将这两个类簇合并并更新簇的中心,通过不断的合并与更新得到最终的话题簇。凝聚式层次聚类能够较准确地对微博话题进行检测,但是凝聚式层次聚类每次合并都要计算簇之间的相似度,算法时间复杂度是 $O(n^3)$,对于大规模数据集凝聚式层次聚类很难在短时间内完成。

3 基于 LDA 模型和多层聚类的微博话题检测

3.1 微博语料预处理

由于刚抓取的微博含有大量噪声,因此需要对微博语料进行预处理。通过对微博语料的观察分析,发现许多微博文本中含有大量的繁体字和链接。如果对这些繁体字和链接不做处理,那么将会对 LDA 模型的训练以及聚类产生很大的影响。文中利用现有的繁简体字对照表对微博文本进行处理,消除繁体字,同时删除微博中所有的链接,使微博文本规范化。

此外,语料中含有大量的重复微博和字数过少的微博。例如,“转发微博”,这类微博不仅对话题检测毫无意义,而且会影响系统性能。因此,去掉重复微博和字数过少的微博也是至关重要的。

微博用户在转发互动中形成的微博大都具有语义相关性,通常是对同一个话题的讨论。对于具有转发关系的微博文本,把原创微博与转发微博进行合并,形成一个语义更加丰富的长文本来替换原始微博,解决微博话题检测的数据稀疏问题。

3.2 改进的 Single-Pass 聚类

传统的 Single-Pass 聚类只使用一次循环遍历所有微博,完成聚类。事实上,有很多微博虽然属于某一个话题,但是由于它发布时间较早,较早完成遍历,这样可能导致这些微博因为与之前得到的话题的相似度略低于阈值而被重新创建了新的话题,从而影响了聚类效果。

```
算法 1:改进的 Single-Pass 聚类算法。
输入:按时间顺序排好序的微博集  $D = \{d_1, d_2, \dots, d_n\}$ 
输出:话题簇  $T_1, T_2, \dots$ 
For count from 1 to n
if (count = 1) then
 $d[count] \rightarrow T_1$  //创建新话题  $T_1$ 
else
maxSim = 0
for i from 1 to 已经创建的话题数
if (sim(  $d[count]$ ,  $T[i]$  ) > maxSim) then
```



```
maxSim=sim( d [ count] , T[ i ] )
clusterNo= i
end if
end for
if( maxSim>= 阈值) then
d [ count] -> T [ clusterNo] //归入话题
update and save T [ clusterNo]
end if
end if
End for
For count from 1 to 没有归入话题的微博数
maxSim=0
for i from 1 to 已经创建的话题数
if( sim( d [ count] , T[ i ] )>maxSim) then
maxSim=sim( d [ count] , T[ i ] )
clusterNo= i
end if
end for
if( maxSim>= 阈值) then
d [ count] -> T [ clusterNo] //归入话题
update and save T [ clusterNo]
else
create new topic
end if
End for
```

文中提出了一种新的改进的 Single-Pass 聚类。该算法在传统 Single-Pass 聚类的基础上,处理了那些漏掉的微博,使聚类更加准确。对于给定的一个微博集 $D = \{d_1, d_2, \dots, d_n\}$,改进的 Single-Pass 聚类的算法如算法 1 所示。

3.3 微博话题检测

文中首先通过 LDA 模型对微博文本进行建模,提取特征,然后采用多层聚类算法对微博文本聚类实现话题检测。多层聚类分两阶段进行,第一步利用改进的 Single-Pass 聚类进行话题初步检测,第二步利用层次聚类对上一步产生的中间结果再次聚类形成最终的话题。改进的 Single-Pass 聚类算法逻辑简单,能够快速处理大规模文本,但是聚类精度一般;凝聚式层次聚类的聚类精度高,但是算法的时间复杂度也较大。

文中利用 LDA 模型有效解决了微博的数据稀疏问题,同时结合改进的 Single-Pass 聚类和层次聚类的优点,使话题检测系统在准确率和时间上都有很大提高。

系统流程图如图 2 所示。

3.4 关键字提取

随着信息时代的到来,每天都有成千上万的信息展现在人们面前,如何快速了解海量信息中谈论的热点话题并且找出自己感兴趣的话题,不论对于个人还

是企业,都是十分重要的。文中利用多层聚类算法检测出的微博话题都是以微博簇的形式存在的,每个微博簇都是谈论某一个话题的微博文本集。虽然可以把谈论同一话题的微博聚集到一个话题簇中,但是要想确定该话题簇具体谈论的话题内容,仍然需要一条一条地阅读微博。因此,检测出微博话题是不够的,还需要用三到五个关键字概括出微博话题的主要内容。本节主要介绍关键字提取,即从已检测出的微博话题中,抽取主要的关键字表示该话题。

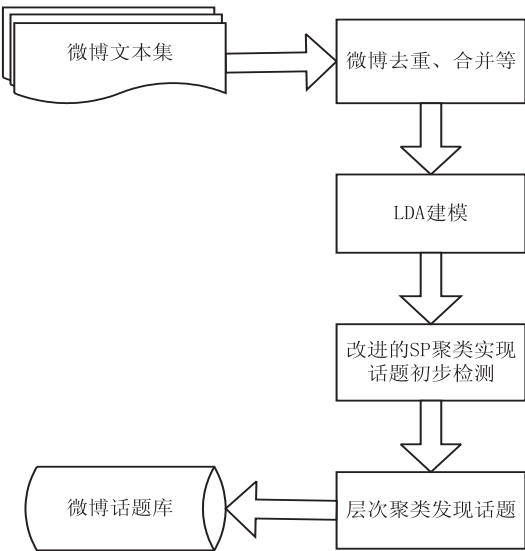


图 2 系统流程图

在关键字提取中,用 TF-IDF 度量每个词语的重要度。经过多次实验后,最终选择 TF-IDF 排名前三的词语作为话题关键字。提取过程如下:

- (1)将每一个话题中的所有微博作为一个整体,分词,去停用词;
- (2)计算第一个话题中去掉停用词后剩下的词语在所有语料中的 TF-IDF 值;
- (3)根据 TF-IDF 值排序,选择 TF-IDF 值排名前三的词语作为该话题的关键字;
- (4)重复步骤(2)和(3),直到所有话题关键字提取完毕为止。

表 1 展示了各话题中的部分微博和 TF-IDF 排名前三的词语。话题一主要以央视曝光星巴克咖啡牟取暴利的行为为背景展开的讨论,抽取出来的话题关键字是“星巴克、咖啡、贵”,这与话题内容基本上吻合。话题二是关于高考改革引发的讨论,主要是关于是否取消英语和数学的讨论,然而抽取出的话题关键字是“英语、数学、高考”,这与话题二的内容也是相吻合的。仔细分析话题三和话题四,话题关键字和微博内容也基本上是吻合的,说明采用 TF-IDF 提取出的话题关键字基本上可以概括出话题的主要内容,而且效果也是不错的。

表 1 微博话题和话题关键字

话题序号	话题关键字	微博内容
1	星巴克、咖啡、贵	星巴克为何能在中国卖高价 【中国星巴克卖得比美日台还贵,凭什么?】逛街时手里拿着一杯星巴克就是“潮”,这家企业在中国大陆卖得超级贵。以小杯来说,在美国卖 2.75 美元(RMB16.9 元);日本卖 320 日元(RMB19.9 元),台湾卖 95 台币(RMB19.8 元),但在大陆一杯却高达 27 元。有人说:爷有钱,喝得起这种高级货,太便宜还不愿买哩 【暴利咖啡】我家有人是星巴克拥趸,真心感觉这家咖啡不便宜 我觉得吧,只要星巴克不逼着我每晚七点喝一杯他们家的咖啡,他爱卖多贵都行,嫌贵不喝就是了
2	英语、数学、高考	高考取消英语的和闭关锁国有何区别 语文高考 180 分,大赞!我觉得整个应试教育的课程中也就语文和英语还是挺有用的 #数学滚出高考#,可能吗?醒醒吧!不想让自己家孩纸继续被数学虐,从娃娃抓起吧,家长们,看过来,有帮助哦! http://t.cn/zWJahio 【英语将不再是高考必考科目】江苏省在酝酿高考改革,英语将改为一年两考,不再计入总分,很多英语成绩不佳的同学欢呼雀跃。10 月 18 日,中新网消息:北京也在酝酿高考英语科目的改革,或将不再计分,而实行等级考试,作为高校录取新生的参考。 http://t.cn/zRfzxU7 ——中国人也许真的不用再学英语了~
3	王菲、离婚、李亚鹏	王菲宣布离婚的微博,1 小时 22 万转发,9 万评论,问题是 2 万点赞的人是什么心态? 重磅消息!!!我一电视台的兄弟跟我说,王菲和李亚鹏离婚了!! 王菲和李亚鹏离婚的主要原因据说是为了在北京买二套房 【#王菲李亚鹏离婚# 李亚鹏,放手是唯一能做的】今日下午 19 时 30 分左右,王菲在微博中发文:“这一世,夫妻缘尽至此。我还好,你也保重。”两人注定不能一起#将爱情进行到底# http://t.cn/h5yV30
4	余姚、水、救援	【蓝媒新闻】目前余姚受灾区所需物资:皮划艇,水,面包,饼干,小孩吃的能消化的食品,蜡烛,常用药,等等 #余姚加油#还有多少个城市还在下雨?又有多少被水淹了……伙伴们你那里还好么? 【蓝媒新闻】【警察蜀黍的脚】这次参加余姚救援的警察,有的腿上敞着伤口泡在雨水里继续救援;有的双脚由于长期浸泡在水中引起很多红斑甚至红肿发脓;有的 7 天长假 6 天是在工作岗位上度过的……见到素不相识的警察,说一声辛苦了吧。这座城市有他们而美丽,向他们致敬! 大灾面前,民众肯定会抱怨政府。余姚的情况,很多人家是多天下不去楼。那条抱怨军车不救援的微博,我会劝慰贴主“军车多半有特定任务,不能为救援帮助停下来。你说的的军人态度问题可能是人家工作了很久表情木然”。但贴主着急有情绪很正常,把人家口语化叙述歪曲成造谣恶毒咒骂的人真的很脏

4 实 验

4.1 实验数据及评价指标

目前,在中文微博话题检测方面还没有统一的微博数据。文中通过网络爬虫,抓取了新浪微博 2 352 个用户发表于 2013 年 6 月 1 号到 2013 年 10 月 31 号之间的所有微博数据。经过语料去重和噪声微博过滤,剩下的微博数据用于实验。

在自然语言处理领域,常用的评价指标有准确率、召回率和 F 值。文中除了使用传统的这三个评价指标以外,还采用漏检率和错检率评价文中的微博话题检测系统的性能。

具体的计算公式如下所示:

$$P = \frac{D}{U} * 100\%$$

其中, P 表示准确率; D 表示话题检测系统正确检测出的属于该话题的微博数; U 表示话题检测系统实际检测出的属于该话题的微博数。

$$R = \frac{D}{T} * 100\%$$

其中, R 表示召回率; D 表示话题检测系统正确检测出的属于该话题的微博数; T 表示语料中所有属于该话题的微博数。

$$F = \frac{2 * P * R}{P + R} * 100\%$$

其中, F 表示 F 值; P 和 R 分别表示准确率和召回率。

$$P_{FA} = \frac{FA}{NT} * 100\%$$

其中, P_{FA} 表示错检率; FA 表示话题检测系统错误检测出的属于该话题的微博数; NT 表示语料中所有不属于该话题的微博数。

$$P_{MISS} = \frac{MD}{T} * 100\%$$

其中, P_{MISS} 表示漏检率; MD 表示话题检测系统没有检测出的属于该话题的微博数; T 表示语料中所有属于该话题的微博数。

4.2 对比实验及实验结果分析

实验一:为了验证改进的 Single-Pass 聚类 and 凝聚式层次聚类对话题检测的影响,文中设置四个系统,四个系统均采用余弦相似度度量微博之间的相似性,具体设置如下:

sys1:只采用 Single-Pass 聚类。

sys2:在 sys1 的基础上融入了层次聚类。

sys3:只采用改进的 Single-Pass 聚类。

sys4:在 sys3 的基础上融入了层次聚类。

实验中,分别用 TF-IDF 和 LDA 模型进行特征选择,由于采用 TF-IDF 进行特征选择时,一些话题根本无法检测出来,一些评价指标都无法计算,无法进行准确地比较。采用 TF-IDF 进行特征选择时,各个系统的话题检测的效果比 LDA 模型的均较差,所以在此不再赘述。

图 3 显示了在采用 LDA 模型进行特征选择的条件下,四种不同的聚类策略进行话题检测的实验结果。

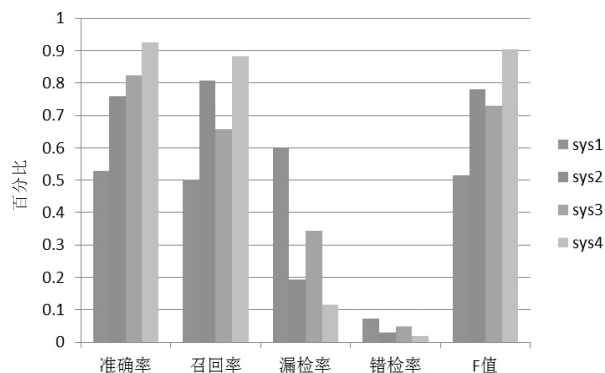


图 3 不同聚类算法下话题检测的性能比较

从图中可以看出,在五个评价指标中 sys1 的性能最差,sys2 和 sys3 的性能居中,sys4 的性能优于其他三个系统,说明采用文中提出的方法完全能够满足话题检测的要求。sys2 和 sys3 比 sys1 在各方面都有所提高,说明层次聚类和改进的 Single-Pass 聚类都能提高话题检测的性能。sys2 在召回率方面优于 sys3,但在准确率方面不及 sys3,说明层次聚类更侧重于召回率的提高,而改进的 Single-Pass 聚类更侧重于准确率的提高。其主要原因是由于改进的 Single-Pass 聚类采

用层叠 Single-Pass 聚类方法,其第二次的 Single-Pass 聚类建立在第一次 Single-Pass 聚类基础上,可以有效处理第一次 Single-Pass 聚类未能处理的微博。而且,层次聚类能够把 Single-Pass 聚类处理完的微博再次整合,提高话题检测效率。其中,sys2 就是文献[8]所采用的聚类算法,由图 3 可以看出,文中方法与文献[8]的话题检测方法相比,各个指标都有提高, F 值提高约 12%。

实验二:为了评估不同阈值对话题检测结果的影响,该实验设置不同的阈值进行话题检测,得到的结果如图 4 所示。

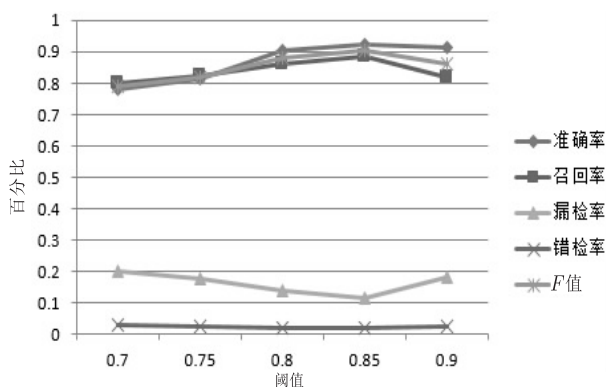


图 4 不同阈值话题检测的性能比较

由图 4 可以看出:随着阈值的不断增大,话题检测的准确率、召回率和 F 值逐渐增大,话题检测系统的性能持续提高;但是当阈值超过 0.85 时,这三个指标开始下降,系统性能也开始下降。

5 结束语

文中根据微博内容的简短性、微博话题的时序性以及微博文本之间存在转发关系等特点,提出了一种基于 LDA 模型和多层聚类的微博话题检测方法。通过合并具有转发关系的微博,以及采用 LDA 模型选取特征,有效解决了微博短文本的数据稀疏问题。通过融合改进 Single-Pass 聚类和层次聚类,能够在保证话题检测性能的前提下更大程度地缩短话题检测时间。最后,通过 TF-IDF 对微博中的词语进行重要度排序,用排名前三的词语作为话题关键字,代表话题的主要内容。

由于微博文本比较随意,口语化较强,网络词语也出现频繁,用现有的分词工具处理微博文本时并不是很理想,导致文中的话题检测性能有所下降。同时,微博文本中会出现大量的同义词,也会影响系统的性能。在下一步的研究中,首先要丰富用户字典,确保分词更加准确;其次要引入同义词字典,处理微博文本中的同义词,进一步提高系统的性能。

通用户的影响力等等。社交系统中的信息非常丰富,数据量非常大,也可以考虑对数据进行分布式处理,这样可以对更多的数据进行挖掘,提高处理的效率。

参考文献:

- [1] Resnick P, Varian H R. Recommender systems[J]. Communications of the ACM, 1997, 40(3): 56-58.
- [2] Linden G, Smith B, York J. Amazon. com recommendations item-to-item collaborative filtering[J]. IEEE Internet Computing, 2003, 7(1): 76-80.
- [3] 王嫣然, 陈梅, 王翰虎, 等. 一种基于内容过滤的科技文献推荐算法[J]. 计算机技术与发展, 2011, 21(2): 66-69.
- [4] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the 14th conference on uncertainty in artificial intelligence. [s. l.]: [s. n.], 1998: 43-52.
- [5] Resnick P, Iakovou N, Sushak M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]//Proceedings of the 1994 computer supported cooperative work conference. [s. l.]: [s. n.], 1994: 175-186.
- [6] Sarwar B, Karypis G, Konstan J. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international conference on world wide web. [s. l.]: [s. n.], 2001: 285-295.
- [7] 陈克寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报, 2013, 36(2): 349-359.
- [8] Chang Pei-Shan, Ting I-Hsien, Wang Shyue-Liang. Towards social recommendation system based on the data from micro-

logs[C]//Proc of international conference on advances in social networks analysis and mining. [s. l.]: IEEE, 2011: 672-677.

- [9] Jiang Meng, Cui Peng, Wang Fei, et al. Scalable recommendation with social contextual information[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(11): 2789-2802.
- [10] 荣辉桂, 火生旭, 胡春华, 等. 基于用户相似度的协同过滤推荐算法[J]. 通信学报, 2014, 35(2): 16-24.
- [11] 郭磊, 马军, 陈竹敏, 等. 一种结合推荐对象间关联关系的社会化推荐算法[J]. 计算机学报, 2014, 37(1): 219-228.
- [12] Jonnalagedda N, Gauch S. Personalized news recommendation using twitter[C]//Proc of IEEE/WIC/ACM International conferences on web intelligence and intelligent agent technology. [s. l.]: IEEE, 2013: 21-25.
- [13] Islam M, Ding Chen, Chi Chi-Hung. Personalized recommender system on whom to follow in twitter[C]//Proceedings of IEEE fourth international conference on big data and cloud computing. [s. l.]: IEEE, 2014: 326-333.
- [14] Deng Yingying, Lu Tun, Xia Huanhuan, et al. AOPUT: a recommendation framework based on social activities and content interests[C]//Proceedings of IEEE 17th international conference on computer supported cooperative work in design. [s. l.]: IEEE, 2013: 545-550.
- [15] 黄创光, 印鉴, 汪静, 等. 不确定近邻的协同过滤推荐算法[J]. 计算机学报, 2010, 33(8): 1369-1377.

(上接第30页)

参考文献:

- [1] 中国互联网络信息中心. 中国互联网络发展状况统计报告[R]. 北京: 中国互联网络信息中心, 2014.
- [2] Peng Feifei, Qian Xu, Li Gaoren. A research of hot topic detection through microblogging[C]//Proc of 4th international conference on intelligent human-machine systems and cybernetics. [s. l.]: IEEE, 2012.
- [3] Ramages D, Dumais S, Liebling D. Characterizing microblogs with topic models[C]//Proceedings of the fourth international AAAI conference on weblogs and social media. Washington, DC: [s. n.], 2010.
- [4] Du Y Y, He Y X, Tian Y. Microblog Bursty topic detection based on user relationship[C]//Proceedings of the 2011 IEEE joint international information technology and artificial intelligence conference. Piscataway: IEEE, 2011: 260-263.
- [5] 孙励. 基于微博的热点话题发现[D]. 北京: 北京邮电大学, 2013.
- [6] 邱洋. 微博数据提取及话题检测方法研究[D]. 大连: 大

连理工大学, 2013.

- [7] 路荣, 项亮, 刘明荣, 等. 基于隐主题分析和文本聚类的微博客中新闻话题的发现[J]. 模式识别与人工智能, 2012, 25(3): 382-387.
- [8] 孙胜平. 中文微博客热点话题检测与跟踪技术研究[D]. 北京: 北京交通大学, 2011.
- [9] 马雯雯, 魏文哈, 邓一贵. 基于隐含语义分析的微博话题发现方法[J]. 计算机工程与应用, 2014, 50(1): 96-100.
- [10] 蒋洪梅. 微博客的特点及其舆论影响力[J]. 新闻爱好者, 2011(5): 85-86.
- [11] 彭泽映, 俞晓明, 许洪波, 等. 大规模短文本的不完全聚类[J]. 中文信息学报, 2011, 25(1): 54-59.
- [12] 马彬, 洪宇, 陆剑江, 等. 基于线索树双层聚类的微博话题检测[J]. 中文信息学报, 2012, 26(6): 121-128.
- [13] 史剑虹, 陈兴蜀, 王文贤. 基于隐主题分析的中文微博话题发现[J]. 计算机应用研究, 2014, 31(3): 700-704.
- [14] Blei M, Ng Y, Jordan I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(4-5): 993-1002.