Journal of Computer Applications
计算机应用,2012,32(8):2346 - 2349
ISSN 1001-9081
CODEN JYIIDU
2012-08-01
http://www.joca.cn

文章编号:1001 - 9081(2012)08 - 2346 - 04
doi:10.3724/SP.J.1087.2012.02346

基于特定领域的中文微博热点话题挖掘系统 BTopicMiner

李 劲[1,2*],张 华[1],吴浩雄[1],向 军[1]
(1.湖北民族学院 信息工程学院,湖北 恩施 445000; 2.华中师范大学 信息管理系,武汉 430079)
(* 通信作者电子邮箱 lj05921@tom.com)

摘 要: 随着微博应用的迅猛发展,自动地从海量微博信息中提取出用户感兴趣的热点话题成为一个具有挑战性的研究课题。为此研究并提出了基于扩展的话题模型的中文微博热点话题抽取算法。为了解决微博信息固有的数据稀疏性问题,算法首先利用文本聚类方法将内容相关的微博消息合成为微博文档;基于微博之间的跟帖关系蕴含着话题的关联性的假设,算法对传统潜在狄利克雷分配(LDA)话题模型进行扩展以建模微博之间的跟帖关系;最后利用互信息(MI)计算被抽取出的话题的话题词汇用于热点话题推荐。为了验证扩展的话题抽取模型的有效性,实现了一个基于特定领域的中文微博热点话题挖掘的原型系统——BTopicMiner。实验结果表明:基于微博跟帖关系的扩展话题模型可以更准确地自动提取微博中的热点话题,同时利用 MI 度量自动计算得到的话题词汇和人工挑选的热点词汇之间的语义相似度达到 75% 以上。

关键词: 数据挖掘;信息检索;微博;话题模型;文本聚类;互信息
中图分类号: TP311.52 **文献标志码:** A

BTopicMiner: domain-specific topic mining system for Chinese microblog

LI Jin[1,2*], ZHANG Hua[1], WU Hao-xiong[1], XIANG Jun[1]
(1. *School of Information Engineering, Hubei University for Nationalities, Enshi Hubei* 445000, *China*;
2. *Department of Information and Management, Central China Normal University, Wuhan Hubei* 430079, *China*)

Abstract: As microblog application grows rapidly, how to extract users' interested popular topic from massive microblog information automatically becomes a challenging research area. This paper studied and proposed a topic extraction algorithm of Chinese microblog based on extended topic model. In order to deal with data sparse problem of microblog, the content related microblog text would be firstly clustered to generate synthetic document. Based on the assumption that posting relationship among microblogs implied topical correlation, the traditional LDA (Latent Dirichlet Allocation) topic model was extended to model the posting relationship among microblogs. At last, Mutual Information (MI) measurement was utilized to calculate topic vocabulary after extracting topics by proposing extended LDA topic model for topic recommendation. Furthermore, a prototype system for domain-specific topical mining system, named BTopicMiner, was implemented so as to verify the effectiveness of the proposed algorithm. The experimental result shows that the proposed algorithm can extract topics from microblogs more accurately. Meanwhile, the semantic similarity between automatically calculated topic vocabulary and manually selected topic vocabulary exceeds 75% while automatically calculating topic vocabulary based on MI.

Key words: data mining; information retrieval; microblog; topic model; text clustering; Mutual Information (MI)

0 引言

如今,在互联网上的很多内容都是用户生成的。尤其是近几年,微博产生了很强的影响力。微博是一种迷你博客,它让用户在任何时间和地点都可以用手机和电子邮件通过网站发表限定字数的信息。信息的产生者可以记录任何发生在自己日常生活中的事,分享对各种话题的看法。由于方便的发表性,自由的文本格式和易于访问的微博服务,很多互联网用户都从传统的联系工具如博客、BBS 转移到了微博。随着越来越多的用户更愿意发表自己使用产品的意见,描述自己经历过的故事,表达自己的政治观点,微博已经成为了一个关于公共事件、新闻故事、甚至个人情绪观点的有价值的数据源。这些数据不仅可以被用来作科学研究而且也可以用作商业用途,例如通过传播的新闻可以分析消费者对某种产品的评价。在 2011 年 7 月 23 日的温州动车事件中,上百万人由传统信息渠道转向通过像微博这样的微博服务来收集时事新闻或自己感兴趣的事件。可是,微博中有海量的短信息,而且这些信息每天都在增加。在中国随着微博服务大量和快速地增长,微博用户也大幅度的增加。中国互联网公司新浪的微博平台已经宣布,它现在已超过 50 万用户,每天约有 25 万条微博更新发布。因此对用户来说从数量巨大的短信息中找到合适的话题新闻并把听众感兴趣的推荐给他们是一件非常困难的事情。

本文对如何从海量微博信息中挖掘热点话题进行了研究,提出了基于话题空间模型的热点话题挖掘算法,同时实现了系统原型。首先通过话题模型对微博信息进行话题抽取,在此基础上将微博向量从基于单词的向量空间向基于话题的

收稿日期:2012-02-15;修回日期:2012-03-30。
基金项目:国家自然科学基金资助项目(61040006);湖北省自然科学基金资助项目(2010CDZ027);湖北省教育厅科技项目(B20101909)。
作者简介:李劲(1973 -),男,湖北恩施人,副教授,硕士,CCF 会员,博士研究生,主要研究方向:基于互联网的数据挖掘和数据管理、面向云计算的 Web 服务及 Web 服务组合; 张华(1978 -),男,湖北恩施人,讲师,硕士,主要研究方向:信息检索、分布式系统及集成; 吴浩雄(1979 -),男,湖北建始人,工程师,主要研究方向:Web 数据挖掘、信息安全; 向军(1978 -),男,湖北来凤人,讲师,博士,主要研究方向:移动计算、实时数据库系统、软件测试。

空间进行映射,从而将微博表示为话题向量; 进一步地,基于话题向量对微博进行聚类分析,从聚类得到的每一个簇中找出热点话题词汇作为热点话题的表示; 最后,挖掘出的热点话题以 RSS 的方式反馈给感兴趣的用户。

1　相关研究

最近两年国内外开始了针对海量微博信息的数据挖掘研究工作,并取得了一定的进展。其中热点话题和新闻的挖掘算法大致可以分为以下几类: 第一类方法是利用分类聚类方法挖掘出当前热点事件。如 Allan 等[1] 利用单路径聚类算法,结合一个新阈值模型实现了一个在线新闻监测系统; 路荣等[2] 利用一个两层的 K 均值和层次聚类的混合聚类方法,结合隐主题模型找出微博中的热点新闻话题。第二类方法是在传统的话题模型——潜在狄利克雷分配(Latent Dirichlet Allocation, LDA) 模型的基础上针对微博消息直接建立话题模型,利用建立好的模型直接抽取话题。如 Ramage 等[3] 构造了一个半监督学习模型L-LDA将用户和 Twitter 特性化来个性化用户信息需求; Asuncion 等[4] 提出了基于分布式算法的改进的 LDA 和分层的狄利克雷过程(Hierarchical Dirichlet Process, HDP) 话题模型; Blei 等[5] 建立了一个新的话题模型——相关主题模型(Correlated Topic Model, CTM) ,该模型通过正态分布建模话题之间的相关性; Sankaranarayanan 等[6] 实现了一个新闻处理系统 TwitterStand 用于捕捉时下热门 Twitter 话题新闻。另外一类基于微博话题挖掘的研究方法是通过分析微博内容自动产生关于微博的总结(Summarization) 。如 Sharifi 等[7] 实现了用一个句子总结微博话题的方法,使用户可以快速并准确地理解一个热门话题; 在他们的研究基础上,Inouye[8] 提出一种用多个句子总结微博上热门话题的方法,克服了单个句子对话题信息量承载不足的缺陷。

为了提高从海量微博中挖掘热点话题的速度和精度,有学者从微博用户传播影响力的角度进行研究,首先找出有影响力的用户,在此基础上挖掘这些用户的微博消息,可以大大提高挖掘的速度和精度。关于这方面的研究有: Yeung 等[9] 提出一种用户采纳行为的概率模型,推断出在微博传播过程中一个用户对另一用户的影响力; Anagnostopoulos 等[10] 在对大量数据进行统计分析的基础上确定了社会影响是个人行为与社会关系相关性的一个重要来源; Crandall 等[11] 确定并模型化了社会影响和个人选择之间的相互作用; Goyal 等[12] 构造了一个根据传播日志静态和动态计算个人影响力的模型。

另外对微博内容进行情感分析和挖掘,可以发现微博用户对热点新闻话题的态度或情感倾向。关于这方面也有一些相关的研究成果,研究方法主要是基于图模型和文本分类技术,例如 Guerra 等[13] 利用随机游走模型和图模型提出一种转换学习方法来进行实时情感的分析; Silva 等[14] 利用基于情绪规则的分类方法对情感进行预测; Wang 等[15] 基于图的分类方法,将粒度细化到 Hashtag 对话题的情感色彩进行了分类。

2　系统架构

BTopicMiner 包括五个基本组成部分: 微博爬虫、索引器、基于 Web 的用户配置界面、热点话题挖掘引擎和用户推荐。BTopicMiner 基本系统架构如图 1 所示。

图中五个基本组成部分的功能如下:

1) 微博爬虫负责自动从互联网爬取微博并进行语义分析及话题新闻的挖掘。微博爬虫的实现是基于国内最大的微博服务商新浪提供的 API 实现,通过新浪微博 API 可以下载微博用户信息和微博内容信息。

2) 索引器词条化微博内容并在离线库中对已经词条化的词项建立索引。系统使用 Lucene API 来执行微博内容的分词和索引构建。此外,Lucene 的 API 提供了接口来统计分词后的词频信息,例如 TF-IDF 得分,这将有助于热点话题的挖掘。

3) 基于 Web 的用户配置界面,允许用户订阅他们感兴趣的热点话题。用户界面管理用户注册过程,并允许用户提供自己喜爱的 RSS 订阅。

4) 热点话题挖掘引擎负责从微博中挖掘话题新闻,挖掘出来的热点话题基于用户兴趣进行排序。热点话题算法在下一章介绍。

5) 用户推荐是负责将挖掘出来的热点话题转换成 XML 格式的 RSS 提要发送给感兴趣的用户。被挖掘出来的热点话题通过 XLST 样式转换成满足 RSS 要求的 XML 格式发送给用户。

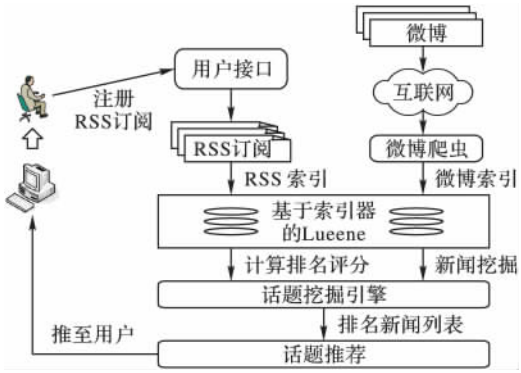

图 1　BTopicMiner 系统架构

3　话题挖掘算法

3.1　基于微博的话题模型

热点话题挖掘算法是系统实现的关键。传统的话题挖掘算法多采用文本聚类方法,其中隐含的假设是: 关于同一个热点话题的文档所用的词汇是相似的,因此如果将文档表示成单词向量,那么关于同一热点话题的文档向量在向量空间中的距离应该是很接近的。基于这样的假设,被聚集在一起的文档应该蕴含着相同的话题。但是基于单词的文档向量表示无法准确地描述出文档的语义,更重要的是: 关于同一话题的文档使用的词汇不一定是相似的。更常见的情况是微博消息所用的词汇完全不同,但却蕴涵着同一话题。为了解决这个问题,在传统的话题模型 LDA 的基础上对微博进行话题建模,从微博中挖掘出有价值的话题。LDA 模型是一种产生式模型,但是和传统的产生式模型有重要的区别。传统的产生式模型认为一个文档只有一个主题(即文档的类别) ,在这个假设的基础上文档的产生过程被描述为

$$p(w) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z) \quad (1)$$

即文档的单词产生过程为: 首先假设文档以概率 $p(z)$ 属于某个主题,以此为条件再以概率 $p(w_n \mid z)$ 产生单词 w_n。但这个模型假设一篇文档只有一个主题是很难成立的。例如一篇关于数据挖掘的论文其中可以有多个主题: 数据挖掘、文本分类、文本聚类等。为了解决这个问题,LDA 模型在文档类别和文档单词之间增加了一个主题层,并将文档单词的产生过程建模为

$$p(w) = \int_{\theta} \left(\prod_{i=1}^{N} \sum_{z_n=1}^{k} p(w_n \mid z_n; \beta) \right) p(z_n \mid \theta) \, p(\theta; \alpha) \, \mathrm{d}\theta \quad (2)$$

即文档的单词产生过程为: 首先以概率分布 $p(\theta)$ 选择参数 θ,再以条件概率 $p(z_n \mid \theta)$ 选择主题 z_n; 在选定主题 z_n 的假设前

提下以条件概率 $p(w_n \mid z_n)$ 选择单词 w_n。式(2)中出现的 α 和 β 为模型参数。

然而 LDA 模型只是对单篇文档建立文档单词产生过程的概率模型,没有考虑文档之间的关系。而对于不同的微博消息,存在着非常重要的关联关系:跟帖关系。即用户可以对某条自己感兴趣的微博消息 M 进行评论(即增加自己的内容)并发表新的微博消息 M',消息 M' 即为消息 M 的跟帖,为了方便描述,将消息 M 称为被引用消息(Cited Message),而消息 M' 称为引用消息(Citing Message)。经过分析,可以发现引用消息和被引用消息有如下重要的性质:1)如果消息 M 没有被引用消息,则消息 M 为原创消息;2)如果消息 M' 有被引用消息,则 M' 只会有一条被引用消息。进一步地,引用消息 M' 有很大可能在话题上和被引用消息 M 的话题相似或者在被引用消息 M 的原有话题基础上增加了新的话题。基于以上分析,将传统的 LDA 模型进行扩展,以对微博消息的这种跟帖关系建模。扩展的 LDA 模型如图 2 所示。

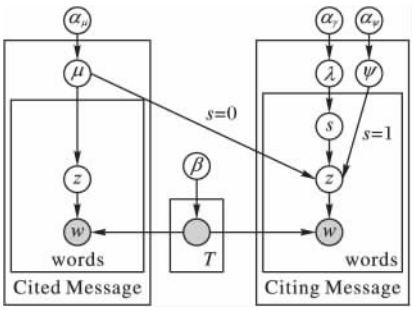

图 2 用于微博话题发现的扩展 LDA 模型

图 2 所示的话题模型描述并建模了微博消息之间的跟帖关系。图中右边为引用消息,引用消息的话题 z 取决于随机变量 s:如果 $s = 1$,表示引用消息为原创消息,该消息的话题由消息本身的话题分布先验概率 α_ψ 和文献本身的话题分布 ψ 决定;如果 $s = 0$,则引用消息是对被引用消息的完全转发,没有增加新的主题,因此其主题完全由被引用消息的分布先验概率 α_μ 和被引用消息本身的话题分布 μ 决定;如果 s 取值为 0 ~ 1,则引用消息的话题 z 由被引用消息和引用消息自身共同决定。参数 λ 决定话题来自被引用消息还是引用消息自身的比例,λ 的分布取决于先验概率 α_λ。因此本文提出的 LDA 扩展模型,充分考虑了微博消息之间的跟帖关系。模型中参数的估计采用的是 Gibbs 采样方法,由于篇幅所限这里不再详述。

3.2 话题抽取的预处理

由于微博消息的文本内容不能超过 140 个字符,因此这种数据的稀疏性会影响话题抽取的效果。在利用扩展 LDA 模型进行话题抽取前,需要对微博进行预处理,所采用的预处理方法为基于单词向量的聚类处理。即首先将微博分词后表示为单词向量,基于单词向量对微博用 K 均值算法进行聚类处理。假设聚类结果为 K 类,将每一类里的微博消息合并成单个文档,则得到了 K 个合成的微博文档,然后再利用 3.1 节提出的扩展 LDA 模型对 K 个合成的微博文档进行话题抽取。这样可以有效地解决微博数据稀疏性问题。

3.3 热点话题词汇的抽取

当利用扩展 LDA 模型对微博进行话题抽取后,可以得到每一个抽取出的话题相关联的词汇和相关联的文档。对于抽取出的话题 T,将该话题生成的单词和文档以产生概率进行排序。对于抽取出的话题 T,该话题生成的单词和相应的生成概率记为 $T\langle w_1: p_1, w_2: p_2, \cdots, w_n: p_n\rangle$;类似地,该话题生成的文档和相应的生成概率记为 $T\langle d_1: p_1, d_2: p_2, \cdots, d_N: p_N\rangle$。其中概率 $p_1, p_2, \cdots, p_n$ 以降序进行排序,同时生成的单词和文档按概率取 TOP N 个。

热点话题词汇抽取基于抽取出的话题所产生的文档进行。假设抽取出的话题集合 $ST = \{T_1, T_2, \cdots, T_K\}$($K$ 为话题个数),集合中话题 T_i($I = 1, 2, \cdots, K$)产生的 TOP N(按产生概率排序)个文档为 $T_i\langle d_1: p_1, d_2: p_2, \cdots, d_N: p_N\rangle$。对于每个话题 T_i 产生的 TOP N 文档集合,可以将集合中的每个文档的类别看成 T_i,在此基础上从这些文档中找出最有代表性的单词。采用的算法为计算文档中单词相对于类别 T_i 的互信息(Mutual Information, MI)。对于话题 T_i,单词 w 相对 T_i 的互信息记为 $MI(w, T_i)$,其值定义如下:

$$MI(w, T_i) = \frac{N_{11}}{N}\mathrm{lb}\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\mathrm{lb}\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\mathrm{lb}\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\mathrm{lb}\frac{NN_{00}}{N_{0.}N_{.0}} \quad (3)$$

其中:N_{10} 为包含单词 w 但不属于话题 T_i 的文档总数,N_{11} 为包含单词 w 且属于话题 T_i 的文档总数;N_{01} 为不包含单词 w 但属于话题 T_i 的文档总数,N_{00} 为不包含单词 w 且不属于话题 T_i 的文档总数;$N_{1.}$ 为包含单词 w 的文档总数,$N_{.1}$ 为属于话题 T_i 的文档总数;$N_{0.}$ 为不包含单词 w 的文档总数,$N_{.0}$ 为不属于话题 T_i 的文档总数;N 为总文档数目。

最后对于每个话题 T_i,用式(3)计算出每个单词的互信息,并按降序排序后,取 TOP N 个单词作为话题词汇。

3.4 热点话题微博推荐

为了将和热点话题相关度最高的微博推荐给用户,需要计算每条微博和热点话题的相关度,并以此进行排序。将 3.3 节抽取出来热点话题 T 的话题词汇集合记为 V_T,$V_T = \{v_1, v_2, \cdots, v_n\}$($n$ 为话题词汇的个数);同时将微博消息 M 包含的单词集合记为 W_M,$W_M = \{w_1, w_2, \cdots, w_m\}$($m$ 为消息 M 包含的单词个数)。微博 M 和话题 T 的相关度记为 $C(M, T)$,$C(M, T)$ 的计算方式如下:

$$C(M, T) = \sum_{i=1}^{m} R(w, T) * tfidf(w) \quad (4)$$

其中:$tfidf(w)$ 为单词 w 在消息 M 中的 tfidf 权重,$R(w, T)$ 为单词 w 和话题 T 的关联度。$R(w, T)$ 的计算方式如下:

$$R(w, T) = \begin{cases} 1, & w \in V_T \\ df_w, & w \notin V_T \end{cases} \quad (5)$$

即当单词 w 出现在话题词汇集合 V_T 中时,认为 w 和话题 T 的关联度的关联度为 1;如果单词 w 没有出现在 V_T 中,则计算单词 w 在话题 T 相关文档集合 $T\langle d_1: p_1, d_2: p_2, \cdots, d_N: p_N\rangle$ 中的文档频率 df_w,df_w 等于 w 在文档集合中出现的文档数除以文档总数。利用式(4)和(5)计算微博和话题之间的相关度后,根据相关度大小进行排序再推荐给用户。

4 实验与结果分析

4.1 实验数据

实验数据从国内最大的微博网站新浪微博进行抓取,通过新浪提供的 API 一共抓取了 50 386 个用户,通过分析用户之间的关注(Follow)关系,去掉一些被关注很少的用户后,筛选得到了 10 488 个我们认为比较重要的用户。在此基础上抓取这些用户过去三个月内所发的微博消息,经过分词,去掉停用词,过滤掉单词个数少于 5 个的消息后得到微博文本共 2 204 520 条。

4.2 实验设置与结果分析

在利用扩展的话题模型进行话题抽取之前,首先基于微博的单词向量进行聚类分析,聚类采用 K 均值算法,K 值指定为 100。每次聚类从 2 204 520 条微博中随机抽取出 50 000 条

微博进行聚类,产生聚类结果后将每类的微博消息合并成一个文档。经过反复抽样、聚类、微博消息合并,最后得到合成的微博文档共 10 220 篇。

下一步利用扩展的 LDA 话题模型对得到的 10 220 篇合成微博文档进行话题抽取,抽取的话题个数指定为 5。对于每个抽取出的话题,取产生概率最高的前 100 篇文档作为话题最相关文档进行话题词汇抽取。话题词汇抽取基于 3.3 节描述的互信息算法,对每个话题抽取出互信息最高的 5 个单词作为话题词汇。

为了方便实验对比,采用人工方式从下载的微博消息中挑选出 5 个热点话题的微博消息,并采用词频统计 + 人工挑选的方法列出每个话题的话题词汇; 同时利用扩展的话题模型对这 5 个热点话题的微博进行话题抽取,并利用互信息抽取出话题词汇,通过人工挑选的热点词汇对自动抽取出的话题词汇进行评估,如表 1 所示。

对于名称确定的评价,只能从主观上进行分析和评价,从表 1 人工挑选的话题词汇可以看出,根据算法得到名称基本可以描述对应的话题。本文提出的话题模型和话题词汇抽取算法可以有效地提取出话题及话题词汇。为了定量地比较自动抽取出的热点话题词汇与人工挑选的热点话题词汇之间的语义相似度,利用 HowNet(http: //www. keenage. com/html/e_index. html) 计算词汇之间的相似度。对于话题 T,将自动抽取出的话题词汇表记为 V_T,$V_T = \{V_1, V_2, \cdots, V_N\}$,人工挑选的话题词汇表记为 W_T,$W_T = \{W_1, W_2, \cdots, W_N\}$,词汇 V_i 和 W_j 之间的相似度记为 $Sim(V_i, W_j)$,则词汇表 V_T 和 W_T 之间的语义相似度用以式(6) 进行计算:

$$Sim(V_T, W_T) = \frac{1}{N^* N}\sum_{i=1}^{N}\sum_{j=1}^{N} Sim(V_i, W_j) \quad (6)$$

自动抽取出的话题词汇与人工挑选的话题词汇(仅列出前 5 个) 之间的语义相似度如表 1 所示。

表 1　人工挑选与自动抽取的热点话题和热点词汇的语义相似度

话题名	热点词汇		语义相似度
	人工挑选	自动抽取	
7-23 动车事故	温州,追尾,献血,动车,事故	温州动车,追尾,问责,伤亡人数,铁道部	0.86
日本地震	地震,救援,核电站,寻亲,辐射	地震,核电站,核辐射,日本,海啸	0.90
乔布斯去世	乔布斯,iPhone,苹果,传奇,IPAD	乔布斯,iPhone,伟人,乔帮主,创意	0.75
郭美美事件	红十字会,郭美美,晒富,信用,诚信	红十字会,郭美美,暴富,信心,捐款	0.81
李娜法网夺冠	法网,李娜,大满贯冠军,威武,历史	法网,娜姐,大满贯,给力,亚洲	0.78

5　结语

本文对如何从海量微博消息数据集中自动检测出热点话题和词汇话题进行了研究。通过分析微博消息的跟帖关系,发现微博的跟帖关系蕴涵着话题之间的关联性,并在此基础上提出了扩展的 LDA 话题模型进行微博热点话题检测。为了解决文本稀疏性问题,首先对文本聚类处理得到合成的微博文档,然后再利用扩展的话题模型抽取话题,最后利用互信息来自动计算热点话题词汇。为了验证所提出的话题模型的有效性,实现了基于特定领域的热点话题自动挖掘原型系统。实验结果表明本文提出的算法可以较准确地自动提取微博中的热点话题, 同时自动计算出的热点话题词汇与人工选取的热点话题词汇的语义相似度超过 75% 。

参考文献:

[1] ALLAN J, PAPKA R, LAVRENKO V. On-line new event detection and tracking [C]// SIGIR '98: Proceedings of the 21th ACM SIGIR International Conference on Research and Development in Information Retrieval. New York: ACM, 1998: 37 - 45.

[2] 路荣,项亮,刘明荣,等. 基于隐主题分析和文本聚类的微博客新闻话题发现研究[C]// 第六届全国信息检索学术会议论文集. 北京: 中国中文信息学会,2010.

[3] RAMAGE D, DUMAIS S T, LIEBLING D J. Characterizing microblogs with topic models [C]// Proceedings of the Fourth International Conference on Weblogs and Social Media. Menlo Park: AAAI Press, 2010: 130 - 137.

[4] ASUNCION A, SMYTH P, WELLING M. Asynchronous distributed learning of topic models [C]// NIPS 2008: Proceedings of the 22th Annual Conference on Neural Information Processing Systems. Atlanta: Curran Associates Inc, 2008: 81 - 88.

[5] BLEI D M, LAFFERTY J D. A correlated topic model of science [J]. Annals of Applied Statistics, 2007, 1(1): 17 - 35.

[6] SANKARANARAYANAN J, SAMET H, BENJAMIN E T, *et al*. TwitterStand: news in Tweets [C]// Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2009: 42 - 51.

[7] SHARIFI B M, HUTTON A, KALITA J K. Automatic microblog classification and summarization [C]// Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. Stroudsburg: The Association for Computational Linguistics, 2010: 685 - 688.

[8] INOUYE D. Multiple post microblog summarization [R]. Colorado Springs, GA: University of Colorado at Colorado Springs, 2010.

[9] YEUNG C-M A, IWATA T. Capturing implicit user influencein online social sharing [C]// Proceedings of the 21th ACM Conference on Hypertext and Hypermedia. New York: ACM, 2010: 245 - 254.

[10] ANAGNOSTOPOULOS A, KUMAR R, MAHDIAN M. Influence and correlation in social networks [C]// KDD'08: Proceeding of the 14th ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 7 - 15.

[11] CRANDALL D, COSLEY D, HUTTENLOCHER D, *et al*. Feedback effects between similarity and social influence in online communities [C]// KDD'08: Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 160 - 168.

[12] GOYAL A, BONCHI F, LAKSHMANAN L V S. Learning influence probabilities in social networks [C]// WSDM'10: Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 241 - 250.

[13] GUERRA P H C, VELOSO A, MEIRA W, Jr, *et al*. From bias to opinion: A transfer-learning approach to real-time sentiment analysis [C]// KDD'11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011: 150 - 158.

[14] SILVA I S, GOMIDE J, VELOSO A, *et al*. Effective sentiment stream analysis with self-augmenting training and demand-driven projection [C]// SIGIR'11: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2011: 475 - 484.

[15] WANG XIAOLONG, WEI FURU, LIU XIAOHUA, *et al*. Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach [C]// CIKM '11: Proceedings of the 20th ACM Conference on Information and Knowledge Management. New York: ACM, 2011: 1031 - 1040.