基于隐主题分析的中文微博话题发现*

史剑虹，陈兴蜀†，王文贤
（四川大学 计算机学院 网络与可信计算研究所，成都 610065）

摘 要：针对高维、稀疏的中文微博数据，提出一种多步骤的新闻话题发现方法。首先结合微博的传播特点，选取出不同时间窗口中具有较高新闻价值的微博文本；再利用隐主题模型挖掘微博内容中隐含的主题信息，并在此基础上进行文本聚类；最后使用频繁项集挖掘技术获取话题关键词集合。该算法能够较好地实现对中文微博数据的降维与话题发现。真实的微博数据集实验结果验证了该方法的有效性。

关键词：中文微博；话题发现；隐主题模型；文本聚类；频繁项集挖掘
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2014)03-0700-05
doi:10.3969/j.issn.1001-3695.2014.03.014

Discovering topic from Chinese microblog based on hidden topics analysis

SHI Jian-hong, CHEN Xing-shu†, WANG Wen-xian
(*Network & Trusted Computing Institute, School of Computer Science, Sichuan University, Chengdu* 610065, *China*)

Abstract: For high dimensional and sparse Chinese microblog data, this paper proposed a multi-step method for discovering topic. Firstly, it combined with the spread characteristics of microblog, it got the microblog content which had a high news value. Then, it used the hidden topics analysis technique to model the text data and got the result of the text clustering by using the hidden topic information. Finally, the keywords which were best represented the topic content would be obtained from the clustered results through frequent itemsets mining. The experimental results verify the validity of the method on Chinese microblog dataset's dimensionality reduction and topic identification.

Key words: Chinese microblog; topic discovering; hidden topic analysis; text clustering; frequent itemsets mining

0 引言

随着互联网技术的迅猛发展，微博(microblog)近年来获得了爆炸式的发展，吸引着越来越多的网民参与。微博是一种基于Web 2.0技术实现的社会媒体(social media)，其允许用户通过Web、WAP以及各种客户端设备及时更新简短文本并公开发布，是一种基于用户关系的信息分享、传播及获取平台。相较于传统的网络文本数据，微博具有如下特点：a)文本长度短，内容通常被限制在140个字符以内，数据稀疏性突出；b)内容语法严谨度低，书写随意，且常伴有新的网络用语；c)信息更新速度快，数据规模大、维度高。微博改变了人们获取信息的方式，其不仅仅是一种单纯的社交工具，同时也是社会舆论传播的重要媒介。因此，如何从多样化的微博数据中快速准确地检测出新闻话题，了解大众关心的问题，对舆情监控、信息安全等都有十分重要的意义。

针对传统的网络长文本信息的新闻话题识别研究已经相对成熟[1]。然而对于微博等短文本数据，由于其文本长度短、内容稀疏，传统的话题识别方法已不能很好地适用。目前，有关短文本数据的话题识别还处于发展阶段，研究者们作了大量的尝试。文献[2,3]分别通过搜索引擎和HowNet来扩展短文本的上下文信息，将短文本扩充为较长的文本，减弱特征词词频过低对聚类结果的影响。上述方法操作简单，但却非常耗时，不适用于大规模数据的操作。文献[4]则利用上下文相关度模型获得文本特征词，进而对以特征词构成的向量空间模型进行增量式聚类获得新闻话题。该方法简化了数学处理，但却忽视了词元的语义特征，实际话题的识别准确率较低。文献[5]利用微博中词的共现度构成主题词共现图，实现对新闻话题的识别，但其在处理大规模数据时空间复杂度太高。文献[6]通过隐主题模型发现词元的语义特征，使用混合聚类的方法实现了对Twitter新闻话题的发现。文献[7]则在LDA(latent Dirichlet allocation)模型的基础上结合Twitter的特征，通过利用用户自定义的背景信息，提出了Twitter-LDA模型，该模型有效地利用了已知信息来提高新闻话题的识别率。但是上述两种模型均是针对Twitter来进行构建的，对中文微博的适用度还不确定。文献[8]针对中文微博提出了MB-LDA模型，该模型有效地将微博的文本内容与用户关系引入主题建模过程中，实验验证了其在小数据样本空间中对中文微博主题挖掘的有效性。

基于以上研究，针对处理大量微博数据时算法的复杂度要求以及中文微博的传播特性，本文提出一种基于隐主题分析的中文微博话题发现方法。实验证明，本方法能够较好地克服微博的数据稀疏问题，发现微博的隐主题信息，在较大规模微博数据中实现对新闻话题的发现。

收稿日期：2013-06-18；**修回日期**：2013-08-09 **基金项目**：国家科技支撑计划课题资助项目(2012BAH18B05)；四川大学青年教师科研启动基金资助项目(2013SCU11017)

作者简介：史剑虹(1990-)，男，陕西西安人，硕士研究生，主要研究方向为信息安全、数据挖掘；陈兴蜀(1968-)，女(通信作者)，教授，博导，博士，主要研究方向为信息安全、计算网络(chenxsh@scu.edu.cn)；王文贤(1978-)，男，讲师，博士，主要研究方向为信息安全、P2P网络.

1　基于隐主题分析的中文微博话题发现

为了解决微博数据规模大、内容短的问题，本文利用微博热度的概念，先从大规模的微博数据中选取出具有较高新闻价值的微博；然后通过 LDA 隐主题建模，将高维的微博数据映射到低维空间，挖掘每条微博隐含的主题信息；再通过 FP-growth 算法对 K-means ++ 算法聚类处理后的微博数据进行频繁项集挖掘，发现热点新闻话题。方法框架如图 1 所示。

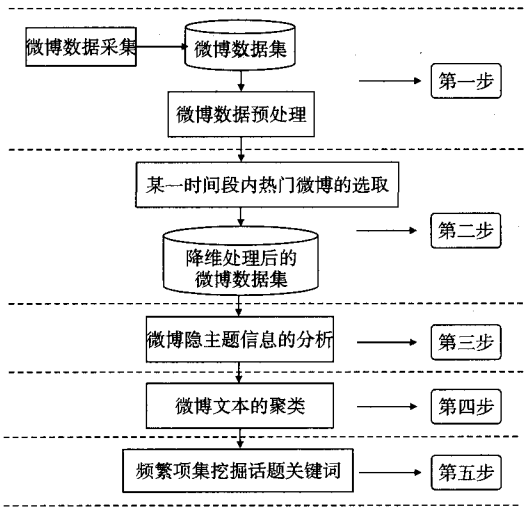

图1　微博话题发现的整体框架

1.1　热门微博的选取

通过分析新闻话题的形成过程及微博的传播特性，发现微博话题的形成有两个重要的特点：a）时域性较强，通常其所讨论的内容较为新颖，且与其相关的内容在之前很少出现，但却在某个时间窗口期内突然大量涌现；b）某一时间段内关注度较高的微博通常会参与到热门话题的讨论中，而关注度的大小取决于微博的转发数和评论数，当某一话题引起讨论时，与其相关的微博的转发数和评论数会显著增加。本文利用这些特点，从大规模的微博数据中筛选出最有可能参与新闻话题讨论的微博，将其定义为热门微博。

1.1.1　微博热度的定义

针对微博的热度定义了两个不同的热度概念，即显式热度（explicit heat）与隐式热度（implicit heat）。

定义 1　显式热度。微博 s 的显式热度 $\mathrm{Exp}H_s$ 为该微博的转发数 C_rel_s 与评论数 C_com_s 的加权和，即

$$\mathrm{Exp}H_s = \chi C\_\mathrm{rel}_s + \delta C\_\mathrm{com}_s \tag{1}$$

式中：χ 与 δ 为调节参数，且 $\chi + \delta = 1$。

定义 2　隐式热度。微博 j 在当前时间窗口中的隐式热度 $\mathrm{Imp}H_j$ 是微博预处理后留下的所有词元的热度值的加权和，即

$$\mathrm{Imp}H_j = \sum_{i=1}^{N} WH_{ij} \tag{2}$$

式中：N 为微博 j 预处理后保留词元的总数，WH_{ij} 为微博 j 的第 i 个词的词元热度。

1.1.2　词元热度的评价

分析微博的相关特征，发现词元与新闻话题之间存在如下关系时，其与热门新闻话题的相关度较高：a）某一词元在一个时间窗口内大量出现，且其出现的次数明显高于其他词元；b）某一词元在某个窗口内大量出现，且出现的频率较前一个时间窗口明显增多；c）某一词元在某个时间窗口内同时被大量用户频繁使用，且其使用频率较前一个时间窗口明显增多。

针对上述关系，分别利用词元的相对词频、词频增加率以及用户增加率来对其进行量化，进行数据归一化处理后的计算公式分别为：

a）相对词频

$$F_{ij} = \frac{f_{ij}}{f_{\max}} \tag{3}$$

式中：F_{ij} 是词元 i 在时间窗口 j 中的相对词频；f_{ij} 是词元 i 在时间窗口 j 中的频率；$f_{\max}$ 是时间窗口 j 中的最高词频。

b）词频增加率

$$FI_{ij} = \frac{f_{ij} - f_{i(j-1)}}{1 + f_{i(j-1)}} \tag{4}$$

式中：FI_{ij} 表示词元 i 在时间窗口 j 中的增加率，$f_{i(j-1)}$ 是词元 i 在时间窗口 $j-1$（即前一个时间窗口）中的频率。

c）用户增加率

$$UI_{ij} = \frac{uf_{ij} - uf_{i(j-1)}}{1 + uf_{i(j-1)}} \tag{5}$$

式中：UI_{ij} 表示词元 i 在时间窗口 j 中被使用到的用户数的增加率，uf_{ij}、$uf_{i(j-1)}$ 分别表示词元 i 在时间窗口 j、$j-1$（即前一个时间窗口）中被使用到的用户个数。

综合考虑词元热度与用户、文档等多方面的联系，对词元的热度表现力 WH_{ij} 定义如下：

$$WH_{ij} = (\gamma \ln F_{ij} + \eta \ln FI_{ij} + \lambda \ln UI_{ij}) \times \mathrm{tfidf}_{ij} \tag{6}$$

式中：WH_{ij} 越大表示该词元是热点词的概率越大；γ、η、λ 均为调节参数，用来平衡相对词频、词频增加率以及用户增加率三者之间的比重关系，且 $\gamma + \eta + \lambda = 1$；$\mathrm{tfidf}_{ij}$ 为微博 j 中的第 i 个词的 TF-IDF 值，其计算公式为

$$\mathrm{tfidf}_{ij} = \frac{tf_{ij} \times \log_2\left(\frac{N}{n_i} + 0.01\right)}{\sqrt{\sum_{j=1}^{t}\left(tf_{ij} \times \log_2\left(\frac{N}{n_i} + 0.01\right)\right)^2}} \tag{7}$$

式中：tf_{ij} 为微博 j 中的第 i 个词的出现频率；N 为微博数据集的总数；n_i 为含有词元 i 的微博个数。

1.1.3　热门微博的选取

因为参与热点话题讨论的微博通常具有较高的转发数和评论数，所以先利用微博的显式热度进行一次筛选，只保留大于一定阈值的微博文本；接着对每条微博的热度进行评分，每条微博的得分为该条微博的隐式热度值；最后对评分后的微博数据按照得分高低降序排列，选取得分较高的前若干条微博作为热门微博，用于后续的处理。

1.2　热门微博的隐主题信息分析

针对微博内容中的隐主题信息，采用 LDA 模型来对微博数据进行隐主题建模。

1.2.1　LDA 模型

LDA 模型是由 Blei 等人[9]在 2003 年提出的，它是一种由文档、主题和词三个层次构成的分层贝叶斯模型，采用产生式全概率模型来对文本进行建模。它将每个文档表示为一系列潜在主题的混合分布，且每个文档均有一个特定的主题分布；主题之间相互独立，并被文本集中的所有文档所共享。文档到主题服从狄利克雷（Dirichlet）分布，主题到词服从多项式（multinomial）分布。LDA 模型的生成过程如图 2 所示。

图 2 中单边圆表示隐含变量，双边圆表示可观测值，矩形表示可重复过程。其中：α、β 均为文档层的参数，α 反映了文档集中隐含主题间的相对强弱，β 表示文档中各隐含主题自身在词语上的比重；θ 表示每篇文本在主题上的概率；φ 代表主

题与单词之间的关系；z 表示文档分配在每个词上的隐含主题比重；w 是文档的词向量表示；M 为文档集中的文档数目；K 表示 Dirichlet 分布中的维度，也就是主题个数；N_d 为该文档中所有词的总数。生成概率模型的计算如式(8)所示。

$$P(\theta,z|w,\alpha.\beta)=p(\theta|\alpha)\prod_{n=1}^{K}p(z_n|\theta)p(w_n|z_n,\beta) \tag{8}$$

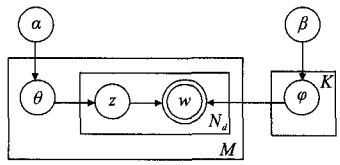

图2 LDA生成过程的概率模型

在任意给定的语料集所对应的 LDA 模型中，生成一篇文档的过程如下：

首先从 Dirichlet 分布 Dir(β) 中抽取出每个主题对应的参数 φ_k，然后按如下方式产生语料集中的第 j 篇文档 x_j：

a) 选择文档：$N \sim$ Poission(ξ)；

b) 选择文档参数：$\theta \sim$ Dir(α)；

c) 选取文档中的每个词 w_{ij}：

(a) 选择一个主题：$z_{ij} \sim$ multinomial(θ)

(b) 从主题 z_{ij} 词中选取出该词：$x_{ij} \sim$ multinomial(θ_{zij})

对于 LDA 模型中的参数很难得到精确的推理，常采用近似的推理方法，而由 Griffith 等人[10]提出的 Gibbs 抽样是 MCMC(Markov chain Monte Carlo)的一种简单的实现方式，比较适合大规模数据的处理，是目前最流行的 LDA 模型抽取算法。本文也采用 Gibbs 抽样来实现。

1.2.2 基于 Gibbs 抽样的 LDA 模型求解

在使用 Gibbs 抽样来获取 LDA 模型中词汇的分布概率时，没有直接评估 φ 和 θ，而是通过先计算词分配给主题的后验概率 $P(z|w)$，再确定其他词的主题分配 z_{-i}，然后估计当前词 w_i 被分配给其他主题的概率 $P(z_i=j)$，不断进行循环迭代，直到状态收敛，从而间接地估计出参数 φ 和 θ。主题的词汇分配概率 $P(z_i=j|z_{-i},w_i)$ 的计算如式(9)所示。

$$P(z_i=j|z_{-i},w_i)\propto\frac{n_{-i,j}^{(w_i)}+\beta}{n_{-i,j}^{(\cdot)}+W\beta}\cdot\frac{n_{-i}^{(d_i)}+\alpha}{n_{-i}^{(d_i)}+T\alpha} \tag{9}$$

其中：$n_{-i,j}^{(w_i)}$ 表示词 w_i 已经被分配在主题 j 中的次数；$n_{-i,j}^{(\cdot)}$ 表示被分配到主题 j 中的所有词的总数；$n_{-i,j}^{(d_i)}$ 表示文档 d_i 中词 w_i 已经被分配到主题 j 中的次数；$n_{-i}^{(d_i)}$ 表示文档 d_i 中被分配到主题 j 中的所有词的总数；W 和 T 为权重，均非负。式(9)中不包含本次迭代时 $z_i=j$ 的分配情况。每轮迭代后，模型中参数 φ 和 θ 的计算方法如式(10)(11)所示。

$$\hat{\varphi}_j^{(w)}=\frac{n_j^{(w)}+\beta}{n_j^{(\cdot)}+W\beta} \tag{10}$$

$$\hat{\theta}^d=\frac{n_j^{(d)}+\alpha}{n^{(d)}+T\alpha} \tag{11}$$

1.2.3 LDA 建模结果分析

利用 LDA 模型对文档集合进行建模后可以获得：

a) 文本—主题矩阵 $M\times K$，其中 M 为文档集合中的文本总数，K 为定义的隐主题个数。它表示每个文本在 K 维隐主题空间上的分量值，即每个文本生成不同隐主题的概率。

b) 主题—词矩阵 $K\times V$，其中 K 代表隐主题的个数，V 代表文档集合中所有不同词的个数。它表示每个词属于不同隐主题的概率。

c) 标记所有文本中每个词元所属隐主题的向量集合 Z。当集合中的同一词元出现在向量集的不同位置时，其可能代表该词被分配给了不同隐主题。由于 LDA 模型中各个隐主题之间是相互独立的，假如同一词元被分配给不同的隐主题，则可认为其具有不同词义，即为多义词。例如表 1 中展示的就是“苹果”一词被分配到不同的隐主题时所表达的不同词义，在隐主题“Topic 19”中“苹果”一词代表的是美国的苹果公司，而隐主题“Topic 37”中的“苹果”则表示的是水果中的苹果。

表1 两个包含“苹果”一词的不同隐主题

隐主题编号	隐主题生成概率最大的 10 个词
Topic 19	三星 销量 GALAXY 公司 iPhone5 华为 苹果 性价比 乔布斯 神坛
Topic 37	西柚 养生 绿茶 苹果 利尿 养生 果农 价格 腹泻 大枣

1.3 热门微博的聚类

研究中常采用具有简单、收敛速度快等优点的 K-means 算法来进行大规模文本数据的聚类处理。然而在传统的 K-means 聚类算法中，聚类中心的选择严重影响到聚类结果的优劣。但是对于聚类中心的选择方法大都倾向于随机选取，这就会使多次聚类的结果表现出较大差异，不能很好地反映数据的实际聚类情况。因而本文中采用优化聚类中心选择的 K-means ++ 算法[11]来进行微博数据的聚类。

1.3.1 隐主题信息的 K-means ++ 聚类

由于只关注于具有相同主题信息的微博聚类情况，因而选用 LDA 建模后得到的文本—主题矩阵 $M\times K$ 作为聚类数据点集合，其反映了每条微博在 K 维空间中的主题权重值。具体的 K-means ++ 算法处理过程如下：

设 $D(x)$ 为文档 x 到已有的聚类中心的最近距离；X 为数据点集合。

a) 从 X 中随机选取一个点作为初始聚类中心 c_1。

b) 计算其他点与 c_1 的距离 $D(x)$，并将其保存后计算这些距离之和 sum($D(x)$)

c) 再选取一个存在于 sum($D(x)$) 中的随机值 random，计算 random $-=D(x)$，直到 random$\leqslant$0，将此点记做下一个聚类中心。

d) 重复步骤 b)c)，直到所有的聚类中心点被选取出来。

e) 执行标准的 K-means 聚类算法对数据进行聚类。

对于最优聚类中心数的确定，可以在已知待处理数据集的实际聚类结果的情况下进行聚类实验，选取出具有最优聚类质量的聚类中心数。

1.3.2 文本相似度的度量

对于文本相似度的度量方法，采用 Jensen-Shannon 散度来进行计算。Jensen-Shannon 散度利用两个文本中词语的概率分布来测量文本之间的距离。文献[12]指出，Jensen-Shannon 散度在测量文本之间的距离时，它的测量精度通常要更优于余弦距离等其他距离的测量精度。对于随机的两个文本 P、Q，其 Jensen-Shannon 散度如式(12)所示。

$$D_{js}(P\|Q)=\sum_{w\in W}P(w)\log\frac{P(w)}{M(w)}+\sum_{w\in W}Q(w)\log\frac{Q(w)}{M(w)} \tag{12}$$

其中：W 是文本集合中所有词元的集合；$P(w)$、$Q(w)$ 是关于 $w\in W$ 的两个概率分布，并且

$$M(w)=\frac{P(w)+Q(w)}{2} \tag{13}$$

而 $D_{js}(P\|Q)$ 值越小，说明文本 P、Q 两者之间越相似。

1.4 聚类结果的频繁项集挖掘

实际上微博数据聚类后一个聚类簇中的微博文本不可能全都属于同一个话题,且从繁杂独立地文本中很难明确地发现新闻话题的核心内容,因而本文对聚类结果进行频繁项集挖掘,这样既可以削弱聚类簇中的噪声项对话题检测的影响,同时能够利用词元间的关联关系清楚地概括出话题主旨。本文采用基于关联规则的 FP-growth 频繁项集算法[13]来获取话题关键词。

FP-growth 算法不产生候选模式而采用频繁模式增长的方法直接产生全部频繁项集。针对聚类后的微博数据,进行 FP-growth 频繁项集挖掘的流程如下:

输入:不同聚类簇中的微博特征词集 $D=\{W_1,W_2,\cdots,W_n\}(i=0,1,\cdots,n)$,最小支持阈值 min_sup。

输出:频繁项集的完全集。

a)构造项头表。扫描某一微博特征词集 W_i,得到频繁项的集合 F 及每个频繁项的支持度。按支持度降序排列 F,记为 L。

b)构造原始 FP-tree。把微博特征词集 W_i 中的频繁项按 L 的顺序重排后,将每个频繁项插入以 null 为根的 FP-tree。如果插入时频繁节点存在,则支持度加 1;否则创建支持度为 1 的节点,并将其链接到项表头中。

c)调用 FP-growth(Tree,null)进行频繁项集挖掘。

```
Procedure FP_growth(Tree,a)
if Tree 含单个路径 P then{
  for 路径 P 中节点的每个组合 b
  产生模式 b∪a,其支持度 support = b 中节点的最小支持度;}else
  {
  for each a_i 在 Tree 的头部(按照支持度由低到高进行扫描){
  产生一个模式 b = a_i∪a,support = a_i · support;
  构造 b 的条件模式基,然后构造 b 的 FP-tree Treeb;
  if Treeb 不为空 then
  调用 FP_growth(Treeb,b);
  }
}
```

在频繁项集挖掘后,对不同聚类簇中的频繁项按照 1.1.2 节定义的词元热度表现力进行排序,只选择热度值较大的前几个词元作为话题的关键词集。

2 实验与结果分析

2.1 数据集获取

由于目前还没有通用的中文微博数据测试集,实验中以新浪微博的 API 接口与 Web 网络爬虫相结合的方法来获取微博数据,爬取时间为 2013 年 3 月 26 日 ~2013 年 4 月 8 日。实验数据由两个不同数据集组成:a)利用新浪微博的搜索功能,以其官方给定的每日前 30 个热门话题的关键词作为搜索词来获得相应的微博数据,其中搜索获得的微博文本中均包含搜索词(认定搜索获得的微博均是与进行搜索的话题相关联的,标定其属于该话题),而每个话题能获得约 800 条相关微博,用它们构成测试集 A;b)对新浪微博进行全网爬取,从爬取的微博数据中每日随机选取约 350 000 条,构成测试集 B。

2.2 实验结果及分析

2.2.1 LDA 模型先验参数的确定

为了评估 LDA 模型中的先验参数 α 对话题识别结果的影响,从测试集 A 中取得 2013-04-01、03、05、08 四天的微博数据,然后设定 $\beta=0.1$,聚类个数 $K=30$,频繁项集挖掘的支持度计数选为 10,比较不同的 α 取值对话题识别结果的影响。实验结果如图 3 所示。

从图 3 可以看出,话题识别率与 LDA 模型的先验概率 α 的取值基本呈正态分布。当 α 取 0.2 ~0.3 之间的值时,话题识别率最高;而当 α 值小于或大于该范围时,则话题识别率依次递减。

2.2.2 算法有效性验证

因为搜索获得的微博文本是话题相关的,所以可以将测试集 A 与 B 中的数据进行一定比例的交叉混淆(避免测试集 A 与 B 中同一天的数据进行混淆,采用的混淆比例为 1:15),然后通过计算每日微博数据中 30 个已知热点话题的识别率来验证算法的有效性。

图 4 显示的是当 LDA 模型参数 $\alpha=0.3$,$\beta=0.1$,频繁项集挖掘的支持度计数选为 10,而聚类个数 K 分别取 30、50、100、150 时,该算法对测试集 A 中每天已知的 30 个话题的识别率。

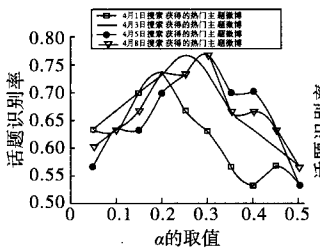

图3 先验参数α对话题识别率的影响

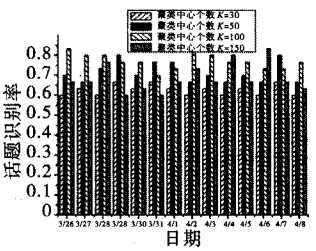

图4 不同K值时的话题识别率

当 $K=100$ 时,话题的识别率最高,平均在 75% 左右,这表明该方法在随机获取的微博数据中同样具有较好的话题识别率,证明了方法的有效性。

2.2.3 话题识别效果

本实验使用数据集 B 中 2013 年 4 月 8 日的 344 758 条微博数据作为测试数据来获取当天的热点话题,设定 LDA 模型先验概率 $\alpha=0.3$,$\beta=0.1$,聚类个数 $K=100$,频繁项集挖掘的支持度计数选为 10,当天最大的热点话题是英国前首相撒切尔夫人逝世。实验结果如表 2 所示。

表 2 2013 年 4 月 8 日热点话题展示

编号	热点话题表示
1	撒切尔夫人、逝世、英国、前首相、铁娘子
2	戴旭、禽流感、引咎辞职、反人类、死者
3	叔侄 冤案 聂海芬 十年 张高平
4	炎黄春秋、乌有之乡、被封、一周年、背景
5	华为、手机、Mate、京东、白色版、上市

表 2 中显示的是频繁项集挖掘后聚类簇中词元热度值排名最高的前五个词元,可以看到该方法准确地识别出了“英国前首相撒切尔逝世”这一当天最大的热门新闻话题;同时该方法还发现了如戴旭针对禽流感的不当言论、浙江叔侄冤案、《炎黄春秋》杂志被封一周年以及华为新款白色版 Mate 手机上市等当天的其他热门新闻话题。

表 3 中列出的是当天聚类得到的最大类中热度得分最高的五条微博的相关信息。

3 结束语

微博热点话题的发现研究有着极大的现实意义和应用需

求,本文通过分析中文微博的特点,结合其他短文本话题发现的研究,提出了一种基于隐主题分析的中文微博话题发现方法。方法中对热门微博的选取,不但极大地缩小了数据处理规模,而且一定程度上减弱了非热门微博对话题识别结果的影响;同时LDA隐主题建模则增强了对文本主题信息的利用,提升了文本相似度度量的精确性。而后通过K-means++聚类算法快速地将微博数据聚集到不同的新闻话题之中,最后的频繁项集挖掘既是对聚类簇中噪声项含量过多造成话题湮没这一缺陷的弥补,同时又能够更加直观地展示出新闻话题的主旨。但是本文也仍有如下方面需要改进:a)每日的新闻话题个数不定,使得K-means++算法的聚类准确度波动较大,进而导致话题识别率不稳定;b)文中对不同用户在话题中扮演的角色尚未考虑,不同用户对于话题形成、传播、演化所起的作用尚待研究;c)文中对于谣言话题尚未研究,需要在日后深入分析微博谣言的相关特性。

表3 2013年4月8日聚类结果最大的一个类中微博得分最高的五条微博内容

用户ID	用户名	微博文本内容
2286908003	人民网	【英国前首相撒切尔夫人去世】据法新社报道,英国前首相撒切尔夫人去世,享年87岁
1644489953	南方都市报	【"铁娘子"撒切尔夫人去世】法新社报道,英国前首相撒切尔夫人去世,享年87岁。玛格丽特—撒切尔,英国保守党第一位女领袖,也是英国历史上第一位女首相,创造了蝉联三届,任期长达11年之久记录的女首相
1977610261	长江证券	【快讯】法新社报道,英国前首相撒切尔夫人去世,享年87岁。专题:http://t.cn/zWxQUOm
1249729383	新浪女性	【追忆撒切尔夫人:从杂货店里走出的"铁娘子"】英国"铁娘子"、前首相撒切尔夫人8日因中风去世,享年87岁。虽然底层出身,但撒切尔夫人依靠自己的顽强奋斗脱颖而出成为英国历史上第一位女首相,她在长达11年的首相任职期间政绩卓著。与丈夫相伴51年,丧偶后健康状况开始下降
2216876105	东来岭人	视频:1982年撒切尔夫人大会堂摔跤片段 http://t.cn/zTb39y4

参考文献:

[1] ALLAN J, CARBONELL J, DODDINGTON G. Topic detection and tracking pilot study: final report [C]//Proc of DARPA Broadcast News Transcription and Understanding Workshop. San Francisco: Morgan Kaufmann Publisher Inc, 1998:194-218.

[2] DANUSHKA B, YUTAKA M, MITSURU I. Measuring semantic similarity between words using Web search engines[C]//Proc of the 16th International Conference on World Wide Web. New York: ACM Press, 2007:757-766.

[3] LIU Zi-tao, YU Wen-chao, CHEN Wei, *et al*. Short text feature selection for microblog mining [C] //Proc of the 4th International Conference on Computational Intelligence and Software Engineering. 2010:1-4.

[4] 郑斐然,苗夺谦,张志飞,等. 一种中文微博新闻话题检测的方法[J]. 计算机科学,2012,39(1):138-141.

[5] 赵文清,侯小可. 基于词共现图的中文微博新闻话题识别[J]. 智能系统学报,2012,7(5):444-449.

[6] 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[J]. 模式识别与人工智能,2012,25(3):382-387.

[7] ZHAO W X, JIANG Jing, WENG Jian-shu, *et al*. Comparing Twitter and traditional media using topic models [C] // Proc of the 33rd European Conference on Information Retrieval. Berlin: Springer-Verlag, 2011:338-349.

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

[9] BLEI D, NG A, JORDAN M. Latent Dirichlet allocation[J]. Journal of Machine Learning Research,2003,3(3/1):993-1022.

[10] GRIFFITH T L, STEYVERS M. Finding scientific topics[J]. PNAS, 2004,101(1):5228-5235.

[11] ARTHUR D, VASSILVITSKII S. K-means++: the advantages of careful seeding[C]//Proc of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 2007:1027-1035.

[12] CAMEL D, YOM-TOV E, DARLOW A, *et al*. What makes a query difficult[C] //Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2006:390-397.

[13] HAN Jia-wei, PEI Jian, YIN Yi-wen. Mining frequent patterns without candidate generation[C]//Proc of ACM SIGMOD International Conference on Management of Data. 2000:1-12.

(上接第670页)

参考文献:

[1] 鲁延京,陈英武,杨志伟. 求解约束优化问题的粒子进化变异遗传算法[J]. 控制与决策,2012,27(10):1441-1446.

[2] 黄民水,吴功,朱宏平. 噪声影响下基于改进损伤识别因子和遗传算法的结构损伤识别[J]. 振动与冲击,2012,31(21):168-174.

[3] 燕乐纬,陈洋洋,周云. 一种改进的微种群遗传算法[J]. 中山大学学报:自然科学版,2012,51(1):50-54.

[4] 李敏强,寇纪淞,林丹,等. 遗传算法的基本理论与应用[M]. 北京:科学出版社,2003:400.

[5] 胡新平,贺王芝,倪巍伟,等. 基于赌轮选择遗传算法的数据隐藏发布方法[J]. 计算机研究与发展,2012,49(11):2432-2439.

[6] 梁宇宏,张欣. 对遗传算法的轮盘赌选择方式的改进[J]. 信息技术,2009,33(12):127-129.

[7] 曹道友,程家兴. 基于改进的选择算子和交叉算子的遗传算法[J]. 计算机技术与发展,2010,20(2):44-47,51.

[8] GAO Shang. Allocation of seats mathematical programming model [J]. Journal of Computational Information Systems, 2011, 7(2):554-561.

[9] 姜启源,谢金星,叶俊. 数学建模[M]. 4版. 北京:高等教育出版社,2011:278-285.

[10] 钱丽丽,邓桂丰. 一个公平分配席位的新方案[J]. 数学的实践与认识,2012,42(18).

[11] ZYCZKOWSKI K, CICHOCKI M A. Distribution of power and voting procedures in the European union[C]. [S. l.]: Surrey, UK Ashgate Publishing, 2010.

[12] Van ECK L, VISAGIE S E, De KOCK H C. Fairness of seat allocation methods in proportional representation[J]. ORiON, 2007, 21(2): 93-110.

[13] 孙文瑜,徐成贤,朱德通. 最优化方法[M]. 2版. 北京:高等教育出版社,2010.