一种时间情境依赖的微博话题抽取方法

庄婷婷　王　平　程齐凯
(武汉大学信息管理学院,武汉,430072)

[摘　要] 微博话题处在动态变化中,为了准确的抽取所需的微博话题,在微博文本内容之外,需要考虑微博的情境依赖关系,而其中尤为重要的是时间情境。本文提出了一种时间情境依赖的时序微博话题检测方法。该方法包括四个步骤:文本预处理、微博文本的特征选取、基于主题模型的文本话题检测、微博话题演化状态的确定。在真实数据上的实验表明,本文提出的方法具有较好的效果。

[关键词] 话题检测　微博　特征选取　主题模型

[中图分类号] G250　[文献标识码] A　[文章编号] 2095-2171(2013)03-0040-07

Temporal Related Topic Detection Approach on Microblog

Zhuang Tingting　Wang Ping　Cheng Qikai
(School of Information Management, Wuhan University, Wuhan, 430072)

[**Abstract**] Topic in microblog data is dynamic. Aiming at detecting microblog's topic more precisely, we should make use of the context of given microblog other than its text content. This paper proposes a temporal related topic detection method. This method can be divided into four phases: preprocessing, feature selection, topic detection, topic evolution analysis. Experimental results on real data show that the methods we propose in this paper perform well.

[**Key words**] Topic detection　Microblog　Feature Selection　Topic Model

1　引言

微博(Microblog)是近年来发展迅速的一种媒体形式,国外的 Twitter、国内的新浪微博、腾讯微博都有着巨大的用户量和大规模的微博数据。微博通过不超过 140 个字的短文本发布信息。由于微博信息的实时性、多样性和移动化,微博已经成为人们获取信息、讨论事件的重要平台。微博的实时性和多样性一方面为信息的发布和获取提供了便利,同时,也带来了信息碎片化、上下文信息缺失等诸多问题。为了解决这些问题,一种可行的技术手段是对微博信息进行话题检测,从话题的角度组织微博信息。

微博话题检测,或者更具体地,微博热门话题检测,是话题检测(Topic Detection, TD)技术[1]在微博数据上的应用。目前,国内对微博话题检测的研究还处在起步阶段。研究者的一般思路是将微博信息看作文本,然后利用

[基金项目] 中国博士后科学基金第六批特别资助项目"网上多源信息的可信度判断与评估模型研究"(2013T60749)的研究成果之一。

[作者简介] 庄婷婷,女,硕士研究生,研究方向为信息检索;王平,男,讲师,研究方向为社会网络;程齐凯,男,博士研究生,研究方向为信息检索与数据挖掘。

文本聚类、话题模型等方法检测微博话题。在这一思路的指导下，现有的研究多是从静态的角度考察微博话题，对一定时间的微博数据进行话题检测。

微博数据具有多个维度的属性，其中之一是时间维度。微博话题具有延续性，现有的话题是过去话题的演化结果，基于这一考虑，话题检测需要将时间维度纳入考虑。需要考虑一种更新式的话题检测方法，即将先前的微博话题数据作为先验知识，在此基础上考察当前的话题。另外，话题的演变形式是多样的，一个话题可能是长期性的活跃话题，也可能是突发话题。人们关注的可能是不同状态下的微博话题，例如：对广告推广者而言，具有延续性的话题可能意义更大；对新闻媒体而言，突发性的话题更具有新闻价值。因此，在微博话题检测时，有必要在考虑时间维度的前提下，对微博话题的演化形式加以区分。

本文提出了一种时序中文微博话题检测方法，该方法的基本思想是将微博话题的出现和消亡看成一个时序演化的过程，首先根据一定的时间窗口将微博文本划分为多个部分，利用时序相关信息确定微博文本的描述特征；基于选定的特征，对各时间窗口的微博文本利用主题模型进行话题检测；进而对前后时间窗口内的微博话题进行演化状态检测。

2 相关研究

话题检测是话题检测与跟踪(Topic Detection and Tracking, TDT)任务的重要组成部分[1]。话题检测是文本挖掘的传统课题，旨在从文本中挖掘出隐含的话题及其结构，以帮助人们了解文本的主题构成。传统的话题监测包括在线话题检测(OTD)、新事件发现(NED)、事件回顾检测(RTD)以及话题层次检测(HTD)等形式，但无论哪一种形式，这些检测任务都主要面向新闻文本以及学术文本进行。话题检测常用的方法包括主题聚类方法、基于共词网络的方法、基于主题模型的方法等。

微博话题检测是近年随着微博的兴起而出现的问题。微博话题检测需要应用一些与传统话题检测技术不同的策略。特征提取方面，由于微博话题差异性大，特征词汇分散性高，传统的TF－IDF策略并不能很好的表示词汇的重要性，为了解决这个问题，郑斐然[2]等引入了词汇增长系数指标，用于选取合适的主题词特征；微博文本的情感词特征较一般新闻文本要多，情感词对文本话题具有一定的指示作用，杨亮等[3]利用微博文本的情感特征，特别是表情符号辅助检测微博话题。微博话题发现方面，主要的方法有聚类和主题模型两种：文献K-Means等经典的聚类方法常被用于微博文本聚类，如文献[4]使用了层次聚类和K-Means聚类结合的聚类方式寻找可能的文档主题；文献[5]使用基于图论的主题识别方案，较为新颖，也可以被视为一种聚类方法；将主题模型应用于微博话题检测也已经得到了很多的研究。研究者引入或者提出了一系列的微博话题检测模型，如Author Topic Model[5]、Tweete-LDA Model[6]等等，这些模型具有很好的理论背景，但在效果上较其它方法并没有根本性提升；主题模型在实际工作中还常常被用于特征降维工作，文献[4]在特征提取中即使用了这种策略。

除了特征选取方面的工作外，研究者也比较重视处理效率的问题，Petrovic等[7]设计了一种基于流数据处理的微博挖掘方案，Mathioudakis[8]给出了一个系统TweeteMonitor以挖掘微博数据流。这些系统或方案应用分布式处理技术、流处理技术加快处理速度，对实际工作有着较大的应用意义。

3 微博话题检测方法

微博数据是一系列不超过一定长度的短文本。尽管微博文本之间存在着转发、评论等关系，但在本文中，为了实现的方便，微博文本被理解为相互独立，也就意味着本文将忽略微博文本的作者信息、文本中的@信息等，而仅考虑微博文本的内容特征和时间特征。不同的微博文本反映了不同的话题。话题检测的任务是从微博文本中找出可能存在的话题。

本文的话题检测处理流程包括文本预处理、主题词特征提取、主题发现、话题状态识别四个步骤：文本预处理将微博文本处理成一定的格式；继而，应用特征提取方法提取有意义的主题词，特征提取的目的是在不过分损失准

确度的前提下进行维数消减，从而降低后续计算的复杂度；方法的第三步是基于提取的特征对文本进行话题探测；基于话题探测的结果，本文给出了一个话题状态的分类方法，基于此确定话题状态。

3.1 预处理

给定微博数据 $D_i=\{D_1,D_2,\cdots,D_k\}$，$D_i$ 是第 i 个时间窗口的微博文本集合，预处理包括两个步骤，一是去除数据中的噪音信息，包括去除广告文本、去除无意义文本等；二是将文本加工成合适的格式，具体包括分词、词性过滤。

微博数据的噪音处理非常重要，微博文本同新闻文本相比内容质量相对较差，包含了大量的广告文本、转发文本和无意义文本。如果不加处理，这些噪音数据将会为主题检测带来较大的干扰。噪音处理的规则是：

(1)去除数据中广告微博，本文使用了朴素贝叶斯分类器以及一个小型的微博广告标注数据集，对微博中的广告文本进行去除。选择朴素贝叶斯分类器的原因在于其相对稳定的效果和较低的计算复杂度。

(2)对每一条微博文本，去除“@”符号之后的所有内容。“@”意味着提到某人或者转发某人信息。在微博数据中，存在着大量的“@用户名”的文本，在不考虑用户互动的情况下，这些文本对主题提取意义相对较小，而用户名的随意性还带来了大量的噪音内容。

经过以上两个步骤，可以在不考虑发布者的情况下去除大多数的噪音信息。接下来需要将微博数据加工成合适的格式。

本文采用 ICTCLAS 分词工具[9]对微博文本进行分词。在已有的中文分词工具中，ICTCLAS 具有最佳的分词效果。另外，ICTCLAS 可以对词汇做词性标注，这也是本文需要的。应用 ICTCLAS，单个微博文本的分词结果是一个带词性标注的词汇序列。不同词性的单词对文本主题的反映能力是不一样的，本文仅仅关注名词和动词，即将分词结果中非名词且非动词的词汇去除。

3.2 主题词提取方法

在话题检测中，主题词提取是非常重要的环节，主题词提取可以在一定程度排除噪音数据的干扰，同时降低计算复杂度。微博话题检测中同样需要应用主题词提取，但同传统的文本特征提取方法又存在着很大的不同，本节介绍微博话题检测中的主题词提取方法。

提取主题词，即确定一定的词汇特征用以代表数据集，本质上是文本维度消减的一种形式。微博话题探测中，由于单条文本平均长度较短，而内容随意性大，传统的基于 TFIDF 的特征词识别方法并不适用，因此，需要采用新的方法来识别主题词。话题意味着人们关注的事件和活动，这种关注可能是长时间的，也可能是短期的、突发的，两种情况分别对应着长期活跃话题和突发话题，两种话题形式是不一样的，其对应的词汇特征也有所不同，在提取主题词时需要照顾到对两种情况的考虑。

郑斐然[2]等引入的主题词增长系数是微博文本主题词提取的一个较好的指标，主题词增长系数用于确定一个时间窗口的主题词汇，并不考虑其它时间窗口的主题词。本文关注的是更新式的话题检测，在描述数据集时，不但要考虑特定时间窗口的主题词特征，还需要以同样的主题词特征描述其它时间窗口的数据。因此，直接使用主题词增长系数并不合适。

给定时序微博数据 $D_t=\{D_u,D_{u+1},\cdots,D_T\}$，对任意连续的时间窗口 t 和 t+1，需要确定一个词汇特征集合描述 D_t 和 D_{t+1}。

(1)对于 D_t 和 D_{t+1} 包含的词汇 i，分别计算其词汇增长速度 G_{ij}，计算方法如下：

$$G_{ij}=\frac{F_{ij}+sp}{mean(F_{ij-})+sp}=\frac{sp+F_{ij}\cdot(j-u+1)}{sp+\sum_u^j F_{iu}} \tag{1}$$

其中，F_{ij} 表示时间窗口 j 中词汇 i 的词频，j−u+1 是回溯窗口的大小，$mean(F_{ij-})$ 表示回溯窗口中词汇 i 的平均频度，sp 是一个平滑系数，通过式(2)得到，其中 length(D_u)表示时间窗口 u 内微博文本 D_u 的词数量，$|V_u|$ 表示 D_u 所包含的词项数。

$$sp=\frac{\sum_u^j length(D_u)}{\sum_u^j |V_u|} \tag{2}$$

(2)为了确定一个词汇 i 在时间窗口内的重要性，使用词汇的相对频度，定义为：

$$RF_{ij}=\frac{log(F_{ij})}{log(max(F_j))} \tag{3}$$

其中，$max(F_j)$ 表示 j 窗口内频度最高的词汇的词频，之所以不使用平均值，是因为文本中词汇频次的分布是一个 power law 分布，也正因如此，需要对公式的上下部做 log 平滑处理。

利用 G_{ij} 和 RF_{ij} 两个指标，为单一窗口内的词汇的重要性进行排序，为了结合利用两个指标，类似于文献[2]的做法，构造一个复合参数：

$$S_{ij} = \alpha log(G_{ij}) + (1-\alpha) log(RF_{ij}) \quad (4)$$

使用 log(·)的目的是对词频做平滑，这是文本分析中的常用策略。参数 α 用于调节两个系数对主题词提取结果的影响，$1 \geqslant \alpha \geqslant 0$。当 α 为 0 的时候，仅有词汇相对词频作用于主题词提取；当 α 为 1 的时候，仅有词汇增长速度作用于主题词提取。

(3)为了给相邻的两个时间窗口的文本提取关键词特征，在计算单独时间窗口的 S_{ij} 得分后，还需要将相同词汇在不同时间窗口的得分进行加权处理，以确定最终得分。给定词汇 i 在时间窗口 t 和时间窗口 t+1 的得分 S_{it} 和 $S_{i(t+1)}$ 计算方法为：

$$S_i = \frac{length(D_t) * S_{it} + length(D_{t+1}) * S_{i(t+1)}}{length(D_t) + length(D_{t+1})} \quad (5)$$

对 D_t 和 D_{t+1} 包含的词汇，使用公式(4)计算其得分，根据一定的数量限制取得分最大的前 n 个词汇作为后续主题检测的特征词汇。

3.3 主题发现

主题发现的方法较多，常用的是聚类的主题模型两种方法。本文使用主题模型探测微博文本中隐含的话题，具体地，使用概率潜语义分析(PLSA)。

PLSA 是 Hofmann 提出的一个文本生成模型[10]。PLSA 可以看作混合主题模型的一个发展，放宽了混合主题模型对单个文本主题归属的限制。PLSA 假设文档 d 和词汇 w 之间存在着一个隐含变量(主题) $z, z \in Z = \{z_1, z_2, \cdots, z_k\}$，Z 是固定大小的主题集合。考虑主题的存在，d 和 w 的联合概率可以表述为：

$$P(d,w) = P(d) * P(w|d) = P(d) \sum_{z \in Z} p(w|z) p(z|d) \quad (6)$$

这一过程可以表述为：

(1)以 P(d)的概率选定文档 d；

(2)以 P(z|d)的概率选定主题 z；

(3)对选定的主题 z，以 P(w|z)的概率生成词汇 w。

进而，文档集的生成过程可以表述为下式：

$$P(D) = \prod_{d}^{D} \prod_{w}^{V} (P(d,w)^{n(d,w)}) \quad (7)$$

其中，V 是文档集的词项集合，n(d,w)词项 w 在文档 d 中的词频。

现在，需要对式子中的参数变量进行估计，由于上式中包含了隐含变量，因此直接应用最大似然估计是无法得到参数的求解的。在数学上，对于这一类问题，可以使用期望最大化(EM)算法求解，EM 算法首先为隐含变量以及其它难以直接计算的参数赋予一个随机值，然后通过比较模型与实际的差别逐步修正参数估计结果，直到收敛。

PLSA 的 EM 求解过程包括两个步骤：

(1)E-Step

$$P(z|w,d) = \frac{P(w,z,d)}{P(w,d)} = \frac{P(w|z)P(z|d)}{\sum_z P(w|z)P(z|d)}$$

(2)M-Step

$$P(d) = \frac{\sum_w \sum_z n(d,w) P(z|w,d)}{\sum_d \sum_w n(d,w) P(z|w,d)}$$

$$P(w|z) = \frac{n(d,w) P(z|w,d)}{\sum_w \sum_d n(d,w) P(z|w,d)}$$

$$P(z|d) = \frac{\sum_w n(d,w) P(z|d,w)}{n(d)}$$

迭代 E-Step 和 M-Step，直到模型收敛，得到的结果即主体模型的参数。

在上述步骤中，主题的个数需要人工设定，这也是 PLSA 及类似模型的一个通病：主题个数设定不同，结果也往往大不相同，一个合适的主题个数设定会带来较好的主题分析结果。

除了 PLSA 外，LDA[11] 及其变种在微博话题发现中有着更多的应用。本文选用 PLSA 主要是出于性能的考虑：PLSA 和 LDA 在效果上类似，却有着更小的计算复杂度，对大文本而言是合适的。

3.4 微博话题状态检测

话题状态检测也是话题检测任务的重要

部分，在 TDT 中，话题状态表现为多种形式，本文仅仅关注三种话题状态，即新增话题、持续话题和消亡话题。

从微博文本中抽取出话题以后，可以对前后时间窗口的话题度量话题间相关性。在词袋模型下，话题或者主题本质上是词的概率分布。因此，度量话题的相关性可以通过度量词项概率分布实现。度量两个分布的相似性方法很多，如 KL 距离、相关系数、余弦相似度等指标。本文使用皮尔逊相关系数度量两个主题的相似性，尽管话题的词项概率分布不是严格的正态分布，但在文本挖掘中，应用皮尔逊相关系数度量话题相似性是比较有效的[12]。

给定话题 t_X 对应的词项分布 X 和话题 t_Y 对应的词项分布 Y，分布 X 和分布 Y 的皮尔逊相关性计算公式如下：

$$\rho(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma_X}\right)\left(\frac{Y_i-\bar{Y}}{\sigma_Y}\right) \tag{8}$$

其中，n 是样本数量，$\bar{X}$ 是 X 的样本平均值，σ_X 是 X 的标准差。

$\rho(X,Y)$ 描述的是两个分布的线性相关程度。$1\geqslant\rho(X,Y)\geqslant 0$，若 $r>0$，表明两个变量是正相关，若 $r<0$，表明两个变量是负相关。

利用皮尔逊相关性计算方法，对前后相继的话题集合 Z_A 和话题集合 Z_B，定义：

对 Z_B 中的话题 z：

(1)话题 z 为新兴话题，如果：$max(\rho(t,z_A))\leqslant 0, \forall_{z_A}\in Z_A$

(2)话题 z 为持续话题，如果：$max(\rho(t,z_A))\geqslant 0, \forall_{z_A}\in Z_A$

对 Z_A 中的话题 t，

(3)话题 t 为消失话题，如果：$max(\rho(t,z_B))\leqslant 0, \forall_{z_B}\in Z_B$

其中，$max(\cdot)$ 表示集合中最大的元素。

4 实验

实验使用的数据集是来自张华平开放的微博数据集[13]，时间跨度为 2012 年 3 月 11 日到 2012 年 3 月 13 日。数据不包括用户信息，仅有微博文本和发布时间。

为了减少噪音信息对话题检测的影响，对数据集做去噪处理，去噪处理前后数据集含有的微博文本条目数目见表 1。

表 1 去噪前后微博条目

时间	去噪前	去噪后
2012.3.11	45056	9872
2012.3.12	56740	10363
2012.3.13	89833	16976

对去噪后的微博文本使用 ICTCLAS 分词器分词并做词性标注，仅保留其中的名词、动词以及自定义词汇。进而，对词汇进行停用词处理，停用词表为《哈工大停用词表》[14]。经过去噪、分词、词性过滤，共得到候选词汇 26081 个。这些词汇作为后续主题词识别的候选词。

本文实验将时间窗口设为一天：将微博文本根据发布日期划分，每一天的微博文本同处在一个时间窗口。

4.1 主题词识别

应用公式(5)对候选词汇进行打分。公式(5)是一个复合公式，其中包含有可供调节的参数 α，不同的参数取值对主题词识别的影响较大。表 2 给出了不同参数下得分最高的词汇(笔者手动删除了一些因分词错误带来的无意义词汇)。

表 2 不同的 α 取值下主题词识别得分最高的前二十个词项

0.1	0.2	0.3	0.5	0.8	1
手机	合并	合并	合并	合并	合并
世界	手机	土豆	祈愿	祈愿	祈愿
时间	世界	手机	公牛	公牛	公牛
东西	时间	世界	植树	流星	树木
女人	希望	时间	失眠	植树	流星
活动	睡觉	优酷	早安	树木	爱护
希望	土豆	希望	土豆	爱护	植树
男人	心情	睡觉	流星	早安	尼克斯
看看	优酷	失眠	爱护	失眠	早安
爱情	活动	中国	树木	尼克斯	失眠
心情	女人	植树	优酷	土豆	见习
照片	中国	心情	年龄	优酷	烟花
天气	爱情	活动	心理	见习	信心百倍
人生	东西	女人	尼克斯	烟花	许愿
中国	人生	工作	睡觉	许愿	土豆
需要	工作	爱情	手机	信心百倍	干爹
关注	事情	人生	时间	年龄	优化
事情	需要	东西	烟花	干爹	年龄
体验	老师	事情	希望	心理	优酷

参数 α 的作用是调节两个系数对主题词提取结果的影响，当 α 为 0 的时候，仅有词汇相对词频作用于主题词提取；当 α 为 1 的时候，仅有词汇增长速度作用于主题词提取。在表 2 中，当 α 取较小的值时，出现在前面的多是词频较高的词汇，而当 α 取接近于 1 的值时，增长速度较快的词越来越多排在了前面。为了在相对词频和增长速度之间取得平衡，本文试验中，α 参数值设为 0.5。

得到每个词汇应用公式(5)的得分后，实验选取得分最大的前两千个词项作为后续话题检测的特征集。

4.2 主题识别和状态判定

基于 PLSA 的主题识别方法要求预设主题数量，实验设预设主题数量为 50。应用 PLSA 对 2012 年 3 月 12 日的微博以及 2013 年 3 月 13 日的微博文本分别做主题识别。主题识别的部分结果见表 3。

表 3 部分主题识别结果

2012 年 3 月 12 日	植树节	植树绿色好友达到移动结局话费相逢红豆友谊充值
	雷锋	感动指数雷锋回归文件做好道德投入演绎不足花朵得了
	限酒令和茅台酒	茅台发表消费放松条件饮料借口失落停止培养养成公款
	土豆优酷合并	土豆优酷合并宣布有限公司米饭时机意料提出北大参考交换
	桃姐上映	电话桃姐电视红色结婚父母女儿香港办法祝福不知老人
	星座	星座射手双鱼网络狮子座双子座巨蟹座狮子天蝎座冠军处女座金牛座
	苹果 ipad3	苹果联系发布长大研究利用产业奖励兴趣祈祷扑克说谎
	两会	习惯建议两会用户思念留下贵州玩具法律通知系列
2012 年 3 月 13 日	土豆优酷合并	土豆优酷合并视频宣布成都错过有限公司加班双方股份
	公务员	国家发展政府服务思想公务员婚姻迷失办公室事实人大代表
	NBA	直播公牛宽容受伤尼克斯移民外表罗斯下手胖子林书豪重视
	星座	星座双鱼射手天秤座狮子金牛座天蝎座狮子座毕业双子座巨蟹座处女座
	茅台	新闻使用消费充满茅台记者下雨心灵气质收入工程条件奢侈品

从上面的主题识别来看，本文提出的方法具有一定的有效性。2012 年 3 月 12 日和 3 月 13 日的一些热点时间都在话题检测结果中有所体现，如土豆优酷合并、植树节、“限酒令”事件、NBA 等等。当然，话题识别结果还包括“星座”等，这些内容也构成了微博上的热门话题。需要注意的是，主题建模方法并不能保证得到的每个主题都有明确的意义。

在确定话题以后，还需要对话题的话题状态进行确定，应用公式(8)，可以发现，在上述主题中，“星座”、“土豆优酷合并”、“茅台”等都属于稳定话题，而 3 月 12 日出现的“植树节”话题则属于消亡话题，3 月 13 日并没有新增主题，这同我们的直观感觉是符合的。

5 总结

本文提出了一种时序中文微博文本话题检测方法，这一方法包括四个步骤：文本预处理、主题词特征选取、话题发现和话题状态识别。实验结果证明，本文提出的方法具有一定的有效性。本文的主要工作在于两个方面，一是特征主题词的选取，由于微博文本的特性，传统的 TFIDF 策略不能用于提取微博文本的特征词，本文提出了一个新的跨时间窗口特征主题词选取策略；同时提出了一个微博话题状态的分类策略和划分方法。

本文的方法还存在很多改进空间，首先，本文的噪音处理策略并不完善，一个好的噪音处理策略将会大幅度提升话题检测效果；其次，本文对微博数据的内部特征利用较少。如何更好的去除微博文本中的噪音数据，如何利用微博数据的内部特征以检测微博话题将是未来研究的重点。

参考文献

[1] 李生，洪宇，张宇. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报，2007，21(6)：71-87

[2] 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

[3] 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

[4] 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[A]//中文信息学会. 第六届全国信息检索学术会议 2010[C]. 北京，2010:291-298

[5] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004:487-494

[6] Quercia D, Askham H, Crowcroft J. TweetLDA: supervised topic classification and link prediction in Twitter[C]// Proceedings of the 3rd Annual ACM Web Science Conference. New York:ACM, 2012:247-250

[7] Petrovic S, Osborne M, Lavrenko V. Streaming First Story Detection with application to Twitter[C]//The Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg:Association for Computational Linguistics,2010:181-189

[8] Mathioudakis M, Koudas N. TwitterMonitor: Trend Detection over the Twitter Stream[C]. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. New York:ACM,2010:1155-1158

[9] 张华平. ICTCLAS 官方网站[EB/OL]. [2013-02-25]. http://www.ictclas.org/

[10] Hofmann T. Probabilistic latent semantic analysis[C]//Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. San Francisco:Morgan Kaufmann Publishers Inc,1999:289-296

[11] Blei D M,Ng A Y,Jordan M I. Latent Dirichlet Allocation[J]. The Journal of Machine Learning Research, 2003(3): 993-1022

[12] 曼宁. 统计自然语言处理基础[M]. 北京：电子工业出版社，2005:24-35

[13] 张华平. 2012 年 3 月 11 日到 3 月 24 日新浪微博的 60 万条实时微博消息，数据堂:科研数据共享平台[EB/OL]. [2013-06-12]. http://www.datatang.com/data/42505

[14] 哈工大信息检索实验室等. 停用词集合(哈工大停用词表、四川大学机器智能实验室停用词库、百度停用词表)，数据堂:科研数据共享平台[EB/OL]. [2013-05-07]. http://www.datatang.com/data/19300

(收稿日期:2013-07-05)