一种网络话题的内容焦点迁移识别方法

周亚东[1)] 刘晓明[1)] 杜友田[1)] 管晓宏[1),2)] 刘 霁[1)]

[1)](西安交通大学智能网络与网络安全教育部重点实验室 西安 710049)

[2)](清华大学自动化系智能与网络化系统研究中心 北京 100084)

摘 要 随着网络信息技术的迅速发展,互联网已经成为人们获取和发布信息的最重要平台之一.在互联网的信息传播过程中,话题相关文本不断更新,而其内容焦点也随着话题发展发生着迁移.识别话题内容焦点有助于有效地挖掘与分析网络信息,是网络舆情分析领域的重要研究问题.文中针对网络流文本,提出了一种网络话题内容焦点的识别方法,首先对话题焦点特征在流文本中的分布情况进行分析,基于分析结果介绍了焦点识别方法 3 个主要步骤的算法模型,分别是基于时间属性的焦点特征词提取、内容焦点特征词的合并和内容焦点的表示.文本基于来自于真实网络的实际数据,对所提方法进行了实验验证,实验结果表明文中所提方法可有效获取话题发展过程中的内容焦点,并能以关键词集和语句集的形式对内容焦点进行表示.

关键词 网络话题;焦点识别;舆情分析;话题模型;社交网络;社会计算

中图法分类号 TP393 **DOI 号** 10.3724/SP.J.1016.2015.00261

A Method for Identifying the Evolutionary Focuses of Online Social Topics

ZHOU Ya-Dong[1)] LIU Xiao-Ming[1)] DU You-Tian[1)] GUAN Xiao-Hong[1),2)] LIU Ji[1)]

[1)](*Ministry of Education Key Lab for Intelligent and Network Security, Xi'an Jiaotong University, Xi'an* 710049)

[2)](*Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University, Beijing* 100084)

Abstract With the rapid development of information technology, Internet has become one of the most important platforms for people to get information. In our daily life, people tend to encounter this phenomenon: the news of certain topic is constantly updated, but the reports focus on different contents, these different focus include: the focus generated by the development of events; new focus caused by the increased user reviews of topics; focus migration due to the impact of other hot news topic in the same period. In this paper, we propose a method for analyzing and identifying the evolutionary focus of topics. The method is consist of three parts, including feature selection based on time attribute, feature combine model and focus presentation. The experimental results show that this method could identify the evolutionary focus of topics effectively.

Keywords online social topics; focus identification; public opinion analysis; topic model; social networks; social computing

1 引 言

随着网络和信息技术的迅速发展,人们每天都在阅读和发布着大量网络文本,例如每天不断更新的网络新闻,博客、微博、论坛用户发布的博文网贴,连续收发的电子邮件等,并且在这些网络文本中蕴含着大量有关舆论热点的有价值信息.从时间的角

收稿日期:2013-09-24;最终修改稿收到日期:2014-08-27.本课题得到国家自然科学基金(61202392,61221063,61375040,60905018,61103240,61172124)、国家"八六三"高技术研究发展计划项目基金(2012AA011003)、高等学校博士点基金(20120201120023)、教育部创新研究团队(IRT13035)和中央高校基本科研业务费专项资金资助.**周亚东**,男,1982 年生,博士,讲师,主要研究方向为在线社会网络、Web 数据挖掘.E-mail:ydzhou@mail.xjtu.edu.cn.**刘晓明**,男,1989 年生,博士研究生,主要研究方向为社会网络.**杜友田**(通信作者),男,1980 年生,博士,副教授,主要研究方向为互联网多媒体理解、机器学习和社会网络.E-mail:duyt@mail.xjtu.edu.cn.**管晓宏**,男,1955 年生,博士,教授,主要研究领域为复杂网络资源分配与调度、网络安全、传感器网络.**刘 霁**,女,1986 年生,硕士,主要研究方向为社会网络.

度分析,这些网络文本形成了一种随时间分布的流文本数据.在阅读这些网络流文本数据时,人们会发现,关于某个热点话题的网络流文本的内容焦点随着时间的变化在动态演变迁移,例如:关于某个事件的网络新闻报道在不断更新,但报道内容侧重不同;针对某个网络话题的网络评论不断被发布,但其关注重点在逐步变化.及时识别和发现网络热点话题的内容焦点的迁移,可以更全面地了解网络话题信息组成结构及其演变趋势,对于分析网络舆情态势、判断舆论预期走向具有十分重要的价值,也是网络舆情分析研究领域中的一个重要研究问题.但在实际数据中,无论是网络新闻文本还是博客、论坛文本都具有文本冗余信息量大、内容焦点特征提取难度大等难题,为识别网络话题的内容焦点提出了较大挑战.

本文以网络流文本为对象,通过分析网络话题内容焦点的迁移特性,提出了网络话题内容焦点的识别方法.网络话题内容焦点可由引发网络话题的热点事件、事件对社会的影响、用户对该话题的评论、后续相关事件发展等组成,也包括关于某话题的焦点起始、达到高峰、逐渐消亡的演变过程以及各个焦点之间的影响关系等内容.

结合以上分析,本文所做研究主要分为以下几方面内容:

(1)针对内容焦点识别的需要,分析焦点特征词在网络话题流文本中的分布情况,归纳分析焦点特征词的主要分布特点,为后续方法提供基础.

(2)针对网络话题流文本数据,提出基于时间属性的焦点特征词提取算法,可提取语义具有代表性且在时序分布上能覆盖全文的焦点特征词,以达到保留原流文本集的内容焦点的主要关键信息和降低处理复杂度的目的.

(3)任一内容焦点可由若干个特征词共同描述,需将同属一项内容焦点的特征词合并,本文借鉴了概率话题模型的建模思想,提出了焦点特征词的合并模型,通过计算任意两个特征词间的相似度,并结合文本时序特性,将描述同一个关注焦点的特征词合并.

(4)根据合并的特征词集合数据,结合流文本数据,将网络话题的各阶段内容焦点表示,提取出各个内容焦点的起始时间、关键词和内容摘要等信息.

文本研究数据来源为"新浪博客"网站和"新浪新闻"网站中的 2009 年和 2010 年两年的网络流文本数据,共人工标定 114 组热点话题的内容焦点数据用于分析研究,后续为叙述便利和篇幅限制,将以 2009 年发生的"杭州飙车案"话题为例介绍算法流程和实验结果.

本文首先分析国内外相关研究工作现状,然后提出网络话题内容焦点的识别方法,接着基于实际网络流文本数据实验验证所提方法的有效性,最后给出结论.

2 相关工作

本文研究了网络话题内容焦点的识别方法,与之相关的研究主要包括话题检测与跟踪、网络话题动态演变特性分析等.以下介绍国内外相关研究的发展现状与趋势.

2.1 话题检测与跟踪研究现状

话题检测与跟踪(Topic Detection and Tracking,TDT)[1-2],起源于美国国防高级研究计划署(DARPA)1996 年展开的一项计划.TDT 研究的数据大多数是真实的新闻报道,新闻报道的事件主要包括时间、地点、人物、事件这几个要素.Carthy 等人[3]利用 WordNet 建立语义链,将文本表示成语义链的集合,通过比较两个文本语义链之间的相似度来判断文本是否属于同一话题,并通过实验得出实体名词有利于提高文本分类精度的结论.Makkonen 等人[4]通过定义语义向量,在 TDT 中引入简单的语义,来提高文本检测的准确率.Pons-Porrata 等人[5]提出了一种新的层次化的文本聚类算法,它以网络新闻文本为研究对象,将文本中的所有与时间表述有关的词都提取出来,并利用这些词重新定义了文本相似度比较函数,并通过实验证明该方法能提高聚类效果.

国内学者在这方面也有较多研究成果.贾自艳等人[6]提出了基于时间距离的相似度计算模型,以文本的创建时间为时间起始点,统一文本中的时间表述方式,通过比较两篇报道间的时间差值,来削弱基于内容获得的相关度.宋丹等人[7]则改进了原有的向量空间模型,将文本表示成四个独立的向量空间,分别为地点向量、时间向量、人物向量和内容向量,同时提出了通用的向量相似度衡量方法,时间向量相似度方法以及地点向量衡量方法,最终将文本的分类问题转化为相似度的计算问题.但是文章中没有介绍如何正确分辨并提取对应的时间类词、地点类词等.

话题模型的层次化和结构化研究是目前 TDT 领域的重要方向,其中,层次化主要指将同一话题下的新闻报道组织为具体的层次结构;结构化侧重挖掘同一话题的不同侧面.国内相关研究在方法上注重自然语言处理技术和统计学相结合,在趋势上,逐

步与数据挖掘、事件抽取以及篇章理解等相关技术相融合.

2.2　信息检索相关领域研究现状

在特征提取方面，Yang 等人[8]比较了几种常用的特征提取方法：文档频率法(DF)、信息增益法(Information Gain)、互信息法(Mutual Information)、C2 统计方法[9]和 Term Strength 方法[10]. 并通过实验表明，在 KNN 和 LISF 分类器下，在保证相同分类精度的前提下，信息增益法和互信息法降维的力度最大. 之后，代六玲等人[11]在中文特征抽取问题上也对这些方法进行了比较，并得到了相似的结论. Mei 等人[12]提出了一种在多维话题模型中，自动且客观的生成标签的方法. Mei[13]还基于 PLSA[14]模型提出了 CPLSA(Contextual Probabilistic Latent Semantic Analysis)，该模型通过对文本的诠释资料(例如：作者、出版社和发表时间等)进行分析，从而判断两个文本主题是否相同.

在话题演变问题上，Pui[15]从文本分类的角度提出文本主体变化的检测方法，通过 DCM(Discrimination Category Matching)方法来建立该主题变化过程的模型. Jure[16]以网络博客为研究对象，分析社会网络，得到了信息传播、话题演变的规律. Jure[17]还以网络新闻文本和博客中的小短句为研究对象，通过分析发现博客帖子达到关注的峰值总要比新闻文本达到关注峰值的时间晚 2.5 h. Myra[18]提出了 MONIC(Modeling and Monitoring Cluster Transitions)框架，通过研究文本聚类中各个子类在不同时刻的不同数据集之间的交并比例关系，分析子类之间的演化过程，包括新生、消亡、吸收和分散.

2.3　其他相关方法研究现状

在自然语言处理领域，向量空间模型(Vector Space Model，VSM)[19]是目前使用最为广泛的文本模型，该模型基于词袋假设把文本表示成由特征权值构成的向量，这种方法已经成功应用在文本自动分类和静态文本聚类等方面.

目前，国内外学者重点对话题模型展开了研究，通过分析话题模型来达到识别话题的目的. 话题模型(Topic Model，Language Model)是信息检索领域中普遍使用的一种文本模型，这种方法构造了一个可以描述文本生成的随机过程，通过 EM[20]算法、蒙特卡洛马尔可夫[21]等参数估计算法反推模型参数，得到描述文本深层主题信息的统计量.

在文本表示中，向量空间模型和话题模型都有应用，例如 Aggarwal 等人[22-23]使用的是基于词频-反向文档频率(Term Frequency and Inverse Document Frequency，TFIDF)的向量空间模型表示法，Liu 等人[24]则使用基于语义平滑模型的话题模型表示法，Walker 等人[25]使用概率话题模型文本表示法.

3　网络话题的内容焦点的识别方法

本节首先对话题焦点特征词在网络流文本集合中的分布情况进行分析，并以此为基础，介绍网络话题内容焦点识别方法中的 3 个主要步骤，分别包括特征提取、特征合并和焦点识别，并将介绍各步骤的理论模型和实现算法.

3.1　话题焦点特征词在流文本中分布特点的分析

为了能够正确识别网络话题内容焦点的特征词，需要从特征候选词集中，选择具有前文提到的语义明确、与话题相关、能反应话题的某个焦点以及能与其他话题区别等 4 个特点的词作为特征词.

以话题“杭州飙车案”为例，随着话题讨论的发展，通过人工分析发现其关注内容迁移分为 6 个焦点(见图 1). 本文绘制了 4 组典型的特征词和噪声词的词频-文档分布图，如图 2～图 5 所示.

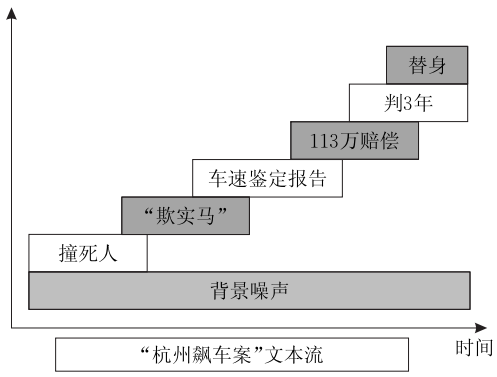

图 1　话题“杭州飙车案”的内容焦点迁移情况

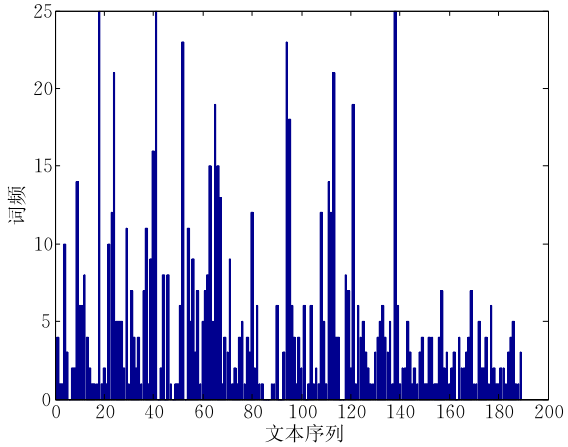

图 2　噪声词“杭州”的词频-文档分布图

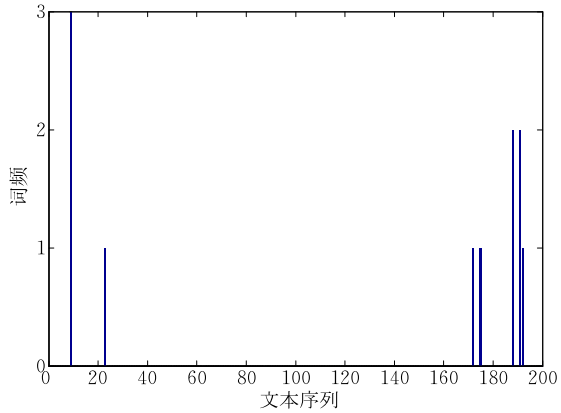

图 3 噪声词“对不起”的词频-文档分布图

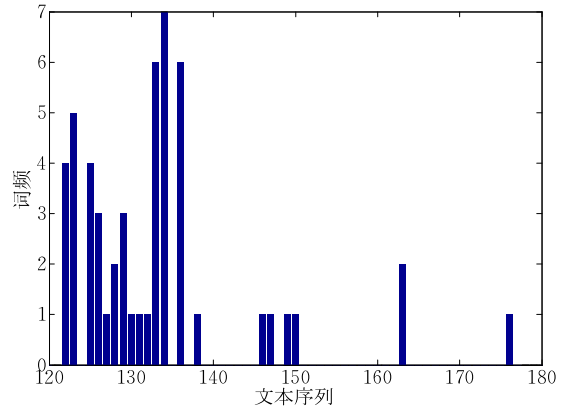

图 4 特征词“113 万”的词频-文档分布图

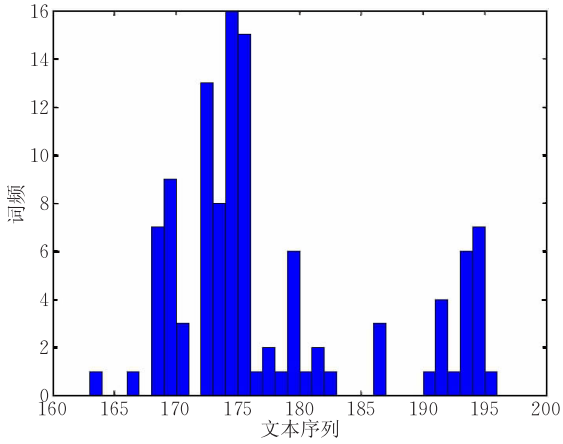

图 5 特征词“替身”的词频-文档分布图

噪声词. 指与话题内容无显著关联(可称为无关噪声词),或者对话题焦点相关内容识别会产生干扰的词语(可称为干扰噪声词).

特征词. 是能够描述事件的最恰当的词语,是代表事件中焦点的标准化术语.

在这些图中,横轴代表流文本集中的各个文本编号(按照文本发布时间先后进行排序),范围从0～195,表示该话题相关的 196 篇网络文本;纵轴代表所示词在各个文本中出现的次数.

通过分析图 2～图 5 可以发现,特征词和噪声词在文档分布上存在较大差异,特征词的分布特性归纳如下:

(1)频度较高. 作为话题内容焦点的特征词,其语义是话题发展过程中某个阶段的重点,因此该词应该在这个阶段具有较高的词频,而在其他阶段不具有高词频.

(2)分布持续. 话题内容焦点存在一定生命周期.因此话题内容焦点的特征词若在 t 时刻出现,那么也很可能在 $t+1$ 时刻出现,呈现出持续性.

(3)持续时间适中. 特征词用于描述某个内容焦点,则其持续时间与该焦点的生命周期相近,并且小于话题的生命周期.

(4)同焦点的特征词时序分布重合. 在话题发展变化过程中,同一项焦点的若干特征将同时出现的,他们共同表达了这个阶段的主要焦点.

这 4 项特性中,前 3 项可用于特征词提取评估函数的建立,第 4 项可用于特征词合并方法的设计.

对于无关噪声词和干扰噪声词,无关噪声词主要是一些常用词,可用于描述话题的个别内容,但与话题主要内容关系不大. 如图 3 中列举的“对不起”,该词语在话题文本中有一定出现频度,但由于与杭州飙车案的焦点内容关系较小,在文档词频上不符合“频度较高”和“分布持续”等特性. 干扰噪声词是与话题内容有关,但含义较为宽泛,会对识别话题焦点造成干扰的词语,但一般在话题相关文本中长时间存在,在文档词频分布中不符合“持续时间适中”特性. 如图 2 中列举的“杭州”,该词语与话题的全部内容均有较大关系,无法用于准确表征某单项焦点,则会对各项焦点的识别产生干扰.

需要注意的是特征词的分布存在一些噪声,如特征词“113 万”在文档 138 之后还有少量出现,这对于评估特征词存在干扰,因此可采用均值滤波法对词频-文档分布进行过滤,以增强特征词与噪声词间的差异.

3.2 基于时间属性的焦点特征词提取算法

通过提取可表征焦点内容的词作为特征,一方面降低了文本表示的维度,使得算法的复杂度降低;另一方面减少了噪声对话题内容焦点识别的影响,提高了识别准确度. 网络话题内容焦点识别针对的数据为带有时间属性的网络流文本集合,其与普通静态文本集合的最大区别在于具有时间属性,因此本文提出了一种基于时间属性的特征提取算法. 该算法主要包括文本预处理和特征提取两个部分.

在文本预处理部分，主要的算法步骤为：

(1) 对关于某话题的流文本集中的每个文本，分别提取该文本的标题、正文和发布时间；

(2) 对每个文本的标题和正文分别进行分词，得到文档词集；

(3) 计算文档词集中每个词在流文本集中的权重，生成词频文档矩阵；

(4) 根据标题的分词结果，生成特征候选词集.

在特征提取部分，主要的算法步骤为：

(1) 分析特征词的特点，并建立评估函数；

(2) 根据评估函数，计算每个候选词的权重；

(3) 将特征候选词按照权重值大小排序，并设定合理阈值，提取特征.

以下，本文将详细介绍特征提取部分中的评估函数和特征筛选方法.

3.2.1　评估函数的建立

根据特征词分布的前 3 项特点，假设某个话题的流文本集为 $D_z=\{d_1,d_2,\cdots,d_{M_z}\}$，则对于特征候选词集中的任意一个词 $w_k\in C$，给出如下的评估函数：

$$f(w_k)=\sum_{i=1}^{N_k}\frac{freq_{i,k}}{\sum_{j=1}^{N_k}freq_{j,k}}\times\frac{1}{(time(d_i')-time(d_{i-1}'))}\times e^{-\frac{(|N_k|-\mu)^2}{2\sigma^2}} \tag{1}$$

式中，N_k 为包含词 w_k 的文本的个数；d_i' 为包含词 w_k 的第 i 个文本；$freq_{i,k}$ 为采用均值滤波法对词频-文档分布平滑去噪后的文本 d_i' 中词 w_k 的出现频度；$time(d_i')$ 表示文本 d_i' 的发布时间；μ,σ 为函数中的经验系数. 该式可体现本文上面提到的焦点特征词分布的几项特点. 评估函数的值与 $freq_{j,k}$ 即词频成正相关；与 $(time(d_i)-time(d_{i-1}))$ 反相关，即出现特征词的两个文本间的发布时间差距越近，评估函数值越大，体现特征词的分布持续性；$e^{-\frac{(|N_k|-\mu)^2}{2\sigma^2}}$ 用于约束词的持续时间，从该函数的增减性可体现候选词的持续时间应该在一个合理的范围内，由于大部分情况下话题的单个焦点持续时间为 2～3 天，因此通常 μ 可取值为 2.5.

3.2.2　特征提取

利用评估函数值确定话题内容焦点的特征词，即特征提取. 本文采用 Top M 原则，该方法的核心思想是把候选特征词按照其评估函数值大小进行排序，然后从中选择 M 个词，成为焦点特征词集合.

3.3　内容焦点特征词的合并方法

网络话题的关注焦点需要若干特征词共同描述，因此需要在提取出的焦点特征词集的基础上，合并描述同一项内容焦点的特征词. 本文借鉴概率话题模型的方法和原理[26-27]，提出了内容焦点特征的合并方法.

3.3.1　特征合并模型的建立

概率话题模型的思想是通过引入隐变量，描述一个文本的生成过程，并且建立了话题与文本间的关系. 本文借鉴其思想，通过引入一个二维隐变量，建立了特征合并模型，可描述网络流文本集的生成过程，并可描述焦点特征词与焦点相关流文本集的关系.

假设包含某一内容焦点特征词 g_i 的网络流文本集 D_i 由文本 $d_{i1},d_{i2},\cdots,d_{iN_i}$ 构成，且每个文本都包含一个或多个该焦点的特征词. 因此，可认为网络流文本集 D_i 可由两部分信息组成：一部分信息 f_i 与焦点特征词 g_i 相关；另一部分信息 $\bar{f}_i$ 与焦点特征词无关，称为背景噪声. 并且可认为文本集 D_i 是由这两部分线性混合组成. 因此，引入一个二维隐变量表示为

$$F=\{f_i,\bar{f}_i\} \tag{2}$$

设 V 为在关于某话题相关的所有网络流文本集中出现过的所有词的集合，如果需要生成一个焦点流文本集 D_i，则步骤如下：

(1) 以 $p(f_i|D_i)$ 的概率从 F 中选择组成焦点文本集 D_i 的一个部分 f_i；

(2) 以 $p(w|f_i)$ 的概率从词集 V 中选择与 f_i 有关的词 w；

(3) 以 $p(\bar{f}_i|D_i)$ 的概率从 F 中选择组成文本集 D_i 的另一个部分 $\bar{f}_i$；

(4) 以 $p(w|\bar{f}_i)$ 的概率从 V 中选择组成与 $\bar{f}_i$ 有关的词 w.

经过上面 4 个步骤，可生成一个文本集 D_i. 从该文本集的生成过程中，可以得出本文特征合并模型的核心公式：

$$p(w|D_i)=p(w|f_i)p(f_i|D_i)+p(w|\bar{f}_i)p(\bar{f}_i|D_i) \tag{3}$$

因为　　$p(f_i|D)+p(\bar{f}_i|D)=1$ (4)

令 $p(\bar{f}_i|D_i)=\lambda$，则 $p(f_i|D_i)=1-\lambda$

$$p(w|D_i)=(1-\lambda)p(w|f_i)+\lambda p(w|\bar{f}_i) \tag{5}$$

其中，λ 表示背景噪声部分信息占文本集 D_i 总信息的比例，根据经验和实验结果分析通常可取值为 0.8. 式(5)为本文特征合并模型的最终表达式. 通过该式中的 $p(w|f_i)$ 可建立特征词 g_i 与文本集 V 中的任一词 w 之间的关系.

此时，判断文本集的特征词 g_i 和 g_j 的语义内容是否描述了同一项内容焦点，可通过计算由特征合并模型建立的两个关系 $p(w|f_i)$ 和 $p(w|f_j)$ 是否相近，如果 g_i 和 g_j 描述了同一项内容焦点，则与其相关的其他词集将会有较大的重叠. 因此，求解式(5)中的 $p(w|f_i)$ 是非常重要的问题，其计算方法在下节详细介绍.

需要注意的是，对于每一个特征词 g_i，我们需要为其构建一个文本集合 D_i，但由于特征词由 Top M 准则选择，因此一个话题的特征词数量最多只有 M 个(在基于实际数据的分析中，我们选取 M 值为 40)，因此在这种情况下，为每一个特征词构造文本集产生的计算量是可以接受的.

3.3.2 合并模型中的参数估计方法

对于式(5)中 $p(w|f_i)$ 的估计问题，可建立该参数估计问题的似然函数，表示为

$$\begin{aligned} L(f_i) &= L(p(w|D_i)) \\ &= \sum_{d\in D_i}\sum_{w\in d}\log((1-\lambda)p(w|f_i)+\lambda p(w|\bar{f}_i)) \end{aligned} \tag{6}$$

令 $n(w,d)$ 为词 w 在文本 d 中出现的次数，则式(6)的似然函数又被表示为

$$\begin{aligned} L(f_i) = \sum_{d\in D_k}\sum_{w\in V} n(w,d)\log(\lambda p(w|f_i)+ \\ (1-\lambda)p(w|\bar{f}_i)) \end{aligned} \tag{7}$$

式(7)的似然函数过于复杂，是和函数的对数，用传统的极大似然估计法进行参数估计比较困难，因此，本文采用 EM 算法进行参数估计. 基于式(5)所示的模型，本文引入了一个隐变量 z，对于任何一个文本中的词 w_{ij} ($w\in V$，文本 d_i 中的第 j 个词)都有一个变量 z_{ij}，用于指出该词是由背景噪声产生，还是由特征词相关内容产生. 因此，z_{ij} 具体表示为

$$z_{i,j}=\begin{cases}0, & \text{如果词 } w_{ij} \text{ 来自背景噪声}\\ 1, & \text{其他}\end{cases} \tag{8}$$

又因为

$$p(z|w,d)=\frac{p(w,z|d)}{p(w|d)} \tag{9}$$

所以，根据式(5)、式(8)和式(9)进一步得到

$$p(z=0|w,d)=\frac{\lambda p(w|\bar{f}_i)}{(1-\lambda)p(w|f_i)+\lambda p(w|\bar{f}_i)} \tag{10}$$

$$p(z=1|w,d)=\frac{(1-\lambda)p(w|f_i)}{(1-\lambda)p(w|f_i)+\lambda p(w|\bar{f}_i)} \tag{11}$$

引入隐变量后，该参数估计问题的似然函数变为

$$\begin{aligned} L_c(f_i) &= \log p(w,z|d) \\ &= \sum_{d\in D_i}\sum_{w\in d} z\log(1-\lambda)p(w|f_i)+ \\ &\quad (1-z)\log\lambda p(w|\bar{f}_i)) \end{aligned} \tag{12}$$

定义 Q 函数为似然函数 $L_c(f_i)$ 对隐变量 z 的条件期望，则该参数估计问题的 Q 函数表示为

$$Q(f_i)=E_{p(z|w,d)}[L_c(f_i)]=\sum_z L_c(f_i)p(z|w,d) \tag{13}$$

因此得出如下优化问题：

$$\begin{aligned} &\max_{p(w|f_i)} Q(f_i) \\ &\text{subject to } \sum_{w\in V} p(w|f_i)=1 \end{aligned} \tag{14}$$

因此，可得拉格朗日函数如下：

$$h(f_i)=Q(f_i)+\eta\Big(1-\sum_{w\in V}p(w|f_i)\Big) \tag{15}$$

对函数 $h(f_i)$ 求一阶偏导数，并令其等于零，得到以下方程：

$$\begin{aligned} \frac{\partial h(f_i)}{\partial p(w|f_i)} &= \sum_{d\in D_k} n(w,d)p(z=1|w,d)\times \\ &\quad \frac{1}{p(w|f_i)}-\eta=0 \end{aligned} \tag{16}$$

从上式中可以发现，$p(w|f_i)$ 和 $\sum_{d\in D_k} c(w,d)p(z=1|w,d)$ 成正比. 因此，针对该参数估计问题，得到其 EM 算法迭代式如下：

E-Step：

$$p(z=1|w,d)=\frac{(1-\lambda)p(w|f_i^{(n)})}{(1-\lambda)p(w|f_i^{(n)})+\lambda p(w|\bar{f}_i)} \tag{17}$$

M-Step：

$$p(w|f_i^{(n+1)})=\frac{\sum_{d\in D_k} n(w,d)p(z=1|w,d)}{\sum_{w\in V}\sum_{d\in D_k} n(w,d)p(z=1|w,d)} \tag{18}$$

从中可以发现，在计算 E-Step 时不需要精确地计算 Q 函数的值，可以通过计算隐含变量的概率分布代替.

3.3.3 特征合并

特征合并是针对焦点特征词集合，两两比较它们的时序分布和相关词集是否都有较大重叠，只有两个条件都满足的特征词，才能够合并. 以下分别给出特征词的时序分布和相关集词的合并条件.

(1) 特征词时序分布的合并条件

对于任一特征词 g，其时序分布 Ts_f 的具体形式为

$$Ts_f=\{ts_1^{(g)},ts_2^{(g)},\cdots,ts_{N_f}^{(g)}\} \tag{19}$$

$$ts_i^{(g)}=\{t_{i1}^{(g)},t_{i2}^{(g)},\cdots,t_{iK}^{(g)}\} \tag{20}$$

式中，$ts_i^{(g)}$ 代表特征词 g 的第 i 个时间分布区间(特征词的时序分布可能会出现不连续情况)，$t_{ij}^{(g)}$ 代表具体的时刻. 对于任意两个特征词 g_i 和 g_j，一般情况下其时序阶段合并条件为

$$\exists ts_i^{(g_i)} \in Ts_{f_i} \text{ and } \exists ts_j^{(g_j)} \in Ts_{f_j}$$

$$\frac{\min\{(t_{iK}^{(g_i)} - t_{i1}^{(g_i)}), (t_{jK}^{(g_j)} - t_{j1}^{(g_j)})\}}{\max\{t_{iK}^{(g_i)}, t_{jK}^{(g_j)}\} - \min\{t_{i1}^{(g_i)}, t_{j1}^{(g_j)}\}} \geqslant threshold \quad (21)$$

其中，$threshold$ 为阈值，在本文中取值 0.5. 当条件式(21)不满足时，考虑特殊情况：

$$\exists ts_i^{(g_i)} \in Ts_{f_i} \text{ and } \exists ts_j^{(g_j)} \in Ts_{f_j} \quad (22)$$

$$ts_i^{(g_i)} \subset ts_j^{(g_j)}$$

(2) 特征词相关词集的合并条件

对于任一特征词 g 通过 $p(w|f)$ 可找到文本词集中与之最相关的 N 个词，记作

$$RW_g^N = \{w_{g1}, w_{g2}, \cdots, w_{gN}\} \quad (23)$$

对于任意两个特征词 g_i 和 g_j，其相关词集的交集满足下式时，可以合并.

$$RW_{g_i}^N \cap RW_{g_j}^N \neq \varnothing \quad (24)$$

其中，N 值可变，在本文中取 $N=10$.

通过上面给出的两个条件，对于焦点特征词集 G，最终可得到特征词的合并结果 U，表示为

$$U = \{u_1, u_2, \cdots u_{|U|}\} \quad (25)$$

$$u_i = \{g_{i1}, g_{i2}, \cdots, g_{ik}\} \quad (26)$$

式中 U 代表对某话题所有焦点的特征词合并结果，u_i 代表其中某一项焦点的特征词集合.

3.4　网络话题内容焦点的表示

网络话题内容焦点的表示是一个从特征词扩充到摘要的过程，核心步骤有两点：(1) 提取可描述关注焦点 s 内容的关键词集合，以及这些词的权重；(2) 检索可描述关注焦点 s 内容的语句，作为该焦点的摘要.

对于提取能描述关注焦点 s 的关键词集合 $terms$，已有存在特征合并后的关于某焦点的特征词集合 $u_i = \{g_{i1}, g_{i2}, \cdots, g_{ik}\}$，目标是提取与 u_i 最相关的关键词并计算权重. 在前文的参数估计和特征合并中，已经知道根据每个特征词的 $p(w|f_i)$，就能找到与特征词 g_i 相关的词，同理，只要找到与 u_i 相对应的 $p(w|u_i)$，就能找到与 u_i 最为相关的关键词. 又因为 u_i 是特征词的集合，所以本文给出如下结论：

$$p(w|u_i) \approx \sum_{j=1}^{M} \alpha_j p(w|f_{ij}), \text{且} \sum_{j=1}^{iK} \lambda_j = 1 \quad (27)$$

此时，假设 u_i 中的特征词都是同等重要的，即 α_j 值都相等，则

$$p(w|u_i) \approx \frac{1}{iK} \sum_{j=1}^{iK} p(w|f_{ij}) \quad (28)$$

若 $p(w_j|u_i) w_j \in V$ 越大，则词 w_j 就与 u_i 越相关. 如果假设需要用 10 个关键词表述一个焦点，那么将 $p(w_j|u_i)$ 按大小顺序排序并取最前的 10 个词作为关键词，则此时关注焦点 s 的 $terms$ 可被表示为

$$terms = \{w_j \in V, p(w_j|u_i)\}_{j=1}^{10} \quad (29)$$

对于检索与关注焦点 s 最相关的语句作为该焦点的摘要 $abstract$，可通过检索该话题相关网络流文本语句中包含关键词集 $terms$ 最多的语句获得.

至此，焦点的关键词集 $terms$ 与语句摘要 $abstract$ 都可以分析得到的，即可对网络话题的内容焦点进行表示.

4　实验结果与分析

4.1　基于时间属性的焦点特征词提取算法实验结果

本实验首先通过人工标注出 114 组热点话题的内容焦点数据中每组热点话题的特征词集合，然后将本文所提的基于时间属性的特征词提取方法与经典的 TF-IDF 方法以及 FSBIFDR 方法[28] 和 STFS 方法[29] 进行特征词提取的性能对比，对比所用指标包括平均的准确率、召回率和 F 值，具体如下所示：

准确率＝正确识别的特征词总数/
　　识别出的特征词总数；

召回率＝正确识别的特征词总数/
　　测试集中存在的特征词总数；

F 值＝正确率×召回率×2/(正确率＋召回率).

以话题“杭州飙车案”为例，人工从话题文本的词汇集共 8271 个词(标题词 641 个)中标注出最能表征网络话题内容焦点迁移的 24 个特征词，接着分别采用基于时间属性的焦点特征提取算法(采用 Top M 原则，且 $M=40$)和经典的 TF-IDF 特征提取算法进行实验对比，如表 1 所示. 由于实验所用分词软件采用了基于知识库的分词算法，因此部分非常见的特征词未能正确划分，为了更好地贴近网络文本的处理流程，本文在做特征词人工标注时直接对分词软件处理后的分词结果进行标注，一些原属于特征词但被错误分词的单字也被标注为特征词. 如“欺实马”被分为了 3 个单字，因此在人工标注中将这 3 个单字分别标为 3 个特征词(其他单字同属这种情况). 对于被错误分词的单字特征词，由于其原属同一个特征词，在时序分布和相关词集两方面相关性较大，因此可在本文 3.3.3 节所介绍的特征合并算法中，被正确合并到同一个话题焦点的特征词集合.

表 1 网络话题“杭州飙车案”特征提取对比实验结果

人工标注			TF-IDF			基于时间属性的特征词提取		
序号	ID	词	ID	词	权重	ID	词	权重
1	7	撞	76	不	0.005218	**181**	**鉴定**	10.679810
2	8	死路	33	胡斌	0.005118	76	不	10.678430
3	9	人	**181**	**鉴定**	0.004993	55	肇事	9.088200
4	128	车速	39	安全	0.003626	95	警方	8.430467
5	142	欺	64	网友	0.003568	**9**	**人**	7.853923
6	80	实	**9**	**人**	0.003548	32	称	7.335126
7	143	马	291	谭卓	0.003527	**175**	**码**	7.151329
8	184	70	**104**	**赔偿**	0.003409	313	交通	6.740934
9	175	码	244	公里	0.003367	14	富家	5.471017
10	196	时速	71	事故	0.003294	**559**	**替身**	5.054636
11	181	鉴定	**175**	**码**	0.003279	22	续	4.669367
12	241	报告	428	上	0.003090	314	肇事罪	4.157694
13	310	双方	52	说	0.003082	**184**	**70**	3.801288
14	442	签订	38	公共	0.003070	525	被告	3.755456
15	443	113 万	98	改装	0.003036	**7**	**撞**	3.745604
16	445	元	37	危害	0.003016	28	子	3.540793
17	104	赔偿	349	罪	0.002957	100	家属	3.029454
18	453	获	96	车辆	0.002906	180	司法	2.998215
19	526	一审	130	富	0.002883	11	图	2.982306
20	491	3	368	法院	0.002859	33	胡斌	2.283121
21	149	年	**184**	**70**	0.002837	**104**	**赔偿**	2.220456
22	25	刑	303	会	0.002835	78	社会	2.131283
23	559	替身	313	交通	0.002731	130	富	1.554044
24	571	出庭	**559**	**替身**	0.002725	**149**	**年**	1.392336
25			95	警方	0.002721	217	死者	1.226147
26			23	肇事者	0.002719	114	父亲	1.207511
27			55	肇事	0.002611	31	专家	1.207075
28			132	法律	0.002592	38	公共	1.117866
29			123	公众	0.002568	**310**	**双方**	0.958237
30			359	部门	0.002557	**491**	**3**	0.850239
31			339	交警	0.002557	**25**	**刑**	0.737005
32			193	超速	0.002517	**443**	**113 万**	0.658967
33			**7**	**撞**	0.002472	**452**	**获**	0.638376
34			345	记者	0.002442	188	成	0.629192
35			299	中	0.002438	64	网友	0.626046
36			180	司法	0.002422	4	跑车	0.569413
37			34	行为	0.002416	108	闹市	0.560337
38			78	社会	0.002412	10	组	0.518214
39			314	肇事罪	0.002406	160	飚	0.502944
40			31	专家	0.002403	**8**	**死路**	0.477285

对所有 114 组话题数据的特征词提取实验结果如表 2 所示，可以看到本文所提特征提取方法的准确率、召回率和 F 值均明显优于 3 种对比方法，更是达到了传统的 TF-IDF 方法的 2 倍.

表 2 114 组网络话题的特征提取对比实验统计结果

	对比指标		
	平均召回率/%	平均准确率/%	平均 F 值
TF-IDF	33.4	16.3	0.22
FSBIFDR	54.1	26.3	0.35
STFS	48.7	23.7	0.32
基于时间属性的特征词提取	69.2	33.8	0.45

4.2 话题内容焦点识别的实验结果

本文中的网络话题内容焦点的表示主要由两部分组成，分别是关键词 *terms* 和摘要 *abstract*. 表 3 展示了话题“杭州飙车案”的内容焦点的这两部分信息.

从表 3 可以看到，对网络话题“杭州飙车案”，运用本文提出的网络话题内容焦点的识别方法，可识别出 6 个阶段性关注焦点，分别为“男子飙车撞死路人”、“车速 70 码成网络新名词”、“悼念死者谭卓”、“签订赔偿 113 万元协议”、“胡斌一审获刑 3 年”以及“胡斌替身出庭”. 可见，该结果与人工标注的结果(见图 1)在第 3 个焦点上出现分歧，但总体上结果吻合. 因此，本文提出的网络话题内容焦点的识别和分析方法是有效的.

表 3　话题“杭州飙车案”的内容焦点对应的关键词、摘要与人工标注对比表

焦点	关键词 *terms*	摘要 *abstract*	人工标注对焦点描述
1	0.024081 跑车 0.020961 撞 0.020872 改装 0.018992 上 0.018825 网友 0.016918 三菱 0.016092 路 0.015767 车 0.015309 辆 0.015086 事故	2009-05-09 男子驾跑车飙车撞死路人续：肇事者遭刑拘 2009-05-07 年轻男子驾三菱跑车飙车撞死路人	男子飙车撞死路人
2	0.126052 车 0.123741 鉴定 0.084944 人 0.084268 70 0.082059 杭州 0.081663 谭卓 0.078944 不 0.078105 警方 0.077937 上 0.077697 码	2009-05-13 网友质疑杭州飙车案 70 码车速发明新名词欺实马 2009-05-14 分析称杭州青年飙车撞人案演变成公共事件 2009-05-14 杭州飙车案警方所称 70 码成最热网络新名词	车速 70 码成网络新名词
3	0.032009 谭卓 0.031051 不 0.030842 车 0.030336 胡斌 0.028634 交通 0.021932 赔偿 0.021161 人 0.020626 飙 0.019932 70 0.019286 公共	2009-05-15 杭州数百市民祭奠飙车案死者谭卓 2009-05-15 杭州飙车案肇事者罪名最终由法院决定 2009-05-16 盛翔：胡斌犯交通肇事罪则谁危害公共安全	悼念死者谭卓
4	0.103439 赔偿 0.051045 元 0.046508 律师 0.042741 胡斌 0.040709 鉴定 0.038283 父母 0.036824 家属 0.035612 113 万 0.034038 交通 0.034010 谭卓	2009-05-20 杭州飙车案肇事方家属提出赔偿 113 万获接受 2009-05-20 杭州飙车案肇事方赔偿受害者父母 113 万	签订赔偿 113 万元协议
5	0.088465 胡斌 0.038932 不 0.036561 被告人 0.035196 车 0.035157 人 0.034731 交通 0.027526 安全 0.027032 后 0.026999 谭卓 0.026359 赔偿	2009-07-20 杭州飙车案被告胡斌一审获刑 3 年 2009-07-20 杭州飙车案被告一审获刑 3 年 2009-07-20 杭州飙车案被告被判 3 年双方亲属均认为不公平	胡斌一审获刑 3 年
6	0.132468 胡斌 0.058990 替身 0.057816 法院 0.036716 赔偿 0.036130 被告人 0.033921 不 0.031762 谭卓 0.029306 西湖区 0.029043 照片 0.028706 元	2009-07-27 杭州法院否认飚车案被告以替身出庭 2009-07-28 盛大林：请杭州司法机关回避胡斌替身门调查 2009-07-27 李克杰：司法机关不能对胡斌替身说置若罔闻	胡斌替身出庭

由于本文所研究的话题焦点迁移识别问题是随着网络舆情发展而产生的新问题，尚未发现有其他的焦点迁移识别算法，难以进行与类似算法的实验对比. 由于较难量化说明关键词和摘要提取的准确率，我们分析了本文所提方法在 114 组热点话题数据上识别的焦点数量结果，如图 6 所示，对于绝大多数的热点话题，本文所提方法均能正确识别出焦点数量.

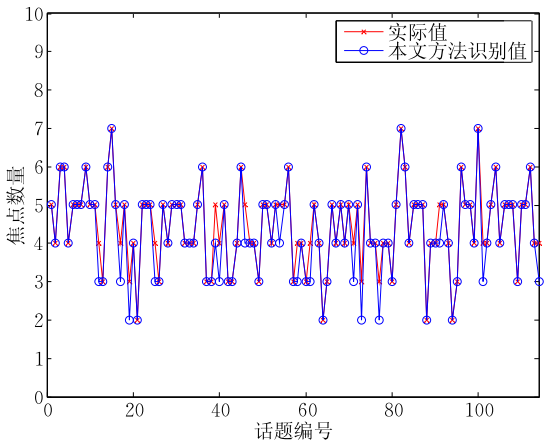

图 6　114 组热点话题中的焦点数量识别结果

4.3　方法中经验参数对实验结果的影响分析

本文所提方法中主要有 3 项经验参数，分别是 μ、σ 和 λ，其中参数 μ 的作用是约束话题焦点的持续时间（单位为日），其取值对特征词提取的性能有较大影响，因此为了进一步分析经验参数的选取对焦点识别方法结果的影响，我们选取了不同的参数 μ 取值，分析其对特征词提取的影响. 如表 4 所示，参数 μ 的取值对特征词提取的影响较为明显，考虑到大部分情况下话题的单个焦点持续时间为 2～3 天，当选择 $\mu=2.5$ 时，特征提取的性能最好. 当参数 μ 过小时，如取 0.2 的时候，无关噪声词就会出现较高的权值；当参数 μ 过大时，取到 5 天的时候，干扰噪声词会得到较高的权值.

表 4　114 组网络话题中参数 μ 对特征提取影响统计结果

	指标		
	平均召回率/%	平均准确率/%	平均 F 值
$\mu=0.2$	12.7	6.1	0.08
$\mu=1$	48.9	23.9	0.32
$\mu=2.5$	69.2	33.8	0.45
$\mu=5$	43.0	20.1	0.28

5　结　论

在互联网的信息传播过程中，随着话题事件的演变关于话题的流文本不断更新发布，而其内容焦点也随着话题发展发生着迁移，因此对网络话题动

态演变特性的分析和理解已成为目前国内外研究的一个热点问题.但由于自然语言问题的复杂,以及目前自然语言处理技术发展的相对不足,为解决这一问题提出了很大挑战.本文在总结和分析前人对网络话题动态特性分析和文本挖掘研究的基础上,通过大量实际数据的分析,发现话题焦点的特征词具有频度较高、分布持续、持续时间适中和同焦点的特征词时序分布重合等特点,并且提出了解决这一问题的主要步骤和相关算法,包括基于时间属性的焦点特征词提取算法和内容焦点特征词的合并方法等.基于实际数据的实验结果验证了所提的特征提取方法相比于其他方法更适合于解决话题焦点特征词提取问题,并且验证了所提的焦点识别方法可有效识别网络话题的内容焦点迁移情况.

参 考 文 献

[1] Allan J. Introduction to topic detection and tracking//Allan J ed. Topic Detection and Tracking. New York, USA: Springer US, 2002: 1-16

[2] Fiscus J G, Doddington G R. Topic detection and tracking evaluation overview//Allan J ed. Topic Detection and Tracking. New York, USA: Springer US, 2002: 17-31

[3] Kaur K, Gupta V. A survey of topic tracking techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 2012, 5(2): 383-392

[4] Mohd M, Crestani F, Ruthven I. Evaluation of an interactive topic detection and tracking interface. Journal of Information Science, 2012, 38(4): 383-398

[5] Pons-Porrata A, Berlanga-Llavori R, Ruiz-Shulcloper J. Topic discovery based on text mining techniques. Information Processing & Management, 2007, 43(3): 752-768

[6] Liu Jin-Ling, Wang Xin-Gong, Zhou Hong. Hot events recognition based on mobile phone short message information flow. Computer Applications and Software, 2012, 29(10): 200-204(in Chinese)
(刘金岭,王新功,周泓. 基于手机短信信息流的热点事件识别. 计算机应用与软件, 2012, 29(10): 200-204)

[7] Chen Zhi-Min, Meng Zu-Qiang, Lin Qi-Feng. Chinese story link detection based on extraction of elements correlative word. Journal of Computer Application, 2013, 33(1): 182-185(in Chinese)
(陈智敏,蒙祖强,林啟锋. 基于要素提取关联词对的中文报道关系检测. 计算机应用, 2013, 33(1): 182-185)

[8] Benevenuto F, Rodrigues T, Veloso A, et al. Practical detection of spammers and content promoters in online video sharing systems. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42(3): 688-701

[9] Aggarwal C C, Zhai C X. A survey of text classification algorithms//Aggarwal C C, Zhai C X eds. Mining Text Data. New York, US: Springer US, 2012: 163-222

[10] Benhardus J, Kalita J. Streaming trend detection in twitter. International Journal of Web Based Communities, 2013, 9(1): 122-139

[11] Yu Miao, Li Yuan. Key technology and systematic framework of Internet public sentiment. Netinfo Security, 2011, 11(1): 21-22(in Chinese)
(于森,李远. 网络舆情的关键技术与系统构架研究. 信息网络安全, 2011, 11(1): 21-22)

[12] Gohr A, Spiliopoulou M, Hinneburg A. visually summarizing semantic evolution in document streams with topic table// Fred Ana ed. Knowledge Discovery, Knowledge Engineering and Knowledge Management. Heidelberg, Germany: Springer Berlin Heidelberg, 2013: 136-150

[13] Mimno D, Blei D. Bayesian checking for topic models// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA, 2011: 227-237

[14] Xing E P, Yan R, Hauptmann A G. Mining associated text and images with dual-wing harmoniums//Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence. Edinburgh, UK, 2005: 633-641

[15] Silvestri F. Mining query logs: Turning search usage data into knowledge. Foundations and Trends in Information Retrieval, 2010, 4(1-2): 1-174

[16] Romero D M, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter// Proceedings of the 20th International Conference on World Wide Web. Bangalore, India, 2011: 695-704

[17] Leskovec J, Backstrom L, Kleinberg J. Meme-tracking and the dynamics of the news cycle//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 497-506

[18] Zimmermann M, Ntoutsi I, Siddiqui Z F, et al. Discovering global and local bursts in a stream of news//Proceedings of the 27th Annual ACM Symposium on Applied Computing. Trento, Italy, 2012: 807-812

[19] Piech C, Sahami M, Koller D, et al. Modeling how students learn to program//Proceedings of the 43rd ACM Technical Symposium on Computer Science Education. Raleigh, USA, 2012: 153-160

[20] Low Y, Bickson D, Gonzalez J, et al. Distributed graphLab: A framework for machine learning and data mining in the cloud. Proceedings of the VLDB Endowment, 2012, 5(8): 716-727

[21] Sen M K, Stoffa P L. Global Optimization Methods in Geophysical Inversion. Cambridge, UK: Cambridge University Press, 2013

[22] Aggarwal C C. Mining sensor data streams//Aggarwal C C ed. Managing and Mining Sensor Data. New York, USA: Springer US, 2013: 143-171

[23] Aggarwal C C, Philip S Y. On clustering massive text and categorical data streams. Knowledge and Information Systems, 2010, 24(2): 171-196

[24] Abdulsalam H, Skillicorn D B, Martin P. Classification using streaming random forests. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(1): 22-36

[25] Walker D D, Ringger E K. Model-based document clustering with a collapsed gibbs sampler//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 704-712

[26] Hofmann T. Probabilistic latent semantic analysis//Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Stockholm, Sweden, 1999: 289-296

[27] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis. Discourse Processes, 1998, 25(2-3): 259-284

[28] Wang S, Li D, Song X, et al. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Systems with Applications, 2011, 38(7): 8696-8702

[29] Liu Zitao, Yu Wenchao, Chen Wei, et al. Short text feature selection for micro-blog mining//Proceedings of the 2010 International Conference on Computational Intelligence and Software Engineering. Wuhan, China, 2010: 1-4

ZHOU Ya-Dong, born in 1982, Ph.D., assistant professor. His research interests include online social networks and web mining.

LIU Xiao-Ming, born in 1989, Ph. D. candidate. His research interest is social networks.

DU You-Tian, born in 1980, Ph. D., associate professor. His research interest covers multimedia understanding on Internet, machine learning and social network.

GUAN Xiao-Hong, born in 1955, Ph. D., professor. His research interests include allocation and scheduling of complex networked resources, network security, and sensor networks.

LIU Ji, born in 1986, M. S. Her research interest is social networks.

Background

With the rapid development of information technology, the network has become one of the most important platform for people to get information. In our daily life, people tend to encounter this phenomenon: the news of certain topic is constantly updated, but the reports focus on different contents, these different focus include: the focus generated by the development of events; new focus caused by the increased user reviews of topics; focus migration due to the impact of other hot news topic in the same period.

In this paper, we propose a method for analyzing and identifying the evolutionary focus of topics. The method is consist of three parts, including feature selection based on time attribute, feature combine model and focus presentation. The experimental results show that this method could identify the evolutionary focus of topics effectively.

This work is part of the "Modeling and Analysis on Information Spreading over Online Social Networks", which is supported in part by the National Natural Science Foundation of China (61202392, 61221063, 61375040, 60905018, 61103240, 61172124), Specialized Research Fund for the Doctoral Program of Higher Education (20120201120023), National High Technology Research and Development Program (863 Program) of China (2012AA011003), the Ministry of Education Innovation Research Team (IRT13035) and the Fundamental Research Funds for the Central University.