微博热词抽取及话题发现研究*

郝晓玲[1] 茅嘉惠[1] 于秀艳[2]

(1. 上海财经大学信息管理与工程学院 上海 200433;
2. 山东理工大学商学院 淄博 255049)

摘 要 旨在实践一种方法,能从大量散乱的微博语料中快速抽取热点词语并从中形成话题。首先参考文本串的词频、内部聚合度、粘联度这三个统计特征判断文本串是否成词,直接从样本语料中抽取出高频词。然后根据这些高频词在不同时间窗口的出现频率筛选出实时热词,最后利用词共现确定热词间的关联度,将热词聚类形成热点话题。实验证明,该算法简单易行,在话题发现中取得了良好的效果。

关键词 微博 微博热词 话题发现 词共现

中图分类号 G206 **文献标识码** A **文章编号** 1002-1965(2015)06-0109-05

DOI 10.3969/j.issn.1002-1965.2015.06.020

Micro-blogging Hot Words Extraction and Topic Detection

Hao Xiaoling[1] Mao Jiahui[1] Yu Xiuyan[2]

(1. School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433;
2. Business School, Shandong University of Technology, Zibo 2550049)

Abstract This paper aims to extract valuable information from massive fragmented content and feed back to the user in a concise form. Firstly, considering three statistical characteristics: word frequency of text string, internal degree of coupling, the external degree of flexibility, we extract high-frequency words from micro-blog corpus, then filter outreal-time hot words according to the frequency of occurrence of these high-frequency words in different time windows, and finally use the word co-occurrence to determine the hot words correlation to get a hot topic. Experimental results show that the algorithm is simple and available, and achieved good results on the topic detection.

Key words micro-blogging micro-blogging hot word topic detection word co-occurrence

0 引 言

微博作为重要的互联网应用,允许用户将自己的最新动态和想法以短信形式发送给手机和个性化网站群,随时随地和世界分享博主所见所想。由于微博140字的字数限制大大降低了信息发布的门槛,原创内容海量生成,每天人们在新浪微博上发布上亿条博文。为及时了解微博圈的动态,有必要对海量的微博信息进行组织和分析,提出有效的算法从这些信息中对话题进行提取,并以简洁的形式提供给用户。因此,如何实现从这些零散多样的内容中提取有价值的主题信息也是重要的研究课题。

1 相关研究综述

目前,微博数据挖掘方面的研究[1-3]主要可分为两大类:微博内容挖掘和用户关系挖掘,热点话题发现属于微博内容挖掘范畴,是指从大量的微博文本中检测出微博用户广泛讨论的话题。涉及两个关键性技

收稿日期:2015-01-14 修回日期:2015-05-12

基金项目:国家自然科学基金项目"面向海量数据语义标注众包的任务管理方法研究"(编号:71401096);教育部人文社会科学基金资助项目"面向用户兴趣基于本体的网络舆情研判体系研究—以论坛为例"(编号:10YJC860010);山东省高校人文社会科学研究计划项目"云计算可持续发展的关键影响因素及对策研究"(编号:J13WG16)。

作者简介:郝晓玲(1975-),女,博士,副教授,研究方向:网络舆情、IT治理、电子商务;茅嘉惠(1991-),女,硕士研究生,研究方向:文本挖掘;于秀艳(1974-),女,博士,副教授,研究方向:IT治理、云计算、数据挖掘。

术:中文分词技术,中文话题发现技术。

中文分词算法主要分为两种:一是基于语言规则的方法,即计算机可以通过自然语言的语法、词性等内部规则分析出文本正确含义并分词,判断文本串是否成词主要依赖词库。主要方法包括:基于统计过滤和规则[4];基于支持向量机[5]与约束条件选取新词。该类方法准确率更高,但对已知词库的依赖性强。二是基于统计学习的方法,利用计算词元在文本中的各种统计特征值直接判断出成词的文本串。该类方法比较灵活、适应能力强,具体可供参照的统计特征很多,如字与字之间的信息熵,字符串的最大组合概率,利用互信息和 t 测试差相结合,字符串出现的频率和置信度,利用词的前后缀信息提取高频词等等[6-9]。对于微博而言,需要在新词不断涌现的情境下发现新的话题。传统方法是,先对文本进行分词,然后猜测未能成功匹配的剩余片段就是新词。其局限性在于:分词的准确性本身就依赖于词库的完整性,如果词库中根本没有新词,分词结果会受到很大影响。顾森[10]提出运用内部凝固程度和自由运用程度来进行新词识别。贺敏[11]判断上下文邻接种类,首尾单字位置成词概率以及双字耦合度等语言特征,分别过滤得到新词。钟将[12]使用互信息和信息熵这两个信息度量反映词语之间的联合度,再创建新的评价函数将两个度量结合起来。

话题发现就是指从大量的微博文本中检测出微博用户广泛讨论的话题。一是基于文本聚类的方法。例如,利用向量空间模型和主题模型等将相同话题的微博聚类后,只将相关性强的文本聚集到一起,再设计算法提取出可以展现话题的主题词[13,14]。黄波[15]用主题模型弥补了传统文本向量化方法的不足,利用 LDA 模型提取出文档间的语义信息,分别对文档集进行 LDA 建模和 VSM 建模,实现文本间相似度的计算,采用 Single-pass 算法和层次聚类的混合聚类方法对文本做话题聚类。也有研究直接基于主题词进行话题发现,计算量较小,效率更高[16]。

2 微博客话题发现模型

基于该领域已有的研究成果,本文提出了针对微博话题发现的综合模型,如图 1 所示。

图 1 话题发现流程

本文参照已有算法[10-12]进行了改进,采用不依赖于知识库的分析方法,对一定规模的语料进行计算,根据词频和信息熵的高低提取出语料中的常见词语。并从以下三方面判断一个文本片段是否能够独立成词:文本片段出现的频数、文本内部聚合度、粘联度。

2.1 文本片段的出现频数 如果一个文本片段在语料中多次出现,那么它有可能是一个词,反之,只是偶然出现的字词组合很难认定为独立的词。本文目的是检测热点话题,出现频率很少的词不太可能为实时热点,可以忽略,同时也可以快速排除大量候选词,加快算法速度。因此本文规定一个文本片段的出现频数应超过某个阈值,否则不作为候选词。

2.2 文本片段的内部聚合度 构成词的字之间必然存在一定相关性,而不仅仅是几个字的随机组合。假设长度为 n 的文本片段 X 由字 $x_1x_2x_3\cdots x_n$ 组成,$\mathrm{Count}(X)$ 表示 X 在训练语料中出现的次数。我们将文本片段 X 看作字符串 X_1 与 X_2 的组合,则 $P(X)=P(X_1)P(X_2\mid X_1)$。对于长度为 n 的文本片段 X 有 $n-1$ 种可能的分割方式。根据最大似然估计的估算公式,$P(X_2\mid X_1)\approx \mathrm{Count}(X_1X_2)/\mathrm{Count}(X_1)$。根据已有研究结论,用互信息度量字串内部紧密性的效果最佳。已知文本片段 X 看作文本片段 X_1 与 X_2 的组合,这个事件的互信息为后验概率与先验概率比值的对数:

$$MI=\log\frac{p(x)}{p(X_1)p(X_2)} \tag{1}$$

为减少计算量,仅取上式中的真数部分作为字串内部聚合度的度量。按 X_1、X_2 所有组合分别计算出这个比值,取其中的最小结果作为文本片段 X 的内部聚合度。之后对其设定阈值,达不到阈值的文本片段不作为候选词。在实际计算时,由于客观条件限制,无法使用大规模训练语料来估计参数,因此使用 X 在样本语料中出现的次数代替 X 在训练语料中出现的次数。实验发现取此近似值不会对抽词效果产生重大影响,但可以极大地简化算法。如果 X 为二字词,简化后的公式为:

$$P(X)=\frac{\mathrm{Count}(X)}{\mathrm{Count}(X_1)\mathrm{Count}(X_2)}\times \mathrm{Length} \tag{2}$$

其中,$P(X)$ 表示文本片段 X 的内部聚合度,Length 表示整个样本语料的长度。$\mathrm{Count}(X)$ 表示 X 在训练语料中出现的次数。

使用 2012 年 9 月的部分微博文本作为实验语料,首先对文本片段的出现频数规定了阈值,出现次数小于 20 的文本片段已经排除。然后计算文本片段的内部聚合度,如表 1 所示。左边是最终抽词结果中内部聚合度较高的 5 个词,可以观察到这些字的搭配相对固定,成词方法有限,因此聚合程度很高。右边列出频数超过 20 次,但内部聚合度较小的文本片段。这些文本片段大都是很常见的字词搭配,它们只是偶然组合的可能性更大,不认为可以独立成词。

另外,程序也抽出了一些地名(如爱尔兰、阿根

廷)和人名(如弗格森、穆里尼奥),还有一些专有名词,这些词的内部聚合度也很高,可见该算法在抽取未登录词时体现出很大优势。

表 1　文本片段内部聚合度的部分实例

文本片段	出现频数	内部聚合度	文本片段	出现频数	内部聚合度
玫瑰	21	38016	最大	189	13.92638
垃圾	23	34710.26	元的	106	1.926008
猥亵	24	30705.23	最多	45	7.174996
矛盾	23	21937.55	最为	37	3.724701
玻璃	30	20880.63	又是	33	9.042543

2.3　文本片段的粘联度　本文使用信息熵来量化文本片段的粘联度。在信息论中,熵被用来衡量一个随机变量出现的期望值。信息熵的计算公式为:

$$H(x) = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i) \tag{3}$$

其中 $p(x_i)$ 表示事件 x_i 发生的概率,$\{x_1, x_2, x_3, \cdots, x_i\}$ 为 x 的集合。b 是对数所使用的底数,通常取2、10 或自然常数 e。这里选择 e 作为底数。信息熵直观反应一个离散事件有多随机,随机性越大信息熵就越大。假定 x 是文本片段左邻字的集合,在语料中该文本片段共出现 n 个不同的左邻字,分别计算这 n 种情况出现的概率并代入公式,就能得到 x 的信息熵,熵越大表示左邻字出现越随机,也说明该文本片段灵活度更高,更可能是一个词。为信息熵设定阈值,将超过阈值的文本片段加入候选词集合。利用微博语料进行分析,可见信息熵最高的文本串是诸如"已经""还是""没有"等词,这些词频繁出现,使用灵活,符合人们的直观感受,如表 2 所示。

表 2　文本片段的左右信息熵实例(一)

序号	文本片段	出现频数	内部聚合度	左信息熵	右信息熵
1	已经	490	125.6774	4.812625	4.67136
2	自己	780	344.0368	4.757555	3.885873
3	还是	270	24.64327	4.636952	4.707418
4	表示	667	334.2687	4.538188	4.00283
5	进行	499	109.0473	4.529864	4.125404

表 3 列出一些信息熵较特殊的文本片段。"比如"几乎只出现在句首,因此没有左邻字;"编辑报道"往往在句末,因此没有右邻字;"微直播"是微博特有词汇,以"#微直播#"的形式单独出现,既没有左邻字也没有右邻字。有些文本片段其实只是半个词,例如第 4 行的"阿里巴",它的左信息熵为 0,因为它的左邻字集合只有一个元素{"巴"},正确的词应该是"阿里巴巴";同样第 5 行的"级地震"也不是一个词,由于非汉字字符会被作为分隔号处理,导致它的左信息熵为 0。虽然这些词内部聚合度很高,但通过计算信息熵可以将其排除。在本文中规定左右熵中有一个为零的文本串就不作为候选词,但左右熵都为零的文本串予以保留。

表 3　文本片段的左右信息熵实例(二)

序号	文本片段	出现频数	内部聚合度	左信息熵	右信息熵
1	比如	38	24.31357	0	3.176083
2	编辑报道	42	520.4276	1.063514066	0
3	微直播	48	257.5554	0	0
4	阿里巴	27	1036.8	1.277034259	0
5	级地震	79	159.7639	0	2.3719

3　微博主题词抽取

3.1　主题词抽取规则　基于前文的算法,本文计算上述三个变量,并为其设置合适的阈值,挑选出语料中可以成词的文本片段,再将其运用于微博关键词检测。本文尝试把微博内容按时间维度区分,探测相邻时间段内出现频率激增的词语,判断出该时间段内可以作为主题词的即时热词。

3.1.1　词的相对出现频率。通过算法抽取出语料中的高频词语,但高频词中往往包含许多无效词语,如"这种""一个""我们"这类词语虽然使用频繁,但并不能代表样本语料的特征。新的话题的形成应该有一定时效性,也就是说主题词在某个时间窗口内集中大量出现,而在之前的时间窗口内不常出现。本文旨在判别当日主题,故把时间窗口的单位长度设定为一天,将每个词的当日相对出现频率定义为:

$$G_i(w) = \frac{\mathrm{Count}(w, T_i)/\mathrm{Length}(T_i)}{\mathrm{Count}(w, , T_1, T_2, \cdots, T_{i-1})/\mathrm{Length}(T_1, T_2, \cdots, T_i)} \tag{4}$$

其中,T_i 表示时间窗口 i,$\mathrm{Count}(w, T_i)$ 表示词 w 在时间窗口 i 中出现的频数,$\mathrm{Length}(T_i)$ 表示时间窗口 i 中语料的长度。在实际计算时,由于只需要将同一时间窗口内的词做横向比较,而同一时间窗口的 $\mathrm{Length}(T_i)$ 都相等,因此将公式简化为:

$$G_i(w) = \frac{\mathrm{Count}(w, T_i)}{\mathrm{Count}(w, , T_1, T_2, \cdots, T_{i-1})} \tag{5}$$

类似"这种""一个""我们"的常用词在每个时间窗口内出现的频率几乎相同,一般不会得到很高的分数;而更能代表当日热点的词应该仅在当日频繁出现,这些词将得到较高的分数。

3.1.2　运用贝叶斯平均作平滑处理。假设词 W_1 在本月仅在当日出现 100 次,词 W_2 在本月仅在当日出现 1 次,那么 W_1、W_2 的得分都等于 1,但是显然只有 W_1 可能是热词,W_2 只是偶然出现一次。因此还需对得分进行平滑,在这里贝叶斯平均的作用就是弱化样本量过小对最终得分的影响。如公式 6 所示。

$$WR = \frac{v}{v+m}R + \frac{m}{V+m}C \tag{6}$$

其中 WR 为平滑后的加权得分，R 平滑前的分数，v 表示某词在整个事件段出现的次数，m 为所有词在整个时间段内的平均出现次数，C 为所有词的平均得分。长期来看，$v/(v+m)$ 这部分的权重将越来越大，得分将逐渐接近真实情况。

3.2 **基于词共现的主题词聚类** 词共现分析是自然语言处理技术中挖掘词与词关联的重要方法，核心思想是词与词之间的共现频率在一定程度上反映词与词之间的语义关联。在大规模的文本语料中，如果两个词频繁地同时出现于同一时间窗口（可能是一篇文档、段落或一句话）内，则这两个词是语义关联的，且两个词共同出现的频率越高，这两个词的关联度越高。在微博文本中，某个话题往往包含多个主题词，词的共现率即指两个词在一条微博内容中同时出现的概率。通过词共现分析，可将代表相同主题的不同关键词关联起来。共现率的计算公式为：

$$P(w_1,w_2)=\frac{\mathrm{Count}(w_1,w_2)}{n(S)} \tag{7}$$

其中 $\mathrm{Count}(w_1,w_2)$ 表示在整个语料 S 中同时包括词 w_1 和词 w_2 的微博数量，$n(S)$ 表示语料 S 中总共包含的微博数量。共现词对的抽取类似于形如 $X\rightarrow Y$ 的蕴涵式，借鉴韩家炜[17] 等观点，本文中的关联规则定义为：给定一个数据集 S，关联规则在 S 中的支持度（Support）是 D 中事务同时包含 w_1、w_2 的百分比，即 w_1、w_2 在所有事务中同时出现的概率；置信度（Confidence）是包含 w_1 的事务中同时又包含 w_2 的百分比、包括 w_2 的事务中同时又包括 w_1 的百分比的平均值。如果同时满足最小支持度阈值和最小置信度阈值，则认为 w_1、w_2 是共现词。结合已有的词共现率公式，得到关于两个词之间共现率的支持度与置信度的计算公式：

$$\mathrm{support}(w_1,w_2)=P(w_1,w_2) \tag{8}$$

$$\mathrm{confidence}(w_1,w_2)=\left(\frac{P(w_1,w_2)}{P(w_1)}+\frac{P(w_1,w_2)}{P(w_2)}\right)*\frac{1}{2} \tag{9}$$

9 月 11 日微博语料中部分共现词对的计算结果如表 4 所示。

表 4 部分共现词对支持度与置信度的计算结果

共现词对 (w_1,w_2)		support (w_1,w_2)	confidence (w_1,w_2)
社保	缴费率	0.857571214	0.007080785
国足	惨败	0.708748616	0.006437078
恐怖	袭击	0.684981685	0.007080785
海域	海监船	0.654856255	0.01577084
穆雷	大满贯	0.603835979	0.007080785
日本政府	钓鱼岛	0.622832719	0.055037013

多次试验之后，认为置信度的阈值设为 0.005 是较为合理的，支持度的阈值设定为 0.5，且支持度的阈值需要依据 n(S)（即语料 S 中总共包含的微博数量）相应调整。相比支持度（即共现率），置信度可以更好地表示两个词之间的关联程度。对每个词 w，找到所有能与其关联、符合阈值的共现词，为每个词建立形如 $\mathrm{Conf}(w)=\{w_1,w_2,\cdots,w_m\}$ 的共现词集合，集合中的词可能属于同一个话题。如果一个词与越多的词相关联、且关联程度越大，则认为该词所含的信息量更大，对话题的表达更有意义。用下式来计算每个词对所属话题的贡献程度：

$$G(w_i)=\sum_{w_j\in \mathrm{Conf}(w_i)}\mathrm{Confidence}(w_i,w_j) \tag{10}$$

假设用 k 个主题词表示一个话题，则提取对该话题贡献程度最大的词（不妨称为关键主题词）及该词与其关联程度最大的 $k-1$ 个词作为话题的表示。具体过程为：

a. 计算所有主题词对所属话题的贡献程度 $G(w)$；b. 选出贡献程度最高的主题词，与该词关联度最高的 k 个主题词构成一个话题；c. 被使用过的 k 个主题词不再作为其他话题的关键主题词，从主题词列表中删除；d. 重复第 b 步，直到所有主题词都被归类。

4 实验设计

4.1 **数据来源** 本文使用的微博语料来自数据堂网站（http://www.datatang.com/），实验语料为新浪微博“名人堂风云影响力榜单——媒体影响力榜”在 2012 年 9 月发布的微博信息集（每位用户采集平均条数为 98.7）。榜单上的大都为媒体用户，因此新闻类微博占多数，最终用于实验的共有 9445 条微博文本信息。

4.2 **实验步骤** 首先进行文本预处理，将每条微博文本按非汉字符号分隔，得到若干文本片段。枚举文本片段中的所有词组合方式，取最大词长为 4，对一个长度为 n 的文本片段，至多可以提取出 $n+(n-1)+(n-2)+(n-3)=4n-6$ 个不重复的文本串。

4.2.1 抽取高频词 对 2012 年 9 月的微博语料抽取高频词，阈值设定为词频大于 20，内部聚合度大于 20，信息熵大于 1，共抽取出 2588 个词。由于所用语料为媒体发布的微博信息，可以看到这类文本在用词上类似于新闻用语。然后分别从 9 月 9 日、10 日、11 日的微博语料中抽取高频词，抽取结果如表 5 所示。

实验语料中每一天收录的微博条数不同，因此只能纵向观察当日词频。在 9 月 10 日和 11 日出现频率最高的三个词都是“中国”“日本”和“钓鱼岛”，可见这是该时间段内较受关注的话题。也有许多高频词，如“一个”“我们”“没有”等词并没有实际意义，需要在接下来的步骤中将其剔除。

表 5　2012 年 9 月 9-11 日微博语料高频词抽取的部分结果

序号	9/9	当日频数	9/10	当日频数	9/11	当日频数
1	中国	261	中国	749	中国	1398
2	地震	152	教师	455	日本	1017
3	一个	120	日本	423	钓鱼岛	923
4	我们	119	钓鱼岛	356	政府	577
5	学生	115	政府	341	报道	328
6	自己	114	老师	314	表示	318
7	银行	114	一个	306	日本政府	318
8	教师	104	我们	281	人民	307
9	上海	98	人民	254	问题	288
10	没有	95	学生	225	经济	281

4.2.2　主题词筛选　分别从 9 月 9 日、10 日、11 日的微博语料中抽取高频词,计算词的相对出现频率,并用贝叶斯平均进行平滑处理后的微博热词,如表 8 所示。

表 8　2012 年 9 月 9-11 日微博语料主题词抽取的部分结果

9 月 9 日	调整分数	当日频数	9 月 10 日	调整分数	当日频数	9 月 11 日	调整分数	当日频数
同比	0.767527	62	上调	0.856816	130	预报	0.852186	216
瑞典	0.728505	44	领土	0.842736	119	达沃斯	0.814674	166
教师	0.725773	104	抚养费	0.836859	82	海域	0.792789	171
机器	0.721564	47	日本政府	0.810552	141	社保	0.786690	101
客服	0.718874	41	汽柴油	0.805347	61	海监船	0.767356	87
上涨	0.710692	73	教师节	0.790958	220	巴西	0.749672	127
教师节	0.710472	41	生二胎	0.789887	53	国足	0.738372	84
光明	0.699958	45	中国政府	0.788979	75	穆雷	0.734297	112
南京	0.694113	75	社会抚养费	0.787780	52	缴费率	0.734259	65
银行	0.687645	114	孙子	0.787780	52	缴费	0.707218	106

4.2.3　基于词共现的主题词聚类　先对主题词做了一些筛选,因为意义重复的主题词会干扰最后的话题结果,如“光明”与“光明乳业”在语料中代表相同的意义,故取含义更详细的“光明乳业”,删除“光明”一词。然后对前 50 个主题词计算词共现率,最终从三天的微博语料中发现的话题如表 9 所示。

表 9　2012 年 9 月 9 日热点话题

编号	9 月 9 日 热点话题表示	9 月 10 日 热点话题表示	9 月 11 日 热点话题表示
1	机器,丁先生,客服,赶到,银行,解释	生二胎,双独夫妻,未经批准,抚养费,计生	预报,周边,海洋,天气预报,海域,钓鱼岛
2	鳄鱼,水面,南京,工作人员	主权,领土,中国政府,钓鱼岛,非法,购岛	合同,日本政府,钓鱼岛,自卫队,外务省,大使
3	价格,同比,上涨,食品,居民,银行	国有化,日本政府,钓鱼岛,购岛,确定	穆雷,大满贯,惨败,国足,巴西
4	光明乳业,变质,上海,致电	上调,油价,汽柴油,零售价	恐怖,袭击,航班,国防部
5	教师节,送礼,上海	小威,阿扎,新浪网球	捅死,性侵,杨某,旋某琦

结合当日的新闻资料看热点话题的检测结果基本正确。如 9 月 9 日国家统计局公布 8 月 CPI,各项价格同比上涨;光明乳业被曝质量问题;南京丁先生被 ATM 机吞一万元,谎称多吐钱客服五分钟赶来。9 月 10 日是教师节,同时这天汽柴油价上调,中国及日本政府就钓鱼岛问题发表声明。9 月 11 日中央气象台开始把钓鱼岛及周边海域的天气预报纳入到国内城市预报,国足挑战巴西队惨败等。在 9 月 10 日与 11 日都提取出两条与钓鱼岛相关的话题,实际上它们是属于同一个主题之下,但是从新闻角度看出发点并不同,如“国有化,日本政府,钓鱼岛,购岛,确定”是指日本单方面确定购买钓鱼岛,“主权,领土,中国政府,钓鱼岛,非法,购岛”是指中国政府对钓鱼岛宣告领土主权,从这个角度认为这样划分话题也是可行的。

5　结　论

本文从文本片段出现的频数、文本内部聚合度、粘联度三个方面提出了高频词抽取算法,并用贝叶斯平均平滑弱化了样本量大小对结果的影响,然后采用基于词共现的主题词聚类方法,挖掘出热点的话题。利用微博的真实语料进行了实验,结果表明,这种抽取方法计算复杂度低,话题发现效果较好。本文研究局限性在于:未考虑微博蕴含着的结构化关系,如用户间的关注与被关注,微博内容的转发、评论等,未来考虑通过权值分配将这些因素融合到热度计算中。此外,本文使用的实验语料是媒体用户的微博信息集,实际上等于进行了一次文本聚类,语料本身的信息密度较大,因此实验结果也比较好。对于文本稀疏性大,信息密度小的文本还需进一步对算法速度和效率进行改进。

参 考 文 献

[1]　丁晟春,孟　美,任李霄. 面向中文微博的观点句识别研究[J]. 情报学报,2014,33(2),175-182.

[2]　孙胜平. 中文微博客热点话题检测与跟踪技术研究[D]. 北京交通大学,2011.

[3]　蒋盛益,麦智凯,庞观松,等. 微博信息挖掘技术研究综述[J]. 图书情报工作,2012, 56(17):136-142.

[4]　黄　轩,李熔烽. 博客语料的新词发现方法[J]. 现代电子技术,2013,36(2):144-149.

[5]　徐远方,李成城. 基于支持向量机和约束条件的新词识别研究[J]. 计算机技术与发展,2014,24(1):98-101.

[6]　任　禾,曾隽芳. 一种基于信息熵的中文高频词抽取算法[J]. 中文信息学报,2006(5):40-43.

[7]　罗盛芬,孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报,2003(3):9-14.

[8]　丁溪源. 基于大规模语料的中文新词抽取算法的设计与实现[D]. 南京理工大学,2011.

[9]　赵　洁,温　润. 基于新词扩充和特征选择的微博观点句识别方法[J]. 情报学报,2013,32(9):945-951.

(下转第 157 页)

(上接第113页)

[10] 顾 森. 基于大规模语料的新词发现算法[J]. 程序员,2012(7):54-57.

[11] 贺 敏,龚才春,张华平,等. 一种基于大规模语料的新词识别方法[J]. 计算机工程与应用,2007,43(21)157-159.

[12] 钟 将,耿升华,董高峰. 一种新词检测方法研究[J]. 数字通信,2013,40(2):1-5.

[13] 路 荣,项 亮,刘明荣,杨青. 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(03):382-387.

[14] 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.

[15] 黄 波. 基于向量空间模型和LDA模型相结合的微博客话题发现算法研究[D]. 西南交通大学,2012.

[16] 李恒训,张华平,秦 鹏,等. 基于主题词的网络热点话题发现[C]. 第五届全国信息检索学术会议论文集,2009.

[17] 韩家炜,孟小峰,王 静,等. Web挖掘研究[J]. 计算机研究与发展,2001(4):405-414.

(责编:王平军)