基于主题词的网络热点话题发现*

张华平[1),3)]，李恒训[1),2)]，秦鹏[1),2)]，莫倩[4)]

[1)] (中国科学院计算技术研究所，北京　100190)

[2)]（首都师范大学计算机联合实验室，北京　100037）

[3)] (北京理工大学，北京　100081)

[4)] (北京工商大学，北京　100037)

Email: pipy_zhang@msn.com

摘 要：网络话题层出不穷，往往会引发重大舆情危机，如何快速高效的从海量信息中发现热点是一重大挑战。本文提出了一种基于主题词的网络热点话题发现算法。其基本思想为：首先综合主题词表和有意义串识别结果生成主题词候选集；然后对候选集进行多重过滤并采用启发式规则对主题词进行权重计算；最后，以主题词为线索进行热点话题提取，采用多特征的话题模型，融合新闻、论坛、博客的相应特征实现了网络热点话题的发现。通过在 TDT4 评测语料和中科院计算所天玑舆情监测系统平台上的实验分别取得了 0.282 的最小识别代价和 93.3%的用户满意度，算法运行效率高于传统方法。实验表明，该算法对网络热点话题发现行之有效。

关键词： 主题词提取；热点话题发现；聚类；舆情；天玑

Internet Hot Topic Detection Based on Topic Words

Abstract: There are mass of information produced by the Internet everyday, in order to get the hotspot from the mass, this paper showed a quick and effective strategy of the Internet hotspot topic detection based on topic words extraction. Its basic content can be summed up as follows: Firstly, we pretreatment the corpus for Chinese word segmentation with ICTCLAS and use the scan algorithm based on the topic word dictionary and meaningful string recognition algorithm to get the candidate topic-word set, then filter the topic words in accordance with certain heuristic rules and calculate the weight, Lastly, considerable and selective use is made of the Meta information of the web pages to hotspot event cluster quickly, because of the different characteristics of the BBS, News and Blog respectively, which obtains a relatively better results in the experiment.

Keywords: topic words extraction; hot topic detection; clustering; public opinions; golaxy

1　引言

互联网的蓬勃发展大大提升了网络媒体的影响力，使其能够引导舆论导向，进而影响受众。互联网的巨大影响在于它的广泛运用。根据 CNNIC2009 年 1 月发布的报告显示：截至 2008 年底，我国互联网普及率以 22.6%的比例首次超过 21.9%的全球平均水平；同时，

*本课题得到国家高技术研究发展计划（863计划） (2007AA01Z438)资助.张华平, 男，1978年生,博士，副研究员，研究生导师，主要研究方向为中文自然语言处理、信息检索与舆情计算；李恒训，男，1985年生，硕士研究生，主要研究方向为中文自然语言处理与信息检索；秦鹏，男，1984年生，硕士研究生，主要研究方向为中文自然语言处理；莫倩，男，1972年生，博士，副教授，研究生导师，主要研究方向为自然语言处理、信息检索与社会计算。

我国网民数达到 2.98 亿，宽带网民数达到 2.7 亿，国家 CN 域名数达 1357.2 万，三项指标均排名世界第一[1]。特别是 Web2.0 技术的到来，充分体现了用户在网络中的参与作用，网民不仅是信息的接受者，还是信息的发布者和推动者。目前，中国网民言论之活跃已达前所未有的程度，不断在网络上形成热点话题，有些甚至形成热点社会事件，显示了其不可忽视的力量。一方面为社情民意的反馈提供了有效的渠道，为促进和谐社会创造了有利的环境；另一方面许多所谓的“网络热点”，集中于社会阴暗面，集中于突发事件的一些负面效应，种种偏激的言论甚至比正面的主流的言论传播速度更快、波及面更广，往往会引发重大舆情危机。如何快速高效的从这些海量数据中发现和分析这些网络话题，及时把握人们普遍关心的问题，是我们面临的一项重大挑战。

热点话题发现能够使公安、宣传等政府相关机构及时了解社情民意，避免发生影响较大的网络群体性事件，对构建和谐社会具有重要意义。此外，这项技术也有助于厂商了解消费者的需求变化，及时调整产品和市场策略也具有重要的参考意义，对于广告、公关、咨询等产业提高市场调研分析能力和效率都有着现实的应用价值和广阔的应用前景。

话题（Topic），就是一个核心事件或活动以及与之直接相关的事件或活动。而一个事件（Event）通常由某些原因、条件引起，发生在特定时间、地点，涉及某些对象（人或物），并可能伴随某些必然结果。与话题相应的一个概念是主题（Subject），主题是“事件所属的类别”，一个事件一定会属于某一个主题。话题与某个具体事件相关，而主题可以涵盖多个类似的具体事件或者根本不涉及任何具体事件。例如：

主题：民族团结与祖国统一

话题：七五新疆打砸抢烧恶性暴乱 (包括：一系列的事件与活动（跟境外势力的政治斗争、新疆各族人民的反映、重要人物的表态、境外的歪曲报道）新闻报道、网民评论及分析)

事件：甲型流感密切接触者预防性用药指南发布//广东东莞维族汉族群架事件（时间：地点：经过：）

根据统计，本文认为热点话题要充分考虑到用户的反馈信息，通常是具有以下四个特征的之一的信息（新闻、关键词等所描述的信息）：

1) 多个新闻源，多次报道或转载的事件。
2) 论坛上的热帖（多次阅读、多次回复）相关的事件。
3) 博客上的热文（多次阅读、多次回复）相关的事件。
4) 查询日志中高频关键词相关的事件。

其中前三条是信息发布者角度的理解热点，而第四条是用户的角度来理解热点。本文充分考虑这四个特征进行热点话题发现。

本文第二节介绍了热点话题发现的相关工作并提出了改进方向；主题词提取是本文热点发现算法的重要基础，在第三节给予着重介绍；热点发现的算法在第四节给出详细阐述；第五节通过实验验证算法取得了良好的效果和用户满意度；最后是对全文进行总结。

2 相关工作

热点话题发现是话题识别与跟踪（Topic Detection and Tracking，简称为 TDT）技术在实际领域中的应用。TDT 旨在发展一系列基于事件的信息组织技术[2]，自从 1996 年前瞻性

的探索以来，通过连续的大规模评测，TDT 已成为国际上自然语言处理尤其是信息检索领域的一个研究热点，大大促进了 TDT 相关技术的发展[3,4]。目前主流的话题发现算法都采用文本聚类技术来实现。在早期的网络话题相关研究中，为了简化问题，一般假定所有的话题没有层次之分，而且一个文档只能与一个话题相关[5]，因此研究人员基本上都采用传统的聚类算法。[Y. Yang, T. Pierce, etc, 1998] 先用增量聚类把新话题中时序相关的文档聚在一起，而后再用 Single-Pass 聚类来找回时间间隔较大的话题相关文档 [6] 。随着研究的深入，从 2003 年开始，层次化话题发现（Hierarchical Topic Detection）作为话题发现与跟踪领域一个全新的研究问题被提了出来，它突破了传统的话题组织忽略话题多粒度现象的不合理之处，采用层次化的结构对话题进行组织。[Margaret Connell, Ao Feng, Giridhar Kumaran,etc,2004]先把文档分成很多小类，再按报道来源、时间和语种将小类合并[7]；[Yu Manquan, Luo Weihua, etc, 2006]先把文档按照时序分组，然后采用自底向上的多层聚类把所有的话题组织成为一个有向无环图结构[8]。

在热点话题发现的研究中应用了许多 TDT 经典算法。基于以往 TDT 评测的结果，本文考虑了如下因素，基于聚类实现话题发现的算法已经比较成熟，这项任务本身已经不是影响热点话题发现的关键性难题，但这并不表明这个问题已经得到了彻底解决。在处理海量网页时，传统的文本聚类速度过低，时间和空间复杂度往往大于 $O(n^2)$，内容庞杂，干扰因素太多，很难精准聚合中心结果，最终结果往往不知所云，离需要的热点差距太远，因此需要对现有的话题发现建模方法的进行改进。

虽然网络热点话题发现的基础仍然可以基于话题发现与跟踪技术来实现，但直接照搬原有的方法是不可行的。这主要是由于所面对的数据不同。以往的 TDT 研究主要关心的是非结构化的文本数据，除了文本内容之外，其他可用信息少之又少，因此研究人员基本上都是基于文本本身的内容来设计和实现算法。网络数据基本上都以半结构化形式存在，而且网页的标题、发布时间、来源信息、用户浏览次数和回复次数等蕴含着大量有效信息。综合来看，如果把话题映射到一个多维特征空间上，那么文本特征只是其中的一维。显然，传统上基于文本内容的向量空间模型和语言模型的文本和话题表示框架，已经不能有效地涵盖其他非文本信息。因此，我们尝试采用多特征的话题模型，通过分析网络数据的特点，把尽可能多的影响对一个话题进行判断的因素都考虑进来。

对于一个实用的系统来说，算法生成的结果是否满足用户需求是最终目标。发现的话题是否属于用户关注的热点，发现的话题数量是否过多，话题层次是否太深，是否由于“垃圾话题”数量太多而影响用户阅读正确的结果都是我们需要综合考虑的问题。

以往研究所处理的数据规模都比较小，算法的运行效率不是大家关心的主要问题。但是对于网络话题发现而言，要处理的是互联网上的海量数据，而且数量还在持续快速增加。因此算法必须满足快速、时间复杂度低、可扩展性、可适应性等要求。

3 热点主题词提取

在本文的方法中，主题词提取是热点话题发现的线索，其质量直接决定热点话题的准确性。主提词具有代表性、简洁性、时效性、信息量大、相关词语关联度高等特点，能够最大程度的以最小的信息量涵盖热点话题的主题和内涵。统计表明，两个主题词一般可以表示一个话题，如法航、失事，三个主题词一般就可以确定一个话题，如公交、成都、燃

烧。利用主题词给用户呈现热点话题，用户看到的话题不只是一个“文本团”，还包括话题的概要信息，并且以主题词为基础的热点聚类计算量小，效率高。

对于主题词的提取，基于词典的方法可以快速提取传统主题词而对网络新词提取效果比较差，有意义串识别在提取新词上效果突出却容易引入主题性较差的串，本文将基于词典扫描的算法和基于有意串识别的算法有机的结合起来，将两种方法获取的候选主题词通过多重过滤和权重计算后进行优化，既保留了新词，而且对于热点词的主题性得到了保证。

3.1 基于词典的主题词提取

主题词提取的一个重要前提工作是对命名实体和未登录词的识别上，通常这类词在话题中对主题的表达有较大的影响。本文首先对语料使用 ICTCLAS 进行分词，ICTCLAS 能够识别出普通的人名地名和复杂嵌套的地名和机构名，对于命名实体有比较高的识别率[9]。ICTCLAS 提供专业词典与用户词典的功能，用户可以添加自己的词典，本文在分词的过程中加入了主题词词典。

基于词典的主题词提取的效果很大一部分取决于主题词词典的质量。主题词词典的构造上，一方面，我们采集百度、搜狗等网站提供的热门搜索词，进行分词和处理，搜索热点词充分体现了用户的参与性，符合本文热点定义的第四个条件；另一方面采集了维普网站上各个分类的文章的关键词，基本涵盖了人文社科、自然科学工程技术，医药卫生等各个领域和专业，这些期刊、杂志等收录了许多人工标注的关键词的文章，并且这些文章都有具体的分类，从学术文章，到科普生活类的文章，包含的词语非常的丰富。在此基础上进一步人工整理和标注得到词典。

3.2 基于局部性有意串识别的新词发现

基于词典的提取算法的不足之处是不能识别新词，而突发性的新词往往包含重要的信息，极有可能是当前的热点。本文的方法是基于局部性原理的有意义串提取方法的基础上进行[黄玉兰，2008]。首先使用 ICTCLAS 对语料 C 进行分词，从分词结果中选出频率大于一定阈值的字符串作为频繁模式集合 FP(C)。FP(C)被称为重复串；上下文分析阶段，计算重复串 S 的 AV 值，滤掉 AV 值过低的字符串；局部性分析阶段，计算重复串 S 的 LE 值；最后，根据 AV 值和 LE 值给 S 打分，衡量它能成为有意义串的可能性，如果该值大于一定阈值，则认为该字符串是有意义串[10]。

有意义串对于新词识别有着非常好的效果，可以通过有意义串识别获取像“猪流感”，“甲型 H1N1”，“杭州飙车”这样的某一时段内的热点新词，但是负面影响是有可能引入大量的相似串和无意义串。本文通过 3.3 节中的多重过滤和 3.4 节中权重计算进行后续处理，得到比较好的新词识别效果。

3.3 候选主题词的多级过滤

候选主题词中可能会引入主题性较差的串，特别是在正文中利用有意义串识别算法进行新词发现时可能会引入无意义串。统计表明无意义串大致分为三类：常见无意义串、相似串以及规则性强的无意义串。为此制定了多级过滤策略：

1)停用词过滤：算法载入一个人工整理的停用词表，用以过滤常见的无意义串。

2)相似词过滤：候选主题词中存在包含关系的词语或者相似度比较大的词语，如“甲型”、“甲型 H1N1”、“甲型 H1N1 流感”等词。通过相似度和词频进行综合判定，在词频相当的情况下，保留长度较大的词语。

3)规则过滤：对于规则明显的无用串，如频繁出现的数词与量词的搭配、一些常见但无意义的前后缀等，编写相应的规则进行过滤[11]。

3.4 多特征融合的主题词权重计算

对于候选的主题词，每个词包括 TF，IDF，词性，词在句中位置信息等有效信息，本文考虑了如下因素：词频越高，说明其受关注的程度越高；IDF 越大，说明词的区分度越大，切合主题的特点；词的长度越长，信息量越大；在词性上，命名实体的信息量高于非命名实体，因此命名实体需要增加权重，动词可以做为衡量标准，其他词性略低；词在文档中首次出现的位置对主题表达也会造成影响，可以作为一个调节标准，所以词的位置靠前权重适当增加。本文综合考虑以上几个因素，构造出权重计算公式，并对候选主题词进行排序。

$$W(t,d)=a*\frac{1+\log_2(tf(t,d))*\log_2(N/n_t)*Weight(POS(t))}{|d|} \quad (0<a,b<1;a+b=1)$$
$$+b*\frac{length(t)}{AvgLen}*Weight(Position(t,d))$$

其中：w(t,d)表示词 t 在文档 d 中的权重;a,b 为调节系数; tf(t,d)表示词 t 在文档 d 中的频率；n_t 表示在整个语料中出现词 t 的文档的数目；N 表示整个语料中的所有文档数目； |d| 表示文档向量的长度。Weight(POS(t))为 t 的词性权重，一般命名实体取 2，动词取 1.5，其他词性为 1;length(t)表示词 t 的长度，AvgLen 表示主题词的平均长度；Weight(Position(t,d)) 表示词 t 在文档 d 首次出现的位置系数，计算方式为在候选词第一次出现前的词的个数除该文档词的总数。

权重的归一化到[0,1]的公式如下：

$$W_i=\frac{W_i}{\max_i(W_i)}$$

4 基于主题词的热点话题发现算法

4.1 多特征选择及预处理

本文根据网页的不同特点将其分为新闻、论坛和博客三种类别分别进行热点话题识别，从网页中抽取的元信息是能够进行快速聚类的有效信息，因此我们首先对网页的多种特征进行选择并根据权重进行排序。

统计表明，新闻文章的标题最具代表性，基本能够反映文章的主题，网民在浏览新闻时最先关注的就是新闻标题,因此新闻的标题可以做为主要语料，正文可以作为辅助语料。新闻的热点往往是多个新闻源，多次报道或转载的事件，并且新闻报道讨论同一事件往往是时间相近的，本文把新闻语料按照发布时间逆序进行排序，最新的新闻排在前面。

论坛的浏览数和回复数能够反应帖子的热度，而热帖相关的话题往往就是热点话题。同时考虑到时间特性，论坛热点往往是某一集中时间段内关注的话题。本文对采集到的论

坛网页根据权重为第一因素，发布时间为第二因素，以权重降序、时间逆序排序。同新闻相似，帖子标题做为主要语料，正文作为辅助语料。其中权重计算公式为：

$$Weight(p) = a*Reply\_Count(p) + b*Browse\_Count(p) \quad (0<a, b<1, a+b=1)$$

其中 Weight(p)为帖子 p 权重，Reply_Count 为帖子 p 的回复数，Browse_Count 为帖子 p 的浏览数。a,b 为调节系数，一般 a 取 0.8，b 取 0.2。

博客与论坛的元信息类似，本文采取与论坛相同的方法进行预处理。

4.2 算法设计

基于主题词的热点话题发现 TWHTD(Topic Words based Hot Topic Detection)算法如图 1 和图 2 所示，假定网页元信息提取和预处理工作已完成。算法从 n 篇文档集合中按照第三节的方法抽取主题词集合 T，为了适应海量数据处理，本文以文档标题代替文档全文进行主题词提取，并对 T 进行权重排序。首先我们默认一个主题词代表一个热点话题，然后对这些热点话题进行凝聚聚类。以 T 中的第一个主题词做为第一个热点话题，以此为线索查找文章标题进行聚类，从新聚类的文章标题和正文中分别抽取两个主题词，五个主题词表征第一个热点话题。从 T 的第二个主题词开始，根据主题词为线索在文章标题中查找进行聚类，从本类文章的标题和正文中分别抽取两个主题词，以主题词及其权重构造的向量采用余弦距离计算本话题与已有热点话题的相似度，若相似度超过阈值 p 则将当前话题合并到已有话题中，并重新对该类抽取五个主题词表征话题。算法迭代执行，达到用户预先设定的热点话题数目后终止。

```
输入：n 个经过排序的文档集合 D,文档 Di 包括标题 Ti 和正文 Ci
输出：k 个热点话题(五个主题词表征)以及每个话题对应的文章。
Function Hotspot_Event_Detection(int k)
    BEGIN
        TopicWords=ExtractTopicWords(D);//提取热点主题词
        SortTopicWords(TopicWords);//排序
        HotTopic(1)=TopicWords(1);//第一个主题词默认为第一个话题
        Cluster(1)=clustering based on the thread of TopicWords(1);
        for(i=1;i<sizeof(Topics);i++)
        BEGIN
            HotTopic(cnt)=Topics(i);//第 i 个主题词做为第 cnt 个话题
            Cluster(cnt)=clustering based on the thread of TopicWords(i);
            ClusterTopicWords=ExtractTopicWords from Cluster(cnt);
            for(j=0 ; j<cnt ; j++)
            if(Sim(Cluster(cnt),Cluster(j))>p)//计算与已有话题相似度
            {
                Cluster(j)=cluster(j)  ∪  Cluster(cnt);
                Continute;
            }
            if(++cnt >= k)//热点话题数 cnt 超过 k 个则退出
            Exit();
        END
    END
```

图 1 TWHTD 算法描述

Fig.1 the description of TWHTD algorithm

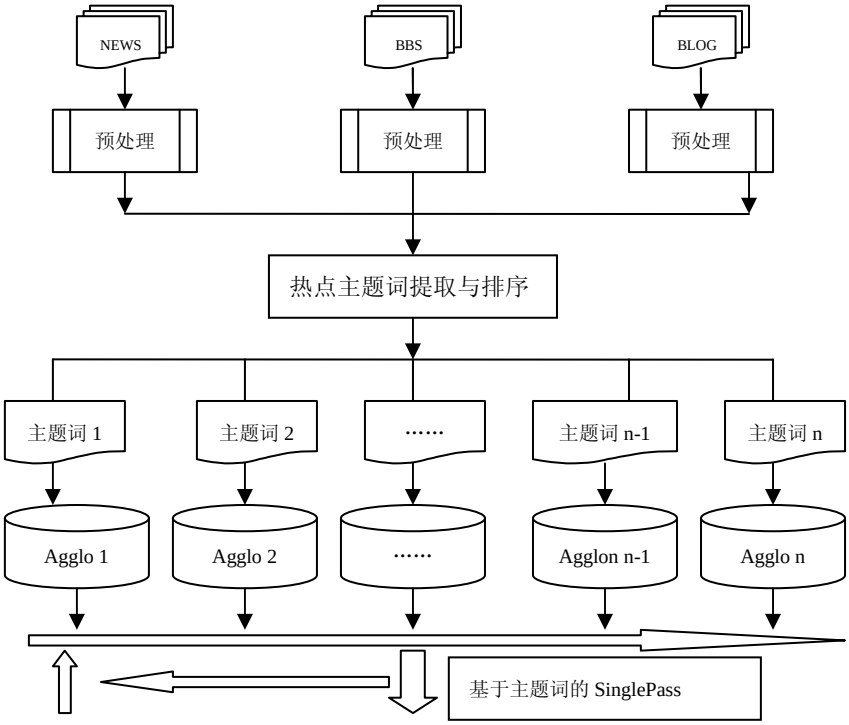

图 2 TWHTD 算法示意图

Fig.2 the overflow of TWHTD algorithm

（图中“Agglo”代表以主题词为线索的聚类）

5 实验结果及分析

5.1 实验设计

TDT4 语料包含时间跨度为四个月(2000.10.1-2001.1.31)的新闻专线和广播语料。它一共包含 98,245 篇报道(英文 28,390 篇，汉语 27,142 篇，阿拉伯语 42,713 篇)。汉语和阿拉伯语语料提供本国语和翻译语料两种形式，翻译语料采用机器翻译软件翻译成英语。广播语料是通过语音识别软件自动转录而成的。整个语料一共 25,960 个文件，共计约 2.5GB。TDT4 语料主要来源于大型的新闻媒体机构，例如“纽约时代新闻”和“新加坡联合早报”等。语料里标注的与话题相关的报道一共 6,000 多篇，蕴涵在 100 个话题中。在实验中，我们利用了 TDT4 的汉语语料。

由于一些原因，TDT 语料并不能完全反映互联网的真实情况，这是因为互联网网页包含大量对话题识别有效的多种特征信息，如论坛帖子的链接关系、标题、浏览数、回复数等信息。所以本文还在中科院计算所天玑舆情系统平台下进行实验。该系统采用分布式框架实时采集各权威站点数据并进行元信息抽取，如新闻采集源包括新浪新闻网、中国新闻网、南方周末、千龙网等，作为 TDT4 的实验的补充。

5.2 基于 TDT4 的实验

本文依据 TDT 评测标准采用漏报率、误报率以及识别代价$(C_{Det})_{Norm}$来评价话题检测的性能，话题 i(i=1,2,..t_n,t_n 为话题个数)的漏报率和误报率定义为：

$$Miss_i=\frac{\text{未检测到的与话题i相关 报道数}}{\text{与话题i相关的报道数}}$$

$$Fa_i=\frac{\text{检测到的与话题i不相关的报道数}}{\text{与话题i不相关的报道数}}$$

$$(C_{Det})_{Norm}=\frac{C_{Miss}*P_{Miss}*P_{target}+C_{Fa}*P_{Fa}*P_{\neg target}}{\min(C_{Miss}*P_{target},C_{Fa}*P_{\neg target})}$$

其中$(C_{Det})_{Norm}$越小表明系统的性能越好，C_{Miss}为漏报一个新话题的代价，C_{FA}为误报一次的代价，P_{target}是目标话题的先验概率，$P_{\neg target}=1-P_{target}$，$C_{Miss}$、$C_{Fa}$和 P_{target}都是预设值，不同的评测中取值不一样，本文中的取值分别为 1.0,0.1 和 0.02。

由于 TDT4 语料只有正文，没有标题、发布时间等网页中可以抽取到的信息，所以实验中我们假定语料已经按照时间顺序排序，并且我们使用主题词抽取的方法从正文中抽取 10 个主题词来表示标题，然后按照与新闻网页相同的方法利用 TWHTD 算法进行实验。本文抽取前 10 个热点，实验结果如表 1 所示。

表 1 基于 TDT4 的热点话题发现测试结果

Tab.1 The Test result of hot topic detection based on TDT4

相似度阈值 p	P_{Miss}	P_{Fa}	$(C_{Det})_{Norm}$
0.35	0.2618	0.0053	0.28777
0.38	0.2655	0.0033	0.28167
0.41	0.2733	0.0021	0.28359
0.45	0.2939	0.0019	0.30321

通过不同的相似度阈值比较，当 p 不断增大时，漏报率在增大，同时误报率在减小，综合考虑识别代价的走势，我们发现 p 取 0.38 时识别代价$(C_{Det})_{Norm}$达到最优值 0.282。

5.3 基于中科天玑舆情监测系统的实验

本文对基于中科天玑舆情监测系统的实验结果由三位专家分别进行了用户满意度抽样调查。系统分别展现了新闻、论坛和博客的热点话题，运行频率为每天一次，发现最近三天的热点话题，由专家进行评判，分为满意、认可和较差三个等级。系统运行的用户满意度测试的统计数据如表 2 所示, 表 3 展示了 2009 年 6 月 8 日系统运行的新闻结果展示及用户满意度。具体数据也可以从中科天玑网站(http://www.golaxy.cn）中在线获取。

表 2 用户满意度统计
Tab.2 the statistics of customer satisfaction index

类型	热点话题数	满意率	认可率	较差率
新闻	300 个	76.67%	18.00%	5.33%
论坛	300 个	77.33%	18.67%	4.00%
博客	300 个	50.67%	38.67%	10.67%
平均	900 个	68.22%	25.11%	6.67%

表 3 一次运行的用户满意度测试
Tab.3 customer satisfaction index of running once

热点话题	聚类文章数	满意	认可	较差
法航 客机 失踪 本钢 残骸	97	√		
綦江 矿难 遇难者 重庆市 获赔	50	√		
盖特纳 财政部长 会见 特使 温家宝	24		√	
甲型 流感 病例 确诊 陈竺	123	√		
小产权房 转正 分而治之 准生证 探路	16		√	
禽流感 演习 防控 蔓延 预案	3	√		
奥巴马 萨科齐 通用汽车 陆军部 众议员	17		√	
黄硕 中国远洋 中核科技 蓝筹 尾盘	17		√	

从用户满意角度我们可以看到算法取到比较好的效果，平均可接受率(满意率+认可率)为 93.33%，但新闻、论坛和博客还存在差别。由于新闻的标题非常具有代表性，且正文容易抽取到，噪音相对较小，实验结果比较好；论坛中通过浏览数和回复数可以判别热贴的程度，并且由此可以很好的进行热点聚类，因此效果最好；博客中由于文章标题主题性不如新闻强，浏览数和回复数相对比较少，而且有的博文大量加有图片表达主题，所以效果相比前两者比较差。

6 总结及下一步工作

热点话题发现有着重要的应用背景，本文提出了一种快速有效的基于于主题词的热点话题发现算法 KWHTD。其基本思想为：首先综合主题词表和有意义串识别结果生成主题词候选集；然后对候选集进行多重过滤并采用启发式规则对主题词进行权重计算；最后，

以主题词为线索进行热点话题提取，采用多特征的话题模型，融合新闻、论坛、博客的相应特征实现了网络热点话题的发现。通过在 TDT4 评测语料和中科院计算所天玑舆情系统平台上的实验表明，该算法行之有效。在接下来的工作中我们将进一步将博客的热点话题识别算法进行优化和提高，在此基础上将着重开始进行热点话题的摘要信息计算和话题追踪的研究。

参考文献

[1]. CNNIC.《第 23 次中国互联网络发展状况统计报告》[R]．北京，2009 年 1 月．

[2]. J．Allan．Introduction to Topic Detection and Tracking in Topic Detection and Tracking：Event- based Information Organization[R]．Kluwer Academic Publishers，2002：1—16．

[3]. 洪宇，张宇，刘挺等．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．

[4]. The 2004 Topic Detection and Tracking(TDT2004)Task Definition and Evaluation plan[R]．version 1.0.5 August 2004．

[5]. J.Allan, J.Carbonell, G.Doddington, J.Yamron,etc．Topic detection and tracking pilot study: Final report[R]．P194-218,1998．

[6]. Y. Yang, T. Pierce, and J. Carbonell．A Study on Retrospective and Online Event Detection[C]．The 21st Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’98), pages 28-36, 1998.

[7]. M. Connell, A. Feng, G. Kumaran, etc. 2004 UMass at TDT 2004[C]. The Seventh Topic Detection and Tracking Conference (TDT2004).

[8]. Yu Manquan, Luo Weihua, Xu Hongbo. Bai Shuo. 2006. Research on Hierarchical Topic Detection in Topic Detection and Tracking [J]. Computer Research and Development. Vol.43 No.3. P489-495.

[9]. 刘群，张华平，俞鸿魁等. 基于层次隐马模型的汉语语法分析[J]. 计算机研究与发展，2004.8.

[10]. 黄玉兰，龚才春，许洪波等. 基于局部性原理的有意义串提取方法[C]. 第四届全国信息检索与内容安全学术会议论文集，2008.11.

[11]. 曾依灵，许洪波，白硕. 网络文本主题词的提取与组织研究[J]. 中文信息学报，2008.5.

[12]. 曾依灵，许洪波. 网络热点信息发现研究[J]. 通信学报，2007.12.

[13]. 毛国君，段立娟，王实等. 数据挖掘原理与算法（第二版）[M]. 北京，清华大学出版社 2007.12.

[14]. 刘菲. 中文文本主题词抽取研究与应用[D].上海，复旦大学 2007.

[15]. 王丫. 网络新闻流中热点话题识别与跟踪算法的改进与验证[D]. 燕山大学 2007.