网络热点事件发现系统的设计*

刘星星[12] 何婷婷[12] 龚海军[12] 陈龙[12]

[1]华中师范大学计算机科学系 武汉 430079

[2]国家语言资源监测与研究中心网络媒体分中心 武汉 430079

xxcolin@163.com tthe@mail.ccnu.edu.cn

navygong@mails.ccnu.edu.cn lestou@mails.ccnu.edu.cn

摘要：本文设计了一种热点事件发现系统。该系统面向互联网新闻报道流，能自动发现任意一段时间内网络上的热点事件，并给出描述事件发展过程的曲线图。针对网络新闻语料具有数据规模大和时间特征明显两个特性，系统将语料按时间（天）分组，对每天的语料采用凝聚聚类得到微类，选取某段时间内的所有微类，再做 Single-pass 聚类得到事件列表，利用事件热度计算公式，把候选事件按热度进行排序。采用本系统对 2007 年新闻语料进行实验，结果表明该系统能取得较好的效果。

关键词：事件发现；凝聚聚类；Single-pass 聚类；热度计算

Design of Network Hot Event Detection System

Liu Xingxing[12] He Tingting[12] Gong Haijun[12] Chen Long[12]

[1]Department of Computer Science, Huazhong Normal University Wuhan, 430079, China

[2]Monitor and Research Center for National Language Resource

Network Multimedia Sub-branch Center Wuhan, 430079, China

xxcolin@163.com tthe@mail.ccnu.edu.cn

navygong@mails.ccnu.edu.cn lestou@mails.ccnu.edu.cn

Abstract: We propose a system of detecting hot event automatically. The system is focused on the stream of news report available on the Internet, which can be utilized to detect the hot event in any period of time and provide a diagram concerning the tendency of the event. Since news corpus is characterized by large scale data and distinct time features, it is divided into hundreds of groups according to the temporal date of day. We amalgamate each group into some macro-clusters using the agglomerative clustering, and select the macro-clusters during a certain period of time, and then we combine all these selected macro-clusters into event lists which sort the events of candidate in terms of the hot degree formulation by taking advantage of Single-pass clustering. Experiments of 2007 news corpus show that our system can get significant improvement.

Key words: Event Detection; Agglomerative Clustering; Single-pass Clustering; Hot Degree Calculate

1 引言

互联网这一新媒体的出现，使我们摆脱了信息贫乏的桎梏，进入一个信息极度丰富的时代。但是在目前信息爆炸的情况下，一方面网络信息的规模急剧膨胀，另一方面信息又凌乱无序，对有价值信息的发现和管理变得越来越困难。因此，一种能自动发现某一段时间内网络上热点事件

* 基金项目：国家社会科学基金 06BYY029；国家自然科学基金 60773167；湖北省自然科学基金计划项目 2006ABC011；973 国家重点基础研究发展计划 2007CB310804；教育部/国家外国专家局高等学校学科创新引智计划 B07042

的工具，成为了人们的迫切需求。热点事件发现系统就是在这种情况下应运而生的。

热点事件是指能引起人们极大关注的事件，它一般具有下面两个条件之一或兼而有之。第一，事件的关注持续时间较长；第二，在某个时间段内，事件的关注程度较高。在国内，有一些机构通常会发布针对某个领域在某一年内的热点事件，但这些事件通常都是由人工筛选得出，或者人工干预的因素很大，并且不能及时给出用户想要知道的任意一段时间内的热点事件[1,2]。

本文在已有的相关研究工作基础上针对实际系统中存在的问题，设计了一种热点事件发现系统。该系统面向互联网新闻报道流，能够快速地自动发现用户选择的任意一段时间内网络上的热点事件，具有一定的实时性，并能给出事件发展曲线图让用户回顾和了解事件的发展变化过程。

本文接下来的结构将分成以下四部分：①相关研究工作；②本文所述系统的结构及采用的关键技术；③实验过程及结果分析；④下一步的研究工作。

2 相关工作

热点事件发现是话题发现与跟踪（TDT）技术在实际领域中的应用。TDT 是一项面向新闻媒体信息流进行未知话题识别和已知话题跟踪的信息处理技术[3]。自从 1996 年前瞻性的探索以来，该领域进行了多次大规模的评测，大大促进了 TDT 相关技术的发展[4][5]。

在热点事件发现的研究中应用了许多 TDT 经典算法。

文献[6]所述系统在参加 TDT2004 层次话题识别任务中取得了第二名的成绩，该系统采用的分层聚类算法，一方面，能降低系统的计算量；另一方面，通过把新闻语料按时间分组，可以避免把内容相似但实际上讨论两个事件的报道聚在一起，而且还可以使那些时间跨度较大的事件，通过组间聚类再合并成一个事件。

一个事件的文档除了内容相关之外，在时间分布上也有一定的联系，因此，在计算文档相似度时，还应考虑时间因素。文献[7]中的相似度计算公式，在传统的余弦基础上，加入时间衰减因子，提高了计算的精度。

通常情况下，一个类中包含不止一个文档，文献[8]直接用类的核心文档代表该类，选择平均连通策略用于计算类间距离，取得了较好的效果。此外类的平均相似度对于事件的热度排序也有重要的作用。

如何对得到的事件进行热度的度量是热点事件发现的一项独有工作，国内外在这方面的研究较少[9]。文献[10]中针对流行词发现的一些方法可以用来借鉴。该方法通过分析历年流行词的走势曲线图，归纳出流行词所具有的特征并进行量化，总结出度量流行词热度的公式，从而可以计算出每个候选词的热度，得到流行词。

3 网络热点事件发现系统

在这一节中，我们首先概要的描述系统的基本思想与流程，然后对系统中采用的关键技术作重点分析。

3．1 系统的基本思想与流程

我们的目的是要开发出面向实际应用的系统，因此，在对现有的成熟算法加以总结的基础上，针对其中的一些问题和实际应用环境的需求，我们对算法做了一些优化，设计了一种新的网

络热点事件发现系统，系统的流程如图 1 所示。

因为系统要处理的是大规模的动态数据流，为了降低系统的时间复杂度，提高热点事件发现的精确性和实时性，系统采用了两层聚类的策略，分为批处理和实时处理两个阶段。每天的批处理过程主要是对该天的语料作第一层聚类，即凝聚聚类，得到每天的微类。实时处理过程则是对某个时间段内所有天的微类，按照时间的先后顺序，做第二层聚类，即 Single-pass 聚类，得到一个事件列表，接着利用事件热度计算公式，对候选事件进行过滤和排序，得到最终的热点事件。

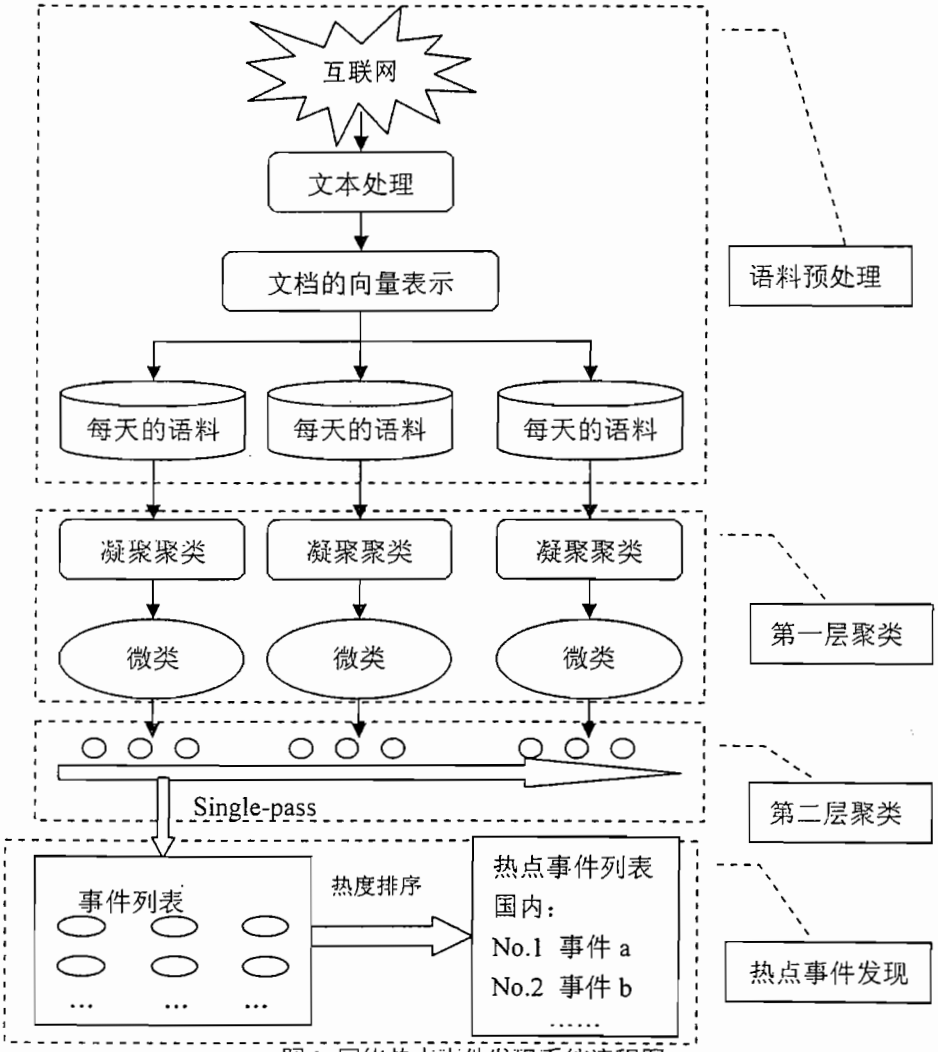

图 1 网络热点事件发现系统流程图

3．2 特征项的权值计算

对于内容这个难以表示的特征，我们采用向量空间模型。对文档进行向量的表示是计算文档之间相似性以及对文档进行聚类的基础，每个文档 d 表示成一个向量 $V(d)=(t_1, w_1(d); \ldots; t_i, w_i(d); \ldots; t_n, w_n(d))$，其中 t_i 为特征项，本系统选取词作为向量的特征项，$w_i(d)$为 t_i 在 d 中的权值。

文档一般可分为标题和正文两个部分，因此在进行权值计算时，对标题和文档中的词分别

赋以不同的权重。另外，文档除了内容词之外，通常还具有人物、时间、地点等命名实体，这些要素对于区分不同的事件起着很重要的作用，因此，系统也对命名实体进行加权。特征项 i 权重的计算采用改进的 TF*IDF 方法，计算公式如下（1）所示

$$w_i = tf_i * idf_i * f_i(w) \quad (1)$$

其中，$f_i(w)$用来分别对命名实体、标题中的词和正文中的词赋以不同的权重，tf 和 idf 均采用经典的算法。

3．3 两层聚类策略

由于系统所要处理的网络新闻语料的规模通常非常大，且时序是其重要的特征。因此系统采用两层聚类的策略，具体步骤如下：

（1）对每天下载的网页，预处理后，形成一个按天分组的语料库；

（2）在每一组内，采用凝聚聚类生成微类，利用阈值控制生成的微类数量；

（3）选取用户选择的时间段内所有天的微类，保持微类间的时间顺序，采用 Single-Pass 聚类算法将各微类进行合并。

组内语料的特点是：包含的文档数量不多，且每个文档包含的信息量较少。由于凝聚聚类的优点是通过控制阈值能较好地区分两个类，缺点是运算量较大，只适用于处理小样本数据，而组内聚类的目的是尽可能将各个组内的文档聚合成微类，因此采用凝聚聚类作为组内聚类算法。

由于语料规模非常大，用户选择的某个时间段内的微类数量可能仍然很多，且微类包含的信息比文档丰富，所以采用 Single-Pass 算法作为组间聚类算法，其原因是由于 Single-pass 聚类属于非层级聚类，其聚类过程是一个迭代过程，算法效率较高，适合于处理数据规模较大的语料[11]。

3．4 事件的热度计算

借鉴流行词发现的一些方法，首先，要归纳出度量事件热度的特征量，然后总结出对事件进行热度计算的公式。

（1）事件的报道频率

事件的报道频率分为时间频率和数量频率。时间频率是指在某一时间段内，以天为时间单位，报道过某一事件的有效天数与该段时间总天数的比值，事件的时间频率越大，越倾向于热度更高的事件；数量频率是指在某一时间段内，关于某一事件的报道数量，与该段时间内报道总数的比值，事件的数量频率越大，也越倾向于热度更高的事件。

（2）事件的平均相似度

加入事件平均相似度（即类的平均相似度），可以减少内部比较杂乱的事件的热度得分，以避免一些内部混乱的事件出现在热点事件排列前列的情况。

根据以上的分析，我们得到度量事件热度的三个特征量：

TR_i，表示在用户选择的 n 天时间内，事件 i 的有效报道天数与所有天数的比值，当某一天关于事件 i 的报道数量大于某一阈值时，即认定该天为事件 i 的有效报道天；

DR_i，表示事件 i 的报道数量频率，即 $DR_i = \frac{\sum_{j=1}^{n} df_{ij} / df_j}{n}$，其中，$df_{ij}$ 表示事件 i 在第 j 天

的报道数量，df_j 表示第 j 天的报道总数；

$F_i(avgsim)$，表示事件 i 的平均相似度。

从而总结出对事件 i 的热度计算公式，如下（2）所示

$$W_i = TR_i \times DR_i \times F_i(avgsim) \qquad (2)$$

此外，通过分析历年的热点事件，得出热点事件必定要经历从上升到稳定再到下降的发展过程，所以利用事件的发展曲线，去除得分虽高，但不符合热点事件特征的事件。例如“姚明征战 NBA07 至 08 赛季”。

4 实验过程及结果分析

4．1 实验数据集

实验选取从五个门户网站（新浪、网易、腾讯、搜狐、Tom）上下载 2007 年 1 月 1 日至 2007 年 12 月 31 日的六个领域（国内、国际、体育、财经、科技、娱乐）的全部网页，共计 652,849 篇，其中，国内领域有 147,126 篇，国际领域有 112,212 篇，体育领域有 128,972 篇，财经领域有 90,252 篇，科技领域有 82,153 篇，娱乐领域有 92,134 篇，每一个领域的文档集合分别单独作为一个实验数据集。

4．2 实验步骤

按照图 1 所示的流程，实验的具体步骤如下：

（1）每天从五个门户网站上下载六个领域的全部网页，抽取网页的文本内容，按标题、日期、正文三部分格式存储文档，以日期作为文件夹名，例如：..\sports\20070101\，存储了 2007 年 1 月 1 日体育领域的全部数据。

（2）语料预处理部分，包括对每天下载的文档进行分词、去除停用词、命名实体的识别、文档过滤以及文档的向量表示等。

（3）每天对预处理过的数据做自底向上的层次式凝聚聚类，通过阈值控制生成的微类数量，然后将文档数大于一定阈值的微类，保存作为下一层聚类的数据。（根据数据的规模，适当改变这两个阈值的大小，以降低下一层聚类的计算量）

（4）选取用户选择的时间段内所有天的微类数据组，去掉各组之间的间隔，但仍保持微类间的时间顺序，采用 Single-pass 聚类算法，将各微类进行合并，得到一个事件列表。

（5）根据每次实验的具体情况，淘汰掉所包含文档数少于 n 或者有效报道天数少于 m 的事件，得到候选事件集，其中 n 和 m 为控制阈值。

（6）对候选事件进行热度计算，根据计算得分和事件发展曲线，对事件进行过滤和排序，输出最终得到的热点事件。

4．3 实验结果与比较分析

对于利用计算机自动获得热点事件，目前没有一个统一的评测标准，本文试图通过与权威网络媒体机构发布的热点事件进行对比，来对本系统进行评测。

因为媒体发布的热点事件都是以年为单位，并且发布的事件数量大多数都是10个，考虑到文章的篇幅，下表1中分别列出了采用本系统进行实验得到的2007年排名前10位的国内热点事件和由人民网等评选的中国网络媒体2007年度国内十大新闻，进行比较，来对本系统进行评测。对于任意时间段内的热点事件，因为没有权威的数据与之进行比较，故没有在此列出。

表1　2007年国内领域十大热点事件对比

序号	本系统得到的排名前10位的国内热点事件	中国网络媒体2007年度新闻风云榜
1	中国共产党第十七次全国代表大会召开	中国共产党第十七次全国代表大会胜利举行
2	山西“黑砖窑”事件	我国首次月球探测工程圆满成功
3	太湖蓝藻暴发	十届全国人大五次会议高票通过物权法
4	台湾“入联”闹剧	我国全面建立农村最低生活保障制度
5	庆祝香港回归祖国十周年	国家预防腐败局成立，中共严查高级干部违纪案表明反腐决心
6	中国铁路第六次大提速	陈水扁当局推动“入联公投”，对台海和平稳定构成严峻威胁和挑战
7	国家调整法定节假日	我国物价增幅创10年新高，政府采取多种措施稳定市场
8	全国人大会议通过物权法	国家调整法定节假日，出台《职工带薪休假条例》
9	药监局原局长郑筱萸案	“黑砖窑”事件震惊全国，中央高度重视要求严查非法用工
10	华南虎照片事件	中共中央政治局集体学习，研究加强网络文化建设和管理

从上表可以看出，由本系统得到的排名前十的国内热点事件中，有6个相似的事件包含在“中国网络媒体2007年度新闻风云榜”中，并且其序号为2的“我国首次月球探测工程圆满成功”和序号为7的“我国物价增幅创10年新高，政府采取多种措施稳定市场”事件也在本系统发现的2007年的热点事件中，排名同样位于前十，只是领域分别为科技和财经。

因为权威媒体在评判热点新闻时较注重宏观层面和潜在长远影响，而本系统在热点事件的判别中则更关注事件在大众中的反应，所以两个表中的结果必然会存在着一定的差异。但本系统最大限度地减少了人工的干预，所以结果会更加客观公正。

4．4 事件的发展曲线

表1仅仅只将事件按照热度进行排序，并不能反映出一个事件的发展过程。因此，引入事件的发展曲线来描述热点事件的发展过程和变化趋势。

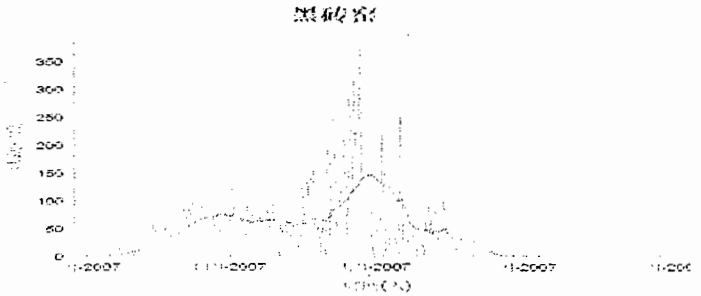

图2 “山西黑砖窑”事件的发展曲线

图 2 所示为“山西黑砖窑”事件的发展曲线，图中横轴表示时间，纵轴“活跃程度”表示每天关于“黑砖窑”事件的报道数量。从图示可以看出 2007 年 2 月至 5 月，已经出现了一些关于“黑砖窑”事件的报道，并呈上升趋势，大约 7 月份时达到最高峰，八月以后开始出现下降趋势。该图基本上反映出了“山西黑砖窑”事件的发展变化过程。

5 总结和下一步的工作

本文设计了一种网络热点事件发现系统，利用该系统能自动发现任意一段时间内网络上的热点事件，且能最大限度地减少人工因素对结果的影响，实验证明本系统能取得较好的实用效果。

门户网站的新闻报道偏重于媒体对事件的关注，而忽略了用户的行为对热点事件产生的影响，因此在下一步的工作中，我们将把语料来源扩大到多种类型（BBS、博客、新闻）的多个站点，既考虑媒体的影响，也关注用户的行为。同时，也将考虑如何依照“HowNet”情感体系结构，把事件的相关报道进行立场分类[12]。此外，在一个事件刚出现时就能预测其以后的发展趋势，这对于舆论的调控将产生重大的现实意义，因此，建立更加完善的事件发展监控体系也是下一步的重点研究目标。

参 考 文 献

[1] Tingting He, Haijun Gong, Wenmin Hu, Guozhong Qu, Yong Zhang.. Hot Event Detection. International Conference on Chinese Computing 2007, Wuhan, China, Oct. 13 to 15, 2007

[2] 罗亚平，王枞，周延泉。基于关注度的的热点话题发现模型。第七届中文信息处理国际会议，湖北，武汉，2007 年 10 月 13-15 日

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

[4] 洪宇，张宇，刘挺，李生。话题检测与跟踪的评测及研究综述。中文信息学报，2007 年 11 月：第 21 卷，第 6 期

[5] The 2004 Topic Detection and Tracking (TDT2004) Task Definition and Evaluation plan

[6] Man-Quan Yu, Wei-Hua Luo, Zhao-Tao Zhou, Shuo Bai. ICT's Approaches to HTD and Tracking at TDT 2004. In Proceedings of Topic Detection and Tracking Workshop

[7] 骆卫华，于满泉，许洪波，王斌，程学旗。基于多策略优化的分治多层聚类算法的话题发现研究。中文信息学报，2006, 1 (20)：29-36

[8] 邱立坤，陶然，龙志祎，程葳。面向互联网的话题发现技术研究。全国网络与信息安全技术研讨会，山东，青岛，2007 年 7 月 17-19 日

[9] YE Hui-min, CHENG Wei, DAI Guang-zhong. Design and Implementation of On-Line Hot Topic Discovery Model. Wuhan University Journal of Natural Science, 2006,11(11):21-26

[10] 何婷婷，朱慧，张勇，任函。基于词语属性的计算机辅助获取流行词语研究。中文信息学报，2006，6 (06)：38-45

[11] Papka R, Allan J. On-Line New Event Detection using Single-Pass Clustering. UMass Computer Science Technical Report TR98-21, 1998

[12] 金珠，林鸿飞，赵晶。基于 HowNet 的话题跟踪及倾向性分类研究。情报学报，2005 年 10 月：第 24 卷，第 5 期