互联网新闻热点事件发现方法研究

孟强

2014 年 2 月

中图分类号：TP391
UDC 分类号：500

互联网新闻热点事件发现方法研究

作 者 姓 名　　孟　强
学 院 名 称　　信息与电子学院
指 导 教 师　　王越院士，罗森林教授
答辩委员会主席　　王华教授
申请学位级别　　工学硕士
学 科 专 业　　信息与通信工程
学位授予单位　　北京理工大学
论文答辩日期　　2014 年 2 月

Research on Methods of Internet News Hot Event Detection

Candidate Name： Meng Qiang

School or Department： Informaion and Electronics

Faculty Mentor： Prof. Yue Wang , Prof. Senlin Luo

Chair, Thesis Committee: Prof. Hua Wang

Degree Applied: Master of Engineering

Major: Information and Communication Engineering

Degree by: Beijing Institute of Technology

The Date of Defence： February，2014

互联网新闻热点事件发现方法研究

北京理工大学

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签　　名：　　日期：

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：①学校有权保管、并向有关部门送交学位论文的原件与复印件；②学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③学校可允许学位论文被查阅或借阅；④学校可以学术交流为目的,复制赠送和交换学位论文；⑤学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

签　　名：　　日期：

导师签名：　　日期：

摘要

互联网新闻已经成为用户获取信息的一个重要来源。新型的网络资源和网络新闻应用不断增加，网络新闻数目呈现爆炸式增长，给用户阅读新闻增加了很多困难，从大量的网络新闻中发现和分析热点事件成为急需解决的重要问题。尽管机器学习、自然语言处理等多方面的技术已经在网络热点事件发现中得到了广泛的应用，但是现有的文本表示模型存在相对局限性，使得文本表示的性能仍不能让用户满意，还有很多问题需要进一步研究。

为了实现更加深入的理解文本的目的，本文基于句义结构模型构建了一种基于聚类的互联网热点事件发现方法。该方法首先对文档进行句义成分分析，计算词的权重后生成语义向量；将语义向量用到热点事件发现系统中，采用 single-pass 聚类思想和凝聚式层次聚类与 K-means 聚类算法相结合的聚类算法，事件发现准确率为 75.2%。此外，构建了一种事件简化表示的方法，抽取事件发展关键点和事件标签，事件发展关键点的准确率为 58.9%。此外，设计并实现了一种热点事件发现和事件简化表示原型系统。

关键词：热点事件；文本聚类；相似度计算；汉语句义结构模型

Abstract

Internet news has become an important source of public access to information. New network resources and network news applications is increasing , the number of network news shows explosive growth , that brings the user a lot of difficulty to read the news . Discovery and analysis of hot events become important issues need to be resolved from a large number of network news . Although machine learning, natural language processing , and many other techniques have been widely used in network hot events discovery , but the existing text represent model has relative limitations exist , user is still not satisfied by the performance of text representation , as well as many issues need further study.

In order to achieve the purpose of a deeper understanding of the text, this article build a cluster-based internet hot event finding method based on chinese sentence structure model. Firstly, this method analysis the document inorde to get sentence meaning element , calculate word weight to generate semantic vectors ; use semantic vector to hot event finding system , use clustering algorithm that combine single-pass clustering ideas , coherency hierarchical clustering and K-means clustering algorithm ,the accuracy is 75.2% . In addition, build a method of simplified representation of the event , extract the key points of events and event tags , key points of events accuracy is 58.9% . In addition, design and achieve a prototype system of hot eventsfinding and event simplified representation.

Key words: hot event; text clustering; similarity calculation; chinese sentential semantic structure model

目 录

图表索引

第1章 绪论

1.1 研究背景和意义

随着互联网的迅速发展，网络上的信息量与日俱增。2013 年 7 月 17 日，中国互联网络信息中心（CNNIC）在北京发布的第 32 次《中国互联网络发展状况统计报告》显示，截至 2013 年 6 月底，我国网民规模达到 5.91 亿，互联网普及率为 44.1%，网络新闻的网民规模达到 4.61 亿，较 2012 年 6 月增长了 6860 万人，年增长率为 17.5%；网民对网络新闻的使用率为 78.0%。网络新闻作为网民的基础应用，已成为网民获取新闻的主要渠道之一，使用率一直保持在较高水平，其使用率增长主要得益于以下几个因素：首先，在移动互联网时代，碎片化时间阅读新闻成为网民的主要活动之一；其次，随着微博、微信等应用的兴起，网民接触新闻的渠道增多，例如，微博对主要新闻事件的快速传播，形成热点话题，并联动主流新闻媒体进行传播，极大促进网民对网络新闻的接触度；最后，各类新闻媒体纷纷发力移动互联网，极大提高了手机网民对网络新闻的阅读频率。在大量网络新闻方便人们获取信息的同时，也给人们带来了阅读量太大的问题。由于网络新闻量很大，与一个事件相关的新闻往往散乱地分布在不同的时间段和地方，自动查找返回所需要的信息耗费大量的时间和资源，并且容易返回大量的冗余信息。并且，其中的相关信息并没有进行有效的组织，只是简单罗列，人们对某些新闻事件难以做到全面地把握，在人员和处理设备有限的情况下势必造成大量数据不能被完全处理。这样不仅浪费已采集的资源，而且一旦丢掉的数据中包含有重要价值的信息，就会造成无法弥补的损失。

话题检测与跟踪（Topic Detection and Tracking，简写为 TDT）正是在这种背景下产生的，它是一种检测新出现的话题、跟踪话题以及跨语言检测，起源于面向事件的检测与跟踪（Event Detection and Tracking，简写为 EDT）[1]。与 EDT 不一样的是 TDT 检测与跟踪的对象不是具体某一时间和地点所发生的事件报道，而是增加为具有更多相关性外延的事件，因此相关的理论与应用研究侧重点也从以前的对于事件的识别跨越到包含事件发生及其事件发展的新闻检测与跟踪。

TDT 是由美国国防高级研究计划局（DARPA）、马萨诸塞大学（University of Massachusetts）、卡耐基-梅隆大学（Carnegie Mellon University）和 Dragon Systems 公

司联合制定和设计完成的。来自这些机构的研究者经过一年的努力对 TDT 进行了前沿性的研究（1996～1997，Pilot study），包括测试目前大量应用于信息检索（Information Retrieval，简写为 IR）和信息抽取（Information Extraction，简写为 IE）等领域的技术是否能够解决 TDT 问题，以及设计和策划具有相同指标的评测标准。虽然很多 IR 和 IE 技术都能够在早期的 EDT 中使用，但是结果并不是很理想，错误率还是很高，不能满足该领域的需要。因此探索研究拓展后的 TDT 研究的特点，寻找适合 TDT 研究的创新性方法对该领域的发展具有重要意义，同时，对自然语言处理和信息处理领域具有重要意义。

在 TDT 的研究中，目前大量应用的主要是信息检索（Information Retrieval，简写为 IR）和信息抽取（Information Extraction，简写为 IE）两个领域的技术，主要涉及的是信息的检测和信息的采集与跟踪[1]。在信息检索系统中，首先需要用户输入自己的需求（Query），然后系统会对现有的信息进行检索，通过关键词匹配等方式获取与用户查询相关的信息，并返回给用户。信息是以现有的信息与用户需求的相关性为尺度进行组织、比较和反馈的，用户的需求是动态输入的。而在信息过滤系统中，首先用户要定义自己的需求，然后系统对动态变化的信息流进行检测比较，如果动态变化的信息流中有用户定义的需求则进行处理，如果没有则不处理。这种信息获取方法更关注用户需求跟踪的时间和空间发展，并实时返回给用户。用户的需求一般是静态的。这两个领域的系统处理信息的方式和技术与 TDT 有很多相似之处，因此很多 TDT 技术都是从这两个领域的信息采集、获取、跟踪技术中引进的，并且还取得了不错的成绩。现在发展比较迅速的个性化信息检索技术和自适应信息过滤技术，都给 TDT 的研究带来了深层次的发展。尽管如此，TDT 在许多方面与 IR 和 IF 存在差异，比如对于 TDT 的新事件检测任务（New Event Detection，简称为 NED），系统在没有任何用户需求的前提下，要对所有的新闻文本进行处理，自动识别和检测，这与信息检索系统中有用户需求这样的前提是不同的，因为完全没有先验知识。而且，话题检测系统一般要保存已经发现的话题信息，这就需要很大的存储。

传统的报纸、广播电视等媒体虽然仍保持大量的用户，发挥重要的传播作用，但是实时性比互联网要稍逊一筹。互联网以其独有的及时性，给大量网民实时提供国内外经济、政治、文化、军事和社会等重要新闻动态。因此，研究和发现互联网新闻热点事件已经成为一个很迫切的需求。新闻话题有其产生的自身规律，会有事件发生源、事件发生时间和地点、事件发展过程、事件消失等几个阶段，具有突发性、延续性等

特点。与传统新闻媒体固定时间发布新闻信息不同的是，互联网新闻可以随时发布新事件，并及时给出分析评论。与此同时，互联网新闻也有其自身的不足，其发布仍需要人工处理；并没有专人实时对同一事件下不同新闻报道进行归类处理，因此不能将报道同一事件的新闻精确区分和聚集。这就需要研究一种能够自动对互联网新闻进行处理的算法和技术，以及时、准确地检测互联网新闻中的热点事件，将属于同一事件的新闻进行聚合，并收集相关的报道，组织为一个有机整体。经过处理的新闻报道形成多个事件的大集合，这些事件自身会依然随着时间的发展而不断变化，网络新闻的内容也会不断更新。但是用户并不需要了解所有的热点事件，或者说用户只需要关注自己关心的事件即可，这就激发了用户个性化服务的产生。同时新闻经过收集处理后，同一事件内的新闻报道数量仍然很多，用户要阅读一个事件下多篇新闻报道，仍然有一定的困难，尤其是大事件发生后报道数量更是巨大。这与研究初始目的仍有一定的距离，因此，在得到事件的集合之后，需要对每一个事件进行进一步处理，得到更加简化的信息。

互联网是完全开放的，可以接受不同信息来源的信息，不仅限于媒体记者，任何个人、企业都可以参与。这给信息收集带来了前所未有的机遇，完全开放的入口带来了信息量的大爆炸，不管是信息的种类、数量都得到了激增。虽然这种开放性带来了非常大的发展，但是也伴随着出现了很多问题：大量虚假信息出现在网络上；大量包含色情的信息出现在网络上；大量包含暴力的信息出现在网络上；大量包含反动内容的信息出现在网络上。而这些内容给互联网用户带来了很大的影响，对互联网用户的身心健康、社会的稳定、国家的稳定造成了重大的影响。因此，政府和互联网用户需要一种针对这些事件和话题的检测和跟踪技术，及时发现并制止这种事件和话题的发生和传播。

发现并监控热点舆论话题有助于让公众及时了解实时的社会焦点，及时地发现社会舆情也能够为政府监管部门制定相关政策提供理论依据，对于构建社会主义和谐社会具有现实意义。加强信息的梳理工作，有效防止非法信息的扩散和传播，确保舆情信息的及时、全面和准确，认真倾听大众心声，随时关注社会上发生的焦点和热点问题，是政府制定路线、方针和政策的基础。社会舆论反映社会各阶层的民意，在全国上下努力构建和谐社会的今天，党中央和各级政府已经将社会舆论作为一件大事给予了高度关注坚持积极的进行信息监测、宏观舆论防范的方法，掌握舆情动态，通过智能监测分析的关键技术，在清理危害社会安全舆论的同时，形成正面的网络舆论环境，

也是促进社会稳定、建设和谐社会的重要手段。这也是本文研究热点事件发现的另一个重要目的和意义。

1.2 研究历史和现状

1.2.1 TDT 发展简史

TDT 是由美国国防高级研究计划局（DARPA）、马萨诸塞大学（University of Massachusetts）、卡耐基-梅隆大学（Carnegie Mellon University）和 Dragon Systems 公司联合制定和设计完成的。来自这些机构的研究者经过一年的努力对 TDT 进行了前沿性的研究（1996～1997，Pilot study），包括测试目前大量应用于信息检索（Information Retrieval，简写为 IR）和信息抽取（Information Extraction，简写为 IE）等领域的技术是否能够解决 TDT 问题，以及设计和策划具有相同指标的评测标准。虽然很多 IR 和 IE 技术都能够在早期的 EDT 中使用，但是结果并不是很理想，错误率还是很高，不能满足该领域的需要。因此探索研究拓展后的 TDT 研究的特点，寻找适合 TDT 研究的创新性方法对该领域的发展具有重要意义，同时，对自然语言处理和信息处理领域具有重要意义。

该领域的研究开始于 1996 年，美国国防高级研究计划局提出一种需求，即需要对新闻报道流自动分析，发现其中的话题。1997 年，做了一些前沿性的研究，主要是建立语料库。1998 年开始在美国国防高级研究计划局的支持下，美国国家标准技术研究所（NIST）开始举办 TDT 国际会议，并且每次会议都有评测，对参评的系统进行评测。主要有 IBM Watson 研究中心、卡耐基-梅隆大学、宾州大学、龙系统公司、BBN 公司、马萨诸塞大学、马里兰大学等著名大学和研究机构参与了国际会议和评测。1999 年国立台湾大学参加了 TDT 国际会议，2000 年香港中文大学也参加了 TDT 的评测会议。中科院计算所、北京大学等研究机构也开始研究 TDT，并取得了一些成果[2]。后来，由于资金等问题，TDT 相关评测没有继续下去。

在 TDT 国际评测的官方文档件，介绍了一些基本概念，包括话题（Topic）、事件（Event）、报道（Story）、主题（Subject）[3][4][5]。其中“话题”不是指一个大的“主题”，如美俄关系、科技发展，而是指有某一个主题的具体的事件，如索契冬奥会、习主席访美等。TDT 把包括一个核心“事件” 以及跟它有关系的事件的总和称为话题，即话题就是关于某个主题的所有新闻报道的综合。而“新闻报道”是指描述某个

事件某一方面的文本，在本文的研究中，可以把一篇新闻就看作是一个报道。

TDT 的国际评测会议把它分成了五个子任务：报道的切分 （ Story Segmentation）、新事件的检测（New Event Detection）、关联检测（Link Detection）、话题检测 （Topic Detection）、话题跟踪（Topic Tracking），其中话题的检测和话题跟踪是 TDT 国际评测会议的核心任务。2004 年又新增了层级式话题发现（Hierarchical Topic Dctection）任务，它是对话题发现任务的一个补充。话题发现任务的主要目的就是将输入中的新闻流进行自动分析，发现其中的不同主题，并将其进行区分，得到不同主题的事件子类，分析其整个过程，可以看出这个过程类似于聚类技术。首先，将同一主题的事件相关报道进行聚合，这就是聚类的过程。另外，事件继续发展，相关的后续报道也会划入到相关的事件中，这就跟聚类中的增量聚类相似。所以，话题检测实际上就是无指导的聚类和增量聚类。因此，在初始阶段，可以采用凝聚式层次聚类、K-means 聚类、单遍聚类等技术[6]。

国外的研究相对比较早，也比较成熟。1999 年 F.Walls 等人[7]就提出了比较成熟的算法，将信息处理领域的算法引入到 TDT 的研究中；2000 年 T.Leek 等人[8]也提出了不同的算法；Bruno Pouliquen 等人使用对数似然检验的方法衡量词项的权重[9]，分别处理不同来源、不同语种、不同大小的文档，也取得了不错的效果；而 IBM 公司用对称 Okapi 公式来衡量两个文档的相似性，在此基础上开发了一种基于两层聚类策略的话题发现系统[10]。2001 年密西根大学研究的 NewslnEssenee，对每一个新闻话题进行自动分析，生成摘要，是一种比较成功的商业应用开发，它的自动摘要是由 MEAD 生成的，而且它对句子的排序利用了基于质心的方法[11]。另一个典型的应用是 2002 年哥伦比亚大学研制的新闻推荐系统-News blaster，它将 TDT 和 DUC （ Document Understanding Conference）领域的研究相结合，分为话题发现部分和摘要生成部分[12]。Google 公司的 Google News 系统实时监测全球 4500 个新闻站点，将所有新闻站点的新闻报道首先按照报道的相似性进行聚合，然后依照用户的兴趣和爱好自动给用户推荐热点事件及其相关新闻报道组成的新闻报道集合[13]。

国内重要的研究机构和大学也意识到 TDT 领域的重要性，开始对 TDT 进行研究，相对于国外基于统计的研究方法，国内更加关注对 TDT 本身的特点进行研究探索，发现其中自有的规律，这几年慢慢成为一个重要研究热点。TDT 处理的新闻报道是面向现实世界中发生的事件进行描述，其中包含的对事件的时间、空间和事件发展的描述的准确性和区分性更加依赖于实体元素[2]；此外，事件的产生和事件的后续发展包

含了事件不同报道之间的时间序列关系，这就要求系统不能只考虑新闻的内容，还应该考虑新闻报道的时间特征，并对事件新闻报道的内容和时间相关性进行分析，得到其中的演化趋势[14]。在此基础上，国内的相关研究也面向建立结构化和层次化的话题模型进行了初步尝试。贾自艳首先将命名实体引入到 TDT 的研究中[16] ，其将文本内容的特征分别标记为人名、地名和主题等，并预先指定每种特征类别的价值系数，特征的最终权重为词频和其所属类别价值系数的乘积。骆卫华[17]和张阔对层次化话题模型进行了研究[18] ，前者首先基于时序关系对报道分组，然后进行组内自底向上的层次聚类，最后按时间顺序采用单路径聚类策略合并相关类；后者则面向报道全集建立层次化的索引树，树中第一层节点对应特定话题，而其子树则描述了该话题的层次体系，其建树过程基于输入的报道相对于树中各层次节点是否为新事件进行组织。这两方面的研究提高了话题发现的效率和话题检测的准确率，做出了很大的贡献，但在话题发展的规律和话题语义域的展现方面研究仍需要探索。国内互联网上出现的百度新闻[19]和 Google 中文资讯[20]，也属于 TDT 相关的应用系统。

由于国内的说法与国外略有不同，中国网络上一般以“事件”来讲述国外所说的“话题”，比如“2012 钓鱼岛事件”，因此，本文采用“事件”这一说法。相对于 TDT，互联网新闻热点事件的发现以及热点事件发展过程展现，属于应用性研究。首先，要对网络新闻进行聚类处理，得到热点事件，再对发现的事件进行排序就能得到热点事件列表，然后抽取热点事件事件发展过程中的关键点信息给用户。所以，研究内容和目标就集中在发现热点事件，并对热点事件发展过程进行抽取，进而进行展现。

1.2.2 语义研究

语义分析是伴随着语义研究的发展而发展的，语义研究可以粗略地划分为语文学（Philology）时期、传统语义学时期和现代语义学时期三个紧密联系的阶段。

1．语文学时期

在语文学时期，语义方面的研究工作主要是针对古文进行注释。在这一时期，人们只是把对语义的研究作为一种了解古代的典籍以及风俗、习惯、制度等的工具，而不是关于语义的独立、全面、系统的研究。

这个时期，我国的相关语义研究叫做“训诂学”，主要是从义理上注释古文，如对《春秋》、《诗经》、《周礼》、《札记》等经典古籍的注释工作都属于这个时代的成果。到了清朝，训诂学除了注重字形和字义的关系，还将古音引入了字义的研究，解决了

文字通假的问题，我国的训诂学达到了鼎盛时期。

2．传统语义学时期

从 19 世纪开始，语义研究进入到了传统语义学时期。语义的研究从只研究古代典籍，转向既研究古文，又研究当代语言作品；从只研究书面语到书面语、口语兼顾；从只着眼于实际工作，到既指导实际工作又从实际工作中总结出理论。这个时期的重要成果是形成了词汇学，并发展成为了语言学的一个重要分支。1893 年，法国语言学家 Michel Bréal 首先使用了“Sémantique（语义学）”这个术语，并在 1897 年的著作《Essai de Sémantique（语义学探索）》中，将语义学从词汇学中分离出来，形成了语言学的一个新的分支。

传统语义学从清朝末年传入中国，在其影响下，我国的学者们编撰了大量的语文性的工具书，包括《词源》、《辞海》、《新华字典》、《现代汉语词典》等等，并出版了大量的现代汉语、古代汉语方面的论文和著作。

3．现代语义学时期

现代语义学兴起于 20 世纪早期，从这个时候起，一些语义学理论模式的出现标志着现代语义学进入了蓬勃发展的阶段。在这个阶段典型的语义学理论包括结构语义学、解释语义学、生成语义学、Fillmore 语义理论、Chafe 语义理论、逻辑——数理语义学等[21]，如表 1-1 所示。

现代语义学在 20 世纪 70 年代后期才逐渐被介绍到国内，此后，我国的语言工作者们开始利用现代语义学理论对汉语进行研究，并针对汉语的特点提出适合汉语的语义学理论。例如，从 20 世纪 80 年代开始，马庆株以语义•语法范畴为中心，以分布和变换为标准，对汉语词类和句法结构进行了深入的研究，将语法意义与语法形式相结合，将变换分析和深层语义分析相结合，形成了多元化的研究方法；袁毓林在 1992 年提出了汉语配价理论[22][23]，将现代汉语名词分为有配价要求和无配价要求两类，从支配能力上对名词进行了划分，在 2002 年构建了汉语的论元语义角色体系，针对汉语常见的语义角色的特征进行了描述[24]；邵敬敏在 1993 年提出了“双向解释语法”（后称为“双向语法”）[25]，在 1997 年提出了“语法的双向选择性原则”[26]（2004 年改称为“语义语法”[27]），主张以语义为语法研究的出发点和重点。陈昌来针对汉语组成语义结构的语义成分[28][29][30]以及与动词相关的句法语义属性进行研究[31][32]，形成了语义平面中以动词为中心的语义结构认识体系；贾彦德先生在其著作《汉语语义学》[21]中全面分析了汉语中的语言规律，结合现代语义学理论模式，对汉

表 1-1 主要的语义学理论

语义学理论	创始人 / 代表人物	时间	说明
结构语义学	Saussure	20 世纪初	是现代语义学的第一个研究流派，提出语义场理论和义素分析法（也称成分分析法）。
解释语义学	Chomsky	1957 提出，1970 左右两次重大修改	创建转换生成语法，将句子的结构分为深层结构和表层结构两个层次，并将语法划分为四个型：无约束短语结构语法（0 型语法）；上下文有关语法（1 型语法）；上下文无关语法（2 型语法）；正则语法（3 型语法）。
生成语义学	Lakoff, McCawley, Ross	20 世纪 60 年代后期	认为语法和语义是不可截然分开的，语义是基础，是中心，它也有生成性。
Fillmore 语义理论	Fillmore	1966 & 1968	提出格语法，以语义为主、句法结构为辅，认为主语和宾语等只是表层中的关系，深层中的动词和名词的语义关系是格关系（case-relation），一切语言中都存在普遍的“格关系”或“格功能”（case-function）。 Fillmore 在 1966 年到 1977 年间共提出了 13 种格，分别是：施事格（Agentive）、工具格（Instrumental）、客体格（Objective）、处所格（Locative）、承受格（Dative）、感受格（Experiencer）、源点格（Source）、终点格（Goal）、时间格（Time）、行径格（Path）、受益格（Benefactive）、伴随格（Comitative）和永存格/转变格（Essive /Translative）。
Chafe 语义理论	Chafe	1970	提出 Chafe 语法（Chafe’s Grammar）以动词的语义特征作为分析的依据，并从不同的层次对语义单位进行分析与划分。
逻辑——数理语义理论	Montague	20 世纪 70 年代	提出 Motague 语法，利用数理逻辑的概念和表示方法，去解释和描述语言现象。

语语义场划分、句义成分、句义结构等进行了细致的研究，形成了汉语的语义理论体系。冯扬在其博士毕业论文中分析了现代汉语语义学，并且提出一种针对汉语句子层面的一种理论分析模型-汉语句义模型，通过对句子结构的分析，得到该句子要表达的意义。该模型即考虑到汉语中的语言规律又结合语义场理论进行句子结构分析，即能得到句子的语义结构也能分析得到句子中词语的语义成分。

1.3 研究内容和结构安排

1.3.1 研究内容

课题研究的主要内容包括：分析网络热点话题发现过程中用到的方法，结合机器学习方法和自然语言处理方法，分析网络热点事件的传播规律、发展趋势和事件简化表示的方法，对热点事件发展过程进行展现。本论文的主要工作包括：

1. 对文本表示模型进行改进。分析目前主流的文本表示方法，利用目前汉语语义研究中新提出的汉语句义结构模型理论，选择更能代表文本的词汇特征，使用更全面准确的信息，从而实现文本的更精确表示。
2. 建立基于聚类的热点事件发现方法。在改进的文本表示方法基础上，提出一种基于聚类的热点事件发现方法。
3. 提出热点事件发展过程展现的方法。在分析网络热点事件的传播规律和发展规律的基础上，提出以事件关键点为主干对事件发展过程进行展现的方法，同时，借鉴网络搜索词标签的方法，以事件标签的方式作为辅助进行热点事件展现的方法。
4. 构建原型系统。按照实验室的原型系统搭建规范设计并实现基于以上方法的热点事件发现原型系统以及事件发展过程展现原型系统。

1.3.2 结构安排

全文共分 5 章，具体内容安排如下：

第一章，绪论。主要介绍课题的研究背景、目的、意义及本文研究工作的定位，并分析 TDT 发展简史、语义发展简史、现状及目前该课题主要的研究方向和内容，最后介绍本文的研究内容。

第二章，相关理论基础和技术研究。属于课题研究理论基础的介绍，主要介绍目前主流的文本表示模型和目前使用比较多的文本聚类技术。从文本表示的准确性和适合使用的场景进行对比分析，对三大类文本聚类技术进行了简单的分析，并对各自的特点详细对比。

第三章，热点事件发现方法及事件发展过程展现。首先介绍基于向量空间模型和句义结构模型的改进文本表示方法，然后对热点事件发现和事件发展过程具体实现过

程进行详细说明，最后给出实验结果及评价分析。

第四章，原型系统设计与实现。设计实现基于以上方法的原型系统，详细阐述系统的设计思想、总体结构及功能需求，并对各个关键模块进行说明，通过实验验证系统的性能。

第五章，结束语。对论文的工作进行总结，并对进一步的研究工作进行展望。

第2章 涉及的相关知识基础

2.1 引言

热点事件发现起源于 TDT 研究，但是更加偏向于应用。随着网络信息资源的爆炸式增长，如何快速、准确的实现网络热点事件的检测和发现日益成为智能信息处理、机器学习和数据挖掘等领域的热点研究问题。热点事件发现，首先需要做的就是将文本转换成计算机可以处理的形式，这也是自然语言处理首先需要解决的问题；其次，基于同一事件下的文档内容都是围绕同一事件进行的，这类文档相似度较大这样的一种假设，发现讲述相似内容的新闻，并将同一事件的新闻聚到一起，实现热点事件发现的目的。目前，文本聚类就是要将相似的文本聚到一起，也就是利用文本聚类的方法可以实现热点事件发现的目的。本章主要介绍以上涉及网络热点事件发现方法的相关理论基础和相关工作，主要涉及到网络新闻的表示方法和热点事件发现中用到的文本聚类技术。

2.2 文本表示模型

文档（Document）由不同的表现形态组成的，具有一定的语义。为了从文档中获取相关信息，所以需要对文档进行分析。从文档中抽取特征词来表示文档信息，将文档转变成可以进行结构化处理的数据，文档分析包括文档的表示和词组的选取。由于关键词词组是能够较为确切体现文档主题的单位，处理非结构化文档将会采用不同的处理方法，与不同的信息处理技术相关联，所涉及的领域广泛，如文档分类、文档聚类分析、文档过滤、文档检测、知识发现、信息检索、关联规则挖掘等等领域。为了解决以上不同信息处理技术或不同的领域出现的问题，首当其冲的是关键词词组的提取，之后再进行其他的处理。总而言之，关键词词组的提取成为对文档进行相关处理的关键一步[33]。

文本表示包括两个问题：表示与计算。表示特指特征的提取，计算指权重的定义和语义相似度的定义。特征提取包括特征的定义和筛选，特征定义和筛选考虑以什么作为文本的特征，并不是所有的词和字都要求或者可以成为特征。特征的权重定义及特征结构上的相似度度量可以选取不同的模型，如向量空间模型、概率模型、语言模

型等。文本表示是文本聚类的第一步，该步骤的变化很多，对最终聚类效果的影响也不尽相同。文本表示本质上是对原始文本进行转换，使之在机器上可形式化描述、可计算。

本节接下来主要介绍信息检索和文本分析处理中经常用到的几个模型，这几个模型根据不同的理论假设推导、定义了不同的特征权重计算方法与语义相似度计算方法，是文本表示模型的重要组成部分。

2.2.1 布尔模型

最初，文本表示这一基础是来自信息检索领域，为了更快的能够检索到需要的文档，而首先提出的布尔模型。布尔模型是基于集合论与布尔代数之上的一种简单模型[34]，主要应用于信息检索中。在布尔模型中，文档D_j中索引特征t_i的权重$w_{i,j}$是二值的，即$w_{i,j} \in \{0,1\}$。一个文档被表示成文档中出现的特征的集合，也可以表示成为特征空间上的一个向量，向量中每个分量权重为 0 或者为 1，这种布尔模型被称为经典布尔模型。经典布尔模型中查询与文档的相关性只能是 0 或者 1，满足查询 query 中的所有逻辑表达式的文档被判定为相关，不满足的就被判定为不相关。

经典布尔模型只能用于信息检索中计算用户查询与文档的相关性，而无法利用该模型计算两个文档更深层面的相似度，无法在更多的文本处理应用中使用。在使用经典布尔模型基础上，研究人员又提出了扩展布尔模型（Extended Boolean Approach），重新定义了 And 与 Or 操作符成为多元操作符，使相关性可以成为[0,1]之间的数。Lee 在 SIGIR94 上的论文中分析了几种扩展布尔模型，比如 fuzzy set，Waller-Kraft，P-Norm 与 Infinite-One，认为 Salton 等提出的 p-norm 模型相对更优，在 p-norm 中，多元 And 查询的相似度定义为：

$$\mathrm{D} = \left\{\left(t_i,\ w_{\mathrm{di}}\right) \middle| i = 1..\mathrm{n}\right\}$$

$\mathrm{Query} = \left\{AND\left[\left(t_i, w_{qi}\right)\right],\ \mathrm{i} = 1..\mathrm{n}\right\}$，共有 n 个特征，$t_i$为特征，$w_{qi}$为查询中$t_i$的权重

$$SIM_D\left(D,(t_1,w_{q1})\mathrm{And}(t_2,w_{q2})\ldots\ldots\mathrm{And}(t_n,w_{qn})\right)$$

$$=1-\left(\frac{\sum_{i=1}^{n}\left((1-w_{\mathrm{di}})^{p*w_{qi}^{p}}\right)}{\sum_{i=1}^{n}w_{qi}^{p}}\right)^{\frac{1}{p}},\ (1\le \mathrm{p}\le\infty)$$

对 p-norm 扩展布尔模型稍微做一下改动就可以将它用于计算两个文本之间的相似度。计算两个文档的相似度之前先将两个文档进行转化，一个用 D 的定义表示出来，另一个用 And Query 的定义表示出来，然后就可以利用 p-norm 扩展布尔模型进行比布尔模型更深层次的相似度计算。

当参数 p 接近于无穷时 p-norm 模型就相当于经典布尔模型，而当 p=1 时 p-norm 模型就相当于向量空间模型。扩展布尔模型是基于集合论与布尔代数之上的一种文本表示模型，其表示与计算可以转化为向量来等价实现，是一种类向量的模型，表示文本简单方便。

布尔模型的优点主要有两方面：一是运算的速率比较快，二是操作简单容易实现。但它也有很多缺点，主要体现在以下三个方面：

1. 不容易控制检索结果。这是利用布尔模型进行检索时的一个相对较大的缺点。对于用户进行一个特定的查询时，它返回的结果可能是检索到众多文献或者文档，但是准确性不高，甚至也可能一篇可以参考的文档也检索不到。这是由于布尔模型检索匹配条件要求过于严格，从而过滤了想要检索到的文章，这就导致了布尔检索的漏检率较高。
2. 只能反映定性的给出相关与否的判断，而不能进行定量的相关程度计算，更不能反映特征项（k）在文献中的重要程度。也就是利用布尔模型的检索只能简单的进行判断，只能给出相关或者不相关这样的结论，不能计算两个文档之间的相似度大小。
3. 不能识别功能词，这也是利用布尔模型检索的缺点，比如“有关…”这类信息等。而这给习惯于自然语言检索的用户造成了很多不方便。

2.2.2 向量空间模型

1969 年，Gerard Salton 提出的向量空间模型 VSM（Vector Space Model ）是自然语言处理领域的一种重要模型，也是信息检索领域中比较经典的一种检索模型，并成功的应用于著名的 SMART 系统中。该模型的主要思想是：将每一个文档都映射为一

组规范化正交词条矢量组成的向量空间中的一个向量或者一个点[35]。

Salton 教授提出的向量空间模型简称为 VSM 模型（Vector Space Model），是信息检索领域中经典的检索模型。向量空间模型将文档表示成一个向量，向量的每一个维度表示文档的一个特征，这个特征既可以是一个字，也可以是一个词，还可以是一个 n-gram 或某个复杂的结构。通过对文档集合进行分析处理就可以得到文档中的这些特征。通常情况下用向量空间模型中的向量表示某个文档时，首先需要进行的就是对文档集合切分（中文分词、英文通过词的分界符识别单词）、去除停用词处理、英文词的词形还原或者提取词干（Stemming），经过若干个处理步骤后，基本上就可以得到一个词的集合，将这些词汇作为文档的特征。一个“空间”就这样产生的，它由所有的这些词汇构成，空间中的一个维度就是由一个词汇构成。每个文档都是由文档中的词汇构成的，而这些词汇又是构成向量空间的维度，因此，每个文档就可以用这个空间中的向量构成，每个向量的维度的权值又可以由词汇的权重来计算。即一个文档对应特征向量空间中的一个向量，对应特征空间中的一个点。表 2-1 VSM 模型中文本与空间的映射表说明 VSM 模型中文档与向量空间之间的映射关系。

表 2-1 VSM 模型中文本与空间的映射表

文档视角	向量空间模型视角
文档	向量或者空间中的点
词	空间中的一个维度
文档集合	分布在空间中的一组向量或者点集
整个词典	构成空间的各个维度
词的权重	空间中点的坐标值

在向量空间模型中，每个文档由组成其的特征词汇 t 来表示，不同的 t 在文档中作用不同，根据特征词汇对文档内容表达的重要性的大小每个特征词汇可以获得一个权重，计算这个权重的经典公式是 TF*IDF 公式。其中 TF 指的是 Term Frequency，表示 t 在文档 d 中出现的频率数，称为词频；IDF 指的是 Inverse Document Frequency，Salton 将 IDF 的计算公式定义为：$IDF_t = \log\left(\frac{N}{n_t}\right)$，N 表示文档集合中出现的所有文档数目，$n_t$表示整个文档集合中包含特征 t 的文档的总数目，称为特征的文档频率。IDF 反映特征词汇在整个文档集合中的分布情况，在一定程度上体现了该特征的区分能力，TF 反映特征在当前文档内部的分布情况，两者相结合的 TF*IDF 可以看成该特征在文档中的重要程度$w_{\mathrm{t,d}}$：$w_{\mathrm{t,d}} = TF_{t,d} * IDF_t$ 。

经过计算，文档就转化成为一个带有不同权重的向量，这些权重是该文档在向量空间中不同维度的坐标值，TF*IDF 公式是向量空间模型中经典的权重度量方法。

经典研究和应用中计算特征值的权重采用 Salton 的 TF*IDF 公式。但随着研究的发展以及 TREC 评测的推动，实际上向量空间模型还存在多个变种[36]，应用中需要针对不同变种的运行结果进行评测，选择最适合的表示方法。

著名的信息检索系统 SMART4 中针对 VSM 模型提出过一套关于特征权重计算变种的命名体系，在该命名体系中综合了 TF*IDF 中的多种变化，将文档某个特征的权重计算归结三个组成部分：TF 正则化因子、IDF 正则化因子和基于文档长度的正则化因子。因此，将文档的特征权重定义为：

特征权重 W= TF 正则化因子* IDF 正则化因子*长度正则化因子

各个组成成分的变种以及其命名[36]参考表 2-2、表 2-3 和表 2-4。在这些命名编码中，n 或者 x 表示 none，s 表示 square，f 表示 frequency，m 表示 max，l 表示 log，p 表示 probability。

表 2-2 基于词频 TF 的正则化因子

编码	公式	描述
s	tf*tf	原始 tf 的平方
b	1.0	Binary 方式，词出现时 tf=1，否则 tf=0
n 或 x	tf	原始词频
a	$0.5+0.5*\frac{tf}{tf_{max}}$	Augmented term frequency.公式中的 0.5 可以用其他数值替换，比如 $0.4+0.6*\frac{tf}{tf_{max}}$
l	$1+\log tf$	Logarithmic term frequency
L	$\frac{1+\log tf}{1+\log avr\_tf}$	Pivoted document length normalization 模型中应用的词频正则化因子

三个组成部分中每部分都有不同的计算方法，这些变化的组合构成了向量空间模型的不同变种。比如在研究中常用的一个权重度量组合 ltc[37]，由上面三个表中对应的编码公式进行组合后的权重公式为：

$$W_{\mathrm{t,d}} = TF_l * IDF_t * LEN_c$$

表 2-3 基于文档频率 DF 的正则化因子

编码	公式	描述
x 或者 n	1.0	不考虑 IDF 因子
t 或者 i	$\log\frac{N}{n_t}$	IDF 最初的定义，变形公式为$1+\log\frac{N}{n_t}$或者$\log\frac{N-n_t}{n_t}$
p	$\log\frac{N-n_t}{n_t}$	Probabilistic IDF factor
s	$\left(\log\frac{N}{n_t}\right)^2$	IDF 的平方
f	$\frac{1}{n_t}$	Collection frequency

表 2-4 基于文档长度的正则化因子

编码	公式	描述
n	1.0	不考虑基于文档长度的正则化因子
c	$\frac{1}{\sqrt{\sum_t w_t^2}}$	文档的欧式长度
u	$\frac{1}{\text{slope}*\#+(1-\text{slope})*\text{pivot}}$	Pivoted document length normalization , slope 和 pivot 都是参数，#表示该向量中特征的数目
s	$\frac{1}{\sum_t w_t}$	用权重之和作为正则化因子

计算过程为：

$w_{\text{t,d}}=TF_t*IDF_t=(1+\log tf)*\left(\log\frac{N}{N_t}\right)$，对于所有的 t

$$W_{\text{t,d}}=w_{\text{t,d}}*LEN_c=w_{\text{t,d}}*\left(\frac{1}{\sqrt{\sum_t w_{t,d}^2}}\right)$$

在实际应用中，某个特征的 DF 值在单个文本的处理过程中是无法得到的，只有整个语料集合处理结束后才可以得到，因此在文本解析处理过程中，IDF 的正则化因子往往采用 n，只在系统需要计算文本与文本相似度的时候才把 IDF 正则化因子引入，这样可以使文档解析处理速度更快、算法更加灵活。所以，应用系统实际处理过程中文档向量的扫描与产生采用 lnn 而不是 ltc。

在文档长度正则化因子上采用 c（ cos ）存在偏好短文档的问题。VSM 模型中，文档的语义越接近其 cos 相似度越高，但是 cos 相似度越高的文档间不一定语义越接近。一个特征比较多的长文档与一个特征比较少的短文档，长文档可能比短文档更接近某个语义，但是由于 cos 正则化的原因会造成在计算上短文档更加接近。Singhal 提出的 pivoted 正则化方法[38]是一个有效的克服这种问题的方法。

文档表示成向量以后，就可以通过计算空间中两个向量之间的几何关系来比较两个文档之间的语义距离或者语义相似度。计算两个向量之间的余弦夹角或者向量之间的欧氏距离就可以衡量两个文档之间的语义相似度，余弦夹角值越大，两个文档的语义相似度越大，余弦夹角值越小，两个文档的语义相似度就越小，文档就越不相似。在 VSM 模型中，通常就是采用这种方式计算两个文档之间的相似度的，经典的计算公式余弦夹角距离（Cosine Distance）如下：

$$\mathrm{sim}(di,dj)=\frac{\overrightarrow{di}*\overrightarrow{dj}}{|\overrightarrow{di}|*|\overrightarrow{dj}|}=\frac{\sum_{k=1}^{m}w_{k,j}*w_{k,j}}{\sqrt{\sum_{k=1}^{m}w_{k,i}^{2}}*\sqrt{\sum_{k=1}^{m}w_{k,j}^{2}}}$$

研究和应用中也会采用其他方法来衡量向量空间中两个文档向量的语义距离或者语义相似度，这些方法都可以作为参考，比如欧氏空间中的 Minkowski 距离：

$$\mathrm{Dis}(d_{i,}d_{j})=\left(\sum_{t}\left(\left|d_{t,i}-d_{t,j}\right|^{p}\right)^{1/p}\right)$$

Dice 距离，用两个文档之间共同含有的词条数目的两倍除以两个文档总的词条数目

$$\mathrm{Dice}(d_i,\ d_j)=\frac{2*|d_i\cap d_j|}{|d_i|+|d_j|}$$

$|d|$表示向量中的词条数，即非零元素数目。

欧几里德距离（Euclidean Distance）

$$\mathrm{Dist}(d_i,d_j)=\sqrt{(w_{i,1}-w_{j1})^2+(w_{i,2}-w_{j2})^2+\cdots+(w_{i,n}-w_{jn})^2}$$

明考斯基距离（Minkowski Distance）

$$\mathrm{Dist}(d_i,d_j)=\left(\left|w_{i,1}-w_{j,1}\right|^{h}+\left|w_{i,2}-w_{j,2}\right|^{h}+\cdots+\left|w_{i,n}-w_{j,n}\right|^{h}\right)^{\frac{1}{h}}\quad(h>0,h\in Z)$$

曼哈顿距离（Manhattan Distance）

$$\mathrm{Dist}(d_i,d_j)=\left|w_{i,1}-w_{j,1}\right|+\left|w_{i,2}-w_{j,2}\right|+\cdots+\left|w_{i,n}-w_{j,n}\right|$$

切比雪夫距离（Chebyshev Distance）

$$\mathrm{Dist}(d_i, d_j) = \max_{1 \le i \le n} |w_i - w_j|$$

2.2.3 句义结构模型

现代汉语语义学的句义结构理论认为，句义可以用具有一定逻辑的结构来表达，并从逻辑结构上将句义划分成话题和述题，而话题和述题又是由谓词和项组合而成，如图 2-1 所示。

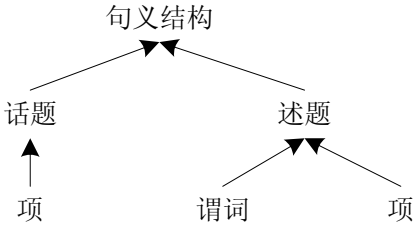

图 2-1 句义结构的组合关系

根据这一语言学理论，BFS-CTC 构建了层次化的句义结构模型[21]，该模型具有可扩展、可计算的特点，图 2-2 所示为句义结构模型的基本形式，包含的要素有：句义的类型、句义中的话题和述题、构成句义的各个成分、谓词时态信息、成分之间的组合关系等。图 2-3 中给出了 BFS-CTC 句义结构模型标注实例。

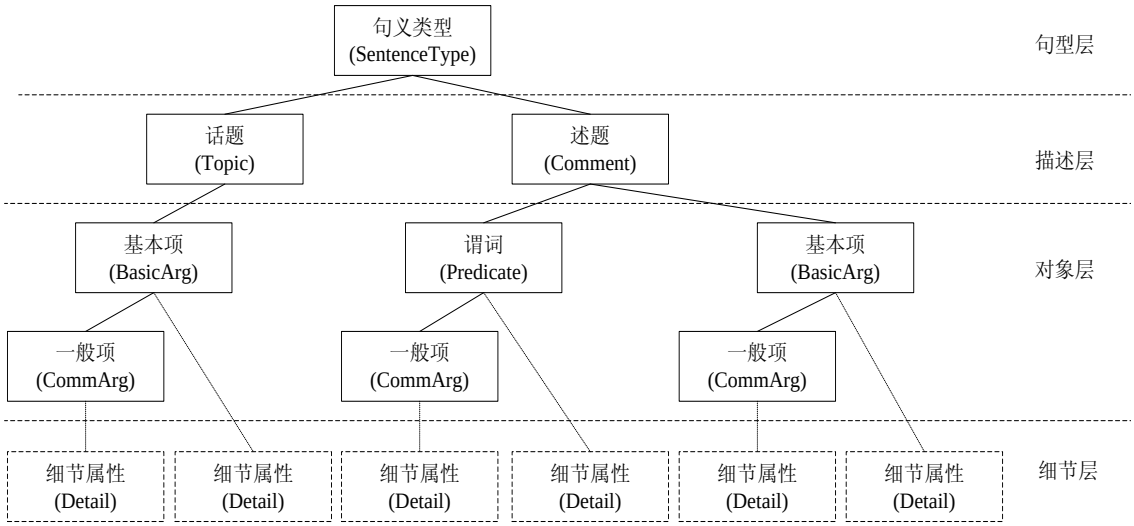

图 2-2 句义结构模型的基本形式

在对象层中，谓词和基本项直接与话题和述题相关联，构成句义结构的基本框架，

而一般项则是用来描述和限定谓词和基本项的（如限定时间、地点、范围等，当然也有一般项修饰一般项的情况）。

细节层是一个可扩展的层次，有些句义中的句义成分（对象）具有细节属性，有些则不存在细节属性，因此并不是所有的成分都需要描述细节属性，目前定义的需要描述的细节属性包括：谓词时态、修饰谓词的时间格的时制类型、时间格范围信息、空间格范围信息。

1．基本项

基本项有 9 种类型（9 种基本格），其中表主体的施事格和参与格都是指变化、动作、行为的发起者，所不同的是施事格反映该动作、行为由一个主体完成，而参与格反映由若干主体共同完成，因此本文中将施事格和参与格合并，统称为施事格；表客体的受事格与客事格分别需要与施事格和遭遇格搭配，但是施事格和遭遇格的区别在于所反映的是否是变化、动作、行为的非自主发起者，其客体都是该动作、行为的承受者，因此本文中将受事格与客事格合并，统称为受事格，既可以和施事格搭配，也可以和受事格搭配。

基本项类型的标记及说明如表 2-5 所示。

表 2-5 基本项类型标记及说明

基本项类型	标记	说明
施事格	AGENTIVE	变化、动作、行为的发起者。
遭遇格	RENCONTRE	变化、动作、行为的非自主发起者。
受事格	OBJECTIVE	变化、动作、行为的承受者。
结果格	RESULT	变化、动作、行为的所产生的结果。
主事格	SUBJECTIVE	与之搭配的谓词是表话题与一个对象之间的关系，或表话题的性质、状态、状况等。
说明格	EXPLAIN	协同谓词对主事格作出说明。
与格	DATIVE	表与事的项，是某些谓词所表行为、动作的间接对象。

2．一般项

一般项有 13 种类型（13 种一般格），由于材料格和工具格都是反映动作、行为等实现所运用的器具、使用的材料等，本文中将其合并，统称为工具格；而依据格、原因格、目的格三者都是反映谓词所表行为的依据、原因和目的的，因此将它们三者合

并，统称为根由格。

一般项类型的标记及说明如表 2-6 所示。

表 2-6 一般项类型标记及说明

一般项类型	标记	说明
范围格	RANGE	反映谓词所表的动作、行为等活动的领域范围。
时间格	TIME	反映谓词所表的行为、运动、情感、状况发生或持续的时间，时间格也有说明项的。
空间格	SPACE	谓词指的行为、运动等出现、发生的处所、方位、长度、高度、面积、体积等。
工具格	TOOL	施事籍以实现某些行为（不限于劳动）等的器具、手段。
方式格	MODE	指出现实现谓词所表行为等的方式、方法、途径。
基准格	STANDARD	测量或比较谓词所表情况的起算标准。
根由格	CAUSE	表实现或发生谓词所表运动、行为等的依据、原因、目的。
属格	ATTACH	也叫所有格，表示一个对象为另一个对象所有。
描写格	DESCRIPTION	项与项在语义上组合在一起，一个修饰另外一个，起修饰作用的项就是描写格。
同位格	PARITY	处于同位格的项与被它修饰的项指的是同一对象，但说明的角度不同，同位格更多处于修饰地位，另一个更多处于被修饰地位。
其它格	OTHER	无法归入到上述格中的其它一般格。

句义结构模型是根据汉语语义学而研究出来的一种语义研究模型，该模型从句子的层次对汉语进行分析，从逻辑上分为话题和述题，进而分为谓词和基本项，再进而分为一般项。具有如下几个特点：可理解性，体现为人可理解和计算机可理解两个方面；可计算性，体现在模型提供的语言特征能够帮助计算机实现对自然语言作品的应用，通过词汇层特征、语法层特征和句义层特征三个层次对计算机理解句子程度进行划分，而句义层特征属于深层的语言特征，它是由句义模型的各分析环节所提供的，并且反映了更多更深入的语言上的特性，如语义的层次、类型、角色、结构等等。由于深层次的理解是建立在浅层次理解的基础之上的，因此句义模型中的特征不仅仅包含了深层的句义层特征，同时也能获取到浅层次的词法层和语法层特征。如果在句义层理解了句子，那么词法层和语法层的特征也都能够获取到。根据特征的性质，可以将这些语言特征分为本征特征、结构特征和规则特征。其中本征特征指自身形式或性

质都不发生变化的特征；结构特征指反映句子在语法和语义上的结构、类型、成分的特征；规则特征是指从大量的句子的句法、句义结构，以及句义分析的结果中能够统计得到的规律性的特征。另外，还具有可裁剪性及释义与句义结构之间相互支持的特性。

综上所述句义结构模型的优点是建立在统计和数学的基础上，同时结合语义的角度进行文本分析和研究，是语义学与计算机学相交叉产生的模型，更加接近人理解文本的形式，更能适合计算机的理解；缺点是特征的计算方式还未得到更加深入的研究，对于属性提取和细节识别还有待解决，对于每个阶段能够获得到的语言特征还没有进行详细的分析，应用方面还有待突破。

2.2.4 LDA 模型

Latent Dirichlet Allocation （LDA）模型是 Blei 等在 2001 年提出的[39]，属于主题模型（Topic Models，是当前文本表示研究的主要范式）的一种。作为一种产生式模型，LDA 模型已经成功的应用到文本分类、信息检索等诸多文本相关的领域[39][40][41]。

LDA 是一个多层的产生式概率生成模型，是典型的有向概率图模型，是一种对文本数据的主题信息进行建模的方法，如图 2-3 所示，包含词、主题和文档三层结构。给定一个文档集合，LDA 将每个文档表示为一个主题集合，每个主题是一个多项式分布，用来捕获词之间的相关信息。在 LDA 中，这些主题被所有文档所共享；每个

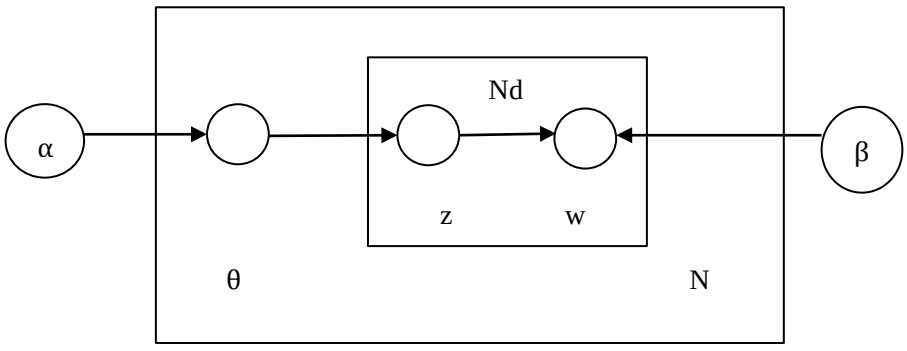

图 2-3 LDA 的图模型表示形式

文档有一个特定的主题比例。LDA 由文档层的参数(α, β)确定，α反映了文档集中隐

含主题间的相对强弱，β代表了所有隐含主题自身的概率分布。Θ代表文档中各隐含主题的比重，z 表示文档分配在每个词上的隐含主题比重，w 是文档的词向量表示。N 为文档集中文档个数，Nd 表示该文档的词总数。

LDA 主题模型涉及到贝叶斯理论、Dirichlet 分布、多项分布、图模型、变分推断、EM 算法、Gibbs 抽样等知识，是一种非监督机器学习技术，可以用来识别大规模文档集（document collection）或语料库（corpus）中潜藏的主题信息。它采用了词袋（bag of words）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。由于 Dirichlet 分布随机向量各分量间的弱相关性（之所以还有点“相关”，是因为各分量之和必须为 1），使得我们假想的潜在主题之间也几乎是不相关的，这与很多实际问题并不相符，从而造成了 LDA 的又一个遗留问题。

LDA 模型较之 LSI/PLSI 等模型有着突出的优点[43]：首先 LDA 模型是全概率生成模型，因此具有清晰的内在结构，并且可以利用高效的概率推理算法进行计算；再者，LDA 模型是通过无监督方法进行训练的，与训练样本数量无关，因此更适合处理大规模文本语料。近几年，LDA 模型、LDA 的扩展模型以及它们在自然语言和智能信息处理中的应用得到充分的重视和深入的研究。

2.3 文本聚类技术

常见的聚类分析的方法有五类：划分方法（Partitionning Method）、层次方法（Hierarchical Method）、基于密度的方法（Density-Based Method）、基于网络的方法（Grid-Based Method）、基于模型的方法（Model-Based Method），其代表算法请参见表 2-7。

2.3.1 基于划分的聚类方法

划分方法（Partitioning Method）的基本思想是：将处理对象集合 D 中的所有对象进行划分，形成一个平面的类组结构，并且满足以下要求：

1．单个类组中至少有一个数据对象；

2．单个对象必须被一个类组包含且只被一个类组包含。（在一些要求不严格的划

分技术中这一条可以放宽）；

3. 根据相异函数 F 来度量每个处理对象之间的相似性，目的是使处于一个类簇中的数据对象之间的相似性最大或者说最相近，而不同类簇中的数据对象要尽量不相似。

表 2-7 聚类算法示意图

基于划分的方法	K-均值	
	K-中心点	PAM、CLARA、CLARANS
	最近邻聚类	
	最大距离聚类	
基于层次的方法	分裂法	BIRCH
	凝聚法	GAC、CURE、Chameleon
基于密度的方法		DBScan、GDBScan、DBCLASD、OPTICS、FDC、
基于网格的方法		CLIQUE
基于模型的方法	统计法	COBWEB、CLASSIT
	神经网络法	SOM

划分方法的典型算法代表有 k–均值算法（K-Means）、k−中心点算法（K-Medoids）、最近邻聚类（Nearest Neighbor）、最大距离聚类（Max-Distance Clustering）等。

K-Means 算法[44]的具体过程如下：

1. 确定要生成簇的数目 k；

2. 按照某种标准产生 k 个聚类子中心作为聚类的种子$\mathrm{S}=\left(S_1,S_2,\cdots,S_j,\cdots,S_k\right)$；

3. 对处理对象集合 D 中的每个对象D_i，计算它们与种子S_j的相似度$\mathrm{sim}\left(D_i,S_j\right)$；

4. 选取其中相似性最大的种子$\arg\max_{S_j\in S}\mathrm{sim}\left(D_i,\ S_j\right)$，将$D_i$归入以$S_j$为聚类中心的簇$C_j$；

5. 再次计算 k 个簇的子中心；

6. 根据需要（中心为该簇内所有点的平均值）重复步骤 2,3,4,5，以得到最好的聚类结果。

K-Medoids 算法的具体过程如下：

1. 确定要生成簇的数目 k；

2. 按照某种标准生成 k 个聚类子中心作为聚类的种子$S = (S_1, S_2, \cdots, S_j, \cdots, S_k)$；

3. 对数据对象集合 D 中的每个对象D_i，计算它们与各个种子S_j的相似度$\mathrm{sim}(D_i, S_j)$；

4. 选取其中相似性最大的种子$\arg\max_{S_j \in S} \mathrm{sim}(D_i, S_j)$，将$D_i$归入以$S_j$为聚类中心的簇$C_j$；

5. 再次计算 k 个簇的中心；

6. 根据需要（中心为该簇内接近聚类中心的数据对象表示）重复步骤 2，3，4，5 若干次，以得到最好的聚类结果。

PAM[45]（Partition Around Medoid）是最早提出的 K-Medoids 算法之一。PAM 方法的思想是：首先随机地选取 k 个数据对象作为簇的子中心，然后对于每一个簇的子中心点，PAM 方法尝试用所有 n − k 个非中心点代替它，从而决定最终用哪个点来代替原来的子中心，实际上就是要在簇的所有点中选取所需代价最小的子中心点作为新簇的中心点。

CLARA[46]（Clustering LARge Application）与 CLARANS（Clustering Large Application based upon RANdomized Search）算法是从 PAM 演化出来的。CLARA 的基本思想是：从所有的数据对象中随机选取出若干个样本实行 PAM 算法，最终选择其中代价最小的作为结果。

CLARANS[46]是 CLARA 方法的改进版，主要提高聚类结果对于样本的依赖性以及聚类结果的伸缩性。其基本思想是：随机从 k 个中心点中选取出一个点，从 n−k 中随机选取出一个点去代替它，如果代替成功则重新开始。如果尝试很多次（参数由人设定）都没有发现最优的结果，就结束。

Nearest Neighbor 算法的具体过程如下：

1. 随机选取一个文件D_i，以该文件为中心建立一个类簇$C_1 = \{D_i\}$；

2. 确定距离（相似度）的阈值 T；

3. 选择下一个要处理的数据对象D_i，当所有数据都计算完，则结束；

4. 计算当前对象D_i与当前所有类簇的夹角（相似性），得到距离最小的类簇及距离值 d，如果 d < T，将该对象合并入该簇，更新类簇的中心点；否则，以该对象为中心新建一个类簇C_j。

5. 返回步骤 2。

Max-Distance Clustering 算法的具体过程如下：

1. 选取一个文件D_i，以该样本为中心建立一个簇$C_1 = \{D_i\}$；

2. 在其它的对象样本中，选取与C_1距离最远的对象，并新建一个类簇C_2，记录该最远距离为 m_dis；

3. 确定距离（相似性）的阈值 T ，一般 m _dis 为 T 的倍数；

4. 在其它的对象样本中，计算该样本与现有的所有簇的最小距离d_i；如果无法再新建簇，转到步骤 6；

5. d = max $\{d_i\}$，如果 d > T，新建一个以该对象为中心的簇，返回步骤 d；如果无法在新建簇，转到步骤 6；

6. 把剩下的对象样本分配到距离最近的那个中心簇。

2.3.2 基于密度的聚类方法

基于密度的聚类算法（Density-Based Clustering Method）的主要思想是：将簇看成是数据空间中被低密度区域划分开的高密度区域[47]。密度是指单位空间内的对象数，簇外部的密度要比簇内小。基于密度思想的聚类算法又被称为局部聚类（Local Clustering）。它的优势在于可以发现任意形状的簇。并且这种方法数据对象集合只处理一遍，所以也被称为单遍扫描聚类（Single Scan Clustering）。

基于密度的聚类中比较有名的算法有 DBScan（Desity-Based Clustering）、OPTICS（Ordering Points to Identify the Clustering Structure）。DBScan 的基本思想是：检测单个数据对象相邻的数据对象数目，并与用户设定的阈值进行对比，当超过该阈值时就判定为该数据对象周围密度足够。

OPTICS 是基于密度思想的一种聚类算法，其特点是并不对类簇进行划分，而是生成一种聚类结构。该聚类结构对所有数据对象进行排序，并对数据对象周围的密度分布进行计算。另外，GDBSCAN（Generalized DBSCAN）、DBCLASD（Distribution Based Clustering of Large Spatial Database）、FDC（Fast Density-based Clustering）等算法都是对 DBSCAN 的进一步扩展与优化。

2.3.3 基于层次的聚类方法

层次方法（Hierarchical Method）的基本思想是：将所有数据对象聚集为小簇，

同时这些小簇再次聚合，生成大簇，最终形成树形结构。树形结构的每个节点是一个簇。根据树的形成是自底向上还是自顶向下，可以分为凝聚式层次聚类和分裂式层次聚类。

基于层次思想常见的算法有：GAC（Group-Average Clustering）、BIRCH[48]（Balanced Iterative Reducing and Clustering using Hierarchy）、CURE[49]（Clustering Using REpresentatives）、Clameleon[50]等。

凝聚式算法的具体过程如下：

1. 将数据对象集合 D 中的每个对象D_i看成一个小簇$C_i = \{D_i\}$；

2. 计算每对簇(C_i, C_j)之间的相似度 sim(C_i, C_j)；

3. 选取具有最大相似度的簇对(C_i, C_j)，将C_i和C_j合并为一个新的簇$C_k = C_i \cup C_j$；

4. 重复以上步骤，当达到终止条件（最终聚合为一个大簇或者用户设定的终止条件、阈值等）时停止。

凝聚式层次算法的示意图如图 2-4 所示。

GAC 与 CURE 都属于凝聚法。GAC 采用分治的策略。其基本思想如下：

1. 将数据对象集合 D 中的每个对象D_i看成小簇$C_i = \{D_i\}$；

2. 把所有的小簇分成大小为 m 的桶；

3. 对于每个桶内部进行凝聚式层次聚类；

4. 根据每个桶的结果进行衡量，最终达到最好聚类结果，否则返回步骤 2。

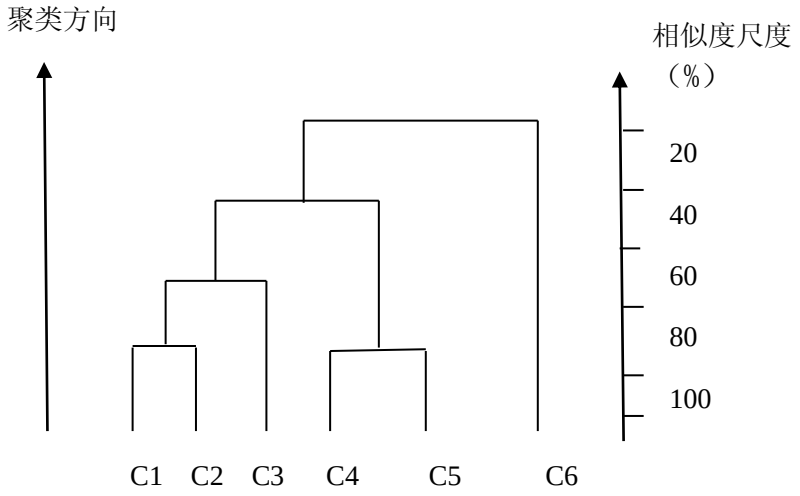

图 2-4 凝聚式算法示意图

CURE 算法对簇的表示方法进行改进，使用固定数目的数据对象作为代表点来表示簇，而使用单个数据对象、单个文本或者所有数据对象的平均值来表示簇。代表点的产生方法是：选取类簇中固定数目的数据对象，这些数据对象具有比较好的分散性，按照收缩因子向类簇收缩。CURE 针对大规模数据采用取样与划分相结合的方法降低计算复杂度，其基本思想是：

1. 从数据对象集合中选取一个随机对象 S；
2. 将样本 S 分割为一组划分；
3. 对每个组进行划分局部聚类；
4. 随机选择其中的孤立点。删除增长缓慢的类簇；
5. 对所有簇进行聚类处理，代表点根据收缩因子向簇的中心移动；

用相应的簇标签标记数据。

分裂式算法的具体过程如下：

1. 将数据对象集合 D 中的所有对象看成一个类簇；
2. 按照某种标准对现有的大类粗进行拆分；
3. 重复上述步骤，直至数据对象集合 D 中的每个数据对象D_i都是一个簇。

分列式层次算法的示意图如图 2-5 所示：

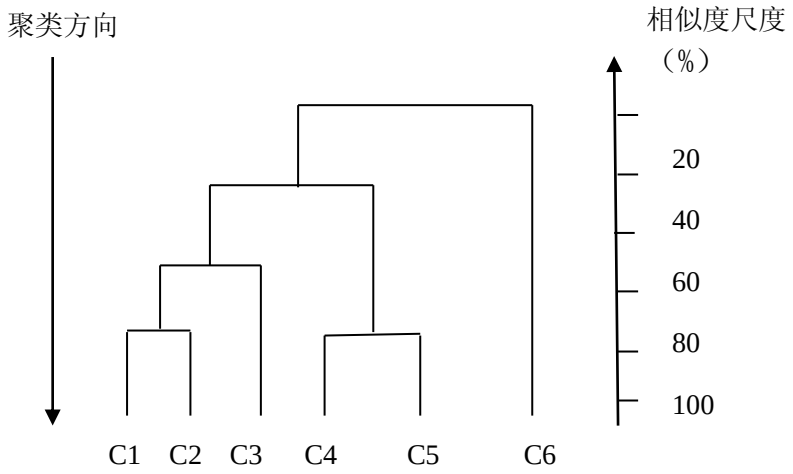

图 2-5 分裂式层次聚类算法示意图

BIRCH 算法属于分裂式层次聚类算法，其基本思想是：两个步骤来处理，第一步，根据所有数据对象生成聚类算法的特征 CF（Clustering Feature）树。在 CF 树中，每个节点代表的是该节点表示的类簇的特征。在 CF 树生成过程中两个参数比较重要：

平衡因子 B 和阈值 T。B 决定 CF 树非叶子节点能含有的最大孩子数，T 表示子簇的最大直径。具体生成步骤如下：对所有的数据对象，将每一个对象插入到最相似的节点中，如果 CF 树节点超过 B 个孩子或者节点阈值大于阈值 T，就将这个节点一分为二。BIRCH 算法的第二个阶段是利用任何聚类方法对第一阶段产生的结果进行聚类。

Chameleon 是层次聚类算法中的一种，主要使用动态模型实现，主要模型构建方法是：首先根据相似度构建一个矩阵，然后从该矩阵中生成 K−最近邻图。在该图的基础上进行划分，将不同的处理对象进行聚类，得到相对较小的子类。最后对这些子类再次进行动态合并，不断地计算，最终找到合理的结果类簇。

2.4 小结

本章首先介绍了目前文本表示的几种模型，详细比较了几种文本表示模型的优点与不足。向量空间模型结合了数学和统计学的优势，利于计算机计算两个文本向量之间的相似度，而汉语句义结构模型能够提供语义域的特征。因此，本文的研究采用向量空间模型与汉语句义结构模型相结合的方法改进文本表示的模型。其次，介绍了文本聚类技术，分类别进行详细的比较，将各个聚类方法的具体实现过程予以对比。本文采用两层聚类的方法，以K-means方法为主进行热点事件发现研究。

第3章 热点事件发现及简化表示方法

3.1 引言

本章针对本文研究的两个方面进行详细介绍。首先是热点事件发现，它与TDT研究中的话题发现具有相似的任务，因此可以借鉴TDT研究中的相关成熟算法，但话题发现与本文事件发现还是有很大的区别，话题发现所处理的语料都是经过人工挑选得到的，结构单一，质量高，噪音数据少。

首先介绍热点事件发现技术发展及主要方法，并对其进行分析，然后在分析的基础上提出热点事件发现算法的原理、原理中涉及的主要过程及主要使用的算法，最后通过实验分析，确定文本表示中汉语句义结构模型中所用的特征，验证事件发展关键点和事件内容简化表示的有效性。

3.2 主要技术和方法分析

热点事件发现是从新闻网页中获取感兴趣的信息，而网页数据是半结构化的，具有复杂性等特点。目前，大多数网页都有如广告、版权、导航链接等噪声，影响Web信息处理的工作质量，因此需要对网页进行分析，快速准确地清楚网页中的噪声内容并从中提取感兴趣的内容，删除不需要的干扰噪声，转化为易于分析处理的形式。在本节首先介绍新闻事件的表示方法，其次介绍事件检测方法。

3.2.1 新闻事件的表示

在2.2节中从自然语言处理领域的角度介绍了文本表示的主流方法，而热点事件发现对于网络新闻事件的处理也是自然语言处理领域中的内容。在热点事件发现这个应用中，采用向量空间模型对文本进行表示的算法和系统占大多数，只有少数几个系统是采用n-gram模型或者LDA模型来对新闻事件进行表示。

所有基于统计的模型都有这种缺陷：不能够有效地区分同一事件下的不同新闻报道主题。以“恐怖袭击”话题为例，即包括阿富汗自杀式炸弹袭击事件，也包括911恐怖袭击事件。在新闻内容上，这些事件的新闻报道中都会不断地出现“恐怖分子”、“自杀式”、“袭击”、“死亡”等词汇，并且这些词汇在新闻报道中出现的频率也都比

较高。因此，基于统计的模型中，这些词汇却成为事件模型的主要特征，从而不能有效地区分同一事件下的不同报道主题。在此基础上，有人提出用命名实体（Named Entities）方法来解决这个问题，将新闻文本表示成三种向量空间，分别为全集特征向量空间、仅有NE的特征向量空间和不包含NE的特征向量空间，经过验证，这种方法可以提高事件之间的区分能力。

在向量空间模型中，每个文档由组成其的特征词汇 t 来表示，不同的 t 在文档中作用不同，根据特征词汇对文档内容表达的重要性的大小每个特征词汇可以获得一个权重，计算这个权重的经典公式是 TF*IDF 公式。其中 TF 指的是 Term Frequency，表示 t 在文档 d 中出现的频率数，称为词频；IDF 指的是 Inverse Document Frequency，Salton 将 IDF 的计算公式定义为：$IDF_t = \log\left(\frac{N}{n_t}\right)$，N 表示文档集合中出现的所有文档数目，$n_t$表示整个文档集合中包含特征 t 的文档的总数目，称为特征的文档频率。IDF 反映特征词汇在整个文档集合中的分布情况，在一定程度上体现了该特征的区分能力，TF 反映特征在当前文档内部的分布情况，两者相结合的 TF*IDF 可以看成该特征在文档中的重要程度$w_{\mathrm{t,d}}$：$w_{\mathrm{t,d}} = TF_{t,d} * IDF_t$ 。

经过计算，文档就转化成为一个带有不同权重的向量，这些权重是该文档在向量空间中不同维度的坐标值，TF*IDF 公式是向量空间模型中经典的权重度量方法。

经典研究和应用中计算特征值的权重采用 Salton 的 TF*IDF 公式。但随着研究的发展以及 TREC 评测的推动，实际上向量空间模型还存在多个变种[36]，应用中需要针对不同变种的运行结果进行评测，选择最适合的表示方法。

根据2.2节中对文本表示模型的讨论和分析，以及热点事件发现研究的特点，本文的事件发现采用向量空间模型和汉语句义结构模型相结合的文本表示方法。为了把报道表示成向量，又能尽量使用语义域的信息，本文仍然采用词作为向量特征，并根据汉语句义结构模型提供的特征来提高向量表示新闻的准确性。

确定了新闻的向量表达方式以后，一篇新闻报道S就可以表示为一个向量：$S_i = \left(w_{1,\mathrm{i}},\ w_{2,\mathrm{i}},\ \ldots,\ w_{n,\mathrm{i}}\right)$，$w_{k,\mathrm{i}}$表示第i个新闻报道中第k个词的权重，n是新闻报道中词的总数。权重方案的选择是本文改进文本向量表示的关键，最简单的计算方法是将词在文档中的出现频率作为权重，但是在实际运用中往往采用其它更有效的计算方法。计算词权重的重要信息包括词频（term frequency）、逆文档频率（inverse document frequency）、汉语语义格（semantic case）。

词频是词t在新闻文档d中出现的次数，即$\mathrm{freq}(\mathrm{d},\mathrm{t})$，词频矩阵TF（d,t)度量词t与给定文档d之间的关联度，Cornell SMART系统使用的公式计算词频权重：

$$\mathrm{TF}(\mathrm{d},\mathrm{t})=\begin{cases}0 & \text{如果}\mathrm{freq}(\mathrm{d},\mathrm{t})=0\\ 1+\log(1+\log(freq(d,t))) & \text{其他}\end{cases}$$

逆文档频率是词t的缩放因子或重要性。如果词t出现在许多文档中，由于其区分能力减弱，所以它的重要性也降低。例如，如果词“数据库系统”出现在很多的数据库系统会议的研究论文中，那么可能会不太重要。根据Cornell SMART系统，IDF（t)由公式定义：

$$\mathrm{IDF}(\mathrm{t})=\log\frac{1+|d|}{|d_t|}$$

其中d是文档的集合，d_t是包含词t的文档的集合。如果$|d_t|\ll|d|$，词t将有很大的IDF缩放因子，反之亦然。但相对于区分度来说，聚类技术更多个关注相似性，因此，在聚类的应用中IDF并不那么重要。

汉语语义格是汉语句义结构模型中对象层属性，其中，谓词和基本项直接与话题和述题相关联，构成句义结构的基本框架，而一般项则是用来描述和限定谓词和基本项的（如限定时间、地点、范围等，当然也有一般项修饰一般项的情况）。词的句义成分不同，在文档篇章语义表达时所占的权重就不一样，这一点与向量空间模型中根据词的频率来确定词的权重有共通之处。因为本文提出不仅仅要考虑词出现的频率，还要考虑词的句义成分来确定词的权重，句义成分的权重如表 3-1 所示，一篇文章中词 t 的句义频率权重为 Cfreq（t,d）。Cfreq（t,d）的计算方法为，当 t 出现时，首先判断 t 的句义成分，然后将其相加。

$$\mathrm{Cfreq}(\mathrm{t})=t_{（1，\ \mathrm{casei}）}+t_{（2，\ \mathrm{casei}）}+\cdots+t_{（\mathrm{m}，\ \mathrm{casei}）}$$

其中 m 是当前文档中词 t 出现的次数，Casei 代表的是句义成分权重，i 取值为（1,2,3）。

表 3-1 句义成分权重示意图

句义成分	权重
基本格	Case1
谓词	Case2
一般格	Case3

根据上文所述，采用如下公式计算词的权重：

$$\mathrm{CTF(d,t)} = \begin{cases} 0 & \text{如果 } \mathrm{Cfreq(d,t)} = 0 \\ 1 + \log(1 + \log(Cfreq(d,t))) & \text{其他} \end{cases}$$

仍然采用$\mathrm{IDF(t)} = \log\frac{1+|d|}{|d_t|}$计算逆文档频率。那么，完整的语义向量空间模型中将$\mathrm{CTF(d,t)}$和$\mathrm{IDF(t)}$组合在一起，形成 CTF-IDF 度量：

$$\mathrm{CTF-IDF} = \mathrm{CTF(d,t)} * \mathrm{IDF(t)}$$

在新的计算权重方法基础上，采用向量空间模型的思想，构建成的语义向量空间模型作为本文热点事件发现算法的文本表示方法。

3.2.2 事件检测

事件检测的主要任务是检测新事件并收集后续相关事件报道。通常，事件检测模块的检测原理主要是根据同一事件内新闻报道的相似性进行新闻聚类的方法，即在线采集和整理新闻报道数据流，根据采集的新闻是否与现有事件相关判断是不是新事件，如果不相关则作为新事件处理，否则将其合并。

事件检测中的主要方法来自于 James Allan 和 Yiming Yang，他们构建了一个在线识别系统（OL-SYS）检测新闻报道流中出现的新事件报道。其中，按照时间顺序流入该在线系统的新闻报道首先与已经存在的事件进行相似性匹配，并与阈值进行对比。超过该阈值的新闻报道则根据该新闻报道建立新的事件模型，否则将其合并到已经存在的事件模型中。在这种系统基础之上，后面又有一些相关的研究，主要是针对两方面进行改进，一是探索研究更好的文本表示方法；一是探索新闻事件时间特征等。事件检测的方法都是采用聚类等方法，将同一事件的新闻报道进行聚合，然后判别是不是热点事件。

事件检测作为热点事件发现的关键环节，直接决定热点事件发现的准确性和全面性，是事件发现的要点。采用新闻流对比事件模型的方法，具有实时性，但是其采用的 singgle-pass 聚类思想只对文档进行一次相似度的比较就决定一个类簇，不够全面和准确，不能处理任意形状的类簇，不能改变类簇的数目，不能对同一事件不同主题的文档进行归类，存在改进的空间。

3.3 热点事件发现原理

3.3.1 算法框架

本文前面已经对热点事件发现的几个关键阶段进行了详细研究，综合前面几个模块提出了热点事件发现算法，原理图如图 3-1 所示。

其中，预处理是指对新闻集进行分词、去除停用词、计算语义向量等几个步骤；相似度计算是指对生成的向量进行余弦度量计算；聚类包含 single-pass 聚类和凝聚式层次聚类与 K-means 聚类相结合的聚类算法；排序是指对发现的热点事件根据热度进行排序。

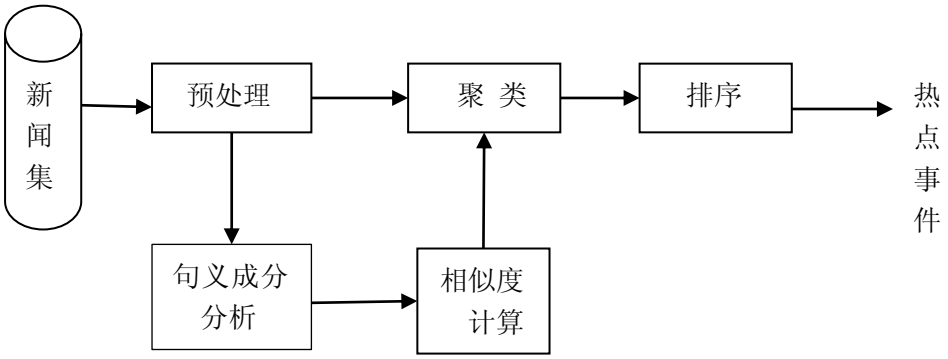

图 3-1 热点事件发现原理图

3.3.2 预处理

热点事件发现是应用性比较强的算法，同时具有个体差异，结合考虑语料的质量、覆盖面以及权威性、实时性，本文选取了国内知名的、流量大的新闻门户网站—新浪网（http://sina.com.cn）、网易新闻（http://news.163.com）、搜狐网（http://sohu.com）和新华网（http://www.xinhuanet.com）作为语料的采集源。该网站新闻具有数量多、更新快、网民浏览量大等特点，且基本上覆盖了大众关注的各个方面，非常适合作为本文热点事件发现的数据来源。

因为本文分析的是网页的正文内容，所以网页采集下来后，要对下载的新闻网页进行正文抽取。一个网页的内容可分为两类，一类是提供给浏览器用的标记信息，另一类是提供给用户阅读的信息。在提供给用户阅读的信息里不但包含新闻的标题和正文内容，也有导航信息、广告信息以及相关链接等无用信息，所以要先把网页中的各

种标记信息和无用信息去掉，保留新闻标题和正文内容。本文选取实验室中收集的网络新闻语料作为实验语料，该语料收集来自以上四个网站，去除掉无用信息之后，以文本形式保存。

西方语言与中文语言最大的不同之处在于在语句（或从句）内词汇之间是否存在分隔符（空格）。因此，要进行中文信息处理的诸多应用，如翻译、文献的索引、分类、过滤以及词频统计等，就必须做好第一步——汉语词汇的切分，也就是分词。分词的自动处理，是在字符串匹配的原理下进行的。到目前为止，在诸多的文献中已经对各种分词方法进行了非常详尽的探讨，所有文献几乎都对其着重点和分词的速度方面进行了阐述，也为分词的精度以及分词的规范上进行了详细地研究和解决。本文采用中科院ICTCLAS分词系统[51]，该系统分词速度单机996KB/S，分词精度98.45%，支持Linux、FreeBSD及Windows系列操作系统，支持C/C++、C#、Delphi、Java等主流的开发语言。

对文本进行分词以后，发现很多文章中的词对于篇章语义表达没有太多的意义，对于文档和用户需求价值不大，需要将这些词项从词汇表中去除，比如连词、叹词、介词、标点符号等。这些词成为停用词（stop word），在新闻中占了一定的比例，对于降低文档的维度有一定的帮助，表 3-2给出了一部分停用词。停用词按照词性区分，主要分为四大类：一是连词；二是叹词；三是介词；四是标点符号。当然还有其他的一些词性的词属于停用词的范畴。本文采用的停用词表结合百度停用词表、四川大学停用词表、中科院停用词表等多个版本，综合考虑进行合并删除等处理后得到，能够涵盖大部分停用词。

表 3-2 部分停用词

连词	叹词	介词	标点符号
和	啊	被	，
以及	哎呀	把	。
并且	哈哈	让	！

3.3.3 相似度计算

文本聚类中，文本表示和文本相似度是最关键的部分。不同的相似度度量方法会出现不同的聚类结果。

在文本聚类中，有三个相似性衡量（或者距离衡量）标准：

1．单个文本和单个文本之间的相似性衡量；

2．单个文本与某个文本簇之间的相似性衡量；

3．某个文本簇与另一个文本簇之间的相似性衡量。

不同的聚类算法应用了这三种类型的度量方法中的某种或者某几种。比如，凝聚式层次聚类算法（Hierarchical Agglomerative Clusetering，文献中常称为 HAC）中使用了 1、2、3 三种；在 K-Means 算法中就用到了 1、3。

这三种距离中不同的定义方式会对结果产生不同的影响，衍生出算法的不同变种。在 2.2.2 节中介绍向量空间模型时，介绍了 cos 相似度等多种方法。而在语言模型中，文档与文档之间的距离可以采用 KL 距离来计算。

衡量文档簇与文档簇之间的相似度，进行簇的选择，实现簇与簇之间的合并或者拆分。文档粗与文档簇之间的相似度也存在多种定义，常见的有 Single link，Average link，Complete link 、质心法、离差平方和、代表点。下面列出常见的簇与簇之间相似度的度量方法。

1．Single link（最小距离）：$\mathrm{D}(C_1, C_2) = min\{D(X,Y)|X \in C_1, Y \in C_2\}$

2．Complete link（最大距离）：$\mathrm{D}(C_1, C_2) = \max\{D(X,Y)|X \in C_1, Y \in C_2\}$

3．Average link（类平均）：

$$\mathrm{D}(C_1, C_2) = \mathrm{average}\{D(X,Y)|X \in C_1, Y \in C_2\} = \frac{1}{|C_1| * |C_2|} * \sum_{X \epsilon C_1} \sum_{Y \epsilon C_2} D(X,Y)$$

4．质心法：使用簇中所有文档向量的算术平均作为簇的中心向量。两个簇用两个中心向量来表示，采用 2.2.2 节中计算文档与文档相似度的方法来计算簇与簇的相似度；

5．离差平方和（也称 Ward）：用类内方差来度量，对簇 C，其类内方差定义

$\mathrm{S}(C) = \sum_{X \in C} \left(X - \overline{X}\right)^T * \left(X - \overline{X}\right)$，

因此$\mathrm{D}(C_1, C_2) = S(C_1 \cup C_2) - S(C_1) - S(C_2)$

6．代表点：与质心距离相近，从该簇中选取若干个最能够代表该簇的点作为代表点，使用这些代表点来衡量簇跟簇之间的相似性，忽略那些对于簇的表现能力差的点，在多次迭代的过程中保持簇的凝聚性。

从上面六种簇与簇距离的定义中，可以用相似度进行类比定义。采用 cos 相似度的 Average link 计算与质心点相似度是等价的。除这六种定义以外，实际应用中还可

以采用其他的定义，比如在针对一个簇 C 定义该簇的熵、内聚度等指标，与 Ward 距离类似，用假定合并后新簇的指标值减去原先两个簇的指标值，得到一个新的距离度量。

文本与文本簇相似度的度量可以参考文本与文本之间相似度或者文本簇与文本簇之间相似度的度量公式来计算。把要计算距离的文本看成一个只包含一个文本的文本簇，那么可以采用度量文本簇与文本簇之间的度量公式来度量；如果簇用中心向量表示，那么又可以当成两个文本向量的距离度量，采用文本之间的度量公式来度量。

在本文研究的热点事件发现系统中，新闻事件表示是采用的语义向量空间模型，因此计算文本与文本相似度采用向量空间模型中计算文本之间相似度的方法。在 2.2.2 节中介绍向量空间模型时已经介绍过多种计算相似度的方法，文本采用的是 cos 余弦夹角值的方法。

3.3.4 聚类

上文中分析了文本聚类技术，不同种类的算法有不同的优缺点。尽管聚类算法体系的划分并不统一，但是分成基于划分的聚类和基于层次的聚类两大类的分类方法基本得到共识。这两种类别也是在文本聚类分析中应用最广泛的类别，层次聚类分析中分为凝聚式层次聚类和分裂式层次聚类，虽然复杂度较高，但是普遍认为效果比较好，但是有个缺点是一个簇一旦被合并或者被分裂就不能再修改。

划分式聚类不同于层次式聚类，其类别结构简单，一般没有清晰的层次关系，算法通过不断的迭代来完成样本数据的最优分配，但其本质是一种贪心算法，容易陷入一种局部最优解。这类算法的典型代表是 K-means 聚类算法、KNN 聚类算法等。K-means 聚类算法简单、高效，一般只需迭代少量几次就能达到收敛。然而这种算法在开始时，需要预先指定 k 值大小和一个初始的划分，从而聚类结果的好坏直接受到 k 值和初始划分的影响[52]。

基于以上考虑，本文采取三种聚类算法和思想相结合的方法，首先，采用 single-pass 聚类算法思想进行噪声去除，对输入系统的新闻文本进行相似度计算后，与阈值T_c（经验值 0.02）相比较去除噪声。超过这个阈值T_c的认为与其他新闻报道相关，低于这个阈值T_c的认为与所有新闻报道都不相关，不是构成事件的报道，不参与下面的聚类处理。然后采用凝聚式层次聚类，获取高质量的初始类中心和类数目，然后用这些质量比较高的类中心为 K-means 的初始类中心进行 K-means 聚类。算法流程

如下：

1．计算每个文本之间的相似度并与阈值T_c比较，留下超过阈值T_c的报道；

2．将所有的点形成一个簇；

3．从现有所有的簇中选择距离最近（或者最相似的两个簇），进行合并；

4．如果达到终止条件，凝聚式聚类结束，否则返回步骤 2；

5．凝聚式层次聚类产生的 k 个簇进行初始化，对于每个数据对象，计算该对象与 k 个簇中心的距离，选择相似度最大的簇将该对象分入该簇；

6．再次计算 k 个簇的中心点，中心点为此簇内所有点的算术平均值；

7．如果簇变化不大或者满足退出条件（达到最大迭代次数），那么结束聚类，否则返回 6。

3.3.5 事件排序

话题关注度是从众多新话题中筛选出热点话题的重要依据。性别不同、职业不同、年龄不同的人有不同的价值观和和世界观，对同一个话题的关注方面、关注程度当然也有所不同。文献[14]和[53]将热点话题定义为一定时期内多个新闻来源同时报道的话题。从社会学角度来看，热点话题的产生与群体行为的选择、社会大众的关注有密切关系[54]。前一方面的关注可以称为是媒体关注度，后一方面的关注可以称为是用户关注度，本文定义媒体和用户同时关注的新闻话题为热点话题。媒体关注度侧重于从新闻发送方的角度来考察新闻话题受到的关注的程度，用户关注程度对于热点话题的形成也有重要作用，它直接决定了一个话题在大众中的传播范围和在社会中的反响程度。用户的关注程度体现在用户的浏览行为，如打开一篇新闻，阅读一则报道的速度，对一则新闻发表评论。用户浏览行为从一定程度上反映了用户潜在的心理状态信息[55]，比如用户对新闻报道的关注程度有多高。目前对用户浏览行为的研究主要用于发现单个用户行为习惯或个人兴趣，从而建立用户模型，进行个性化服务。

由于本文主要研究探讨热点事件发现算法，并不是做上线应用系统，而且网络浏览量和评论数目等信息一般不容易获取，本文中并不过多的考虑用户浏览行为，即不计算用户关注度，仅计算媒体关注度。因此，在排序的时候，本文仅仅以新闻报道的数目多少作为事件热度多少的判别标准。主要基于以下两个考虑：一是热点事件的特征，热点事件具有报道数目多的特点，所以报道数目越多的事件，事件热度也相对较高；二是用户关注度数据不易获取，在热点事件推荐中应用较广，而在热点事件发现

中应用并不广泛。综上所述，本文以事件报道数目的多少作为热度多少的判别标准，这个是热点事件发现中比较重要的一个环节，但是却并不是最复杂的一个环节，而且热度公式可以随时在算法中进行修改。

3.3.6 实验及分析

3.3.6.1 实验目的和数据源

为了验证本文提出的语义向量的有效性并确定句义结构成分权重的有效性，本文设计两个计算语义权重的实验；为了验证本文提出的热点事件发现设计一个实验。

搜狗文本分类语料库来源于Sohu新闻网站保存的大量经过编辑手工整理与分类的新闻语料与对应的分类信息。其分类体系包括几十个分类节点，网页规模约为十万篇文档，该语料即可用于分类。该语料库统计的意义：提供一个较大规模的标准中文文本分类测试平台，为各种从事中文文本分类工作的研究者提供一个标准的较大规模的研究平台。该语料库提供多种版本供下载使用，分mini版、精简版、完整版，分类编码如表 3-3所示。

表 3-3 搜狗文本分类语料库

分类编码	文本类别	分类编码	文本类别
C000007	汽车	C000008	财经
C000010	IT	C000013	健康
C000014	体育	C000016	旅游
C000020	教育	C000022	招聘
C000023	文化	C000024	军事

表 3-4 热点事件数据源

事件列表	文档数	事件列表	文档数
美国通用汽车破产	50	秦俑一号坑发掘	52
苏丹红事件	53	索马里海盗	51
香港回归十周年	52	香港小姐选拔	51
艳照门事件	51	英国报销门	50
许宗衡事件	55	邓玉娇事件	51

由实验室收集的网络新闻语料可以作为热点事件发现、事件关键点抽取和事件标签实验的数据，由于新闻语料是人工收集标注的，所以符合大众关注度和媒体关注度

两个标准，从中挑选出 10 个热点事件 516 篇文档作为实验数据源。数据分布如表 3-4 所示。

3.3.6.2 实验环境和条件

实验在Windows操作系统，Intel（R） Core（TM）2 Duo CPU，1.97GHz，2.00GB内存，300G硬盘的环境下进行，编程语言为C++，编译环境为Microsoft Visual studio C++ 2010。

应用软件：ICTCLAS分词软件，汉语句义结构模型自动构建系统，凝聚层次聚类和K-means聚类开源软件。

3.3.6.3 评价方法说明

热点事件发现算法中，本文以单个事件的准确率、召回率、F值及整体准确率作为评价指标。假设事件A，其准确率计算方法为：

$$\text{precision} = \frac{\text{被识别为事件A且识别正确的文档数}}{\text{被识别为事件A的文档数}}$$

召回率计算方法为：

$$\text{recall} = \frac{\text{被识别为事件A且识别正确的文档数}}{\text{实际为事件A的文档数}}$$

F值计算方法为：

$$\text{Fscore} = \frac{2 \times \text{precision} \times \text{recall}}{precision + recall}$$

最后综合所有事件的识别结果，得出算法的整体准确率：

$$\text{precision}_{全} = \frac{\text{识别正确的文档数}}{\text{文档总数}}$$

3.3.6.4 实验过程和参数

为了验证语义向量在文本聚类中的的有效性并确定句义成分的权重值，设计了如下实验。

1. 利用搜狗实验室的聚类数据源，选取六大分类（汽车、军事、教育、财经、

文化、体育）共600篇文档作为实验数据源；

2. 将采用经典的K-means聚类方法，作为句义成分权重实验方法；
3. 对句义成分权重Case1、Case2、Case3进行步进选择。设置Case1、Case2、Case3的初始值为1，因为基本格在汉语句义结构模型中担当的任务最为简单，因此本文以基本格作为基准。

首先，进行Case1的权重实验。利用语义向量和K-means聚类方法进行聚类实验，计算聚类准确率。Case1的值从1开始，依次增加0.5，比较不同Case1值得到的准确率，以准确率最高者作为Case1的权重值；其次，在Case1的基础上进行Case2的权重实验，从1开始，依次增加0.5进行聚类实验，比较不同Case2值得到的准确率，以准确率最高者作为Case2的权重值；然后，前面实验的基础上以0.1为步进上下浮动0.5进行权重实验，通过聚类实验，比较不同的Case1和Case2得到的准确率，以准确率最高者作为最终的权重值。

在实验过程中发现，抽取的600篇文档作为聚类实验是不合适的，通过分析发现，该语料库是建立在文本分类基础上（主要有六类汽车、军事、教育、财经、文化、体育）的，在类与类之间相似度很小，但是类内文档之间的相似度也会出现很小的情况。因此，实验结果具有不稳定性。基于以上这种情况，本文使用3.4.1节中表 3-3的热点事件发现数据源进行语义权重实验，采用上面所述的实验方法，最终得到实验结果Case1的值为2.1，Case2的值为1.7，Case3的值为1。

为了验证热点事件发现算法的有效性，设计了如下实验。

1. 对表 3-4所示的热点事件发现数据源进行预处理，并输入到汉语句义结构模型自动构建系统中，进行句义成分分析；
2. 根据步骤1生成的汉语语义成分，计算每个词的语义成分权重值，并将每篇文档构建成语义向量；
3. 根据3.3.3节的算法计算各个文档之间的语义向量相似度；
4. 利用3.3.4节的聚类算法对步骤3中的语义向量进行聚类得到事件列表，并按照文档数目进行排序。

3.3.6.5 实验结果及分析

为了评价本文提出的热点事件发现算法的有效性和创新性，文本复现了以经典向量空间模型作为文本表示方法的热点事件发现算法，采用相同的聚类思想和方法，并

表 3-5 热点事件发现实验结果

事件列表	准确率（%）	召回率（%）	F值（%）
美国通用汽车破产	78.4	80.0	79.2
苏丹红事件	76.5	73.6	75.0
香港回归十周年	81.3	75.0	78.0
艳照门事件	76.7	67.7	70.2
许宗衡事件	66.7	69.1	67.9
秦俑一号坑发掘	69.8	71.1	70.5
索马里海盗	74.2	84.3	78.9
香港小姐选拔	77.4	80.4	78.8
英国报销门	79.6	78.0	78.7
邓玉娇事件	72.0	70.6	71.3
加权平均	75.2	74.7	74.9

表 3-6 向量空间模型实验结果

事件列表	准确率（%）	召回率（%）	F值（%）
美国通用汽车破产	76.5	78.0	77.2
苏丹红事件	74.5	71.7	73.1
香港回归十周年	79.2	73.1	76.0
艳照门事件	72.0	60.7	65.9
许宗衡事件	59.6	61.8	60.7
秦俑一号坑发掘	69.8	71.2	70.4
索马里海盗	74.1	84.3	78.8
香港小姐选拔	77.4	80.4	78.4
英国报销门	69.4	68.0	68.6
邓玉娇事件	72	70.6	71.2
加权平均	72.4	71.9	72.1

进行了相同数据源的实验。经典向量空间模型为数据源进行实验，实验结果如表 3-6所示。热点事件发现实验结果如表 3-5所示。对实验结果进行分析，每个事件的准确率是不一样的，有的高有的低，这是因为聚类结果不同造成的。通过两次实验结果进行对比，可以看出虽然每个事件的准确率都不同，但是对比加权平均以后的准确率、召回率和F值，加入了语义成分的向量模型，能够提高热点事件发现算法的准确率，平均提高3个百分点，这说明本文提出的文本改进表示方法使有效的。

同时，加入了句义分析之后，算法的工作量增加，复杂度也增加了，时间复杂度也增加。这对于实时性要求高的系统来说，是需要进行改进的。本文进行的实验，都是在给定语料库的基础上进行的，并没有从新闻媒体上进行实时新闻下载，也就没有考虑实时性。虽然热点事件不是新闻推荐，不用实时处理，但是如果要进行上线应用的开发，就必须考虑实时性，考虑用户的体验。

3.4 事件简化表示

3.4.1 事件发展关键点抽取

经过聚类以后，对事件进行热度排序，我们将得到一个按照热度降序排列的热点事件列表，列表中的每一个事件都对应有一个文档集合，该集合由所有报道该事件的新闻文档所组成。接下来的工作任务是以友好的方式将热点事件的信息呈现给用户，本文借鉴和改进了 TDT 中的相关研究和一些成功系统的显示模式[56]，用事件发展关键点、事件相关词群等方面全方位地阐述热点事件。

通过对往年热点事件的观察和分析，得出热点事件是传播广泛，一段时间内都受到大众的关注，并且蔓延持续一段时间的。这一事件是在某一时间段内的受关注程度明显高于该事件平时受关注的程度，这样的事件就是“热点事件”。有的热点事件的被关注程度总有一个明显的提升过程，有的热点事件是每个时间段的被关注程度都不低，但在一段时间的被关注程度有大幅度的提高，然后达到最高点，如 2012 年的“高房价”。有的热点事件是新出现的，所以前期被关注程度很低，甚至接近于零，当受到大众的关注后达到被关注程度的最高峰，如 2013 年的“嫦娥登月”。

虽然每个热点事件发展过程是不一样的，但是还是可以发现一些规律，热点事件的产生、发展演变并不是平稳过渡的，是呈现高低起伏的。一方面是事件的报道频率出现变化；一方面是事件的时间属性；一方面是事件的关注人数。由于本文研究的热点事件发现算法是建立在搜集网络新闻的基础上，那么并不是说一个热点事件新闻集内所有新闻报道内容都有新信息的出现。一方面，因为新闻采集自不同的网络源，可能会出现不同网络新闻源出现相同内容报道的情况；另一方面，众多新闻报道并不是因为事件发展出现变化才会出现。因此，在经过排序后，不需要将所有的新闻文档提供给用户，这样不方便用户了解事件的整个发展过程。

因此本文定义事件发展关键点为事件发展过程中能够代表事件发展的新闻报道，

即用户通过阅读时间发展关键点的新闻报道，就可以了解整个事件的概要信息，而不需要阅读全部新闻报道。

二十世纪50年代末，Luhn开创了自动文摘领域，自动文摘技术逐渐发展起来。多文档自动文摘的研究工作最早开始于二十世纪80年代，当时的研究还不具有普遍性，主要是面向科技论文的摘要，而真正的任意域的多文档文摘的研究是在1997年开始的[57]。目前，在自动文摘领域，获取各种类型文摘的主流研究方法有两种。一类是基于抽取的方法，另一类是基于泛化生成的方法。他们的主要区别在于摘要结果中的句子是否来自于原文档。

自动文摘领域从开始起步的时候就广泛采用了基于抽取的方法。至今，这种方法仍然是绝大多数文摘技术采用的主流方法，只是在具体的技术上有所改进和发展。比如从最初的依靠文本浅层特征来抽取句子发展到现在的采用更为复杂的句子抽取策略，从开始的抽取重要句子到现在的抽取段落这样更详细的文摘单元。还有很多的诸如基于知识库的机器学习方法[58]，基于文本修饰结构分析的方法[59]以及基于文档主题结构分析的文摘方法[60]等等。基于抽取的文摘研究方法实现起来容易，高效快捷，而且实用于全领域，不受领域限制。但是由于它只分析文档表层特征，很难理解文档的语义信息。这样就可能造成生成的文摘质量不稳定，在全面性、连贯性上表现不佳。

本文提出的事件发展关键点与多文档自动摘要技术有相通之处，但是却与多文档自动摘要的粒度不同，多文档自动摘要抽取的是句子，粒度小，特征不足，所以准确率不高。又因为，一般的热点事件报道集合中文档的数量非常多，计算复杂度也非常大，因此，本文并不采用这种自动摘要的方式来抽取事件的摘要，但是却利用相近的原理。

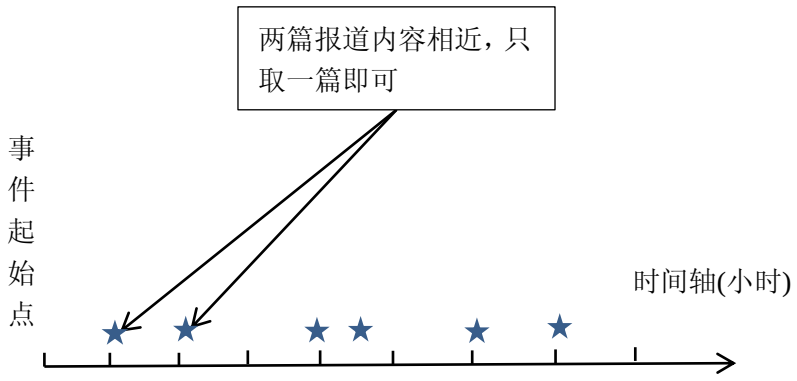

图 3-2 事件发展关键点示意图

本文假设两篇报道相似度高则代表两篇报道内容相同或相近，不需要都呈现给用户，只需要将两篇报道不同角度不同方面的新闻呈现给用户。因此，首先将相同事件的新闻报道按照时间顺序进行排序，这样就是将事件发展过程安插在时间轴上；通过计算不同新闻报道之间的相似度，抽取相邻时间点之间相似度小的新闻报道作为事件发展关键点，如图 3-2所示（图中星星表示新闻文档）。

本文设定抽取事件发展关键点时相似度大小的参考值（reference value）为0.6，超过该值的文档则认定为内容相近，不必都呈现给用户；未超过该值的文档则认定为内容不相近，需要呈现给用户的文档。

3.4.2 事件标签抽取

本文这样定义事件的标签，一般由对事件信息贡献量最多的若干词组成，由于事件的文档集合包含了事件的全部信息，所以事件的标签必定出自于事件的文档集合。本文可以借鉴多文档关键词抽取的相关成熟算法，但本文定义的事件标签与多文档的关键词还是有很大的区别。事件的发生都带有明显的时间顺序，用户在了解一个事件时，除了关注其发生了什么事外，也更加关注事件的各种参与者，以及事件发生的时间、地点、机构等，这些命名实体有些可能由于出现频率较低，而被排除作为事件的关键词。因此，本文的事件标签抽取方法要在相关成熟算法的基础上作必要的优化和改进。

事件的标签是与事件核心内容最相关的文档集合，结合3.2.1节中词汇权重计算方法来分别计算事件内所有词的权重和术语的权重，并按权重值降序排列，选取前10个词，作为事件的标签。

上文中对事件发展关键点进行抽取以后，将热点事件关键点的新闻报道简化呈现给用户，让用户从具体真实的报道中了解事件。同时，本文还提供一种事件标签的方式，从事件的文档集合中抽取出与事件相关的关键词集，这些关键词除了包含有信息量丰富的内容词外，还包含有事件发生的人物、时间、地点、机构等命名实体，通过这些关键词及其相关信息，如热点事件发展关键点，往往就已经能够把握住事件的关键信息。

事件的相关标签具体的计算方法描述如下：

1. 利用3.2.1节中权重计算方法来计算事件发展关键点相关文档中的所有词项，统计出排名前10的词项；

2. 利用汉语句义结构模型，统计时间格、空间格词项，统计出各自排名前5的词项；
3. 以步骤1和步骤2中抽取的20个词作为事件的标签。

得到事件的标签后，我们还可以根据需要丰富相关词群的信息，如利用信息检索技术，查询事件的文档集合中哪些文档出现了关键词，或者统计关键词随时间的频率变化曲线。用户通过这些附加信息，也更能充实对事件的了解。

3.4.3 实验及分析

3.4.3.1 实验目的和数据源

为了验证事件简化表示中事件发展关键点的有效性和准确性，设计一个实验，实验数据源如3.3.6.1节中所述热点事件发现数据源。

3.4.3.2 实验环境和条件

如3.3.6.2节中所述。

3.4.3.3 评价方法说明

在事件发展关键点中，本文以单个事件发展的关键点的准确率、召回率、F 值及整体准确率作为评价指标。其准确率计算方法为：

$$\text{precision} = \frac{\text{被识别为事件发展关键点且识别正确的文档数}}{\text{被识别为事件发展关键点的文档数}}$$

召回率计算方法为：

$$\text{recall} = \frac{\text{被识别为事件发展关键点且识别正确的文档数}}{\text{实际为事件发展关键点的文档数}}$$

F值计算方法为：

$$\text{Fscore} = \frac{2 \times \text{precision} \times \text{recall}}{precision + recall}$$

最后综合所有事件发展关键点的识别结果，得出算法的整体准确率：

$$\text{precision}_{\text{全}} = \frac{\text{识别正确的文档数}}{\text{文档总数}}$$

3.4.3.4 实验过程和参数

为了验证事件发展关键点的有效性，设计如下实验。

1. 采用3.3.6.1节中描述的数据源，对其中的10个热点事件分别进行事件发展关键点的人工标注；
2. 在3.3.6.2节中描述的实验环境和条件下，按照时间顺序计算同一热点事件文档集中的文档之间的相似度；
3. 设置初始参考值为0.6，抽取相邻文档之间相似度未超过阈值的文档作为事件发展关键点。

其中，参考值0.6为经验值。

3.4.3.5 实验结果及分析

事件发展关键点实验结果如表3-7所示，以人工标注的事件发展关键点作为正确的事件发展关键点。

表 3-7 事件关键点抽取实验结果

事件列表	准确率（%）	召回率（%）	F值（%）
美国通用汽车破产	54.2	60.4	57.1
苏丹红事件	65.6	62.3	63.9
香港回归十周年	61.9	60.0	60.9
艳照门事件	50.0	52.4	51.2
许宗衡事件	59.1	63.7	61.3
秦俑一号坑发掘	67.8	64.5	66.1
索马里海盗	62.2	66.3	64.3
香港小姐选拔	50.7	59.2	54.6
英国报销门	58.5	50.3	54.1
邓玉娇事件	67.8	63.6	65.6
加权平均	59.8	60.2	60.0

3.5 小结

本章首先介绍了热点事件发现的几个步骤，预处理中经过分词、去除停用词，生

成语义向量空间模型；计算各个向量之间的相似度，通过两种聚类方法相结合的方式对新闻报道进行聚类，产生事件集，并进行排序；通过研究热点事件发展的自然规律，提出事件发展关键点来简化表示事件发展过程，同时辅助以事件的标签方式来对事件进行展现。

第4章 原型系统设计与实现

4.1 引言

根据之前确定的方案，本章设计并实现一个基于向量空间模型和汉语句义结构模型的改进文本表示模型基础上的热点事件发现原型系统和热点事件发展过程展现原型系统，同时通过实验，验证该系统的有效性、稳定性，实现系统的可靠性设计和对大批量数据资源的处理能力。

4.2 系统总体设计

4.2.1 技术路线和设计原则

设计并实现一种热点事件发现原型系统。该系统以网络搜集的新闻作为语料。首先，通过中科院ICTCLAS2010分词系统对新闻语料进行分词处理；然后将分词后的文件进行去除停用词处理，同时，将分句后的文档输入到句义结构模型自动构建系统中，生成文档句义成分；计算词的权重并构建成语义向量；经过计算语义向量之间的相似度，再利用两次聚类对文档集进一步判断，再经过排序后，输出热点事件；最后，抽取事件发展关键点和事件标签。

本系统采用VS2010编译环境，采用C++语言开发，各个模块之间均相互独立，通过文件进行数据的交互，保证了系统的可扩展性。同时，充分考虑系统的容错性和对异常错误的处理能力，实现系统的可靠性设计和对大批量数据资源的处理能力。

4.2.2 目标和功能需求

系统能够实现语义成分的自动识别、改进文本表示模型的自动获取、热点事件自动发现、事件发展关键点和事件标签自动抽取的功能。

1．语义成分的自动识别。能够对输入系统的经过分句的汉语句子进行分析，自动将编过号的句子进行汉语句义结构模型自动构建，实现对每个汉语句子的句义结构和句义成分的自动识别，并以词为最小单元的形式保存在文件中。

2．改进的文本表示模型即语义向量自动获取。能够对经过语义成分分析后的文

件进行词的权重分析，生成语义向量数据。

3．热点事件发现。使用语义向量进行相似度计算，然后通过凝聚式聚类算法进行凝聚聚类，生成K个簇；利用凝聚式聚类算法产生的K个簇初始化中心点，输入到K-means聚类算法中，实现热点事件发现，并进行事件排序。

4．事件发展关键点及事件标签自动抽取。实现对每个热点事件发展关键点的自动抽取，并在事件关键点文档中抽取出现频率高词汇和时间地点相关词汇作为事件标签。

4.2.3 系统的总体架构

热点事件发现和事件展现系统的原理图如图 4-1所示。首先，输入新闻语料集，进行预处理，分词并去除停用词；然后将分句后的文档进行句义成分分析，并根据识别后的句义成分计算词的权重，构建语义向量；之后对语义向量进行聚类分析，得到事件列表并进行排序；最后，对识别出的热点事件，进行事件发展分析，抽取出事件发展关键点和事件标签，实现事件简化呈现给用户。

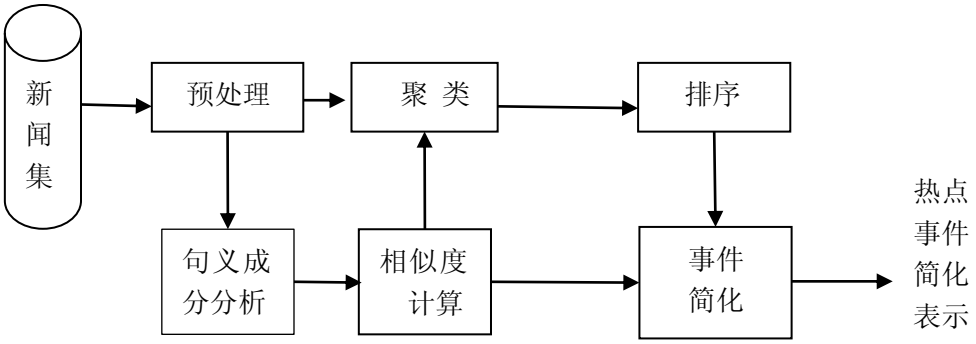

图 4-1 热点事件发现及事件展现原理图

热点事件发现及事件展现系统从功能上可划分为五个模块：预处理模块，相似度计算模块，聚类模块，事件排序模块，事件展现模块。各个模块之间相互独立，具有低耦合性，便于系统以后的扩展和升级。

系统的结构功能图如图 4-2所示。

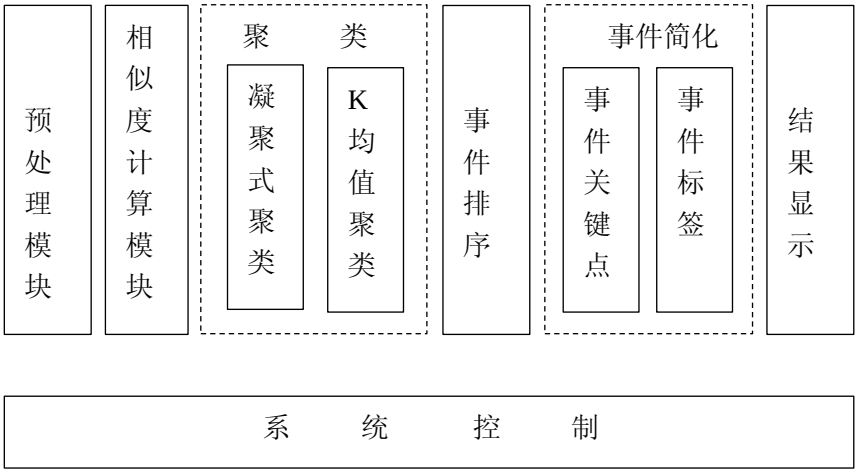

图 4-2 系统功能结构图

系统的主程序流程图如图 4-3所示。

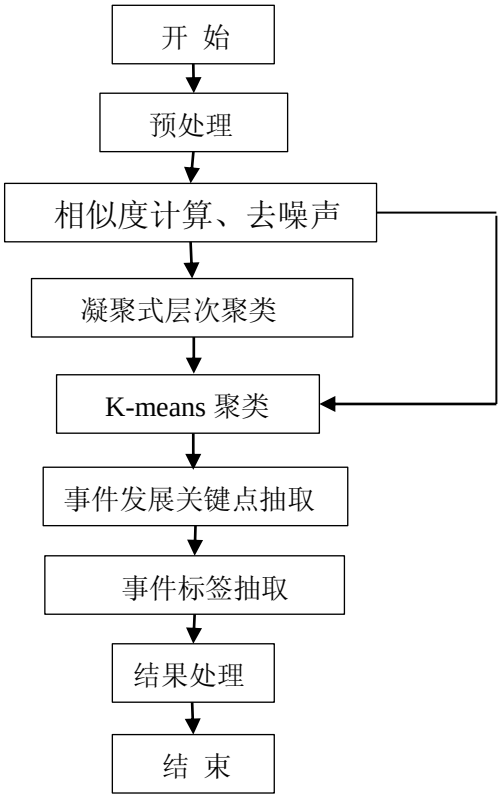

图 4-3 热点事件发现及事件展现主程序流程图

4.3 关键功能模块实现

4.3.1 文本表示

预处理模块中主要实现三个功能，一是对输入的新闻文档进行分词处理；二是将分词以后的文件进行去除停用词处理；三是将去除停用词后的文件输入到汉语句义结构模型自动构建系统中进行句义成分识别。这些工作都是为了改进文本表示模型，从而提高文本表示的准确性。

文本表示模块的主要功能是根据汉语句义成分自动识别结果，构建语义向量，利于计算机识别热点事件。语义向量对热点事件发现系统的效果起着决定性的作用，因此，它是系统中十分重要的模块。该模块实现了汉语句义结构模型自动构建，自动识别句义成分，自动计算语义向量各个维度的权重。

该部分的相关函数如下：

```
void Init();//初始化权重计算器
void GenerateTerms(const vector<string>& docs,vector<string>& terms);//分词处理
void cFileOri.GetFileDataCRF()//自动句义成分识别
void GetTermcCaseVector(i,data[i]); //获取第i个文档的句义成分
void GenerateTermWeight();//计算词的权重
void GetWordFrequency(string& input,map<string,int>& freq);
int CountWords(string& word, const vector<string>& words);//统计词数
int GetTermIndex(const string& term);//查询词语对应的下标
double ComputeTermWeight(int term, int doc);//计算词语在指定文档中的权重值
double GetTermFrequency(int term, int doc);//获取词语在指定文档的词频
double GetInverseDocumentFrequency(int term);//计算倒排文件频率
```

其中，几个函数先是实现识别句义成分，然后计算逆文档频率。根据这两种权重计算得出词的向量权重。

4.3.2 文本聚类

文本聚类模块是热点事件发现算法的关键实现模块，其结果直接关系到事件发现的准确率。其具体方法是凝聚式层次聚类与K-means聚类相结合，先利用凝聚式层次聚类产生聚类簇，之后利用簇的数目K和簇中心初始化K-means聚类算法，实现热点事件发现的功能。算法流程如3.3.4节中所描述，流程图如图 4-4所示。

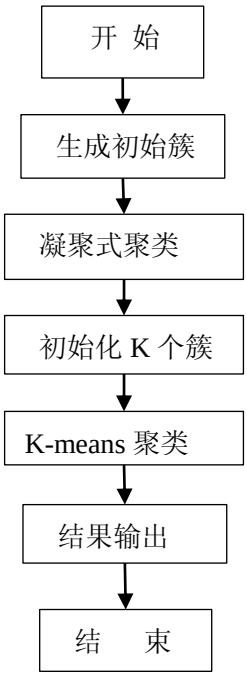

图 4-4 聚类流程图

4.3.3 事件简化表示

热点事件发现完成之后，下一阶段就是将事件呈现给用户，实现该应用的最后一步。这是热点事件发现应用的可视化部分，也是热点事件发现的一个关键部分，实现新闻文档的精简本就是热点事件发现研究方向产生的背景，也应该是发现热点事件后的主要工作。将热点事件的新闻报道按照时间顺序进行排序，这样就是将事件发展过程安插在时间轴上；通过计算不同新闻报道之间的相似度，抽取相邻时间点之间相似度小的新闻报道作为事件发展关键点。流程图如图 4-5所示。

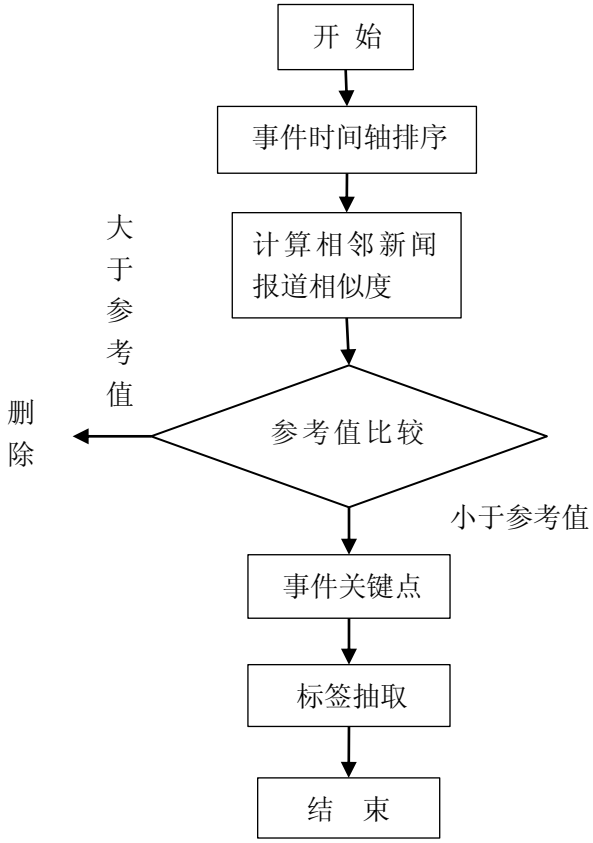

图 4-5 事件简化表示流程图

4.4 实验及分析

4.4.1 实验目的和数据源

通过实验验证热点事件发现方法及事件简化表示原型系统的性能，包括系统的稳定性、容错性、可靠性以及处理大批量数据文件的能力等。

实验数据将3.3.6.1节的新闻数据源作为实验的基础语料，其中包括10个热点事件，共516篇文档。其事件分布如表 3-4所示。另外，随机添加3.3.6.1节中搜狗语料库中的500篇文档，作为补充。

4.4.2 实验环境和条件

实验环境和条件详见3.3.6.2节。

4.4.3 评价方法说明

如3.3.6.3节和3.4.3.3节所示。

4.4.4 实验过程和参数

本文设计了两个实验，分别对热点事件发现系统和事件展现系统进行验证。首先进行热点事件发现原型系统的验证，实验过程和步骤如下。

1．对表3-3所示的热点事件发现数据源和随机添加的搜狗语料库500篇文档进行预处理，并输入到汉语句义结构模型自动构建系统中，进行句义成分分析；

2．根据步骤1生成的汉语语义成分，计算每个词的语义成分权重值，并将每篇文档构建成语义向量；

3．根据3.3.3节的算法计算各个文档之间的语义向量相似度；

4．利用3.3.4节的聚类算法对步骤3中的语义向量进行聚类得到事件列表，并按照文档数目进行排序。

本实验中，不再进行句义成分权重实验，而直接使用3.4.4节中句义成分权重实验的结果Case1的值为2.1，Case2的值为1.7，Case3的值为1作为句义成分权重。

然后，对热点事件展现原型系统进行验证，实验过程和实验步骤如下。

1．采用3.3.6.1节中描述的数据源，对其中的10个热点事件分别进行事件发展关键点的人工标注；

2．在3.3.6.2节中描述的实验环境和条件下，按照时间顺序计算同一热点事件文档集中的文档之间的相似度；

3．设置初始参考值为0.6，抽取相邻文档之间相似度未超过阈值的文档作为事件发展关键点；

4．使用3.4.2节中的事件标签方法对步骤3中抽取到是事件关键点进行事件标签抽取。

4.4.5 实验结果及分析

采用改进文本表示模型语义向量作为新闻文档的计算模型，利用凝聚式层次聚类和 K-means 聚类两种聚类相结合的方式进行聚类分析，并将事件通过事件发展关键点和事件标签进行展现，具体实验结果如图 4-6 所示。

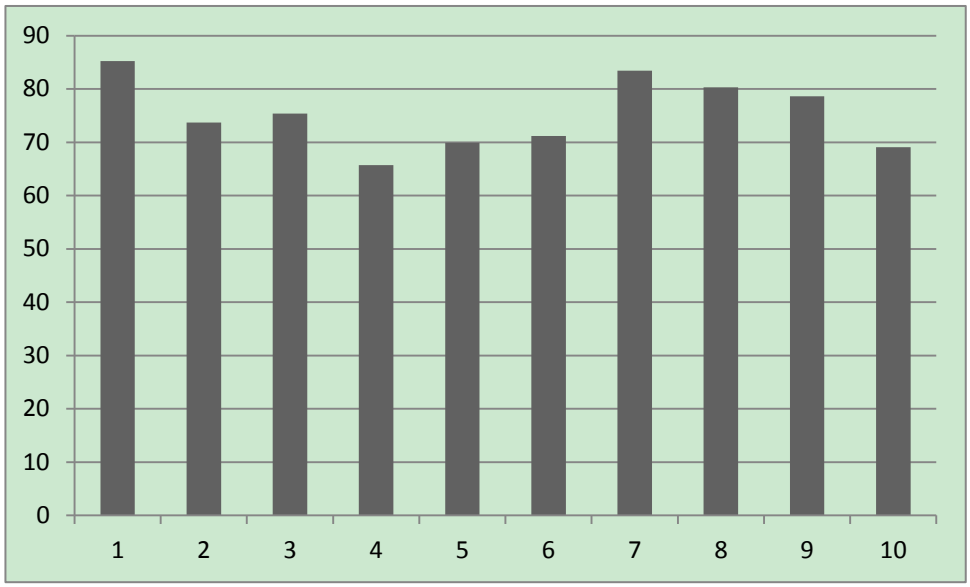

图 4-6 热点事件发现实验结果

实验显示，加入搜狗语料库 500 篇文档模拟非事件新闻，实验结果与没有加入非事件干扰的实验结果相近。说明系统可以处理正常的新闻文档，即使加大了处理文档的数目，系统仍然具有良好的稳定性。

事件展现模块的实验结果如图 4-7 所示。

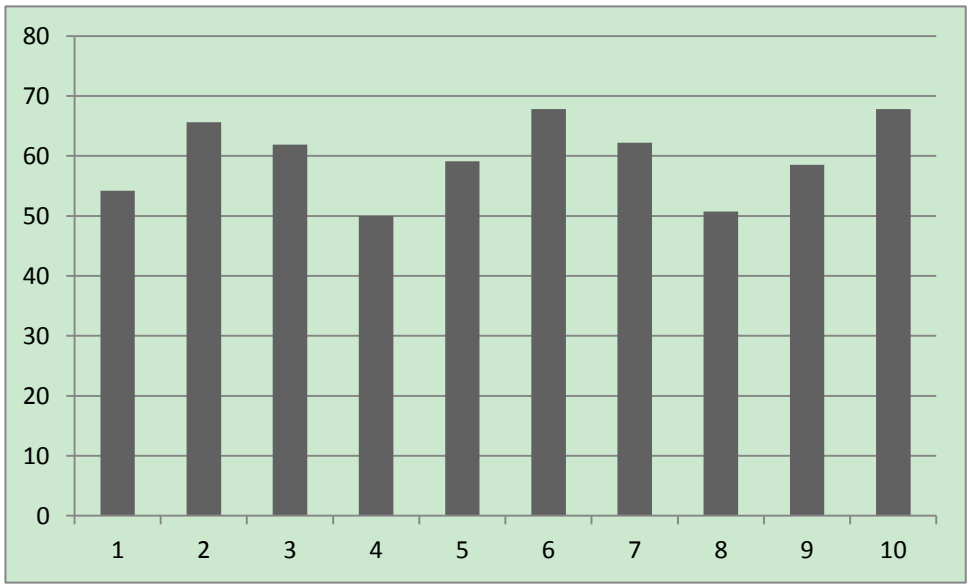

图 4-7 事件关键点抽取实验结果

4.5 小结

本章设计并实现了一种基于汉语句义结构模型分析词担任的句义成分，进而实现

语义域改进向量空间模型的原型系统，提高了文本表示的准确性。加入搜狗语料库500篇文档作为非事件新闻对系统的有效性和处理数据的功能进行了验证，经过实验验证，系统可以很好的处理新闻文档，能够自动删除与事件无关的新闻，并且可以处理大量的新闻文档，其准确率达到75.2%。并且通过对事件发展过程的分析，实现了事件发展关键点和事件标签表示事件发展过程的系统。通过实验验证验证了系统功能的有效性和可靠性，可以作为实际应用系统的基础。

第5章 结束语

5.1 全文总结

本文通过对话题发现与跟踪（TDT）、文本聚类、相似度计算以及事件发展自然规律等进行研究和分析，利用汉语语义学在语义表达上的优势，改进了传统的基于统计的文本表示方法，提出了一种事件发展过程展现方法，设计实现了以上两种方法的原型系统。

论文的主要工作包括：

1.提出了一种基于文本聚类的热点事件发现方法。

本文将向量空间模型和句义结构模型结合到一起，通过对新闻语料进行词频、逆文档频率和语义格分析，以语义分析辅助统计代数分析，提高文档表示的准确性，在应用到热点事件发现方法时，热点事件发现准确率达到了 75.2%。提出了一种事件简化表示方法，使用以事件发展关键点为主，以事件标签为辅对事件进行简化表示，其中事件发展关键点准确率为 59.8%。研究结果表明，将汉语句义结构模型和向量空间模型融合，提高了热点事件发现的准确率；事件发展关键点及事件标签对事件进行表示的方法使事件结果易读、精确、简练。

2.设计并实现了互联网新闻热点事件发现原型系统。系统具有自动分析文档词频、逆文档频率函数、句义结构模型自动构建、热点事件发现、事件简化表示的功能，并有自动评价的功能。系统每个模块采用文件进行数据交互，便于分析每个模块的准确率，具有一定的容错性、可扩展性和可复用性

3.分析了主流的文本表示方法和主要的文本聚类技术，并对比了各自的优缺点。

目前主要的文本表示方法，包括布尔模型、向量空间模型、句义结构模型及 LDA 模型。理论基础上，布尔模型最初是应用在搜索引擎上的，对文档篇章语义的表达并不准确，也不全面；向量空间模型是结合代数论，基于统计词频和逆文档频率的，对于文档的语义表达也不准确，不全面；句义结构模型是基于汉语语义学的，是在语义学上构建的模型，在数学表达和计算性上有些不足；LDA 模型与向量空间模型有些相像，但是它是基于一篇文档有多个主题这样一种假设，对于主题发现和文档分类方面有很大优势，但是对于文档的语义表达并未涉及。

5.2 工作展望

本文针对热点事件发现及事件发展过程展现做了一些改进的工作，虽然完成了课题的任务，达到了预期的目的，但仍然存在一些不足的地方，有待进一步完善。下一步的主要工作包括：

1. 数据源构建的问题。目前实验采用的数据规模虽然能够满足现阶段算法的研究需要。但是，目前的数据量与评测语料相比仍然有限，需要不断积累扩大数据规模。而且，研究过程中发现，现在使用的语料并没有涵盖所有新闻语料的类型，需要进一步收集不同长度的语料以及不同内容的语料，以保证语料的合理性和有效性，更好的支持后续的相关研究工作。另外，网络新闻的收集系统也可以与本文研究的原型系统做统一。

2. 汉语句义结构模型的效率和准确率。实验室在汉语句义结构模型方面做了很多的工作，已经能够应用在其他系统上，就如本文所作的一些探索。但是，该模型自动构建的效率还有待提高，能做到实时处理就能很大的提高系统的效率。

3. 热点事件发现系统的自动检测与自动跟踪。本文主要研究的是事件的自动检测，并没有对事件后续发展进行跟踪。因此，基于互联网更新速度快这样的考虑，应该研究事件的自动跟踪。

4. 事件简化表示的问题。该方法主要考虑的是不同新闻报道之间的相似度差异，可以尝试考虑其他的策略，如时间轴发展规律；事件发展关键点，目前采用的是新闻报道这一层次，可以考虑在这些新闻报道上应用多文档自动摘要技术，形成更加简单、易读的结果。

参考文献

[1] Allan J, Papka R, Lavrenko V. On-line New Event Detection and Tracking. In the proceedings of the 21th annual international ACM SIGIR conference on Reasearch and development in information retrieval[C]. Amherst, 1998, 37-45.

[2] 骆卫华, 刘群, 程学旗. 话题检测与跟踪技术的发展与研究. 全国计算语言学联合学术会议论文集[C]. 北京, 2003, 560-566.

[3] James Allan, Jaime Carbonell, George Doddington, et al. Topic Detection and Tracking Pilot Study: Final RePort. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop[C], San Francisco, 1998: 194-218.

[4] Doddington G, Fiscus J, The 2002 Topic Detection and Tracking (TDT2002) Task Definition and Evaluation Plan[R]. Technical Report, 2002.

[5] Mayne C L. Multilingual Topic Detection and Tracking: Suceessful Researeh Enabled by Corporaand Evaluatlon. In: Language Resources and Evaluation Conference[C], 2000: 1487-1494.

[6] James Allan. Topic Detection and Ttacking: Event-based Information Organization [M]. Springer, 2002.

[7] Wa11s F, Jin H, Sista S, et al. Topic Detection in Broadcast News. In: Proceedings of the DARRA Broadcast NewsWorkshop[C], Hemdon, USA, 1999: 193-198.

[8] Leek T, Jin H, Sista S.et al. The BBN Crosslingual Topic Detection and Tracking System. In: Working Notes of the Third Detection and Tracking Workshop[C]. Vienna, 2002: 332-346.

[9] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Emilia Kasper, Irina Tenmikova. Multilingual and Cross-lingual News Topic Tracking. In Proceedings of the 20th Intenational Conference on Computational Linguisties[C]. Association for Computationl linguistics, 2004:959.

[10] Allan J. Introduction to Topic Detection and Tracking[M].Springer US, 2002:1-16.

[11] Gabriel Pui, Cheong Fung, Jeffrey Xu Xu, Hongjun LU, Philip S Yu. Parameter Free Bursty Events Detection in Text Streams. In: Proceedings of the 31th inrernational conference on Very large data bases[C].endowment, 2005:181-192.

[12] McKeown K R, Barzi1ay R, Evans D, et al. Tracking and Summarizing News on a Dally Basis with Columbia's Newblaster. Proeeedings of second international conference on Human Language

Technology Research[C]. Vancouver, 2002: 162-168.

[13] Google News. http://news.google.com.

[14] 刘远超, 王晓龙, 徐志明, 关毅. 文档聚类综述[J]. 中文信息学报. 2005,20(3):55-62.

[15] 李保利, 俞士汶. 话题识别与跟踪研究[J]. 计算机工程与应用, 2003, 39(17):6-l0.

[16] 贾自艳, 何清, 张俊海等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机研究与发展, 2004, 41(7): 1273-1280.

[17] 于满泉, 骆卫华, 许洪波, 白硕. 话题识别与跟踪中的层次化话题识别技术研究[J]. 计算机技术与发展, 2006, 43(3): 489-495.

[18] Zhang Kuo, Li Juan Zi, Wu Gang. New Event Detection Based on Indexing tree and Named Entity. In: 30th annual international ACM SIGIR conference on Research and development in information retrival[C]. Amsterdam, 2007:215-222.

[19] 百度新闻. http://news.baidu.com.

[20] Google 中文资讯. http://news.google.cn.

[21] 贾彦德. 汉语语义学[M]. 北京: 北京大学出版社, 2005.

[22] 袁毓林. 现代汉语名词的配价研究[J]. 中国社会科学, 1992, 3.

[23] 袁毓林. 一价名词的认知研究[J]. 中国社会科学, 1994, 4: 241-253.

[24] 袁毓林. 论元角色的层级关系和语义特征[J]. 世界汉语教学, 2002, 3: 10-22.

[25] 邵敬敏. 量词的语义分析及其与名词的双向选择[J]. 中国语文, 1993, 3: 181-190.

[26] 邵敬敏. 论汉语语法的双向选择性原则[J]. 中国语言学报, 1997, 8.

[27] 邵敬敏. “语义语法”说略[J]. 暨南学报, 2004, 1: 100-106.

[28] 陈昌来. 论语义结构中的与事[J]. 语文研究, 1998, 2: 22-27.

[29] 陈昌来. 汉语语义结构中工具成分的性质[J]. 世界汉语教学, 1998, 2: 22-26.

[30] 陈昌来. 现代汉语句子[M]. 上海: 华东师范大学出版社, 2000.

[31] 陈昌来. 现代汉语动词的句法语义属性研究[M]. 上海: 学林出版社, 2002.

[32] 陈昌来. 现代汉语语义平面问题研究[M]. 上海: 学林出版社, 2003.

[33] 章成志. 基于机器学习的文本聚类描述算法研究. 第二届全国信息检索与内容安全学术会议[C]. 2009: 216-225.

[34] Baeza-Yates R and Ribeiro-Neto B. Modern Information Retrieval[M]. New York, ACM press, 1999.

[35] 王永成等. 中文信息处理技术及其基础[M]. 上海交通大学出版社, 1990.

[36] Lee J H. Combining multiple evidence from different properties of weighting schemes. In:18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C].ACM, 1995: 180-188.

[37] Yang Y, Carbonell J G, Brown R D, Pierce T, Archibald B T, Liu X, Learning Approaches for Detecting and Tracking News Events[J]. IEEE Intelligent Systems, 1999, 32-43.

[38] Singhal A, Buckley C, Mitra M. Pivoted Document Length Normalization. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrival[C].ACM SIGIR, 1996: 21-29.

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

[40] Castillo C, Donato D, Becchetti L, Boldi P, Leonardi S, Santini M, Vigna M. A Reference Collection for Web Spam[C]. ACM SIGIR Forum, 2006, 40(2): 11-24.

[41] 潘文锋. 基于内容的垃圾邮件过滤研究[D]. 北京:中科院计算技术研究所, 2004.

[42] Yue LU, Chengxiang Zhai. Opinion Integaration Through Semi supervised Topic Modeling. Proceedings of the 17th International Conference on Word Wide Web[C], Beijing, China: 121-130.

[43] 曹娟, 张勇东, 李锦涛, 唐胜. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报, 2008, 31(10): 1780-1787.

[44] Forgy E W. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications[J], Biometrics, 1965, 21: 768-780,.

[45] Kaufman L, Rousseeuw P J, Finding Groups in data: An Introduction to Cluster Analysis[M]: John Wiley & Sons, 2009.

[46] Ng R T, Han J. CLARANS: A method for clustering objects for spatial data mining[J].Knowledge and Data Engineering, IEEE Transactions on , 2002 , 14(5): 1003 -1016.

[47] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques[M]. 范明, 孟晓峰 译, 机械工业出版, 2001.

[48] T Zhang, Ramakrishnan R, M Livny. Birch: A new data clustering algorithm and its applications[J]. Data Mining and Knowledge Discovery, 1997, 141-182.

[49] Guha S,Rastogi R, Shim K. CURE: An Efficient Clustering Algorithm for Large Database.In: ACM SIGMOD Record[C].ACM, 1998,27(2):73-84.

[50] George K, Han Eui-Hong, Vipin K. Chameleon: Hierarchical clustering using dynamic modeling

[J]. Computer, 1999, 32(8): 68-75.

[51] 刘群, 张华平, 俞鸿魁, 程学旗. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004, 41(8): 1421-1429.

[52] 陈浩, 何婷婷, 姬东鸿. 基于 k-means 聚类的无导词义消歧[J]. 中文信息学报, 2005, 19(4): 10-1.

[53] Bun KK, Ishizuka M. Topic Extraction from News Archive Using TF*IDF Algorithm. The Third International Conference on Web information System Engineering[C]. SingaPore, 2002.73.

[54] Liang TP, Lai HJ. Discovering User Interests from Web Browsing Behavior: An Application to Internet News Service. In Proceeding of the 35th Hawaii International Conference on System Science[C]. Hawaii, 2002: 2718-2727.

[55] http://maroo.cs.umass.edu/pub/Web/search_authors_results.php?id=1918&frame=James&mname=&Iname=Allan.

[56] 刘星星. 热点事件发现及事件内容特征自动抽取研究[D]. 华中师范大学, 2009.

[57] 秦兵, 刘挺, 李生. 多文档自动文摘综述[J]. 中文信息学报. 2005.19(06):13-19.

[58] 孙春葵, 钟义信. 关于自动文摘系统中文摘句式的一种机器学习方法[J]. 计算机工程与应用, 2000(5): 18-23.

[59] 沈玮杰. 基于文献结构的自动文摘初探[J]. 现代图书情报技术, 2002(3): 23-34.

[60] Sirnone Teufel, Marc Moens. Summarising Scientific Articles-Experiments with Relevance and Rhetorical Status[J]. Computational Lingulstics, 2002.28(4): 409-445.

[61] 于江德, 樊孝忠, 庞文博. 事件信息抽取中语义角色标注研究[J]. 计算机科学, 2008, 35(3): 155–157.

[62] 冯志伟. 自然语言处理的历史与现状[J]. 中国外语, 2008, (1): 14-22.

[63] 冯志伟. 自然语言处理的学科定位[J]. 解放军外国语学院学报, 2005, 28(3): 1-8.

[64] 冯扬. 汉语句义模型构建及若干关键技术研究[D]. 北京理工大学, 2010.

[65] Doddington G, Fiscus J. The 2002 Topic Detection and Tracking (TDT2002) Task Definition and Evaluation Plan[R]. Technical Report, 2002

[66] 贾自艳, 何清, 张俊海等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机研究与发展, 2004, 41(7): 1273-128.

[67] 萧国政, 何炎祥, 孙茂松. 中文计算技术与语言问题研究[M], 北京: 电子工业出版社, 2007, 402-408.

[68] 孙学刚, 陈群秀, 马亮. 基于主题的 Web 文档聚类研究[J]. 中文信息学报. 2003, 17(3): 21-26.

学习期间发表的学术论文与研究成果清单

[1] 罗森林, 白建敏, 韩磊，孟强等. 融合句义特征的多文档自动摘要算法研究[J]. 北京理工大学学报: 自然科学版. 录用待发表.

致谢

两年半的研究生生活和学习已经步入尾声，在完成这篇论文的同时，回首一路走来的这些时光，我需要感谢太多的人！

首先，衷心的感谢我的导师——罗森林老师。在我的硕士学习中，他十分关心我的课题研究，给了我很多好的指导和建议，并且让我知道了在学习过程中应该具备何种态度和素质。罗老师给了我宽松自由的科研环境，提供了最好的研究条件，给与了难得的锻炼机会。罗老师严谨的治学态度、广阔的思路、求实的作风以及孜孜不倦的追求一直都陶冶和熏陶着我，让我懂得了很多做人做事的道理。与此同时，老师还特别关心我的生活和身体，这让我非常感动。在这里，要对老师特别表示感谢。

感谢我的父母，是他们对我的爱激励着我一步步地前进，走到了今天。无论遇到什么问题，他们都始终支持我、关心我，做我坚强的后盾，是他们使我有勇气去面对一切，顺利完成硕士阶段的学习。

感谢韩磊师兄、王倩师姐以及实验室其他同学们。两位师兄、师姐对我的课题及学习生活上的很多事情都给予了帮助，经常占用自己的时间与我一起讨论课题，提出许多宝贵的意见，耐心解答我提出的各种问题，在此对他们表示感谢。

最后，感谢所有关心我的朋友们。因为你们，我度过了开心快乐，又收获颇丰的硕士时期，与大家的友谊我会好好珍惜。祝你们事事顺心！