杭州电子科技大学

硕士学位论文

网络舆情热点检测与跟踪技术研究

姓名：徐宁

申请学位级别：硕士

专业：计算机软件与理论

指导教师：王小华

201111

摘　要

网络舆情热点检测与跟踪技术主要利用计算机对海量的网络信息进行处理，提取热点主题并且对热点主题进行跟踪，它能把握整个舆情的动向，并且能够为下一步的舆情处置行动提供参考，是网络舆情分析的关键技术。本文主要针对 BBS 论坛，对网络舆情热点检测与跟踪技术进行了研究，提出了新的网络舆情热点检测与跟踪方法，并获得了满意的实验结果。

首先，本文对网络舆情热点检测与跟踪技术的现状和发展进行了简要的回顾。文中分别对网络舆情信息挖掘的相关技术和热点检测与跟踪算法作了分析。网络舆情信息挖掘主要包括网络舆情信息采集、网络舆情信息预处理、网络舆情信息分析等部分，并且介绍了现有的国内外网络舆情系统，对网络舆情热点检测与跟踪技术的研究主要包括话题检测与跟踪任务和话题检测与跟踪的关键技术等。

其次，本文提出了基于共词分析的网络舆情热点检测方法。传统的共词分析方法一般运用在某一专业的学科领域中，通过判断学科领域中主题间的关系，进而展现该学科的研究结构。本文提出将共词分析运用到网络舆情热点检测方法中，而BBS是网络舆情的主要载体之一，该方法将共词矩阵和聚类方法相结合，从而在BBS环境下发现舆情热点主题。实验证明本文提出的算法在BBS环境下的应用具有稳定性和高效性，并具有一定的可信度。

再次，本文在总结了现有的主题关注度提取方法的基础上，分析了它的优缺点，并提出了一种基于关注度的热度提取方法，即综合考虑论坛帖子权重值和主题的媒体关注度对主题热度的影响。紧接着主要根据主题距离构建出主题进化图，将相对熵的概念引入到主题距离提取的方法上，并介绍了一些相对熵的应用。通过相对熵的阈值判断，从而发现各个时间戳中主题的延续性。

最后，分别使用大规模数据语料和真实论坛语料对本文提出的基于共词分析的网络舆情热点检测算法和基于热度分析的网络舆情热点跟踪算法进行了实验，并对测试结果进行了分析。实验结果表明，本文的算法对处理网络舆情热点检测与跟踪问题具有一定的可用性。

本文最后对论文所做的工作进行了总结与评述，并提炼了网络舆情热点检测与跟踪技术中值得继续研究的若干问题，为以后的研究指明了方向。

关键词：热点检测，跟踪技术，网络舆情，共词分析，中文信息处理

ABSTRACT

The technologies of Internet public hot opinions detecting and tracking are to deal with the mass of Internet information by making use of computers. The hot themes are detected and tracked from a great deal of Internet information. Then the evolvement tendency of the whole public opinions can be predicted, which is useful for determining to take appropriate actions. Base on BBS information, a novel method on detecting and tracking of Internet public hot opinions is proposed in this thesis. The experimental results show that the proposed method is feasible and satisfied detecting.

Firstly, this paper has a brief review on the present situation and development of hot topic detection and tracking technology of Internet public opinion. This paper also analyzes related technologies and hot topic of detecion and tracking of internet information. Internet public opinion information extraction mainly includes information source, information collection, information and internet information and so on. Key technologies of public opinion information analysis include topic detection, tracking, and text tendentiousness analysis.

Secondly, a method based on co-word analysis for Internet public hot opinions detecting is proposed in this thesis.Generally, the traditional co-word analysis is applied mainly in a certain professional field. The discipline structure can be presented by analyzing the relationships of key words in a domain. In this thesis, we apply the co-word analysis to detect Internet public hot opinions from BBS information. In the method, one co-word matrix and a clustering method are combined to extract the hot themes from BBS data. The experimental results show that our proposed method is efficient and promising.

Thirdly, a hotspot focus retrieval method based on attention is presented in this thesis after analyzing the advantages and disadvantages of existing retrieval methods of themes' attention. That is, how the weight of forum information in BBS and the attention from news media affect the theme hotpot focus is considered. Then, the evolution diagram of a theme is constructed according to the transition distances of themes. At the same time, the relative entropy is introduced and applied into the computation of transition distances of themes,so the continuity of a theme in different timestamps can be explored according to the threshold of relative entropy.

Finally, the experiments on the proposed methods in this thesis are performed based on the large corpus and real forum data.The experimental results show that the our

methods is feasible for Internet public hot opinions detecting and tracking.Then a summarization on the main work is presented.The unsolved problems in Internet public hot opinions detecting and tracking are also analyzed and considered as our future research work.

Keywords: hot points detecting, tracking technology, internet public opinion, co-word analysis, Chinese information processing

第一章 绪论

本章主要从网络舆情产生的背景，研究的紧迫性分析入手，分别从主题识别与跟踪等方面介绍了国内外舆情分析与处理技术现状和未来的发展趋势，阐述了本课题研究的必然性和合理性，最后对本文的章节安排进行了说明。

1.1 课题研究背景与意义

随着互联网在全球范围内的飞速发展，网络媒体作为一种新的信息传播形式，已深入人们的日常生活。网络媒体具有进入门槛低、信息规模超大、信息发布与传播迅速、参与群体庞大、实时交互性强等综合性特点。如今网友言论活跃已达到前所未有的程度，不论是国内还是国际重大事件，都能马上形成网上舆论，通过这种网络来表达观点、传播思想，进而产生巨大的舆论压力，达到任何部门、机构都无法忽视的地步。可以说，互联网已成为思想文化信息的集散地和社会舆论的放大器。但是，与传统媒体不同的是，互联网是一个开放性和互动性的平台。在网络上，任何人都可以在博客、论坛、网络社区或者自建站点上发布言论和观点，他们既可以针对某一个社会现象或某一条新闻事件发表自己的看法，又可以通过互联网在网民之间形成互动场面，赞成方的观点和反对方的观点同时出现，相互探讨、争论，相互交汇、碰撞，使得各种观点和意见能够快速地表达出来。不同网民之间通过网络交流经验和共享资源的现象变的十分普遍。

2011 年 7 月 19 日，中国互联网络信息中心（CNNIC）在京发布了《第 28 次中国互联网络发展状况统计报告》[1]。《报告》数据显示，截至 2011 年 6 月底，中国网民规模达到 4.85 亿，较 2010 年底增加 2770 万人，增幅 6.1%，互联网普及率进一步提升。同时，BBS、博客用户的快速增长离不开网民增长的带动作用，随着搜素引擎的快速发展，作为互联网入口地位的确立，BBS、博客用户间的分享行为变得更加活跃，分享渠道变得更加的丰富，特别是我国微博用户数量从 2010 年底的 6311 万爆发增长到 1.95 亿，以高达 208.9%的增幅成为用户增长最快的互联网应用模式。互联网业务的迅速发展，一方面加快了信息传播的速度，拓展了信息传播的渠道，对经济的发展，社会的进步，科技的普及起到了积极的作用，另一方面以即时，互动为主要特点的互联网传播方式，更为公众表达舆情、参与经济社会及政治生活，提供了一个方便快捷的平台。

所谓舆情，是舆论情况的简称，是在一定的社会空间内，围绕中介性社会事件的发生、发展和变化，作为主体的民众对作为客体的社会管理者及其政治取向产生和持有的社会政治态度。它是较多群众关于社会中各种现象、问题所表达的

信念、态度、意见和情绪等等表现的总和[2]。而网络舆情是社会舆情的一种表现形式，它是民众在互联网上发布和传播的能够反映民众舆情的文字、图像、音频、视频等，往往是以文字的形式为主[3]。网络舆情的特点与网络传播方式的特征息息相关，主要表现为如下特点：传播信息的多元性、传播方式的自由性、传播主体的隐匿性、传播事件的突发性等[4]。同时，我们也必须清楚地认识到，由网络舆论的自由化所带来的一系列的消极影响：比如一些网民通过互联网散布谣言、进行偏激的评论等，因此，监测网络舆情，因势利导，提高新形势下的舆情信息的分析能力，及时准确地掌握社会舆情动向对于引导舆情以及做好舆情预警有着重要的意义。

热点话题检测与跟踪技术是目前在网络舆情分析中具有重要作用的一种信息处理技术。因为在海量的网络信息中，与同一话题相关的信息不管在时间上还是在空间上往往都比较分散。不同的语料源可能在相同的时间上对同一事件进行了报道，或者同一语料源可能在不同的时间段对同一事件进行了报道。各种报道的视角和分析都有可能不同，而舆情事件又具有不断演化的特性。这样，面对互联网上众多的语料源，人们对各种舆情事件难以做到全局性的把握。因此人们在寻求一种可以自动把各个孤立分散的语料按其话题的属性进行有效组织的技术。热点话题检测与跟踪技术正是在这种应用背景下产生，它已经成为网络舆情分析的重要技术手段。

此外，研究这项技术也有助于企业了解客户的需求变化，可以及时调整产品和市场策略，同时对于公关、广告等产业提高市场的调研分析能力和效率都有着现实的应用价值和广阔的前景。

1.2 课题研究现状

热点话题检测与跟踪是一种新颖的信息处理技术。它将一系列相关的报道按其所属话题进行有效的组织，以实现在语料流中对新话题或新事件的自动检测以及对已知话题的后续报道的跟踪。对于网络舆情的研究最早的是国外的话题检测与跟踪技术，其中研究比较好的是美国的 TDT 系统。美国有一个研究项目被称为 TDT(Topic Detection and Tracking，TDT)，它最初研究目的是能够发现和归纳来自于新闻流中的重要的信息[5]。TDT 中的话题检测与跟踪技术的思想源于 1996 年，来自 DARPA、卡内基一梅隆大学、Dragon 系统公司以及麻萨诸塞大学的研究者们开始确定话题识别与跟踪研究的内容，并开发了应用于解决新闻流问题的一些技术[6]。它能够实现话题组织、话题发现，并可以识别出各种新话题、突发事件以及关于某些特定事件的新报道，它主要可以用以帮助用户解决海量信息的智能分析处理。这可以广泛应用于政府、媒体、企业和证券市场等领域。此外，它还可以帮助用户找出感兴趣话题的所有报道，研究话题的发展历程等等。

随着网络舆情有效监测的需求和重视程度不断的提高，舆情分析的相关技术也日益成为研究的热点。舆情分析技术主要包括以下几个方面[7]。

1、话题检测，其实是一种面向信息安全的技术，它主要依靠舆情信息的关注度、评论稀疏程度等参数，检测出最近发生的事件。

2、文本倾向性分析，对于文本舆情，主要根据文本的上下文信息提炼出文章的情感方向，给文章中的每个词汇用打分的方式进行分析统计，最后通过打分的结果来评价文本的倾向性。

3、话题跟踪，利用相似度分析下一个时间戳中新报道的话题，通过阈值判断当前话题是否被跟踪。可以及时了解和掌握后续发展动态。

4、统计分析技术，利用统计学、概率论的知识对舆情语料中的各个属性进行统计分析。通过统计分析技术可以直接获取文档摘要。

5、关联分析，挖掘出在海量数据中存在的隐性关系，通过分析这些隐性关系集合，得出它们的相关性。一般将关联分析切分成频繁项的挖掘和规则的挖掘两部分。

6、新事件产生，对新事件的产生进行时空的分析，通过对新事件各个时间戳进行跟踪，可以了解整个事件发生的场景，而且可以对该事件之后的发展进行合理的预测。

7、舆情统计报告生成，根据舆情分析引擎处理后生成报告，用户可根据指定条件对热点话题进行倾向性进行查询，并浏览信息的具体内容，以此来提供决策支持。

舆情分析技术是一门新兴的技术，国内外的许多研究机构都陆续展开了对该领域的研究，但还处于起步阶段，成果还是很有限。

1.3 本文的研究内容

针对网络舆情的探究背景，本文主要研究和探讨了网络舆情热点检测与跟踪技术，本课题重点开展以下几方面的研究工作：

热点主题发现技术：是本课题研究的重点之一，热点主题发现功能实现的主要算法就是文本的聚类，当一个类的规模很大的时候就可以认为它是一个主题，本文通过结合共词矩阵与 Bisecting K-means 聚类算法，提出了基于共词分析的文本主题词聚类的方法，并将该方法运用到真实的 BBS 语料环境中，从而得到舆情热点主题。

热点主题跟踪技术：是本课题研究的重点之一，本文主要通过主题词回溯算法提取各个时间戳的话题相关报道篇数，然后根据 TF*PDF 的思想计算各个时间戳的媒体关注度，通过结合帖子的权重和媒体关注度两个参数，提出了主题热度的计算方法，接着将 KL 相对熵的方法运用到主题距离的计算上，最后通过主题

距离阈值的判断来构建主题进化图。

信息采集技术：主要使用基于 Larbin 的网络爬虫来实现舆情信息的采集。利用网页中的 URL 链接来访问页面，周期性地采集网页中的正文信息，为后面的舆情信息处理提供数据。

信息预处理技术：包含页面过滤、中文分词、停用词过滤、以及关键词权重计算策略等。页面过滤是对从网络上直接采集过来的数据通过网页的标记特点进行过滤掉一些网页上的页面噪声。中文分词是进行中文信息处理最基本的一步。关键词权重是通过计算词频和文档频率，以此来提取文摘和关键词，这也是直接表达舆情信息的关键步骤。

1.4 本文的组织

第一章　绪论：简单介绍了本课题的研究背景、意义、现状、内容等，并同时介绍了本论文的研究内容及各章节的安排。

第二章　相关工作及研究进展：主要介绍了网络舆情挖掘的基础知识，包括种子页面自动生成、主题爬行策略等。舆情信息预处理，同时介绍了几个国内外常用的网络舆情系统，并作了比较等。分析了舆情信息分析技术的研究现状，重点介绍了舆情热点检测与跟踪技术在网络舆情中的重要作用，并列举了几种常用的热点检测与跟踪技术。

第三章　基于共词分析的网络舆情热点检测：针对现有算法存在的不足，主要通过结合共词矩阵与 Bisecting K-means 聚类算法，提出了基于共词分析的文本主题词聚类的方法，并将该方法运用到真实的 BBS 语料环境中，给出了详细的思路与算法步骤。

第四章　基于热度分析的网络舆情热点跟踪：基于 TF*PDF 和 KL 相对熵的思想，主要介绍了各个时间戳主题热度提取算法与各个主题演化距离提取算法，并给出了详细的思路与算法步骤。

第五章　实验及性能评价：用经过处理的带有时间属性的论坛语料对本文提出的基于共词分析的网络舆情热点发现算法和基于热度分析的网络舆情热点跟踪算法进行了测试，并对实验结果进行了评价。

第六章　总结与工作展望：主要是对本文所做的工作进行了总结，并展望了未来的工作。

第二章 相关工作及研究进展

在网络日益普及的今天，网络舆论的影响力渐渐地在广大民众之间日渐突显，网络舆情信息分析技术开始逐渐得到国内外研究者的关注。我国对舆情的研究分析是近几年才发展的，得到了一些国外舆情分析的借鉴。研究发现将舆情信息分析技术与舆情背景有效的结合，能够更好地发现网络舆情热点，控制舆情的发展方向。本章重点介绍了舆情分析的关键技术及其它在国内外的研究现状，主要分析了以下几个方向：网络舆情信息挖掘的概述、现有网络舆情系统介绍、常用网络舆情热点检测与跟踪技术。

2.1 网络舆情信息挖掘概述

网络舆情信息挖掘研究可分为 4 个层次：网络舆情信息采集、网络舆情信息预处理、网络舆情信息分析、网络舆情信息输出。这 4 个层次的研究从整体上反映了网络舆情挖掘的研究现状和成果，具体如图 2.1 所示。

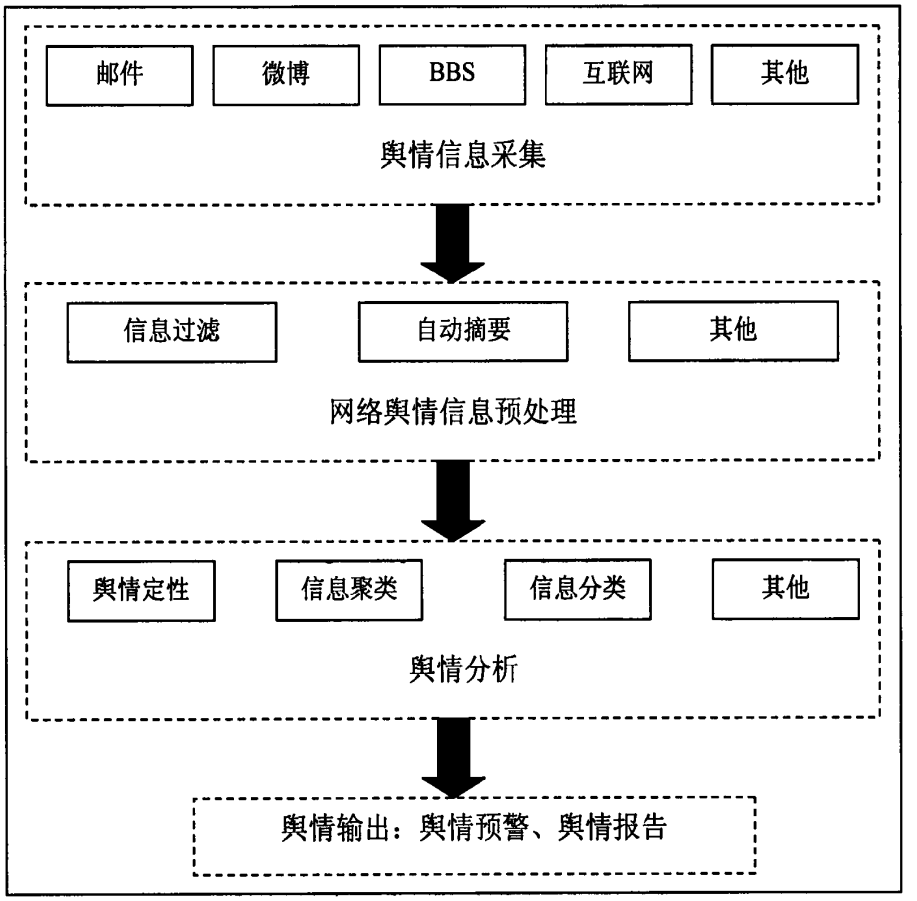

图 2.1 网络舆情挖掘框架图

（1）网络舆情信息采集。网络舆情信息来源是网络舆情信息采集的基础，

目前，网络舆情主要的信息来源一方面来自于传统媒体如报刊、广播、电视的报道，比如光明日报、新华日报、电台之声等。另一方面来自于网络原创，尤其是大型 BBS 上某些网友自行发布的新闻和言论，比如网易论坛、新华网论坛、新浪博客、天涯杂谈、高校 BBS 等。目前，网络舆情信息采集主要是 web 信息挖掘，所谓 web 信息挖掘它是指使用数据挖掘技术在互联网数据中发现潜在、有价值的信息[8]。它主要通过 web 信息采集器（Web Crawler）将每一个需要采集的网页构建出可以互相关联的链接关系，然后自动下载舆情网页中的主要内容。简单的说，它主要是从一个或若干个最开始的 URL 集散发，获得初始网页上的 URL 地址，根据 URL 处理器将这些 URL 全部放入准备采集的序列里。Web 信息采集器将会根据 HTTP 协议标准，在序列中依次拿出 URL，通过 URL 从一个页面找到另一个页面，直到没有满足条件的新的 URL 为止[9]。目前，Web 信息挖掘的方法基本上可以分为以下几种：基于迁移的信息采集、基于 Agent 的信息采集、基于主题的 web 信息采集、基于用户个性化的 web 信息采集等[10]。在实际的系统运行中，往往会将其中的几种 web 信息挖掘的方法相互结合，这样可以取得更好的效果。

基于 web 的信息采集主要是将一些种子页面扩展到多个 URL 页面，搜索网络空间，然后进行采集。这类信息采集过程往往采用的是并发模式，因此要花更多的精力在主题爬行策略与重复 URL 消除上[11]。由这类信息采集器并发构成的搜索引擎，适合搜素范围比较广的话题。其主要模块流程如图 2.2 所示。

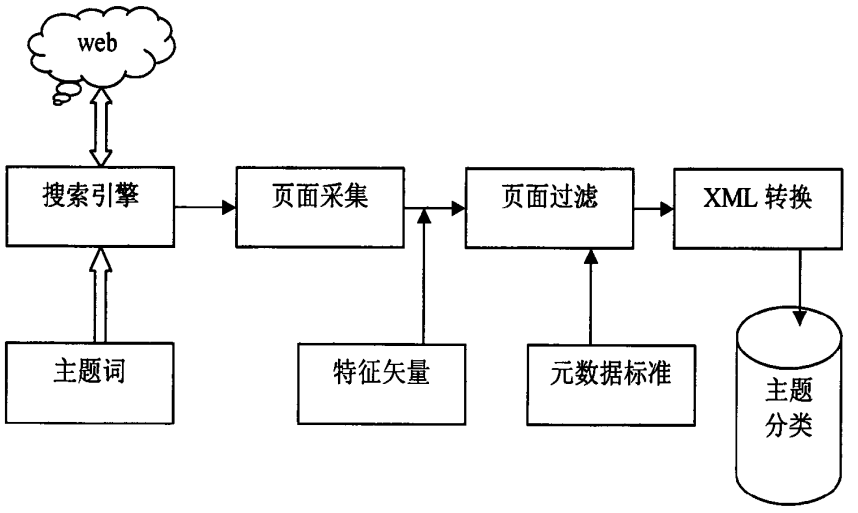

图 2.2 基于 web 的信息采集

1）种子页面生成。种子页面是主题爬行的起始页面，它可以是一个网站的首页或者是某个网站的下属子页面，种子页面的选择要求具有较高的主题相关性和主题链接的中心度[12]。种子页面的生成是将一些主题词通过搜索引擎进行检

索，然后从检索的结果中提取出前 N 个页面作为种子页面，最后把这些种子页面加入种子库中。然后种子库会根据前一周目标网站的信息量和本周目标网站的信息量来设定一个更新阈值，同时每周对种子库进行检查，以此来实现种子库的动态维护和自动扩展更新功能。

2）主题爬行策略。主题爬行策略的目标是保证爬行器尽最大可能多的下载与主题相关的网页，同时也要避免与主题无关或质量较低的网页，以此来提高主题资源的发现率和覆盖率。在制定主题爬行策略的时候要考虑很多因素，主要包括待爬行 URL 的访问次序等。主要的主题爬行策略算法有：主题优先策略、增量式的 URL 搜索策略、Fish 搜索策略、Shark 搜索策略等[13-15]。

3）网页去重策略。Web 上的网页存在大量重复的情况，一部分是对原始的网页进行完全的拷贝，另一部分主要是将原始网页中的主题内容放在不同的模板中进行转载。网页去重是信息采集中必须提前及时解决的问题，否则在后期信息处理中会占用大量资源，目前比较常用的网页去重策略是采用MD5 算法[16]。MD5 的作用是让大容量信息在用数字签名软件签署私人密钥前被压缩成一种保密的格式，它需要获得一个随机的信息并产生一个 128 位的信息摘要。该算法主要针对某一字符串可产生一个唯一的标识，只要字符串不同，产生的标识也不同。比如舆情中的热点话题、热点主题词等具有唯一性的属性值，若此属性值重复，则根据投票规则对现有信息进行更改，从而在保证信息完整性和准确性的同时，达到去重的目的。但是，MD5 是一种比较耗时的计算，并且占用的空间比较大[17]。

（2）网络舆情信息预处理。网络舆情信息预处理主要是针对从舆情站点上下载到本地的舆情页面，通过页面的预处理可以为下一步的工作打下基础。其中预处理主要包括自动摘要生成、关键信息提取、文本标签净化等。

1）自动摘要是指利用计算机自动地从原始文献中提取文摘，文摘是全面准确地反映某一文摘中心内容地简单连贯的短文[18]。通过自动文摘可以提高舆情工作者的信息利用率。用户只需要浏览自动摘要就能够了解文档中与查询词相关的部分，进而判断是否值得详细阅读整篇文档。目前，不同的摘要方法大致可以分为两大类：基于语法语义分析的理解式摘要和基于统计的机械式摘要。Web 文档自动摘要的抽取主要包括主题词提取、句子重要度计算和文摘句抽取三大环节，一篇文章中反映主题的词语往往被多次提及，词频是词语重要度的主要体现之一。句子重要度也是自动摘要的关键，对于它的计算需要给文档中的每个句子赋予权值，句子的重要性主要受句中所包含主题词的重要性和句子在文档中的位置的影响。对于文摘句的抽取，通常是将句子按权重从大到小进行依次排序，选取一定比例的句子加入文摘。

2）关键信息提取主要是提取出舆情页面中的核心内容，在提取页面核心内

容之前必须先对舆情页面进行净化处理，如去除页面中的广告信息，导航条，超链接等，通常网页关键信息提取是根据舆情网页的总体布局来提取的，一般页面中的核心内容都是用紧凑的文字信息进行表述，而非核心内容部分一般在文字中穿插着一些超级链接等。根据页面本身的结构属性可以有效的抽取出页面的关键信息。目前主要采用的方法是基于可视布局信息的网页噪音去除算法[20]和基于主题的网页噪音去除机制[21]。

（3）网络舆情信息分析。网络舆情信息分析是指将从舆情站点上所采集到的舆情语料进行关键词提取、聚类分析、信息跟踪、信息趋势分析等，然后通过比较舆情分析的阈值指标，从中挖掘出关键信息，对挖掘出来的信息进行有效的组织，最后结合舆情背景生成舆情结论，通过这个过程可以为下一步的舆情信息处理提供科学依据。其中主要的评价指标有关注聚焦度、内容危险度、主题敏感度等，关注聚焦度是指在持续数个数据统计期间内，某个信息点在热点排行榜上保持较高排名，这个信息点也被称为焦点。内容危险度是指舆情事件内容可能造成的危害程度，也称为重度。在网络舆情判别中会预先设定部分目标词，只要出现与预设目标词一致的信息点，比如“H1N1”、“疫情”等，不论该信息点的排名如何，就将该信息列为“重点”。预先对不同的预设目标词进行重要程度的判断，赋予不同的数值用来表示需要引起注意的轻重程度，称为“重度”[22]。主题敏感度是指过去某一时间段内，在热点排行榜上排名上升较多的信息点。在网络舆情信息分析中，倾向性分析也同样起着至关重要的作用，文本倾向性分析又称为情感分析，它主要实现文本中意见、情感和态度等信息的分析提取，这就需要对文本内容进行智能化的理解。文本倾向性分析的主要流程为：语料库准备，文本预处理，主客观句识别，倾向性判别等[23]。

（4）网络舆情信息输出。网络舆情信息输出主要指经过舆情信息分析后产生相应的舆情报告的过程，通过比较舆情信息的各项指标，对舆情信息进行舆情预警，然后对不好的舆情信息进行正确的引导。舆情预警主要是指从感知到舆情中的不良信息的征兆开始到可以预测出它可能带来的不良后果的时间段内，及时的采取有效的方法进行解除危机。它的预警流程主要包括以下三个环节：针对各种类型的危机事件，制定危机预警方案；密切关注事态发展，在第一时间大量的采集、汇集各个网站上的信息源；与舆论危机涉及的政府相关部门保持紧密的沟通，建立和运用信息沟通机制[24]。舆情预警指标是根据实际的情况来设定阈值，这个阈值是根据舆情主题下的信息条目随时间变化的百分比而设定的，通过阈值判断不仅可以反映当前主题事件是不是突发性事件，而且可以预测不良的舆情信息，当舆情分析的结果超出了预先设定的预定值时，可以及时的发出舆情预警。预警区一般分为[25]：蓝色良好区、绿色安全区（正常区）、橙色警戒区（基本安

全）和红色警戒区。

预警信息的应用可以帮助舆情分析人员及时发现一些网络热点或负面信息的预兆，同时注意群众的心理情绪和突发性的动态，如果引导不善，负面的网络舆情将对社会公共安全形成较大威胁。舆情报告是在得出舆情预警之后，根据相应的工作机制，加强对网络舆论的及时监测，进行有效引导，对网络舆论危机进行积极化解[26]。

2.2 现有网络舆情系统介绍

国外对于网络舆情的研究起步较早，在1996年，新加坡SBA（The Singapore Broadcasting Authority）设立监控网络有限信息中心，监控范围主要包括色情、政治、宗教、种族，内容提供商被要求用代理服务器对某些网络舆论信息来源进行过滤[27]。随后美国国防高级研究计划局（DARPA）在2002年提出了TIA(Total/Terrorism Information Awareness)计划，这个目的在于利用计算机的相关技术来分析和处理网络上海量数据的舆情问题[28]。尼尔森公司在2005年推出了“BuzzMetrics”服务软件，它主要在结合经验、数据、技术的基础上，帮助企业对在线言论及传播行为进行分析。英国的Autonomy互联网舆情监控分析系统在2006年被研发出，该系统主要注重语义分析，它主要是基于概念的数学算法，可支持海量的信息检索和自然语言检索，能够自动识别海量信息中的概念，并且自动实现上下文摘要、检索结果自动分组、信息关联等操作。使用语义分析的系统可以发现舆论的态度与倾向，可以更好地表现舆情的观点[29]。目前，虽然国内的网络舆情体系还不是很成熟，而且与国外舆情监测的模式有些不同，但是国外舆情监测的思想和方法对于我国这方面的研究起到一定的借鉴作用。

人民日报社网络中心舆情监测室是国内最早从事互联网监测、研究的专业机构之一，该机构已形成了一套比较完整的舆情监测理论体系、作业流程、应用技术等。随后国内建立了许多针对网络舆情的研究机构，如2005年11月，复旦大学成立的舆情研究实验室，2008年，中国传媒大学成立了网络舆情信息研究所等。目前，新华南方智库中心开发了一款新华南方智库全球网络舆情监测系统，该系统融合了数据挖掘技术、数据库技术、搜索引擎技术、多项自然语言处理技术等，是一个集“网络监测、信息采集、搜索过滤、智能分析、舆情处理、自动预警、报告生成”为一体的，数字自动化舆情监测和辅助决策参考的系统。如今已有许多网络舆情监测系统投入使用，但是由于业务模式和关注的方向不同，各个系统功能之间存在着很大的区别，有基于网络舆情管理的经验，在本公司已有的核心技术的基础上研发的，如中科点击科技有限公司开发的军犬网络舆情监控系统和中科新天科技有限公司开发的新天网络舆情监控系统等，有依托搜索引擎技术和文本挖掘技术研发的舆情系统，如谷尼国际软件有限公司开发的Goonie

网络舆情监控系统，有在本公司原有的智能技术的基础上改进而来的舆情系统，如北大方正技术研究院自主研发的方正智思网络舆情分析系统。具体比较如表2.1所示。

表2.1 网络舆情应用系统

网络舆情系统	开发者	应用时间及URL	主要功能
新天网络舆情监控系统	中科新天科技有限公司	2010年6月. http://www.newsksoft.com.cn	该系统集成了模式识别、自动标引、文本挖掘、机器学习、分布式计算、自然语言处理、统计语言学、关联分析等众多研究领域的最新成果，是一个面向网络舆情应用的完整的解决方案。
军犬网络舆情监控系统	中科点击科技有限公司	2009年1月. http://www.54yuqing.com	该系统在“第一时间”、“一站式”把境内、境外网站，对各种网络载体（如新闻、论坛、博客、微博）等全面布控检测，经系统对海量数据进行智能分析、稳准狠快地把互联网读懂、读薄。
Goonie网络舆情监控分析系统	谷尼国际软件有限公司	2008年1月 http://www.goonie.cn	该系统主要增加了自定义URL来源及其采集频率的设置、支持多种网页格式、多种字符集编码、对整个互联网进行采集、基于内容相似性去重等功能。为最后的决策者提供了强有力的舆论导向，对维护社会稳定、促进国家发展具有重要意义。
方正智思舆情分析系统	北大方正技术研究院	2007年 http://wwwfounderegov.com	该系统着重强调加强互联网、手机短信等新型传媒的信息搜集和分析，以计算机智能处理技术辅助舆情信息汇集整理和分析，对新出现的社会舆论热点、焦点去伪存真，为确保我国互联网大众传媒的舆论导向的正确性起到一定的辅助作用。

除了上述系统外，实际应用的网络舆情监测系统还有邦富互联网舆情监控系统、翼腾网络舆情信息监控系统等。这些系统大多是通过对BBS论坛、博客、新闻跟贴、转贴等实现并加以强化，对现实生活中某些热点、焦点问题所持的有较强影响力、倾向性的言论和观点进行分析与快速处理。通过国内网络舆情系统的比较分析，这些系统主要包括了以下三个基本功能：一是舆情信息采集，基本

实现了针对性的数据收集；二是舆情处理分析，主要是基于文本搜索和自然语言处理的基础；三是舆情结果的呈现，基本实现舆情报告的生成，因势利导舆情的发展方向。而如今现有的网络舆情系统及相关算法的研究都不是很成熟，还存在许多问题[30]。

2.3 网络舆情热点检测与跟踪技术的研究现状

目前，网络舆情信息主要通过网络媒体进行传播，而 BBS，微博和网站等都是网络媒体的主要载体，由于网络媒体的信息量大、范围广，这给网络舆情进行正确的分析造成了一定程度的困扰。针对现有的现象，很多研究机构都对网络舆情分析技术进行了深入的研究，本课题主要研究的是基于网络舆情的热点检测与跟踪技术，它主要的技术是在话题检测与跟踪技术的基础上展开的。目前，对于话题检测与跟踪技术的研究主要包括：话题检测与跟踪技任务[31]；话题检测与跟踪的关键技术[32]；话题检测与跟踪的评测标准[33]；现有网络舆情系统中话题检测与跟踪技术的应用。

2.3.1 话题检测与跟踪任务

话题检测与跟踪技术（Topic Detection and Tracking，TDT）的研究开始于 1996 年，当时美国国防高级研究委员会（DARPA）提出需要一种能自动确定新闻信息流中话题结构的技术。随后，来自 DARPA、卡耐基—梅隆大学（Carnegie Mellon University）、Dragon 系统公司以及马萨诸塞大学（University of Massachusetts）的研究者开始定义话题识别与跟踪技术的内容，并且开发用于解决问题的初步技术，TDT 的任务以及评测体系就是由他们联合在 1997 年制定和设计完成的。其实网络舆情热点检测与跟踪技术是一种面向信息安全的技术，它的本意是如何检测出新近发生的事件并追踪其后续发展动态的信息智能获取技术。舆情话题的检测和追踪可以使用户对互联网上的当前话题及时快速的了解和掌握，这样对加强互联网信息的管理和监控和全面掌握社情民意发挥了较好的作用。

目前，主题检测与跟踪技术研究集中于五个子任务展开，各个子任务的解决将有助于最终研究目标的实现。这五个子任务包括：对报道的切分、新事件的识别、报道关系识别、话题识别和话题跟踪[34]。

（1）报道切分。报道切分（Story Segmentation Task，SST）的主要任务是找出所有报道的临界点，在此基础上把输入的初始数据流分割成各个独立的报道。如一段节目中通常会包含股市行情、娱乐八卦、体育竞技等很多条分类报道，它主要要求系统能够模拟人一样将某一时间戳中的语料切分成多个不相关的话题。该系统的性能完全依赖于数据源的形式以及做出决策准许的最大延迟时间，通常采用最大熵和决策树混合的模型来处理这个问题，它利用各种与信息源相关的特征，比如句子长度、播音语速、信息员在节目中的位置，以及字或词的 N

元文法。报道切分是其他四个任务的预处理，而其他任务是在报道切分的基础上进行的[35]。

（2）新事件的识别。新事件检测（New Event Detection Task，NED）的主要任务是在报道信息流中识别出对一个新话题的首次报道，比如某个娱乐丑闻、政治活动等等。NED 的主要任务是从具有时间顺序的报道流中及时的查找与发现未知话题出现的第一篇相关报道。目前，在新事件识别任务中采用典型的方法是：用以向量或概率分布形式表示的特征集合代表每一篇报道，每遇到新来的报道，就将其特征集合与过去所有报道的特征集合进行比较，据此便可以判断该报道是不是在描述一个新的话题。James Allan 和 Yiming Yang 提出的 NED 的方法，主要是通过建立一个在线识别系统来检测报道流中的最新出现事件，主要通过计算进入该系统的报道与每个已知事件的模型之间的相似度来实现，根据预先设定的阈值来判断该报道是否为新事件的第一次报道[36]。CMU 使用 Single-pass 算法进行新事件的探测，该算法虽然计算简单、运算速度快，但它的执行结果由于过分依赖于语料被处理的顺序，从而影响它的探测性能[37]。

（3）报道关系识别。报道关系识别（Story Link Detection，SLD）的主要任务是判断两则报道是否在讨论同一个话题，SLD 也不具备先验学习的知识，每一对参加报道关系识别的两篇报道都没有相应的先验知识辅助系统来进行评判。因此 SLD 系统必须设计检测模型，通过检测模型之间的比较，提取出彼此之间的相似性。目前，国内外对这方面的研究也比较关注。James Allan 和 Schultz 用向量空间模型来描述报道的特征空间，并利用计算特征之间的余弦夹角来判断两篇报道之间的相似性[38]。同时，Umass 对事件探测与跟踪也进行了深入的研究，它主要根据词法特征自动生成多个分类器，并且每类事件通过包含查询语法和阈值的分类器来表现，最后根据标准化后的相似分值确定事件的类别归属[39]。也有直接通过分析报道内容的结构关系和语义分布规律，提出了基于语义域语言模型的关联检测方法，并融合依存关系辅助其语义描述，在此基础上建立了话题模型，并将该模型参与报道相关性的计算中[40]。

（4）话题识别。话题识别（Topic Detection Task，TD）的主要任务是从输入的文本语料中发现以前未知的新话题。通过上述可知话题识别对话题没有先验知识，比如舆情中的关键词“H1N1”和“地震”，这些都是未知话题，TD 需要检测出并建立新的话题簇。话题识别一般可以看作是一种聚类的方法，检测的目的就是要按照报道所要表达的话题将其相似性高的报道进行聚类。聚类是信息组织的重要手段，它是根据对象自身信息之间的相似性进行归类并划分成簇。空间表示，距离计算以及算法的选取是聚类技术的三大关键要素。在距离计算中，常见的有余弦距离计算和欧式距离计算，其他的距离计算还有明氏距离、马氏距离、

兰氏距离等等[41]。简单的说，话题识别是对新事件识别任务的一个自然的扩展。目前，国内外研究者常采用的算法有：增量K-means聚类、单遍聚类、agglomerative聚类等[42]。Dragon Systems提出了一种基于简单K-means聚类的话题检测算法，该算法一次性将报道划分到最近邻的话题类簇中，而这个聚类过程在K-means算法中要多次进行，如果没有找到合适的话题类簇那么就创建新的话题类簇[43]。

（5）话题跟踪。话题跟踪（Topic Tracking Task，TT）的主要任务是首先得到某一个时间戳的报道，紧接着将这些报道进行大量的训练，最终得到一个话题模型，然后将下一个时间戳输入的报道与话题模型通过相似性计算，从而进行话题分类，找出所有讨论目标话题的报道。话题跟踪的性能主要受到以下一些因素的影响：训练用的报道的数量，训练及测试预料使用的语言，文字记录的质量等。目前，主流的话题跟踪算法都是在改进分类算法的基础上实现的，主要有Rocchio等分类算法。国内外有多种不同的方法在这项任务研究中被尝试使用。如James Allan[44]和Michael[45]用Rocchio算法来实施话题跟踪，Rocchio算法的基本思想是话题模型的经验性构造策略，即假设相关报道的某些特征可以帮助对该话题的正确描述，因此这些特征在话题模型中的权重被相应的加强，反之则消弱。也有直接利用链接分析的话题跟踪方法，是在内容计算的基础上引入链接分析技术[46]。该方法的核心思想是根据种子报道和内容相似度较高的网页，可以为它们所指向的网页进行“投票”，然后设定一个阈值，如果非种子报道网页的内容相似度高于这个阈值，则该报道认为是与这个话题相关的报道，可以为它的相关链接所指向的网页加分，分值越高，说明被跟踪的概率越大。Statoshi Morinaga和Kenji等人提出了利用一个有限混合模型动态追踪话题发展趋势的方法，该模型集合了话题发现、新事件发现以及话题追踪于一体，可实现实时动态话题趋势分析[47]。文献[48]提出了基于反馈学习自适应的中文话题追踪技术来解决追踪过程中话题的动态演变性，由于话题从无到有到漂移的过程中，话题特征也在随之变化，所以在自适应过程中需要考虑利用调整特征权重的方法来优化追踪效果，因此将话题向量更新与调整特征权值的方法相融合的自适应方法可以更好的提高话题追踪的性能。

2.3.2 话题检测与跟踪的关键技术

目前国内外对热点话题检测与跟踪技术的研究主要集中在话题模型建立、特征项选择、话题相似度计算、话题聚类与分类策略四个方面。

（1）话题模型建立。话题模型建立是话题检测与跟踪技术中的首要任务，一般要确定一个话题或者识别出一个新话题时，都需要建立一个话题模型。目前常用的话题模型主要有语言模型、向量空间模型等[49]。语言模型是用来计算一个句子的概率的概率模型。它最开始因语音识别领域而诞生的。语言模型具体描述如

下：

假设某报道中出现的词 w_n 各不相关，则某则报道S和话题C相关的概率是：

$$p(c|s)=\frac{p(c)p(s|c)}{p(s)}\approx p(c)\prod\frac{p(w_n|c)}{p(w_n)} \tag{2.1}$$

其中，$p(c)$是任意一则新报道与话题C相关的先验概率，$p(w_n|c)$表示词 w_n 在某个话题C中的生成概率。

向量空间模型（VSM）主要是将文本内容处理成向量空间中的向量运算，通过文档的特征项将文档表示成文档空间向量，在对文档处理时，就可以运用余弦距离、欧式距离等方法处理向量之间的相似性。它的基本结构如下表示：

文档d转化成向量空间模型：$d=d(t_1,w_1;t_2,w_2.....;t_n,w_n)$，其中$t_1$，$w_1$是文档d的特征项。在实现话题检测与跟踪算法时，一般都先将话题模型转化成中心向量模型，通过中心向量间的相似度计算，可以将各个话题进行分类，从而实现该算法。

（2）特征项选择。文本的特征项通常是由字、词、词组等文本的基本语言单位所组成的集合，这些基本的语言单位可以统称为项。在文档中需要给这些特征项赋予一个权值，通过权值的大小来辨别这些特征项的重要程度。目前常用的权重计算方法有布尔函数、TF-IDF、特征频度等方法[50]。布尔函数描述的是如何基于对布尔输入的某种逻辑计算确定布尔值输出。一般当一个特征项通过布尔函数表示的时候，它只有两种情况：存在或不存在。具体公式如下所示：

$$w_{ij}=\begin{cases}1, tf_{ij}\geq 1,\\ 0, tf_{ij}=0.\end{cases} \tag{2.2}$$

特征频度权重计算方法主要是统计特征项在文档中的次数多少，该方法忽略了特征项在整个语料集的分布情况。TF-IDF 权重计算方法不但考虑了特征项在单个文档的信息，还考虑了文档频率。通过 TF-IDF 可以得出某个关键字在某篇文章粥的重要性，计算方法如下：

$$w_{ij}=tf_{ij}\times idf_{ij} \tag{2.3}$$

其中，tf_{ij}表示特征项i在文档d_j中出现的次数。idf_{ij}是逆文档频率，指出现特征项i的文档个数的倒数，计算方法如下：

$$idf_i=\log(N/n_i) \tag{2.4}$$

其中，N表示整个文档集中文档的个数，n_i表示文档集中出现特征项i的文档个数。

（3）话题相似度计算。话题相似度计算是话题检测与跟踪技术中的关键技术，它可以将某一报道通过相似度计算可以进行归类。从而识别出它是新话题，还是归属于原来的话题。目前常见的计算文本向量之间相似度的公式主要有内积、余弦相似度、Correlation 距离、Spearman 距离、Euclidean 距离等[51]。此外，IBM

的 Okapi 公式也得到了较多的应用[52]。

（4）聚类与分类策略。文本聚类主要是依据同类的文档相似度比较大，而不同类的文档相似度较小的思想。它作为一种无监督的机器学习方法，所以不需要任何的文本训练过程。常用的文本聚类方法主要有两大类，即系统树状图的等级聚类方法和基于平面划分的动态聚类方法[53]。等级聚类方法主要通过建立并逐步更新距离系数矩阵，找出并合并最近的两类，直到全部对象被合并为一类为止。动态聚类算法主要是在将一个样本随机的进行切分，通过不停的迭代每次选出聚类中心，当聚类中心变化小于某一个阈值时，则完成了聚类过程。K 均值聚类算法就是典型的动态聚类算法。动态聚类算法的流程图如图 2.3 所示。

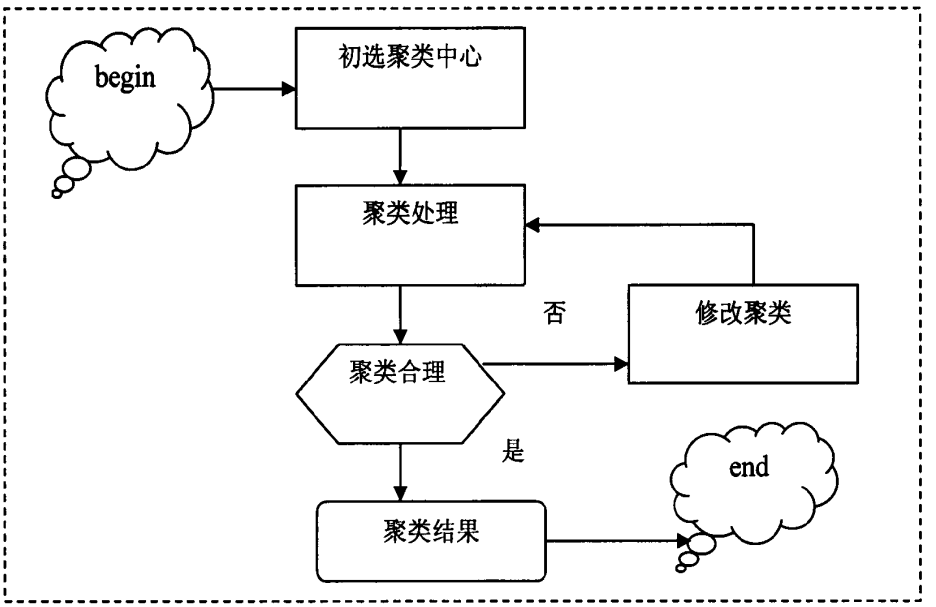

图 2.3 动态聚类算法的工作流程图

话题检测与跟踪技术的研究中，分类策略的合理选择也可以完成相应的任务，与聚类策略相比，分类是一种有监督的学习过程，它需要一批样本数据进行事先的训练，然后得到数据分类模型，最后根据后续的文本内容自动进行归类。常用的分类方法有 KNN 方法、Naïve Bayes 分类方法、SVN 方法等。

2.3.3 话题检测与跟踪评测标准

在话题识别与跟踪领域中，最常用到的系统性能评价指标是归一化识别代价 $(C_{Det})_{Norm}$，它是由系统的识别漏报率和误报率计算得到的，计算公式如下[54]：

$$(C_{Det})_{Norm} = \frac{CDet}{\min(C_{Miss} \bullet P_{Target}, C_{FA} \bullet P_{non-target})} \tag{2.5}$$

$$C_{Det} = C_{Miss} \bullet P_{Miss} \bullet P_{Target} + C_{FA} \bullet P_{FA} \bullet P_{non-target} \tag{2.6}$$

其中，C_{Det} 是系统的错误识别代价。它由公式（2.4）计算所得，C_{Miss} 和 C_{FA} 分别是漏报和误报的代价。P_{Target} 是先验的话题在新闻报道中的出现概率，P_{Miss} 和

P_{FA}是系统识别的漏报率和误报率，$P_{non-target}=1-P_{target}$。

2.3.4 基于话题检测与跟踪的网络舆情系统的应用

目前主要有 Goonie 网络舆情监控分析系统和方正智思网络舆情监控分析系统等对突发事件进行跨时间、跨空间综合分析，获知事件发生的全貌并预测事件发展的趋势，在主题检测与追踪技术上这些系统主要运用自动分类与自动聚类算法，使得它们能够自动检测信息片断集合中的各个未知主题，并能在线检测出新主题，并且在各种信息来源中追踪那些讨论目标主题的相关信息片段。也有类似蚁情的舆情分析系统，它主要适用于企业进行产品口碑跟踪，技术和商业情报的收集，以及社会维稳进行舆论研判、引导与管理。在实现舆情热点检测与跟踪上，它主要实现了独特的现实与虚拟社区的信息切换过程。总之，从各个产品的应用来看，基于话题检测与跟踪的网络舆情系统的应用越来越受到企业、政府的关注。

2.4 本章小结

本章介绍了网络舆情热点检测与跟踪技术的相关知识，涵盖了网络舆情信息挖掘的各个方面，热点检测与跟踪的关键技术，并分析此技术在在网络舆情中起着重要的作用，为本文后续的网络舆情热点检测与跟踪算法研究作了相应的知识准备。

第三章 基于共词分析的网络舆情热点检测

本文前述章节主要介绍网络舆情热点检测与跟踪技术所涉及的一些基本理论，本章着眼于基于共词分析的文本主题词聚类与主题发现算法的研究和实现，提出了一种新的网络舆情热点检测的方法。

3.1 算法总体框架

在本节中，对基于共词分析的文本主题词聚类与主题发现算法的主要流程作基本介绍。算法的总体流程如图 3.1 所示。

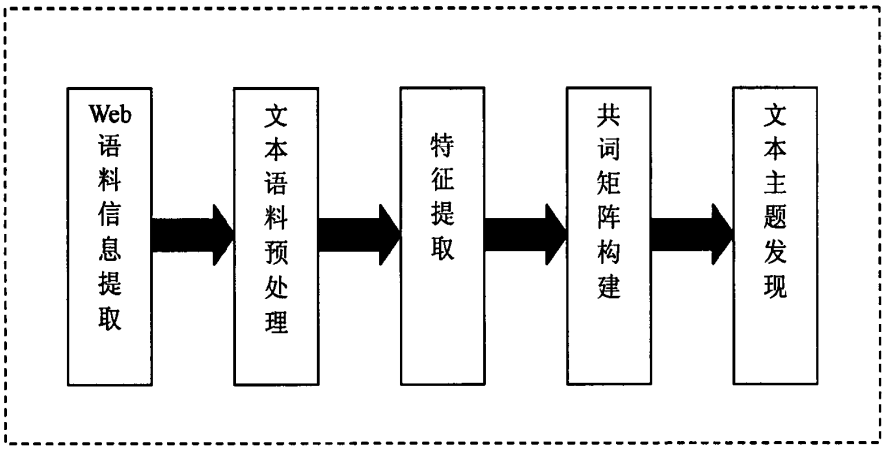

图 3.1 算法总体流程图

其中本章节具体改进的算法是将共词矩阵与 Bisecting K-means 聚类算法相结合，并将他们运用到真实的 BBS 语料环境中，主要是将设经过预处理之后带有年、月、日时间标签的论坛语料归入各个时间戳，经过分词、停用词过滤等预处理操作之后，统计出词语 i 的词频和文档频率，候选关键词提取主要是根据这两个参数得到的。然后通过词与词之间在文档中的共现程度构建出一个共词矩阵，并且根据词与词之间共现频数与彼此间距离成反比的思想，对构建出的共词矩阵进行标准化。这样不仅可以很直观的衡量各个词之间的距离，而且可以作为 Bisecting K-means 聚类算法的入口，从而进行分层聚类，通过大量的实验，发现在 BBS 论坛的环境下可以得到各个时间戳的 BBS 特定版块中的热点话题。

3.2 学科领域中的共词分析方法

共词分析法最早在 20 世纪 70 年代中后期由法国文献计量学家提出来的。共词分析经过 20 多年的发展，方法已经被广泛应用到各个领域[55]。到目前为止，共词分析方法产生了大量的应用成果，在人工智能、信息科学、信息管理系统和信息检索等领域都得到了很好的应用。共词分析法主要利用语料库中的词汇对或

者名词短语共同出现的情况，来确定该语料库所代表学科中各个主题之间的关系，一般认为词汇对在同一篇文献中出现的次数越多，则代表这两个主题的关系越紧密[57]。因此我们可以基于这样一种假设：文献的关键词是关于文献内容的充分描述，如果两个不同的关键词出现在同一篇文献中，则认为这两个关键词之间有一定的联系[58]。基于这种概念，研究主题可以用几个特定的关键词来表示。通过计算出一组语料库的主题词两两之间在同一篇文献中出现的频率，便可以得到一个由这些词组成的共词网络，网络内节点之间的远近便可以反映主题内容的亲疏关系。钟伟金等人详细分析了共词分析法的过程与方式、共词分析法的类团分析的原理与特点。Qin(2001)对共现分析的研究步骤分阶段进行讨论，考察共现分析研究过程中一些既定假设，并对研究结果的显著性进行分析[56]。

共词分析方法发展至今，它主要是通过对能够表达某一学科领域主题或研究方向的专业术语共同出现在同一篇文献或论文中的现象的分析，判断学科领域中主题间的关系，进而展现该学科的研究结构[57]。例如：基于共词分析的科技文献的研究、基于共词分析的学科主题的动态跟踪等等。一般研究共词分析的步骤如图3.2所示：

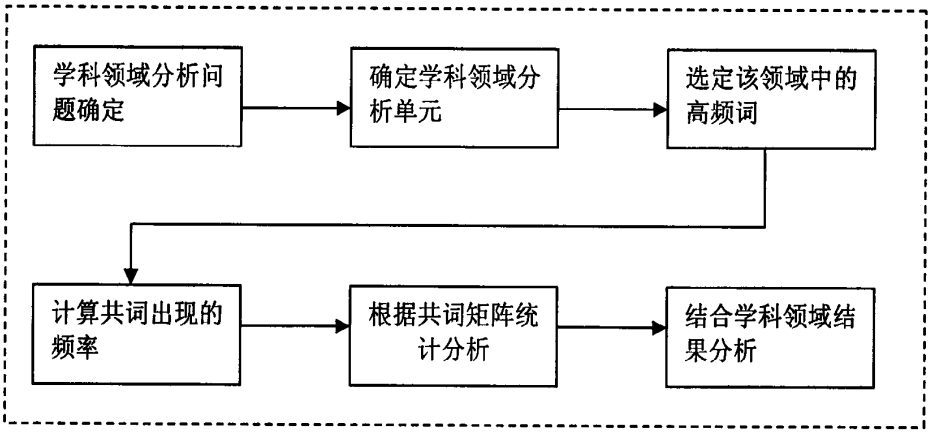

图 3.2　共词分析研究步骤

在相互关联的共词网络中，某一学科领域下的一个主题和多个主题形成关联，相互间构成了一个关系网，在这个关系网中很难分辨出哪些主题词组成的类团，为此我们要借助数据挖掘中的聚类分析法。聚类是信息组织的重要手段，它是根据对象自身信息之间的相似性进行归类并划分成簇。空间表示，距离计算以及算法的选取是聚类技术的三大关键要素。共词聚类分析方法是共词分析中的一种。由于某一学科下的主题词具有一定的受控性与规范性，它们代表的学科研究方向是学科内研究的热点问题。共词聚类分析法正是基于这一原理，把学科研究领域内的关系密切的主题词聚类成团，并对类团进行深入分析[58]。

3.3 BBS 中的关键词共现分析

本文认为关键词共现分析方法可以作为网络舆情热点挖掘的有效工具，而 BBS 是网络舆情主要的载体之一，它可以实现以下的目标：

(1)能将网络论坛内的各个知识点关联在一起，反映出其中的热点主题。

(2)能够追踪研究各个时间戳主题之间的演变过程，可以分析其所处自身发展阶段。

(3)分析结果可为科研人员提供重要参考资料，有助于科研人员从全局上把握整个舆情的热点分布。

共词分析方法用于网络舆情热点挖掘的主要优点：共现分析是一种定量与定性合理结合的分析方法。单纯定性或定量分析网络舆情语料，都不能很好的解决网络舆情热点挖掘的问题，而共词分析则结合了定量与定性两种思路：首先将舆情内容转化为用数量形式表达的信息，运用数学方法定量计算研究结果后再定性分析，以理解其中蕴含的新的、有意义的知识。克服了定性研究的主观性和不确定性，能更深刻、更精确地挖掘文本知识。

关键词共现分析方法在网络舆情热点检测中具有重要地位，与其他挖掘方法相比具有客观性、及时性等优点，本文提出的一个完整的关键词共现分析方法并将它运用到真实的 BBS 论坛环境下，共词聚类的结果是文本语料内容现状的客观的真实的反映，直接通过共词矩阵得出各个关键词之间的距离，并将这个距离作为聚类分析中的一个参数。它主要包含如下几步，数据选择与提取、候选关键词的提取、构建共词矩阵、共词矩阵标准化、聚类分析、可视化展示。具体步骤如图 3.3 所示。

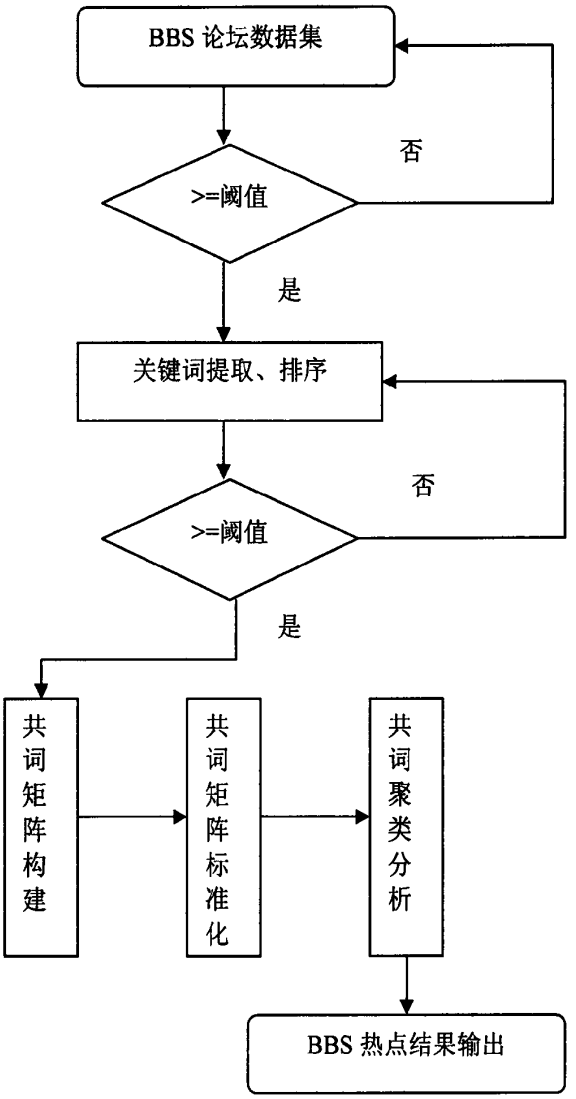

图 3.3　基于 BBS 上的共词分析的研究步骤

3.3.1 数据选择与提取

数据来源的好坏直接关系到实验结果的准确度，因而所选数据源务必最大限度的包含所有与主题相关的数据。网络舆情的数据主要来源是通过 BBS 论坛、博客、新闻跟帖、转帖等实现并加以强化。本文的数据来源主要通过 BBS 论坛，并且针对不同网站的 BBS 论坛所对应的类似版块进行针对性的分析。在对数据源的选择上主要基于以下两点。

(1)所选 BBS 论坛必须具备较高的知名度，内容版块全面，网民数量集中；

(2)所选 BBS 具有实时性、传播性，可以及时反映民众的情绪，真实的体现民众的社会政治态度。

通过以上几点考虑，本文选择使用网络爬虫技术对天涯论坛、新浪论坛、网易论坛中的军事版块进行了文本语料抽取。网络爬虫是搜索引擎中的关键技术，

它的主要功能就是根据某些搜索策略为搜索引擎从互联网上下载的各种信息。网络爬虫的基本功能如图 3.4 所示。

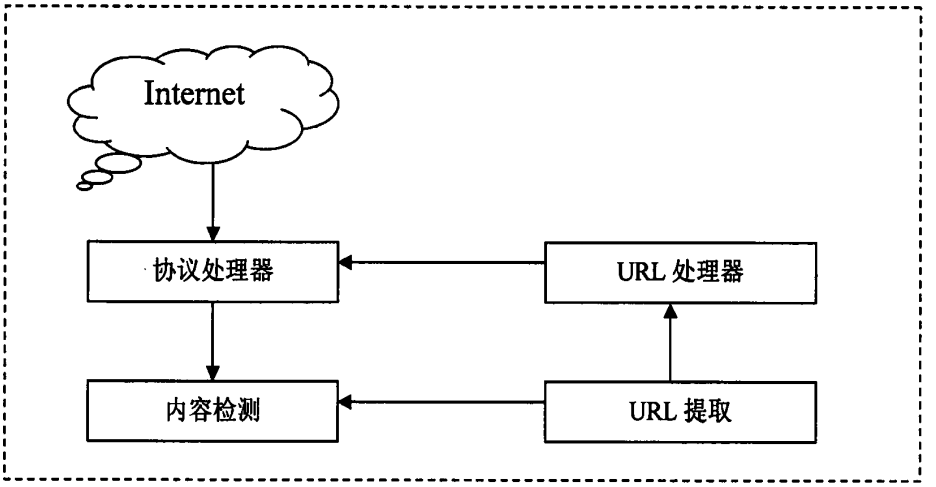

图 3.4 网络爬虫基本功能

本文的网络爬虫是基于 Larbin(一种开源的网络爬行器)的一种应用，按照 Labin 自身的配置，运行速度比较快，占用的内存空间比较小[59]。Web 上的网页按照内容一般分为有主题网页、广告网页和多媒体网页。主题网页中以文本为主的内容块是我们抽取的重点。在 HTML 规范中定义了一套标签来规划网页内容的布局（如：<BODY>、<B>、<TABLE>等），通过标签的规则本文过滤出主题内容和跟帖信息。抽取出语料的格式为：

……

中新社东京 3 月 24 日电福岛第一核电站排水口附近海水中放射性碘的浓度 24 日再次超过百倍地高出法定值，且有两种新的放射性元素——碲和钌被发现。日本文部科学省当天也称，距福岛核电站附近海域放射性碘超标，放射性铯也在大幅增加。东京电力公司 24 日宣布，23 日对福岛第一核电站排水口南侧 330 米处所取海水样品的检测结果显示，海水中碘 131 的含量达到 5.9 贝克勒尔，超过法定浓度的 146.9 倍。同一地点的海水，21 日检测的结果是碘 131 含量超标 126 倍。此外，当地还发现放射性钌 106 的含量超出规定的 3.7 倍。而该公司对排水口北侧 30 米处的海水进行检测时发现碘 131 超标 66.6 倍，铯 134 超标 29.9 倍，放射性碲 132 超标 7.8 倍，碲 129 超标 4.2 倍。其中，碲和钌都是首次被发现。东京电力公司还表示，核电站北侧海水被发现比排水口附近的放射性物质含量更高，并解释说这可能是与放射性物质扩散时的风向、洋流等情况有关。日本文部科学省 24 日也宣布，福岛第一、第二核电站近海 30 公里海域的放射性物质调查显示，8 个观测点中有 3 处放射性碘 131 的含量超过规定值，其中浓度最高的地方达到每升 76.8 贝克勒尔，高出每升 40 贝克勒尔的法定浓度近一倍。同时，8 个观测点的放射性铯 137 的浓度达到每升 11.2～24.1 贝克勒尔，虽然未超出规定的每升 90 贝克勒尔，但在该海域此前的调查中，铯 137 的含量仅为 0.0015 贝克勒尔。参与天涯用户调查送红包。

……

紧接着将BBS论坛中的大量的帖子通过网络爬虫下载到本地，然后从中提取出热点帖子，从而进行热点话题的提取。总的来说，热点帖子的提取是热点话题提取的基础。对于热点帖子的分析，它主要是受论坛帖子的回复数和浏览数这两

个参数的影响，这两个参数可以比较直观的反应当前帖子的一个热度，而往往我们要提取的热点话题都是基于热帖的相关话题。与此同时，考虑到帖子的时空分布特性，某一集中时间戳内权重较大的帖子所讨论的话题往往就是论坛的热点话题。本文对采集到的论坛网页进行分析，将帖子的权重值作为首要因子，以权重值的大小进行降序排序。其中权重的计算公式为：

$$weight(m) = a * ReplayCount(m) + b * BrowseCount(m) \quad (3.1)$$

其中a，b取值范围在0到1之间，且a+b=1。$weight(m)$表示帖子m的权重，$replayCount(m)$为帖子m的回复数，$browseCount(m)$为帖子m的浏览数。a，b为调节系数，一般将a取0.8，b取0.2。

然后本文根据帖子的权重值依次将帖子进行从大到小进行排序，通过阈值判断取出每天排名前 50 的帖子。这样分析的结果是提取出来的帖子基本上是热点帖子。这为下一步热点话题的提取打下基础。

3.3.2 候选关键词的提取

（1）文本预处理。首先本文对提取出的论坛语料进行分词，本文算法采用了中国科学院计算技术研究所自行研制的分词系统 ICTCLAS（Institute of Computing Technology，Chinese Lexical Analysis System）对论坛语料进行分词[60]。然后对分词后的语料进行停用词的过滤，本文算法载入了一个根据中科院的 ICTCLAS 分词系统的二级标注匹配，从一批 4G 大小的天宇语料中提取出并列连词、连词、叹词、介词、人称代词、处所疑问代词、时间疑问代词、谓词性疑问代词、疑问代词、处所指示代词、时间指示代词、谓词性指示代词、指示代词、代词、助词、语气词构成的一个词表，并结合哈工大停用词表进行扩展。该词表包含 900 多个停用词，以及一些标点符号、数字等特殊符号，共计 1176 条。

（2）基于 TF-IDF 的关键词提取。TF-IDF（term frequency-invers document frequency）是一种用于文本挖掘与信息搜索的常用加权技术。它是一种统计分析的方法，用以评估一字词在整个语料库中的重要程度。字词的重要性随着它在文件中的出现次数成正比增加，但同时会随着它在语料库中出现的文档频率成反比下降。它倾向于过滤到常见的词语、保留重要的词语。TF（term frequency）其实就是某个关键词出现的频率，DF（document frequency）就是说某个关键词在 N 篇文档中出现的次数。得到 DF 后我们便可以通过公式计算 IDF（invers document frequency）。本文所采用的计算 IDF 的公式如下：

$$IDF = \log((M+1)/DF) \quad (3.2)$$

其中 M 表示语料库中的文档总数，DF 为关键词的文档频率。 它的主要思想是：如果包含词条 t 的文档越少，也就是 DF 越小，则 IDF 越大，则说明词条

t 具有很好的类别区分能力。主题词的特征包括词频、文档频数和反文档频率（IDF），如：钱学森 TF=4439；DF=1237；航天 TF=1183；DF=408；接种 TF=7621；DF=1176；等，通过计算 TF*IDF 的值将关键词进行按降序排列。在大量的关键词中本文通过 TF-IDF 的大小设定了一个阈值，通过阈值判断选出排名靠前的关键词，这样分析的结果是提取出来的关键词基本上属于热点主题词。这为下一步的工作打下基础。

3.3.3 基于共词分析的文本主题词聚类

（1）共词矩阵的构建。关键词共现矩阵是进行聚类分析的基础。为方便词对共现频率的运算，本文设计了一个共词矩阵。由此，我们统计一组语料库的主题词两两之间在同一篇论坛帖子中出现的频率，便可以构成一个由这些词对关联组成的共词矩阵。如果两个主题词在众多的文献中出现频率高，则说明它们关系密切，在共词分析中，对于 N 个关键词的共词分析，便形成了 N*N 的共词矩阵。具体算法流程见图 3.5。

（2）共词矩阵标准化。由于主题词在同一篇文章出现的频率与他们之间的距离成反比。与此在度量距离的时候，本文利用共词统计算法，通过计算主题词共现的词频，然后将共现词频进行加 1 取倒数进行处理。具体公式如下：

$$l = \frac{1}{s + 1} \tag{3.3}$$

其中 S 代表词与词之间的共现频率，l 代表主题之间的距离。

（3）主题词聚类。本文在度量距离计算是根据共词分析生成的共词矩阵，根据主题词对在同一篇帖子出现的频率越高反映出词对关系更加紧密，则他们之间的距离越小的思想原则，通过共词矩阵算法统计主题词在同一篇文章中两两同时出现的次数（在同一篇帖子中不累加，依次对不同帖子共同出现进行累加），然后通过上一步共词矩阵标准化，便可以得到主题间的距离。也就是说主题距离与相关帖子的篇数息息相关。在聚类算法中，本文将主题的距离值作为 Bisecting K-means 聚类算法中的一个参数，对形成的 N*N 的共词矩阵以数组的形式进行输入并进行聚类。其主要思想是根据主题距离值发现描述对象之间的关系信息，将描述对象进行分组。其目标是，组内的对象相互之间是相似的，而不同组中的对象是不同的。组内的相似性越大，组间的差别就越大，这样结果就越好。Bisecting K-means 聚类算法在分裂出一个簇时，它不需要去计算所有簇的中心距离，它只需要计算每个点与其中的两个簇的中心距离[61]。它其实是一种分裂式层次聚类算法。

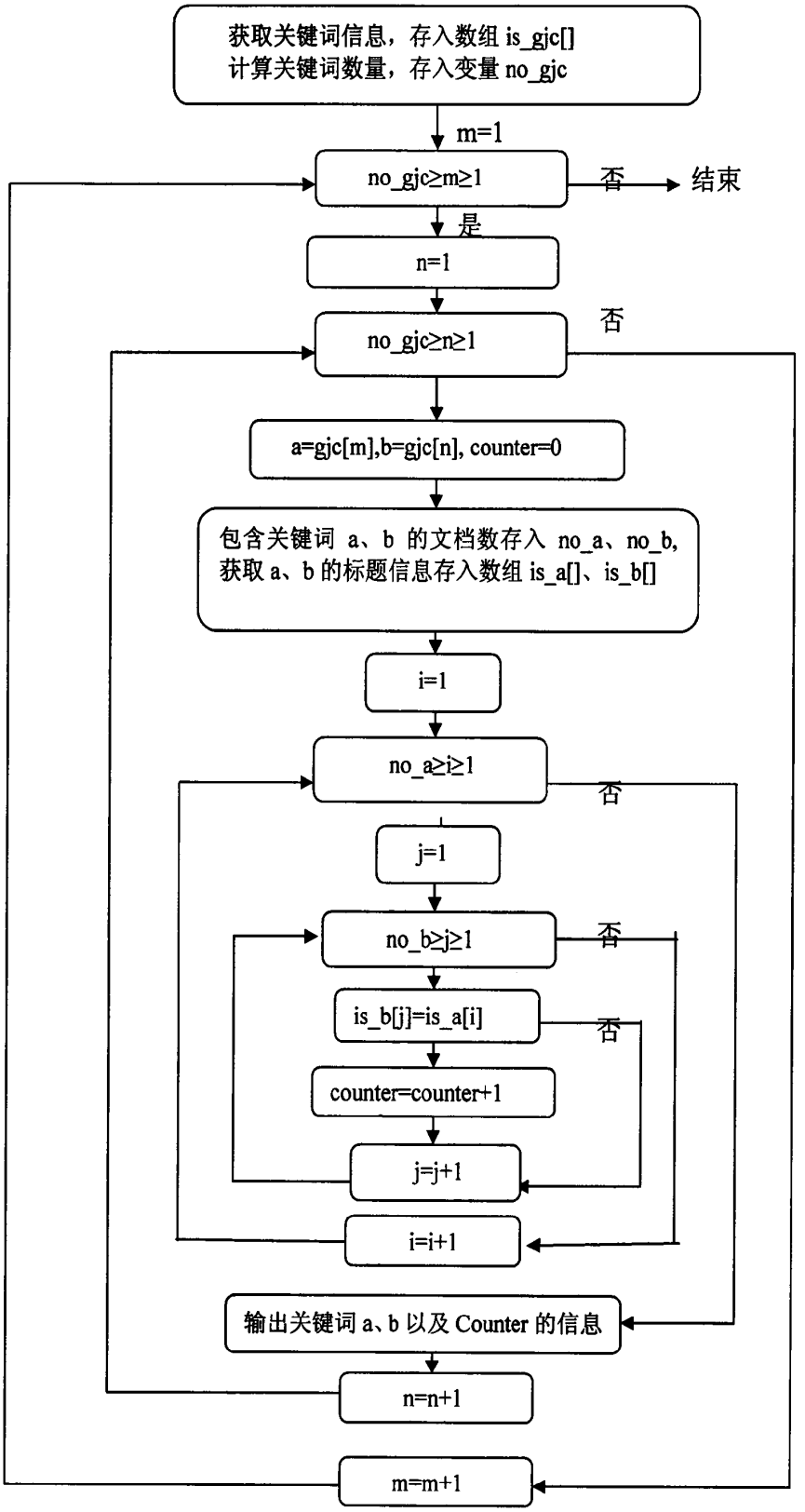

图 3.5　共词矩阵计算程序

其 Bisecting K-means 算法基本流程如下：

步骤 1：利用随机分裂的方法选择一个群集进行切分，对于共词矩阵以数组的形式进行输入，其中关键词之间的距离比较直观的体现在共词矩阵上。

步骤 2：通过中心点偏移策略从现有的簇中选择其中一个簇通过使用标准的 K-means 算法拆分成 2 个簇，依次循环。

步骤 3：重复上一步步骤，待中心点偏移量达到比较小、趋于稳定的时候，从多个循环组中选出相似度比较高的聚类结果。

步骤 4：如果关键词已经拆成 K 个簇则结束算法。

具体算法如图 3.6 所示。

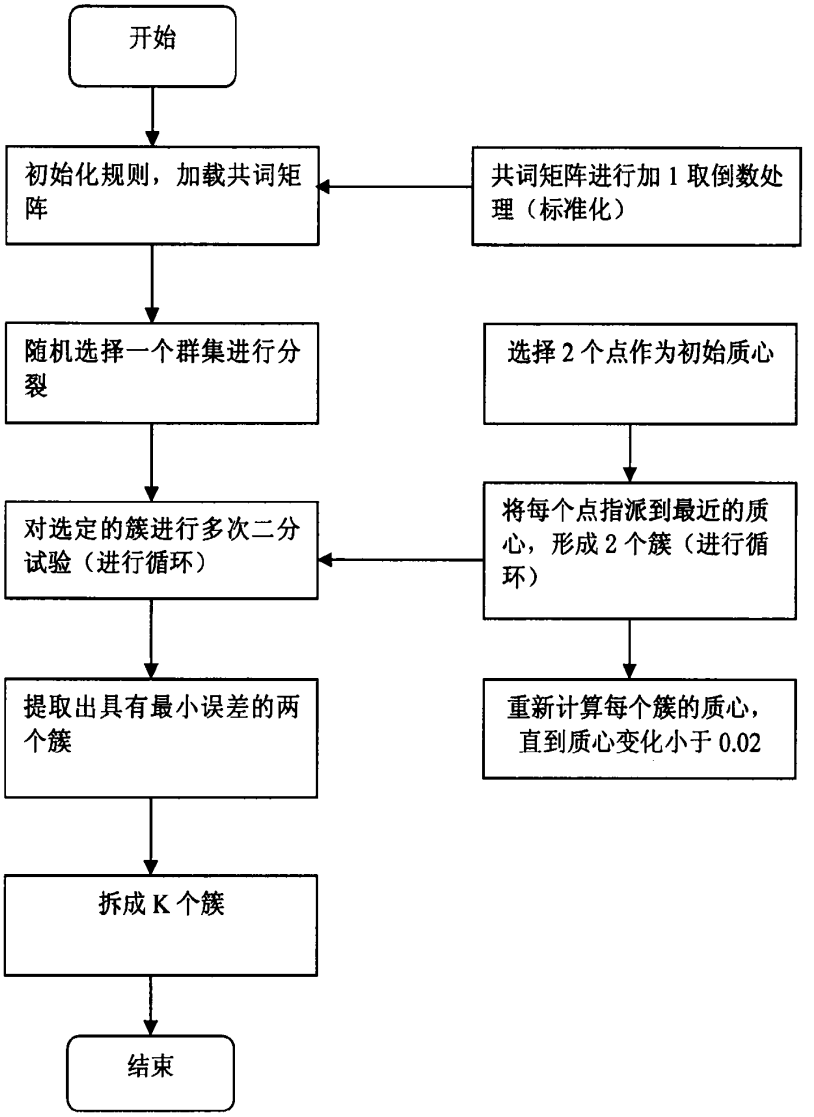

图 3.6　聚类算法

其中本文使用的邻进度函数是余弦函数，质心和目标函数分别选为质心，最

大化对象及质心的余弦相似度和。在选取初始质心时，进行多次运算，每次适用一组不同的随机初始质心，然后选取具有最小目标函数和的簇集。

3.3.4 测试实例及其分析

本章节采用新浪网的论坛语料作为测试实例，将新浪网上采集到964个网页作为语料，去除网页中的链接，导航等信息，处理成纯文本形式，只包含标题和正文，因为它在反映网络真实环境的同时又具有一定的系统性。网页日期从2009年11月1日到2009年11月15日。实例测试结果如下所示，表3.1为基于共词矩阵的Bisecting K-means聚类结果。

表3.1 Bisecting K-means聚类结果

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7	Cluster8	Cluster9
报名	流感	钱学森	总统	受灾	李义	气温	陈玉蓉	公务员
征兵	疫苗	钱老	奥巴马	灾区	开县	寒潮	手术	考试
征集	病例	航天	访华	宾川县	非法	冷空气	儿子	确认
女兵	H1N1			救灾	期徒刑	天气	移植	报考
毕业生	疫情			抗震救灾	犯罪	降温	妈妈	考生
应届	重症			倒塌	判处		肝脏	网上
青年	卫生部				黑社会		母子	考核
体检	医疗				告人		母亲	
应征	应急				团伙			

根据实验的结果，我们发现关于同一事件的主题词，使用Bisecting K-means聚类算法可以很好的将同一事件的词串聚到了一块。我们参照新浪网上对该时段的热点排行，表3.2便是该时段新浪网上热点的排行。

表3.2 热点排行

序号	热点排行
no1	国家卫生部：H1N1与接种疫苗毫无关联
no2	中国航天之父钱学森今日在京逝世
no 3	我国将受到气温突降的恶劣环境
no 4	云南宾川县今晨发生5.0级地震
no 5	国家公务员考试报名已启动
no 6	国防部征兵办就女兵征集工作试行办法答问
no 7	“暴走妈妈”今日割肝救子
no 8	奥巴马即将访华
no 9	重庆开县涉黑团伙首犯李义被判20年

从实例的测试实验聚类结果和表2的比较可以得出结论，本测试实例的热点

主题词的聚类结果很好的反映了新浪网上排行的前 10 的热门标题，说明基于共词分析的 Bisecting-kmeans 聚类算法在网络舆情热点发现中具有一定的可行性。

3.4 本章小结

本章节首先把论坛语料通过 ICTCLAS 进行自动分词，对分词后的语料进行停用词过滤，根据 TF-IDF 关键字提取的方法提取新闻语料库的关键词。然后通过共词分析来构建共词矩阵，在共词矩阵的基础上用 Bisecting K-means 算法进行聚类。通过测试，测试结果证明了该方法的准确性和可行性，对网络舆情热点主题发现也有一定的作用。

第四章 基于热度分析的网络舆情热点跟踪

本文上一章节主要介绍网络舆情热点检测所涉及的基于共词分析的文本主题词聚类与主题发现算法的研究和实现，本章着眼于基于热度分析的热点跟踪算法的研究和实现，提出了一种新的主题热度计算方法和网络舆情热点跟踪的方法。

4.1 算法总体框架

在本节中，对基于热度分析的网络舆情热点跟踪算法的主要流程作基本介绍。算法的总体流程如图4.1所示。

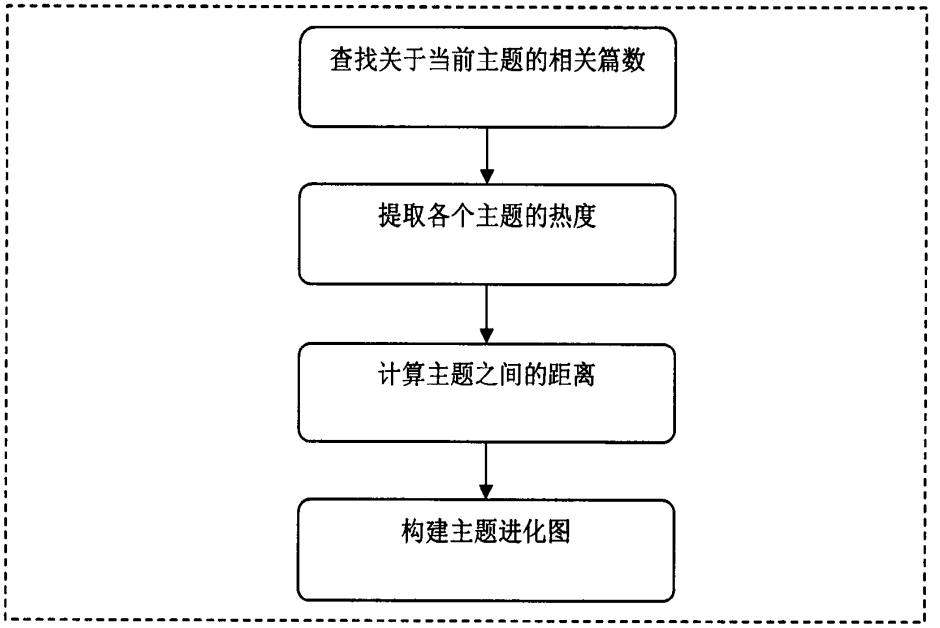

图 4.1 算法总体流程图

其中本章节具体改进的算法是根据 TF*PDF 和媒体关注度的思想，然后结合了帖子的权重值，提出了主题热度的计算方法，然后将 KL 相对熵的思想运用到主题距离的提取上。通过阈值判断后续时间戳的主题是否跟踪上前一个时间戳的主题。最后构建主题进化图。

4.2 现有主题关注度分析

4.2.1 TF*PDF 算法

TF*PDF是Bun和Ishizuka提出的一种算法，它主要是用于跟踪某一个研究领域中的网络突发事件[62]。该算法的研究目的与TDT很相似，但是他们有着不同的

研究重点。TF*PDF算法是用来识别特定时间内新闻流中可以区分不同新闻话题的特征项，我们需要注意的是，这些特征项都是由词组成的，但是不包括停用词。它的思想是一个热点新闻话题必然会被多个新闻来源报道，并且关于这个话题的新闻报道频度和数量都相对比较高。对于一些网络新闻来说，TF*PDF的意思是有多少个新闻网站报道了该新闻话题，并且每个网站对于这篇新闻话题有多大的宣传力度。简单地说，如果报道一个新闻话题的网站数越多，那么这个新闻话题所受到关注的程度就越高；如果在单个网站中报道这个新闻话题的相关新闻越多，则说明这个新闻话题越频繁，受到这个网站的关注程度就越高。根据这种思想构建TF*PDF的计算公式如下：

$$W_j = \sum_{c=1}^{c=D} | F_{jc} | \exp\left(\frac{n_{jc}}{N_c}\right) \tag{4.1}$$

$$| F_{jc} | = \frac{F_{jc}}{\sqrt{\sum_{k=1}^{k=K} F_{kc}^2}} \tag{4.2}$$

其中，W_j是特征项 j 的权重；F_{jc}是特征项 j 在新闻频道 c 中出现的频率；n_{jc}是在新闻频道 c 中出现特征项 j 的数目；N_c新闻频道中全部新闻数；k 在新闻频道中全部特征项个数；D 是全部新闻频道数；$|F_{jc}|$是特征项 j 的标准特征频率。

4.2.2 媒体关注度

根据 TF*PDF 的思想，引出了话题的“媒体关注度”的概念，话题的媒体关注度是指话题受到媒体所关注的重要程度，话题的媒体关注度对于热点话题的提取来说是一个很重要的因素，因为只有网站上报道了这个话题，网络用户才有机会去浏览和讨论这个话题，引发群众关注的概率才会更大，该话题才有可能成为热点话题[63]。话题媒体关注度主要利用了以下两个主要因素：一个是该站点上关于某个话题报道的数量，如果话题报道的数量越多则说明这个话题越热。另一个是话题报道所出现的频度。以此来定量的描述一定时间内站点上话题受媒体关注的程度：

$$T_m(j,t) = M_j(t) * \exp(\frac{D_j(t)}{N(t)}) \tag{4.3}$$

$$M_j(t) = D_j(t) \Big/ \sqrt{\sum_{c=1}^{c=C} D_c(t)^2} \tag{4.4}$$

其中，$T_m(j,t)$为时间段 t 内站点上有关于话题 j 的媒体关注度，t 可以使任意长的时间，如一天、一星期等；$D_j(t)$是站点上话题 j 的相关报道数量；$N(t)$是

站点上的报道总数；C 为该站点上的话题总数。$M_j(t)$表示站点上话题 j 的标准话题频度。公式中所取的指数主要是用来提高它的权值。从数学的角度分析，话题媒体关注度的取值是在 0 到 e 之间。

4.3 基于关注度的热度计算

主题热度提取在网络舆情热点跟踪中具有重要地位，它具有客观性、及时性等优点，它可以比较直观的反应该主题在各个时期民众讨论、关注的程度。而单单从关注度的角度并不能从全局把握主题的热度，本文在现有主题关注度分析的基础上提出了一个完整的主题热度提取方法，具体包含如下几步，根据主题词回溯提取反应当前主题的相关篇数、热度计算。

4.3.1 主题相关篇数提取

对于热点主题来说，该主题相关报道的数目和出现频度都是重要的影响因素，如果一个主题的相关报道数在一段时间内密集出现，在一定程度上会说明这段时间内该主题是一个热点。本章节提出了利用当前主题下的主题词进行回溯来统计当前主题的相关报道数的方法。首先通过判断某一报道中是否包含主题词，然后计算该报道所包含主题词的个数，通过包含主题个数和主题词总数的比例，给当前报道赋一个权重。例如：某一时间段的主题词为“流感”、“疫苗”、“病例”、“H1N1”、“疫情”、“重症”、“卫生部”、“医疗”、“应急”，它包含有9个主题词来反映当前的主题，某一篇报道中包含“流感”、“病例”、“医疗”、“卫生部”、“疫情”这篇报道中包含了5个主题词，那么我们对应这篇报道的报道篇数赋予5/9的权重值。

4.3.2 热度计算

本文提出了主题热度，即主题的重要程度，主题的热度越高，说明用户对该主题的兴趣越大，越容易形成热点主题，反之亦然。直观的讲，一个主题在语料中所占的文档数目越多且用户对该主题浏览、回复越多，说明该主题受到的热度越大。主题的特征主要包括主题的相关文档数、该时段的主题个数、帖子权重值，如：时段 1 下的主题 1，S=38.3；C=5；weight(1)=1002；等。而主题的媒体关注度基本考虑到了主题的相关文档数和该时段的主题个数，因此对于主题热度提取来说主要受到论坛帖子权重值和主题的媒体关注度的影响，如果一个主题的媒体关注度的值很大，说明主题受到媒体所关注的程度高，如果论坛帖子的权重值越高，从另一个角度说明主题受到网民的议论程度越高。

对于主题来说，主题的相关文档数目和帖子的权重，都是重要的影响因素，如果一个主题出现的相关文档数目很多，则在一定程度上说明该主题是一个热点

主题。因此根据上述 TF*PDF 算法和话题媒体关注度的分析，本文结合帖子权重和媒体关注度两个参数，提出了热点主题的热度计算方法：

$$Hot(m) = T_m(j,t) * (\sum_{i=1}^{i=s} \frac{weight(i) * s(i)}{a}) \tag{4.5}$$

其中$T_m(j,t)$表示话题 j 在时段 t 内的媒体关注度，$weight(i)$表示帖子 i 的权重，s(i)表示根据关键词回溯取得的跟主题相关的帖子篇数，且$0 \le s(i) \le 1$，系数 a 是由于帖子权重值较大，主要用这个系数来降低帖子权重值，这样方便后面的运算。一般 a 的值是跟语料库中帖子的权重有关，本文用到的是天涯论坛、新浪论坛、网易论坛的语料，根据实验中帖子的权重分析，将 a 取值为 1000 进行计算。

4.4 基于相对熵的主题进化

相对熵（relative entropy）又称为 KL 散度（Kullback-Leibler diverqence），所谓相对熵其实它主要用来表示两个随机分布之间的差异程度，当他们之间的差别增大时，其相对熵也随之增大，所以它是用来衡量随机分布之间的近似程度，其中相对熵的计算公式为：

$$D(\theta_2 \| \theta_1) = \sum_{i=1}^{|v|} p(w_i \mid \theta_2) \log \frac{p(w_i \mid \theta_2)}{p(w_i \mid \theta_1)} \tag{4.6}$$

其中我们约定$0 \log(0/q) = 0$，$p \log(p/0) = +\inf inity$，相对熵具有非负性和非对称性两大属性。

目前，有很多的研究都是基于相对熵展开的。例如，文献[64]中使用了相对熵来度量决策表条件属性的重要性。也有通过对多个相对熵加权得到综合相对熵在网络异常判断中的应用等等[65]。

在主题演化过程中，除相同主题和不同主题之外，还客观存在着相似主题类型，相似主题也就是主题变异的通常类型。虽然相似主题介于相同主题和不同主题之间，但相同主题和不同主题都可以认为是相似主题的两种极端情况。当两个主题达到完全相似时，就成为了相同主题；当两个主题达到完全不相似时，就成为了不同主题。相同主题代表了主题间继承关系，相似主题是主题继承与变异的统一体，不同主题则代表了主题的突变。现有的主题跟踪相似算法主要包括非相似指数、影响和出处指数等。1991 年 Michel Callon 提出了一种利用非相似指数测度聚类之间的非相似程度，它主要依据主题间共有的相同词语数量来分析一定的主题演化关系，非相似指数只是停留在对浅层相同主题的分析中。1992 年英国社会学家 John Law 和 John Whittaker 提出了影响指数和出处指数的概念，以此来揭示不同时期主题网络的相似关系，影响指数揭示了前期主题网络中的词语在

后期给定的主题网络中所占的比例，表明前一段时间段聚类中的主题对后续时间段聚类中主题的影响程度，出处指数代表了后续主题网络中的词语在前期主题网络中所占的比例，它揭示了后续聚类主题来自前期聚类中的哪些主题网络[66]。

本文主要提出了利用 KL-divergence 相对熵来计算两个主题之间的距离在利用相对熵时，我们首先将各个主题词和它的概率分布形成 HashMap 映射，查找两个主题之间共同出现的词，要是出现相同的主题词，便利用 TF-IDF 的权重属性，通过相对熵的计算公式得到两个主题之间的距离。当这个距离大于阈值 M 时，这就表示该主题演化到下一个时间段。阈值 M 是根据实验经验值设置所得。具体算法流程图如图 4.2 所示：

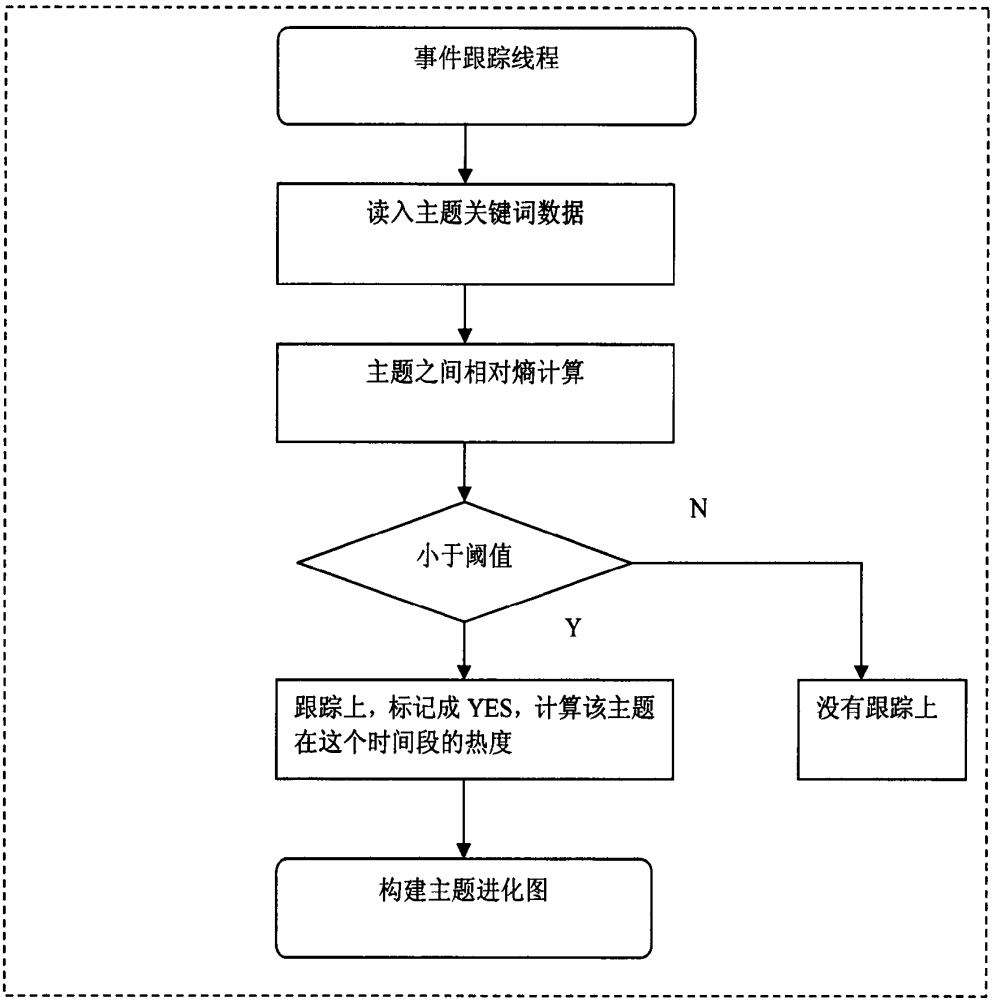

图 4.2 主题进化距离算法

通过主题进化距离，本文构建了主题进化图，通过主题进化图可以更加直观的体现各个时间段主题之间的关系。图 4.3 是一个主题进化图的例子。

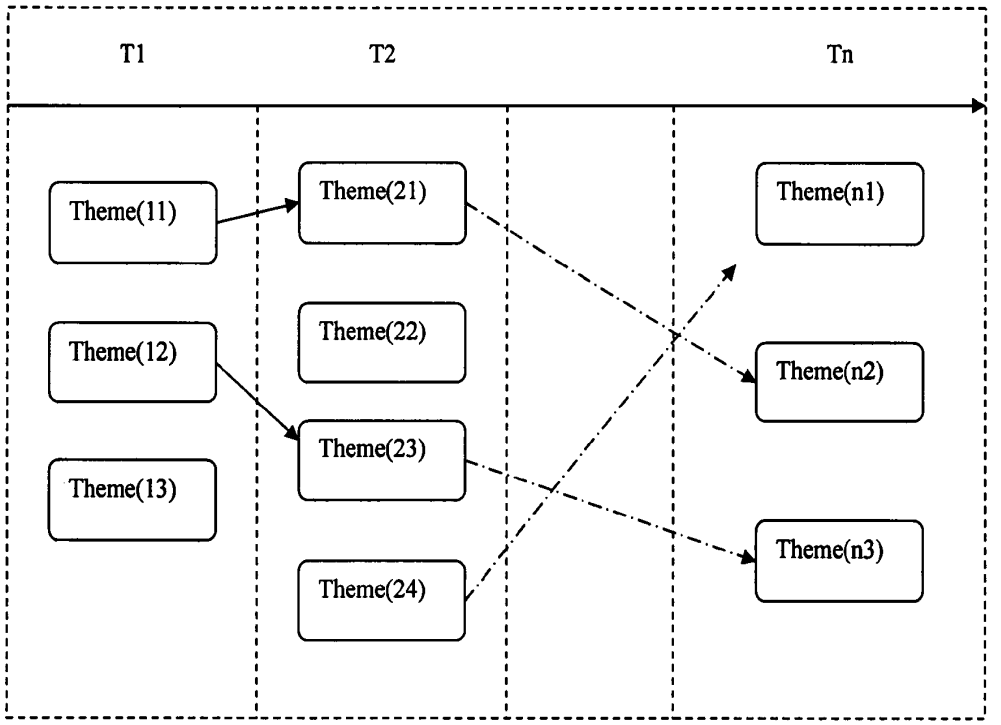

图 4.3 主题进化图

4.5 测试实例及其分析

本章节采用天涯论坛的军事板块上的语料作为测试实例，去除网页中的链接，导航等信息，处理成纯文本形式，只包含帖子正文和回复。时间从 2011 年 3 月 8 号到 4 月 6 号，每 3 天作为一个时间段，总计 1500 篇。在实例中主题进化的阈值取值为 50，通过测试实例我们可以发现在这段时间内主要有两大事件具有连续性，这两件事件主要是日本海啸灾后重建问题和国际军事事件卡扎菲与利比亚当局执政的军事斗争。在测试实例中，我们可以发现日本海啸事件从时段 t_2 开始爆发，而卡扎菲与利比亚的军事斗争事件是在时间段 t_4 被触发，两个事件一直持续到实验结束，说明这两大事件在这段时间内一直被网民讨论。具体的主题进化图的实例结果如图 4.4 所示

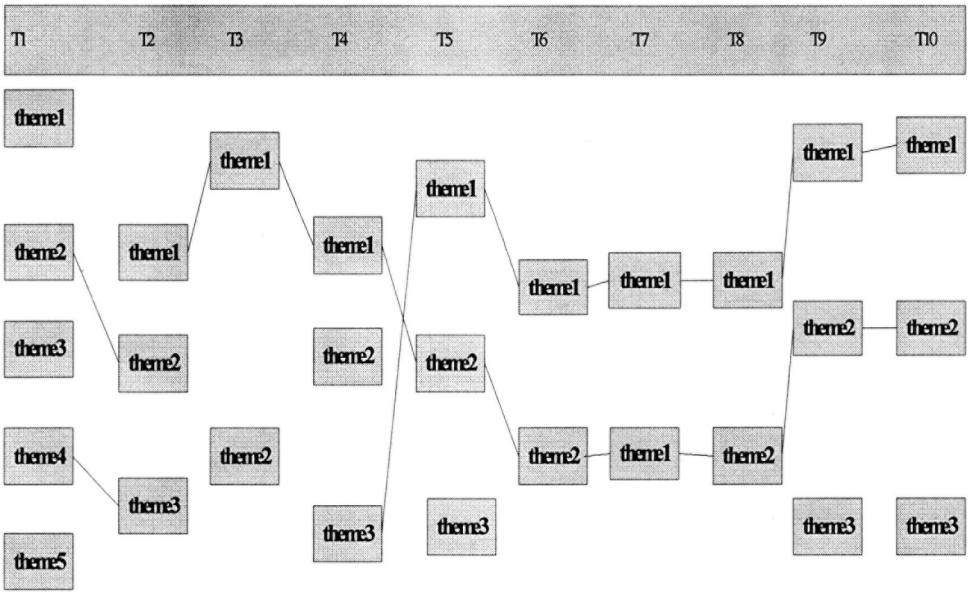

图 4.4 主题进化图实例

从测试实例的结果可以得出结论，主题进化图结果可以很好的反映了各个时间戳中主题之间的联系，说明 KL-divergence 相对熵运用在计算主题之间的距离上具有一定的可行性。

4.6 本章小结

本章节首先通过主题词回溯的方法提取关于相关主题的篇数，通过对结合回帖数、浏览数以及媒体关注度的分析提取主题的热度，最后通过计算主题之间的距离构建主题进化图，可以直观的反映各个时间段主题之间的关系。通过实验测试，测试结果证明了该方法的准确性和可行性，对网络舆情热点跟踪起到一定的作用。

第五章 实验及性能评价

本章是根据上文章节的实现，采用新浪论坛、网易论坛、天涯论坛的国际军事版块的语料集对本文实现的网络舆情热点检测与跟踪技术进行实验测试，以真实的论坛数据为对象对本文的算法做全面的实验评估。

5.1 实验环境及系统体系结构

5.1.1 实验环境

系统环境配置：CPU 为 Core2 T6500，内存为 2G，硬盘为 Seagate 300G，7200r/m，操作系统为 Windows 7，实现程序语言为 Java，运行环境为 eclipse 3.6。

5.1.2 系统体系结构

本节概述本文研究的整体框架，系统的体系结构如图 5.1 所示，以下对各模块的主要功能作简要说明。系统主要分为 4 个模块，分别是系统控制模块、论坛语料预处理模块、舆情热点主题提取模块、舆情热点主题跟踪模块。具体如图 5.1 所示。

（1）论坛语料预处理模块。本文所用的语料主要来自于新浪论坛、网易论坛和天涯论坛的国际军事版块上所采集的 2500 篇纯文本语料，时间跨度从 2011 年 3 月 26 日到 2011 年 5 月 6 日，每 3 天为一个时间戳。由于论坛每一天的发帖数量很大，而论坛的浏览数和回复数能够反应帖子的热度，本文将上一章节提出的帖子权重公式中的参数 a 和 b 分别取值为 0.8 和 0.2，根据帖子的权重值，进行从大到小排序，将一天权重排名前 50 的帖子提取出来。而爬虫爬下来的原始语料的格式不适合进行时间标签的实验，因此需要对这些文本语料通过编程进行初步的分割预处理，使其满足某天的论坛报道总数在一个文件里，经过上述预处理后的语料就形成了带有日、月和年时间标签属性的文本语料。例如 201141.txt 共包含权重值排名前 50 的帖子，每行代表一篇帖子，共 50 行。

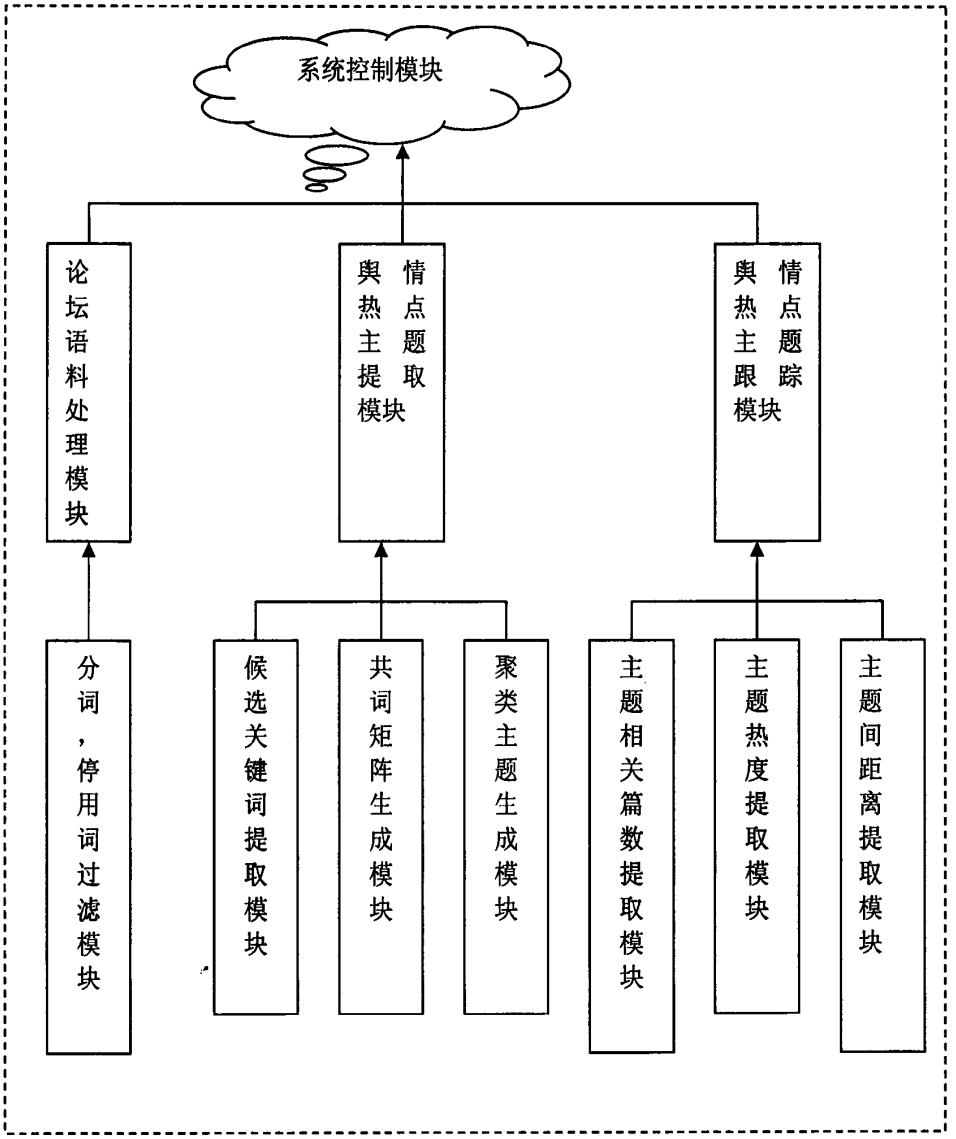

图 5.1 系统主要模块

（2）舆情热点主题提取模块。该模块主要是通过计算上述语料中词语的词频和文档频率，这两个参数都是对于关键词的提取有主要影响的变量，计算其 TF-IDF 的权重值，根据权重值大于 0.004 的策略，从中提取出关键词，关键词的重要性随着它在文件中的出现次数成正比增加，但同时会随着它在语料库中出现的文档频率成反比下降。然后本文通过词共现，构建了共词矩阵，在共词矩阵标准化的基础上，通过 Bisecting K-means 聚类算法，便可以得到每个时间戳的帖子话题。

（3）舆情热点主题跟踪模块。该模块主要是实现热点主题的跟踪，分别计算从舆情热点主题提取实验中得到的各个时间戳的主题的热度，本文结合帖子的权重和帖子的关注度两个参数提取出各个时期主题的热度。然后通过

KL-divergence 相对熵来计算两个主题之间的距离，如果相对熵的值小于某个阈值，那么就表示这个主题在一个时间段到另一个时间段内发生了主题的演化。

（4）该模块系中央控制模块，用于控制参数的设置和结果的输出。

5.2 舆情热点主题提取实验

5.2.1 帖子提取

本实验使用的网络爬虫是基于 Larbin(一种开源的网络爬行器)的一种应用，根据帖子权重值的排序，由高到低提取出每天权重值排名前 50 的帖子，帖子权重值计算公式如下：

$$weight(m) = 0.8 * replayCount(m) + 0.2 * browswCount(m) \quad (5.1)$$

其中取得 3 月 26 号这一天的帖子结果如表 5.1 所示。其中表中右侧显示的数值表示的是该帖子的权重值。

表 5.1　4 月 26 日帖子提取结果

……
1589.4[时事聚焦]日本死亡人数必超 2 万_国际观察_天涯社区.txt
2821.4[时事聚焦]西方打卡扎菲，为何不干预已打死 500 万人的刚果内战(转载)_国际观察_天涯社区.txt
3702.4[时事聚焦]菲律宾派出轰炸机追向中国渔政船(转载)_国际观察_天涯社区.txt
3778.4[时事聚焦][讨论]　从“惊人的 7000 亿”看真相？还是谣言？_国际观察_天涯社区.txt
4972.0[时事聚焦]东京核辐射指数比中国一些城市还低（图）(转载)_国际观察_天涯社区.txt
5313.8[时事聚焦]你们低估日本人的素质了，日本灾区挨饿只是因为核泄漏_国际观察_天涯社区.txt
7041.0[时事聚焦]歼 20 首飞催生百万军迷 日本担忧激化尚武传统(转载)_国际观察_天涯社区.txt
10259.2[时事聚焦]中石油给日本捐款 3000 万，昨天又向日本无偿提供 3 万吨汽油，真是大手笔_国际观察_天涯社区.txt
……

本实验的时间跨度从 2011 年 3 月 26 日到 2011 年 5 月 6 日，设成每 3 天为一个时间戳。这样我们就可以得到 14 个时间戳，如表 5.2 所示。

表 5.2 时间戳

3.26-------3.28	3.29-------3.31	4.1-------4.3	4.4-------4.6
4.7-------4.9	4.10------4.12	4.13-------4.15	4.16-------4.18
4.19-------4.21	4.22-------4.24	4.25-------4.27	4.28-------4.30
5.1-------5.3	5.4-------5.6		

5.2.2 TF-IDF 热点关键词提取

上述处理后的语料经过 ICTCLAS 分词系统进行分词，分词后的语料形式如下所示：

……

利比亚 叛军 节节 挺进 ， 临近 首都 的 城市 继续 有 难民 出逃 。 本台 特派 记者 在 利比亚 边境 ， 采访 到 出逃 的 利比亚人 ， 不过 她 不是 因为 害怕 叛军 ， 而是 要 逃离 卡扎菲 的 镇压 。哈纳 和 她 的 孩子 ， 原本 住 在 邻近 首都 的

城市 祖瓦拉 。当地 以 柏柏尔 人 为主 ，随着 判 军 逼近 ，爆发 了 反对 利比亚 领导人 卡扎菲 的 抗议 示威 。利比亚 难民 哈 纳 ：卡扎菲 派 了 几十 辆 坦克 进来 ，她 说 ， 我 看 到 很多 参与 抗议 的 人 被 打 死 。 卡扎菲 的 军人 夜里 挨家挨户 搜捕 。 哈纳 说 ， 有 祖瓦 拉 居民 被 抓 去 首都 ， 强充 卡扎菲 的 支持者 。 利比亚 难民 哈 纳 ： 有时 卡 扎 菲 军队 抓住 他们 的 孩子 ， 强迫 他们 游行 ， 还 有 时 人们 喊 口号 支持 卡扎菲 时 ， 不远处 有 枪口 对准 他们 。 哈纳 最 后 带 着 三 个 孩子 ， 假装 求医 ， 穿 过 两 个 亲 卡扎菲 的 城市 ， 来到 利比亚 突尼斯 边境 ，住 进 阿联酋 提供 的 难民营 。哈纳 讲述 时 情绪 激动 ，说 一半 祖瓦 拉 居民 已经 被 打 死 ， 但 我们 无法 独立 证实 她 的 说法 。 不过 可以 肯定 的 是 ，哈纳 说 自己 不 会 再 回国 ，除非 是 没有 卡扎菲 的 利比亚 。 凤 凰 卫视 周轶君 、 莫子豪 、 林逸 生 利比亚 边境 报道

……

论坛语料经过分词预处理之后我们进行停用词过滤。根据词性统计出来的停用词表很好的去除了一些对文章干扰不大的词。然后我们根据 TF-IDF 关键词提取的原理，对论坛语料库进行了处理，提取出这批语料中的关键词，并且根据 TF*IDF 的值按照降序排列出来。本文根据关键词 TF-IDF 的权重值大于 0.004 的策略选取每个时间戳的词串，结果几乎为该时段的重要事件的关键词。其中包含了像“战争”、“放射性”、“空袭”等重大事件用语，还包括了重大事件中的各种人物名：如“卡扎菲”、“奥巴马”等。表 5.3，表 5.4，表 5.5，表 5.6 分别是各个时间戳关键词的实验结果。

表 5.3　关键词提取结果示例

3.26-----3.28	3.29-----3.31	4.1-----4.3	4.4-----4.6
卡扎菲 Ttdf= 0.02134	国际 Ttdf=0.01056	自卫队 Ttdf=0.01429	西方 Ttdf=0.01845
战争 Ttdf= 0.01898	西方 Ttdf=0.01016	卡扎菲 Ttdf=0.01248	卡扎菲 Ttdf=0.0161
天皇 Ttdf=0.01657	谣言 Ttdf=0.00993	战争 Ttdf=0.01047	联合国 Ttdf=0.0117
西方 Ttdf=0.01430	三一 Ttdf=0.00967	联合国 Ttdf=0.00942	战争 Ttdf=0.01119
全球 Ttdf=0.01357	重工 Ttdf=0.00907	放射性 Ttdf=0.00916	人民 Ttdf=0.01036
人民 Ttdf=0.01185	卡扎菲 Ttdf=0.00843	护罩 Ttdf=0.00910	发展 Ttdf=0.01026
政治 Ttdf=0.01014	人民 Ttdf=0.00832	西方 Ttdf=0.00851	国际 Ttdf=0.01017
国际 Ttdf=0.00952	灾区 Ttdf=0.00791	救援 Ttdf=0.00759	政治 Ttdf=0.01003
突尼斯 Ttdf=0.00802	放射性 Ttdf=0.00746	索马里 Ttdf=0.00710	广播 Ttdf=0.00884
联军 Ttdf=0.00721	危机 Ttdf=0.00678	冷却 Ttdf=0.00699	全球 Ttdf=0.00797
零部件 Ttdf=0.00704	海啸 Ttdf=0.00654	欧盟 Ttdf=0.00699	听众 Ttdf=0.00779
产能 Ttdf=0.00704	空袭 Ttdf=0.00632	融合 Ttdf=0.00688	政权 Ttdf=0.00717
反对 Ttdf=0.00677	全球 Ttdf=0.00629	时间 Ttdf=0.00681	放射性 Ttdf=0.0066
总统 Ttdf=0.00634	政治 Ttdf=0.00580	国际 Ttdf=0.00679	自卫队 Ttdf=0.0065
空袭 Ttdf=0.00630	联合国 Ttdf=0.00568	人民 Ttdf=0.00652	伊朗 Ttdf=0.0063
放射性 Ttdf=0.00603	时间 Ttdf=0.00544	危机 Ttdf=0.00642	总统 Ttdf=0.00621
海啸 Ttdf=0.00602	电子 Ttdf=0.00543	总统 Ttdf=0.00639	时间 Ttdf=0.00605
灾区 Ttdf=0.00581	战争 Ttdf=0.00536	大屠杀 Ttdf=0.00617	教授 Ttdf=0.00590
⋮	⋮	⋮	⋮

表 5.4 关键词提取结果示例

4.7-----4.9	4.10-----4.12	4.13-----4.15	4.16-----4.18
西方 Ttdf=0.0121	战争 Ttdf=0.01161	核电厂 Ttdf=0.01356	属国 Ttdf=0.01424
战争 Ttdf=0.00949	航母 Ttdf=0.01077	核电 Ttdf=0.01259	奥巴马 Ttdf=0.0100
抗战 Ttdf=0.00883	人民 Ttdf=0.01023	西方 Ttdf=0.01097	西方 Ttdf=0.00981
人民 Ttdf=0.00862	墨西哥 Ttdf=0.01007	战争 Ttdf=0.01041	发展 Ttdf=0.00919
包机 Ttdf=0.00860	国际 Ttdf=0.00986	航母 Ttdf=0.00903	卡扎菲 Ttdf=0.00861
发展 Ttdf=0.00855	联合国 Ttdf=0.00948	容器 Ttdf=0.00898	亚洲 Ttdf=0.00843
卡扎菲 Ttdf=0.0084	西方 Ttdf=0.00940	卡扎菲 Ttdf=0.00818	国际 Ttdf=0.00796
过度 Ttdf=0.00786	卡扎菲 Ttdf=0.00877	发展 Ttdf=0.00763	战争 Ttdf=0.00754
政治 Ttdf=0.00754	总统 Ttdf=0.00853	放射性 Ttdf=0.00731	政治 Ttdf=0.00726
国际 Ttdf=0.00667	油价 Ttdf=0.00827	人民 Ttdf=0.00723	人民 Ttdf=0.00719
总统 Ttdf=0.00577	发展 Ttdf=0.00788	国际 Ttdf=0.00697	危机 Ttdf=0.00697
奥巴马 Ttdf=0.00576	安理会 Ttdf=0.00662	总统 Ttdf=0.00695	意大利 Ttdf=0.0067
全球 Ttdf=0.00548	公民 Ttdf=0.00604	政治 Ttdf=0.00685	总统 Ttdf=0.00660
火山 Ttdf=0.00542	政治 Ttdf=0.00594	政权 Ttdf=0.00641	放射性 Ttdf=0.00652
时间 Ttdf=0.00521	成品油 Ttdf=0.00547	时间 Ttdf=0.00584	组织 Ttdf=0.00619
放射性 Ttdf=0.0052	控制 Ttdf=0.00536	救援 Ttdf=0.00573	时速 Ttdf=0.00616
谣言 Ttdf=0.00502	海啸 Ttdf=0.00497	联合国 Ttdf=0.00555	时间 Ttdf=0.00535
海啸 Ttdf=0.00488	西非 Ttdf=0.00476	全球 Ttdf=0.00555	列车 Ttdf=0.00526
⋮	⋮	⋮	⋮

表 5.5 关键词提取示例

4.19-----4.21	4.22-----4.24	4.25-----4.27	4.28-----4.30
航母 Ttdf=0.02033	弹射 Ttdf=0.01130	西方 Ttdf=0.01142	航母 Ttdf=0.01641
潜艇 Ttdf=0.01523	欧盟 Ttdf=0.00883	航母 Ttdf=0.01000	两百 Ttdf=0.0163
发展 Ttdf=0.01035	发展 Ttdf=0.00843	人民 Ttdf=0.00918	卡扎菲 Ttdf=0.0127
人民 Ttdf=0.00907	人民 Ttdf=0.00842	资本 Ttdf=0.00878	普查 Ttdf=0.01031
国际 Ttdf=0.00805	贫民窟 Ttdf=0.00819	富豪 Ttdf=0.00807	人民 Ttdf=0.00942
战争 Ttdf=0.00794	岛屿 Ttdf=0.00819	南海 Ttdf=0.00803	西方 Ttdf=0.00929
卡扎菲 Ttdf=0.0070	卡扎菲 Ttdf=0.00768	时间 Ttdf=0.00771	赔款 Ttdf=0.00921
政治 Ttdf=0.00628	奥巴马 Ttdf=0.00757	卡扎菲 Ttdf=0.00745	战争 Ttdf=0.00866
危机 Ttdf=0.00578	航母 Ttdf=0.00742	组织 Ttdf=0.00714	南海 Ttdf=0.00806
棒子 Ttdf=0.00566	国际 Ttdf=0.00701	控制 Ttdf=0.00688	发展 Ttdf=0.00791
隐形 Ttdf=0.00565	西方 Ttdf=0.00694	国际 Ttdf=0.00684	时间 Ttdf=0.00778
西方 Ttdf=0.00542	南海 Ttdf=0.00626	战争 Ttdf=0.00636	国际 Ttdf=0.00710
防线 Ttdf=0.00538	救生 Ttdf=0.00622	全球 Ttdf=0.00595	奥巴马 Ttdf=0.00678
总统 Ttdf=0.00529	控制 Ttdf=0.00614	奥巴马 Ttdf=0.00589	生命 Ttdf=0.00646
支援 Ttdf=0.00513	独裁者 Ttdf=0.00593	分子 Ttdf=0.00555	贷款 Ttdf=0.00614
时间 Ttdf=0.00506	意大利 Ttdf=0.00590	危机 Ttdf=0.00536	政治 Ttdf=0.00573
联合国 Ttdf=0.00481	时间 Ttdf=0.00570	核电 Ttdf=0.00534	数据 Ttdf=0.00572
外交 Ttdf=0.00465	编队 Ttdf=0.00565	政治 Ttdf=0.00506	壁画 Ttdf=0.00534
⋮	⋮	⋮	⋮

表 5.6 关键词提取示例

5.1-----5.3	5.4-----5.6		
本拉登 Ttdf=0.0173	本拉登 Ttdf=0.0264		
西方 Ttdf=0.01400	卡扎菲 Ttdf=0.01774		
纽约 Ttdf=0.01178	奥巴马 Ttdf=0.01186		
战争 Ttdf=0.01005	孔子 Ttdf=0.01059		
奥巴马 Ttdf=0.0090	组织 Ttdf=0.00993		
组织 Ttdf=0.00901	总统 Ttdf=0.00832		
空袭 Ttdf=0.00816	战争 Ttdf=0.00825		
卡扎菲 Ttdf=0.0080	发展 Ttdf=0.00813		
总统 Ttdf=0.00766	人民 Ttdf=0.00750		
发射 Ttdf=0.00758	西方 Ttdf=0.00729		
南海 Ttdf=0.00757	卡尔 Ttdf=0.00696		
航母 Ttdf=0.00737	海葬 Ttdf=0.00643		
倒塌 Ttdf=0.00669	分子 Ttdf=0.00587		
时间 Ttdf=0.00655	伊朗 Ttdf=0.00560		
联合国 Ttdf=0.00644	大屠杀 Ttdf=0.00559		
全球 Ttdf=0.00641	国际 Ttdf=0.00537		
潜艇 Ttdf=0.00628	空袭 Ttdf=0.00535		
人民 Ttdf=0.00599	制造业 Ttdf=0.0050		
⋮	⋮		

5.2.3 构建共词矩阵

根据共词矩阵算法，统计出已经提取出的关键词的词串在文档中两两出现的次数，这样 N 个关键词在共词分析中，便构成了 N*N 的一个共词矩阵。其中一组 3 月 26 号到 3 月 28 号这个时间戳的实验结果如下表 5.7 所示。

表 5.7 共词矩阵

共词矩阵	卡扎菲	战争	天皇	西方	全球	…
卡扎菲	0	15	0	19	5	…
战争	15	0	0	20	10	…
天皇	0	0	0	0	0	…
西方	19	20	0	0	8	…
全球	5	10	0	8	0	…
人民	16	20	1	16	8	…
政治	13	10	0	11	5	…
国际	14	18	0	17	11	…
突尼斯	4	2	0	2	1	…
发展	15	13	1	11	7	…
联军	12	6	0	10	1	…
183	8	10	1	7	3	…
零部件	0	0	0	0	1	…
…	…	…	…	…	…	…

由于主题词在同一篇文章出现的频率越高则表示他们之间的距离越小。本文对共词矩阵进行进一步的处理，首先将共词矩阵中两两出现的主题词的共现次数加 1，然后取其倒数，将这个值作为两个主题词之间的距离。而自身对自身的距

离最远，本文设置为 1。具体公式如下所示。

$$l = \frac{1}{s + 1} \tag{5.2}$$

其中 L 代表两个主题词之间的距离，S 为共词矩阵中两两出现的主题词的共现次数。表 5.8 为标准化后的共词矩阵。

表 5.8 共词矩阵标准化

共词矩阵	卡扎菲	战争	天皇	西方	全球	…
卡扎菲	1	0.0625	1	0.05	0.166667	…
战争	0.0625	1	1	0.047619	0.090909	…
天皇	1	1	1	1	1	…
西方	0.05	0.047619	1	1	0.111111	…
全球	0.166667	0.090909	1	0.111111	1	…
人民	0.058824	0.047619	0.5	0.058824	0.111111	…
政治	0.071429	0.090909	1	0.083333	0.166667	…
国际	0.066667	0.052632	1	0.055556	0.083333	…
突尼斯	0.2	0.333333	1	0.333333	0.5	…
发展	0.0625	0.071429	0.5	0.083333	0.125	…
联军	0.076923	0.142857	1	0.090909	0.5	…
183	0.111111	0.090909	0.5	0.125	0.25	…
零部件	1	1	1	1	0.5	…
…	…	…	…	…	…	…

本文将每个时间戳的关键词，都利用上述方法构建出共词矩阵。为下一步的聚类提供合理的输入。

5.2.4 Bisecting K-means 聚类结果

根据 Bisecting K-means 算法，本文对共词矩阵进行聚类，其中列出了 4 组实验结果，具体如表 5.9-5.12 所示：

表 5.9　3.26-3.28 聚类结果

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
卡扎菲	芯片	天皇	突尼斯	全球
战争	供应	放射性	联军	国际
西方	数据	海啸	资本	发展
人民	油价	灾区	部落	20%
政治	拥有	矿泉水	伊斯兰	控制
反对	上涨	宫内	萨科齐	危机
总统	零部件	救援	欧盟	
奥巴马	产能		霸权	
空袭	电子		意大利	

表 5.10　3.29-3.31 聚类结果

Cluster1	Cluster2	Cluster3	Cluster4
谣言	灾区	国际	政治
三一	放射性	人民	战争
重工	海啸	危机	外交
布拉莫斯	灾民	全球	空袭
GDP	救援	发展	卡扎菲
电子	生命	时间	西方
熊猫	支援	控制	联合国
壳牌			联军
制造			奥巴马
北非			总统

表 5.11 5.1-5.3 聚类结果

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
本拉登	西方	发射	海地	萨达姆
纽约	战争	南海	俄国	叛逃
奥巴马	组织	航母	东家	政权
总统	空袭	潜艇	火力	憎恨
倒塌	卡扎菲	分子	精确度	阿尔贾纳比
911	联合国	南沙	突防	
联邦	反对		空客	
活捉	政治		杀伤性	
被杀	伊朗			

表 5.12　5.4-5.6 聚类结果

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
本拉登	卡扎菲	大屠杀	制造	分子
奥巴马	组织	海盗	矿业	航母
总统	战争	信徒	全能	射程
海葬	西方	新教	同源	潜艇
美利坚	伊朗	冰岛	制造业	南海
藏身	空袭	伊斯兰	发展	
被杀	国际	控制	卡尔	
911				

其他几组时间戳的聚类结果也是利用上述方法进行表述。

5.2.5 实验小结

从上述实验结果来看，对于给定的论坛语料，经过相应的预处理后，热点主题词的聚类结果能一定的反映各个时间戳热点话题。对网络热点主题发现有一定的作用。

5.3 舆情热点主题跟踪实验

5.3.1 主题篇数提取

通过关键词回溯算法，本文提取出主题篇数的权值并且统计出相关文档总数，具体实验结果如表 5.13 所示：

表 5.13　主题篇数提取

时间戳	聚类 1	聚类 2	聚类 3	聚类 4	聚类 5	相关文档总数	总文档数
3.26-3.28	38.3	10.3	15.78	12.1	40.9	117.38	144
3.29-3.31	13.48	20.23	42.62	32.9		109.23	150
4.1-4.3	35.71	22.75	10.68	46.83	16.02	131.99	150
4.4-4.6	36.6	48.21	10.72	20.06	13.39	128.98	144
4.7-4.9	36.15	10.5	46.48	25.27		118.4	150
4.10-4.12	32.47	20.45	15.32	42.59		110.83	150
4.13-4.15	15.3	40.2	10.55	46.8		112.85	150
4.16-4.18	10.73	40.15	40.9	10.2		101.98	144
4.19-4.21	21.75	38.2	35.5	15.73		111.18	150
4.22-4.24	25.22	36.96	19.4	22.5	12.8	116.88	151
4.25-4.27	20.02	20.09	17.04	25.56	15.8	98.51	128
4.28-4.30	25.3	25.62	20.3	15.22	16.46	102.9	149
5.1-5.3	40.4	25.29	28.5	10.12	15.67	119.98	146
5.4-5.6	40.56	20.03	15.85	16.68	30.74	123.86	149

5.3.2 主题热度计算

结合帖子权重和帖子的关注度计算便可以得到的某个时间戳各个主题的热度，具体实验结果如表 5.14 所示：

表 5.14　主题热度

时间戳	聚类 1	聚类 2	聚类 3	聚类 4	聚类 5
3.26-3.28	1.1147	0.2468	0.3928	0.2936	1.2122
3.29-3.31	0.2496	0.3921	0.9585	0.6935	
4.1-4.3	0.6865	0.4011	0.1737	0.9695	0.2701
4.4-4.6	0.7147	1.0204	0.1749	0.3492	0.2225
4.7-4.9	0.7085	0.1734	0.9759	0.4606	
4.10-4.12	0.6794	0.3950	0.2859	0.9534	
4.13-4.15	0.2630	0.8157	0.1757	0.9923	
4.16-4.18	0.1953	0.8964	0.9179	0.1850	
4.19-4.21	0.4287	0.8402	0.7669	0.2978	
4.22-4.24	0.5398	0.8550	0.3995	0.4730	0.2525

4.25-4.27	0.5237	0.5259	0.4355	0.6983	0.3999
4.28-4.30	0.6376	0.6471	0.4947	0.3585	0.3909
5.1-5.3	0.9095	0.5133	0.5914	0.1851	0.2978
5.4-5.6	0.8974	0.3861	0.2971	0.3144	0.6368

5.3.3 主题距离计算

通过 KL 相对熵计算公式计算得到主题间的距离，在实验中本文设定参数 infinity=10000000 进行计算，得到各个主题之间的距离，表中 Ti 代表各个时间戳，其中 i 取值在 0 到 14 之间，统计各个时间戳的文本中所有字符的相对频率，我们将这些相对频率代表各个字符的概率，在算法中，首先去匹配主题到下一个主题之间是否存在相同的主题词，要是存在相同的主题词则进行计算相对熵的值，要是彼此不存在共同的主题词，则赋值为 1.0E8，最后将主题进化距离进行归一化，将每个值除以 1.0E8，本文规定两个主题之间距离为 1 时表示两个主题距离最远，当值为 0 时，表示两个主题相同，而他们之间的距离最近，其中列出了 4 组时间戳中各个主题间距离的实验结果，具体如表 5.15-5.18 所示。

表 5.15　T1 到 T2 时间戳中各个主题间的距离

主题	T2 主题 1	T2 主题 2	T2 主题 3	T2 主题 4
T1 主题 1	1	1	1	0.32431
T1 主题 2	1	1	1	1
T1 主题 3	1	0.41125	1	1
T1 主题 4	1	1	1	1
T1 主题 5	1	1	0.22976	1

表 5.16　T2 到 T3 时间戳中各个主题间的距离

主题	T3 主题 1	T3 主题 2	T3 主题 3	T3 主题 4	T3 主题 5
T2 主题 1	1	1	1	1	0.20068
T2 主题 2	1	0.59968	1	1	1
T2 主题 3	0.13861	1	1	1	1
T2 主题 4	0.23127	1	1	1	1

表 5.17 T12 到 T13 时间戳中各个主题间的距离

主题	T13 主题 1	T13 主题 2	T13 主题 3	T13 主题 4	T13 主题 5
T12 主题 1	1	1	0.38782	1	1
T12 主题 2	1	0.36891	1	1	1
T12 主题 3	1	1	1	1	1
T12 主题 4	1	1	1	1	1
T12 主题 5	1	1	1	1	1

表 5.18 T13 到 T14 时间戳中各个主题间的距离

主题	T14 主题 1	T14 主题 2	T14 主题 3	T14 主题 4	T14 主题 5
T13 主题 1	0.30561	1	1	1	1
T13 主题 2	1	0.10518	1	1	1
T13 主题 3	1	1	1	1	0.30241
T13 主题 4	1	1	1	1	1
T13 主题 5	1	1	0.89241	1	1

5.3.4 构建主题进化图

本文通过多次实验，从经验值的角度将阈值取在 0.65，当相对熵的值小于阈值时，则表明下一个主题跟踪上。其中主题进化图如图 5.2 所示：

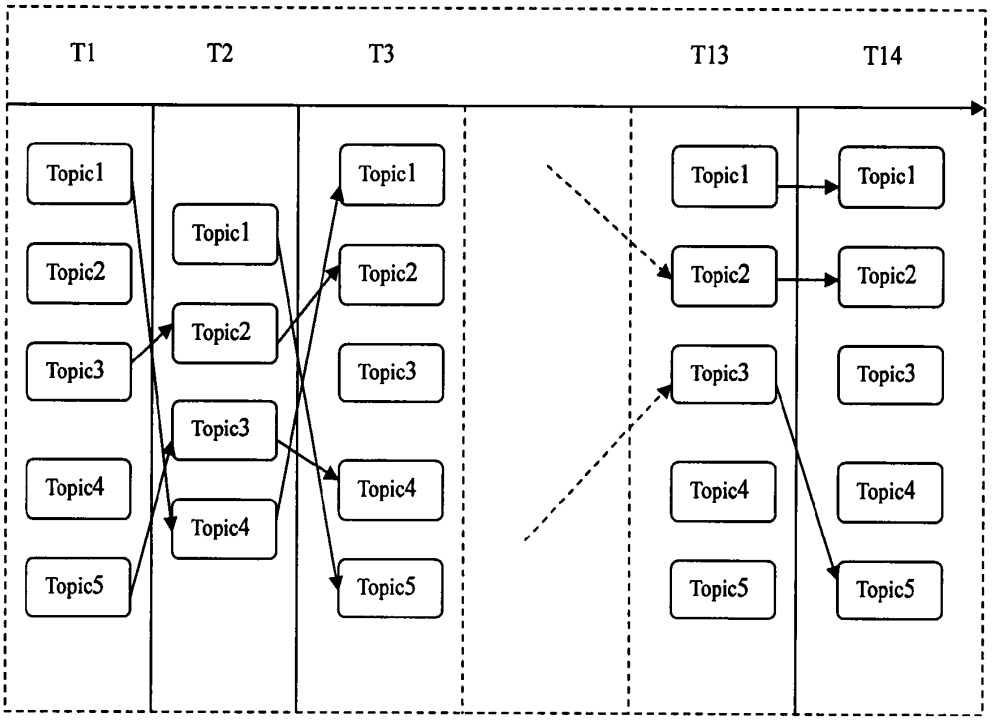

图 5.2 主题进化图实验结果

为更加直观的反映主题热度，本文分别构建了基于主题篇数和基于主题热度的分析表，具体如图 5.3 和图 5.4 所示。

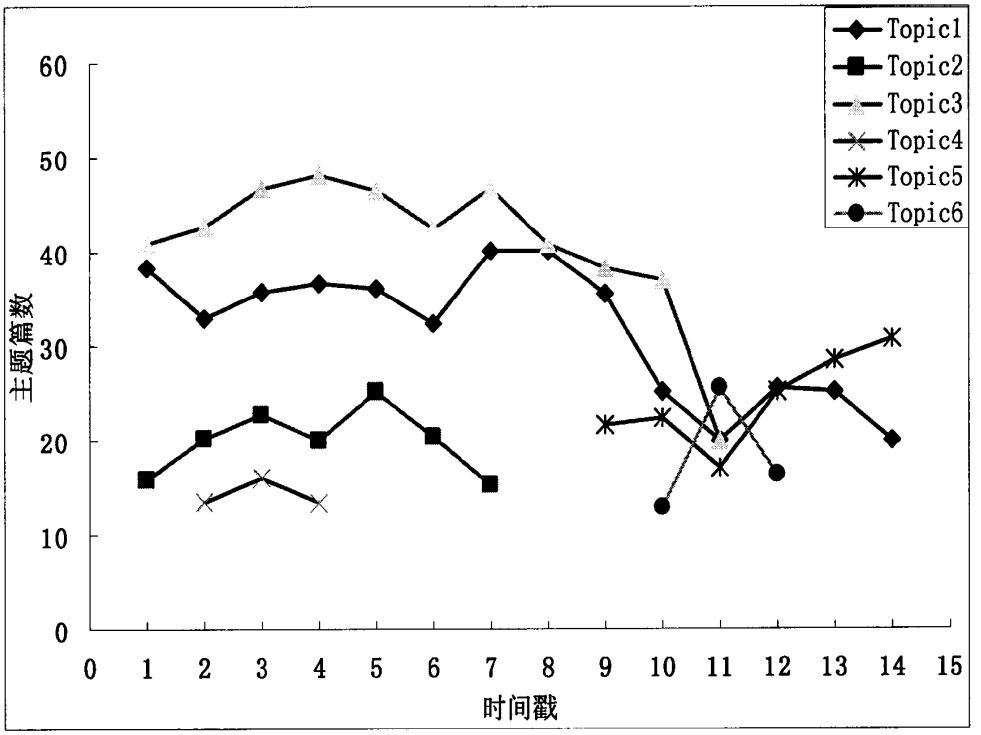

图 5.3 基于主题篇数分析图

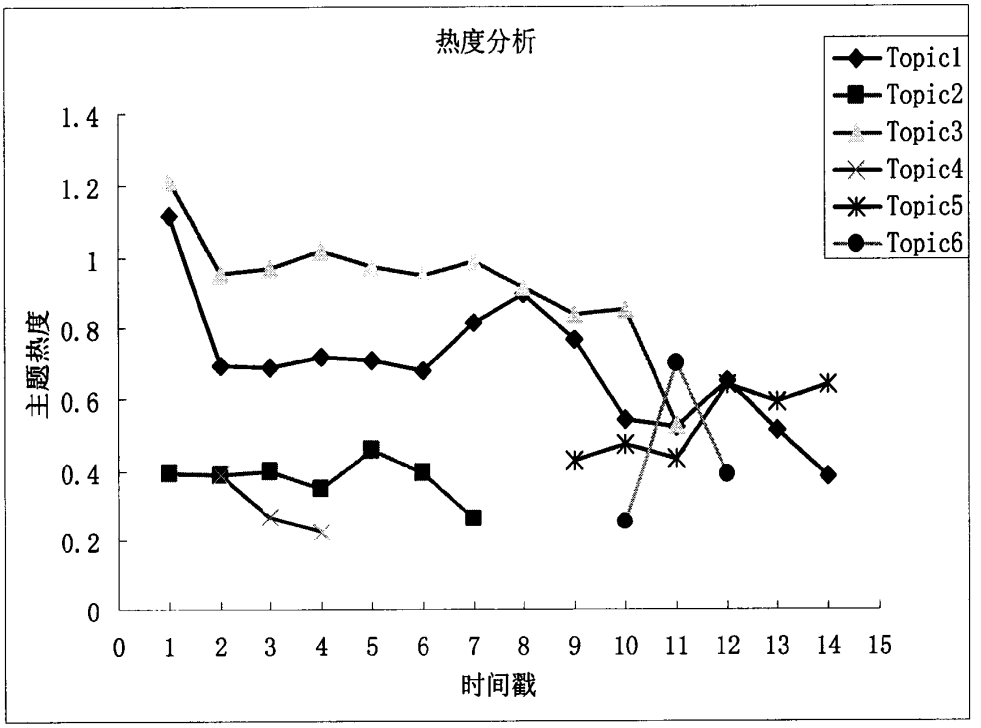

图 5.4 基于主题热度分析表

从上图的分析表中可以很直观的看出各个主题在不同时间戳上的连续性和主题热度的变化趋势。其中 Topic1 是主要讨论关于卡扎菲的军事斗争事件，我们可以发现该事件一直持续到实验结束，而且它呈现出热度先升后降的趋势，说明在此期间一直被网民所讨论着，讨论的程度在有段时间内达到了顶峰，然后开始慢慢的下降，Topic2 主要讨论的是关于全球军事危机的事件，从上图可以发现该事件在有一段时间内热度比较平稳，然后渐渐的在网民的讨论中消失，Topic3 主要讨论的是日本海啸后重建的问题，该事件持续了 7 个时间戳，热度比较平稳，然后慢慢的淡出了网民的讨论，Topic4 主要讨论的事件是由于日本海啸事件引起的电子产品市场涨价的问题，该事件只持续了 3 个时间戳。Topic5 主要讨论了国际航母发展的事件，从上图可以发现热度处于上升的趋势，表明网民越来越关注讨论该事件。Topic6 是骆家辉访美受到美国总统奥巴马接待的事件，该事件只持续了 3 个时间戳，很快的在网民的讨论中消失。

5.3.5 实验小结

从上面的实验结果可以看出，基于热度分析网络舆情热点跟踪可以很直观的体现各个主题之间的连续性和它们的变化趋势，它主要分析了各个时间戳主题的热度，体现了网民的关注程度，在此基础上通过距离阈值的判断，实现了舆情热点的跟踪，综合分析该算法对网络热点跟踪有一定的作用。

5.4 本章小结

本章对本文实现基于主题热度分析的跟踪算法做了相应的实验。实验表明，本文提出的算法具有稳定性，准确性以及高效性。同时，实验结果也显示出了本文的算法的局限性，例如，阈值的合理选择，候选关键词的合体提取，这为今后的研究工作指明了方向。

第六章 总结与工作展望

在信息社会中，随着计算机技术、通信技术的快速发展，网络已成为庞大的公共信息集散地和民众参政议政最常用的平台，是社情民意中最活跃和尖锐的一部分，最直接和快速地反映出社会各个层面的舆情状况与发展态势，并且随着网民数量持续增长，受到相关部门的高度关注和重视。信息指数的增长为用户获取有效的资源带来了非常大的困扰，如何有效的控制海量信息，有效的进行文本主题的发现将成为网络信息时代亟待解决的问题。它具有广泛的应用前景：对于个人，它能更方便地发现和组织当前重要资讯，对于企业；它能及时掌握企业相关领域、战略伙伴及其竞争对手的最新动态；对于国家，它能及时了解当前社会的重要资讯，流行趋向，舆论动态。

网络舆情热点检测与跟踪技术是一项新兴的信息处理技术，它是一门融合了文本挖掘、文本聚类、自然语言处理等的综合性研究课题。另外，该技术能够有效地发现网络舆情热点，并且可以控制舆论导向。

本论文主要是将共词矩阵与 Bisecting K-means 聚类算法相结合，提出了基于共词分析的网络舆情热点发现的方法，并详细描述了该算法实现的步骤与流程图，同时根据 TF*PDF 和话题关注度的思想，提出了主题热度的计算方法，并且将 KL 相对熵的思想运用到计算主题之间的距离的方法上。最后，用大规模数据和真实网络论坛数据其进行了测试。

6.1 本文的主要研究工作及成果

（1）论文首先通过研究分析，论述了网络舆情热点检测与跟踪技术的研究背景、研究目的、研究意义和研究现状；

（2）论文对网络舆情热点检测与跟踪技术的相关知识做了主要介绍，有网络舆情挖掘的基础知识，包括网络舆情信息采集：种子页面自动生成、主题爬行策略等。舆情信息预处理，同时介绍了几个国内常用的网络舆情系统，并作了比较等。重点分析了网络舆情热点检测与跟踪技术研究现状，主要包括：话题检测与跟踪任务，话题检测与跟踪的关键技术、话题检测与跟踪的评测标准等。

（3）论文提出了基于共词分析的网络舆情热点发现的方法：通过结合共词矩阵与 Bisecting K-means 聚类算法可以很好的进行网络舆情热点发现，主要分为共词矩阵生成模块和基于聚类算法提取主题模块。提出了主题热度的计算方法，并详细描述了主题热度提取算法的步骤和流程图，它主要结合了帖子的权重与主题媒体关注度的思想。最后将 KL 相对熵的思想运用到计算主题之间的距离的方法上，构建出主题进化图。实验证明，该算法具有稳定性和高效性，并具有

一定的可信度。

（4）论文中我们用经过处理的带有时间属性的论坛 BBS 语料对本文提出的基于共词分析的网络舆情热点发现算法和基于热度分析的网络舆情热点跟踪算法进行了测试。详细而充分的实验数据表明，本文所提出的算法具有很高的可信度。

6.2 存在的问题及对将来工作的展望

虽然，本文从对基于共词分析的网络舆情热点发现算法和基于热度分析的网络舆情热点跟踪算法的分析研究中取得了一定的成果，但同时也引出了许多新的问题，这些问题仍然需要进一步的研究。

（1）由于时间限制，本文仅从网络舆情热点检测与跟踪算法上作了较深入的研究，但是由于文本数据的特殊性，文本数据的预处理过程，包括分词、文本的计算机表示形式、如何确定候选关键词、更有效的排序方法等方面对热点主题的准确提取也有很大的影响，本文对此没有作更深入研究，所以，作为下一步工作，拟对此作深入研究。

（2）本文提出的基于相对熵的主题进化距离算法，对于主题进化距离阈值的设置均需有一定的专业知识，如何设置更准确的阈值参数也是下一步工作的一个方向。

（3）本文提出的两种算法，还是存在一些算法本身的缺陷，如何弥补这些缺陷也需要进行更深一步的研究。

（4）基于共词分析的网络舆情热点发现算法和基于热度分析的网络舆情热点跟踪算法还存在很多问题有待解决，例如，在候选关键词提取方面，有些词语可能在某一时间段内权重值趋于稳定，但是由于网络的不可预见性，这些词语也有可能会成为候选关键词。同时需要更好的去改进聚类算法，这样可以发现是否可以更好的发现热点话题。在将来的工作中，我们可以对上述问题进一步地研究。展望将来，网络舆情热点检测与跟踪技术将会得到广泛的应用，其技术的进步也会极大地提高热点话题发现的效率和效果。

由于个人经验、能力有限，再加上时间、精力有限，本文所做的工作中难免存在很多不完善甚至错误的地方。

致 谢

在即将完成这篇致谢词的时候，也就是代表研究生生活即将走向尾声，依依不舍之际，也心怀感恩，有那么多的人在研究生阶段给予了我帮助与关怀，现在回想起来，心里很温暖。论文在王小华教授的指导与鼓励下顺利完成，首先衷心感谢我的导师王小华老师。

自硕士入学以来，我的导师王小华老师给人的印象就是具有渊博的学识，在学术上很严谨，对待学生很和蔼可亲。有这样一位好导师我感到很幸运，他给予我许多关心和指导，当在学习上遇到困难或者是对某些问题有疑惑的时候，有了他的指点，很多问题都能迎刃而解。每周讨论会上，他每一次的指点总能使我们的研究更加深入，让我增长了不少知识，锻炼了我独立从事科研工作的能力。不光是他严谨治学的态度，他真诚的待人处事方式也给我留下了深刻的印象。在硕士论文的写作过程中，王老师百忙中抽出许多时间阅读我的论文，给我提出宝贵的修改建议，使我的论文能够顺利完成。我在此向王老师表示衷心感谢，感谢您对我的细心栽培，您的鼓励和教诲，我将终生铭记。

感谢陆蓓老师、吴海虹老师、王荣波老师、谌志群老师、姚金良老师，正是有了他们在我的学习论文研究工作中的无私教导与帮助，我才能在两年半的时间里在学术上有了很大的提升，扩大了我的视野，理清了研究思路，使我顺利完成学业。他们不光在学习上帮助我，还为我们提供了良的学习环境，使我们更加有归属感，能在这样的环境下学习生活，我感到无比的幸福。

感谢实验室宁长英、汪澄、陈法叶、张犀等同学，不论是学习、工作还是生活中碰到困难，他们都会主动帮忙，给我鼓励，并提出保贵的意见，使得我顺利渡过难关。

特别要感谢我的父母，他们虽然不懂我研究的课题，但是他们一直默默地支持着我、鼓励着我走下去，让我有信心可以完成学业，感谢他们这么多年来的养育之恩。

感谢所有关心我、帮助我的老师、同学、朋友，感谢研究生生活中给我家一般温暖的人们，跟你们一起生活的日子很难忘。

参 考 文 献

[1] 中国互联网络发展状况统计报告[EB/OL]. http://research.cnnic.cn.

[2] 王来华,刘毅.中国 2004 年舆情研究综述[J].新华文摘，2005，18.

[3] 刘毅（2007）.网络舆情研究概论.天津：天津人民出版社.

[4] 姜胜洪.我国网络舆情的现状及其引导[J].广西社会科学,2009，1：1-4.

[5] Yiming Yang，Jaime Carbonell，Ralf Brownetal.Learning Approache for Detectingand Tracking New Events，IEEE Intelligent Systems Speeial Issue on APPlications of Intelligent Information Retrieval，1999.

[6] 王科,刘渊,罗万伯,高行宇,高常波（2004）.基于中文文本主题跟踪的网络信息分析[J],四川大学学报,2004，26（2）：68-71.

[7] 齐海凤（2008）.网络舆情热点发现与事件跟踪技术研究.哈尔滨工程大学硕士学位论文.

[8] 胡健,董跃华,杨炳儒等.基于关键词的 WEB 文献自动跟踪系统的实现方法[J].南昌大学学报,2008，32（3）：21-24.

[9] 王志明,沙莎等.WEB 文本挖掘技术在新闻主题检测中的应用研究[J].长沙大学学报,2007，21（5）：33-37.

[10] 李胜韬,余智华,白硕等.Web 信息采集研究进展[J].计算机科学,2003，30（2）：151-157.

[11] 李盛韬（2002）.基于主题的 Web 信息采集技术研究.中国科学院计算技术研究所硕士论文.

[12] Leuski A，Allan J.Improving realism of topic tracking evaluation[A].Proceedings of SIGIR[C]，2002.65-71.

[13] 刘剑宇.Web 挖掘技术在网络舆情预警中的研究与应用[J].四川警察学院学报,2009，21（3）：77-81.

[14] Chakrabartia S,Doma B.Automatic resource complation by analyzing hyperlink stucture and associated text .[2004-05-26].http://cindoc.csic.es/cybermetrics/pdf/.

[15] EhrigM,MaedcheA.Ontology-focusedcrawlingofWebdocuments.[2004-05-10].http://www.uni-karlsruhe.de.

[16] Lavrenko V，Allan J，DeDuzman E，LaFlamme D，Pollard V，Thomas S.Relevance models for topic detection and tracking[A].Proceeding of the Human Language Technology Conference(HLT)[C]，2002.104-110.

[17] HAN jiawei，M.Kamber.Data Mining：Concepts and Techniques.NewYork，Morgan Kaufmann Publishers，2001:103-105.

[18] 陈志敏,沈洁,林颖等.基于主题划分的网页自动摘要[J].计算机应用,2006，26（3）：641-644.

[19] 李舒晨,刘云,李勇.网络舆情分析中网页预处理方案的实现[J].电脑与电信,2008，10：30-33.

[20] 孙春葵，李蕾，杨晓兰.基于知识的文本再要系统研究与实现[J].计算机研究与

发展,2000,37（7）：874-881.

[21] 刘挺，吴岩，王开铸.自动文摘综述[J].情报学报,1998，16（1）：63-69.

[22] 谢海光,陈中润.互联网内容及舆情深度分析模式[J].中国青年政治学院学报,2006，3：95-100.

[23] 唐果,陈宏刚.基于 BBS 热点主题发现的文本聚类方法[J].计算机工程,2010，36（7）：46-49

[24] 吴绍忠,李淑华.互联网络舆情预警机制研究[J].中国人民公安大学学报,2008，3：38-42.

[25] 许鑫,张志成.互联网舆情分析及应用研究[J].情报科学,2008，26（8）：1194-1200，1204.

[26] 钱爱兵.基于主题的网络舆情分析模型及其实现[J].现代图书情报技术,2008，4：49-55.

[27] Lzumi Aizu.Social，Legal and Policy lssues around Internet in Asia [EB /OL].http: //www.anr.Org.

[28] Wayne C..Multilingual Topic Detection and Tracking:Successful Research Enabled by Corpora and Evaluation[C].In：Language Resources and Evaluation Conference(LREC),2000：1487-1494.

[29] 王丫.网络新闻流中热点事件识别与跟踪算法的改进与验证.燕山大学硕士论文.

[30] 王娟.网络舆情监控分析系统构建[J].长春理工大学学报，2007（12）：201-204.

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

[32] 刘伟.基于句子索引图的新闻流话题检测与跟踪研究.中山大学硕士论文.

[33] 洪宇,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.

[34] 李保利，俞士汶.话题识别与跟踪研究[J].计算机工程与应用,2003，17（7）：63-66.

[35] Wayne C..Multilingual Topic Detection and Tracking:Successful Research Enabled by Corpora and Evaluation[C].In：Language Resources and Evaluation Conference(LREC),2000：1487-1494.

[36] Y Yang,T Pierce,J Carbonell. Proceeding of the 21 st annual international ACM SIGIR conference on Research and development in information retrieval[C].1998：28-36.

[37] M.Spitters,W.Kraaij.A Language Modeling Approach to Tracking.News Events[A].Proceeding of Topic Detection and Tracking Workshop[C].2003：516-524.

[38] J M Schultz and Mark Liberman. Topic Detection and Tracking:Event-based Information Organization [C].Kluwer Academic: Massachusetts,2002：225-241.

[39] Ron Papka,J.Allan.On-line New Event Detection Using Single-pass Clustering Technical Report UMASS Computer Science Technial Report,University of Massachusetts,USA,1998：21-98.

[40] 洪宇,张宇,范基礼等.基于子话题分治匹配的新事件检测[J].计算机学报,2008，31（4）：687-695.

[41] 钟伟金,李佳,杨兴菊.共词分析法研究（三）--共词聚类分析法的原理与特点[J].情报杂志,2008，7：118-120.

[42] James Allan. Topic Detection and Tracking:Event-based Information Organization [M].Kluwer Academic Publishers,2002.

[43] Frey,D Gupta,Allan J Monitoring the News：a TDT Demonstration Systems.In the Proceeding of HLT 2001.San Diego,CA,2001：18-21.

[44] James Allan,Ron Papka,Victor Lavrenko.Online New Event Detection and Tracking[A].In：the proceeding of SIGIR[C].University of Massachusetts：Amherst,1998：37-45.

[45] J M Schultz and Mark Liberman.Proceeding of the DARPA Broadcast News WorkShop[C].San Francisco：Morgan Kaufmann.1999：189-192.

[46] 宋丹,林鸿飞,杨志豪.基于内容计算和链接分析的 Web 话题跟踪方法[J].情报学报,2007，26（4）：555-560.

[47] 邱立坤,龙志炜.层次话题发现与跟踪方法及系统实现[J].广西师范大学学报,2007，25（2）：157-160.

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

[49] Lavrenko V,Allan J,DeGuzman E.Relevance Models for Topic Detection and Tracking[A]. Proceedings of HLT-2002[C].San Diego.2002：104-110.

[50] Lewis D,Feature Selection and Feature Extraction for Text Categorization[A].Proceedings of Speech and Natural Language Workshop[C].San Franciso:Morgan Kaufmann.1992：212-217.

[51] 苏新宁.信息检索理论与技术[M].北京：科学技术文献出版社.2004,279-287.

[52] 刘清.Rough 集与 Rough 推理[M].北京：科学出版社.2001,11-37.

[53] 孙越恒,曹桂宏,侯越先.对称和非对称词语聚类模型的比较研究[J].计算机工程,2009，35（10）：56-61.

[54] J Allan,V Lavrenko,Margarete.A month to Topic Detection and Tracking in Hindi.ACM Transactions on Asian Language Information Processing.2003,2(2)：85-100.

[55] 钟伟金,李佳.共词分析法研究（二）--类团分析 [J].情报杂志,2008，6：98-100.

[56] 钟伟金,李佳.共词分析法研究（一）--共词分析的过程与方式 [J].情报杂志,2008，5：70-72.

[57] Qin He. Knowledge discovery through co - word analysis[A]. Library Trends, 1999：133 -159.

[58] 邓中华,孙建军.网络环境下共词分析方法的应用研究[J].图书馆杂志,2008，12：17-21.

[59] 龚秋艳,陈良育,曾振柄.简单高效的 URL 消重的方法[J].计算机应用,2010，30：36-42.

[60] Hua-Ping Zhang,Qun Liu,etal.Chinese name entity recognition using role model.Special issue Word Formation and Chinese Language processing of the International Journal of Computational Linguistics and Chinese Language Processing,vol.8 No.2:p.29-60,2003.

[61] Steinbach M, Karypis G, Kumar V. A comparison of document clustering

techniques[A].KDD Workshop on Text Mining[C].No.3:p.53-65 ,2000.

[62] Bun KK，IshizukaM.ToPic Extraction from News Archive Using TF*PDF Algorithm[A].In:Proceedings of the 3rd Intemational Conference on Web Information Systems Engineering(SISE2002)，SingaPore，2002.73-82.

[63] 王永恒.海量短语信息挖掘技术的研究与实现[D]. 长沙：国防科学技术大学，2006.

[64] 林仁炳,王基一.连续属性离散化算法的时间复杂性分析[J].计算机与现代化,2005，9： 40-42.

[65] 张亚玲,韩照国,任姣霞.基于相对熵理论的多测度网络异常检测方法[J].计算机应用,2010，30（7）：99-104.

[66] 赵凡.基于共词分析的学科主题动态跟踪相似方法探讨[J].情报杂志,2009，28（9）：66-70.

附录

作者在读期间发表的学术论文及参加的科研项目

学术论文：

[1] 王小华，徐宁，谌志群. 基于共词分析的文本主题词聚类与主题发现[J]. 情报科学.

参加项目：

1. 大规模汉语文本知识挖掘关键技术研究（08JC740011）2008.12 立项，教育部人文社会科学研究项目。起止日期：2009.1-2010.12.

2. 面向网络舆情管控的趋势挖掘技术（Y1100176） 浙江省自然科学基金项目。起止日期：2010.8-2012.7，GK100906004.

杭州电子科技大学

硕 士 学 位 论 文
详 细 摘 要

题　　目：网络舆情热点检测与跟踪技术研究

研 究 生　徐宁

专　　业　计算机软件与理论

指导教师　王小华教授

谌志群副教授

完成日期　2011年12月

随着互联网在全球范围内的飞速发展，网络媒体作为一种新的信息传播形式，已深入人们的日常生活。网络媒体具有进入门槛低、信息规模超大、信息发布与传播迅速、参与群体庞大、实时交互性强等综合性特点。如今网友言论活跃已达到前所未有的程度，不论是国内还是国际重大事件，都能马上形成网上舆论，通过这种网络来表达观点、传播思想，进而产生巨大的舆论压力，达到任何部门、机构都无法忽视的地步。可以说，互联网已成为思想文化信息的集散地和社会舆论的放大器。但是，与传统媒体不同的是，互联网是一个开放性和互动性的平台。在网络上，任何人都可以在博客、论坛、网络社区或者自建站点上发布言论和观点，他们既可以针对某一个社会现象或某一条新闻事件发表自己的看法，又可以通过互联网在网民之间形成互动场面，赞成方的观点和反对方的观点同时出现，相互探讨、争论，相互交汇、碰撞，使得各种观点和意见能够快速地表达出来。不同网民之间通过网络交流经验和共享资源的现象变的十分普遍。

所谓舆情，是舆论情况的简称，是在一定的社会空间内，围绕中介性社会事件的发生、发展和变化，作为主体的民众对作为客体的社会管理者及其政治取向产生和持有的社会政治态度。它是较多群众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等等表现的总和。而网络舆情是社会舆情的一种表现形式，它是民众在互联网上发布和传播的能够反映民众舆情的文字、图像、音频、视频等，往往是以文字的形式为主。网络舆情的特点与网络传播方式的特征息息相关，主要表现为如下特点：传播信息的多元性、传播方式的自由性、传播主体的隐匿性、传播事件的突发性等。同时，我们也必须清楚地认识到，由网络舆论的自由化所带来的一系列的消极影响：比如一些网民通过互联网散布谣言、进行偏激的评论等，因此，监测网络舆情，因势利导，提高新形势下的舆情信息的分析能力，及时准确地掌握社会舆情动向对于引导舆情以及做好舆情预警有着重要的意义。

热点话题检测与跟踪技术是目前在网络舆情分析中具有重要作用的一种信息处理技术。因为在海量的网络信息中，与同一话题相关的信息不管在时间上还是在空间上往往都比较分散。不同的语料源可能在相同的时间上对同一事件进行了报道，或者同一语料源可能在不同的时间段对同一事件进行了报道。各种报道的视角和分析都有可能不同，而舆情事件又具有不断演化的特性。这样，面对互联网上众多的语料源，人们对各种舆情事件难以做到全局性的把握。因此人们在寻求一种可以自动把各个孤立分散的语料按其话题的属性进行有效组织的技术。热点话题检测与跟踪技术正是在这种应用背景下产生，它已经成为网络舆情分析的重要技术手段。

此外，研究这项技术也有助于企业了解客户的需求变化，可以及时调整产品

和市场策略，同时对于公关、广告等产业提高市场的调研分析能力和效率都有着现实的应用价值和广阔的前景。

热点话题检测与跟踪是一种新颖的信息处理技术。它将一系列相关的报道按其所属话题进行有效的组织，以实现在语料流中对新话题或新事件的自动检测以及对已知话题的后续报道的跟踪。对于网络舆情的研究最早的是国外的话题检测与跟踪技术，其中研究比较好的是美国的 TDT 系统。美国有一个研究项目被称为 TDT(Topic Detection and Tracking，TDT)，它最初研究目的是能够发现和归纳来自于新闻流中的重要的信息。TDT 中的话题检测与跟踪技术的思想源于 1996 年，来自 DARPA、卡内基一梅隆大学、Dragon 系统公司以及麻萨诸塞大学的研究者们开始确定话题识别与跟踪研究的内容，并开发了应用于解决新闻流问题的一些技术。它能够实现话题组织、话题发现，并可以识别出各种新话题、突发事件以及关于某些特定事件的新报道，它主要可以用以帮助用户解决海量信息的智能分析处理。这可以广泛应用于政府、媒体、企业和证券市场等领域。此外，它还可以帮助用户找出感兴趣话题的所有报道，研究话题的发展历程等等。

随着网络舆情有效监测的需求和重视程度不断的提高，舆情分析的相关技术也日益成为研究的热点。舆情分析技术主要包括以下几个方面。

1、话题检测，其实是一种面向信息安全的技术，它主要依靠舆情信息的关注度、评论稀疏程度等参数，检测出最近发生的事件。

2、文本倾向性分析，对于文本舆情，主要根据文本的上下文信息提炼出文章的情感方向，给文章中的每个词汇用打分的方式进行分析统计，最后通过打分的结果来评价文本的倾向性。

3、话题跟踪，利用相似度分析下一个时间戳中新报道的话题，通过阈值判断当前话题是否被跟踪。可以及时了解和掌握后续发展动态。

4、统计分析技术，利用统计学、概率论的知识对舆情语料中的各个属性进行统计分析。通过统计分析技术可以直接获取文档摘要。

5、关联分析，挖掘出在海量数据中存在的隐性关系，通过分析这些隐性关系集合，得出它们的相关性。一般将关联分析切分成频繁项的挖掘和规则的挖掘两部分。

6、新事件产生，对新事件的产生进行时空的分析，通过对新事件各个时间戳进行跟踪，可以了解整个事件发生的场景，而且可以对该事件之后的发展进行合理的预测。

7、舆情统计报告生成，根据舆情分析引擎处理后生成报告，用户可根据指定条件对热点话题进行倾向性进行查询，并浏览信息的具体内容，以此来提供决策支持。

舆情分析技术是一门新兴的技术，国内外的许多研究机构都陆续展开了对该领域的研究，但还处于起步阶段，成果还是很有限。

本文首先介绍了网络舆情热点检测与跟踪技术的提出及其相关概念，重点介绍了网络舆情挖掘的一个重要分支----舆情热点检测与跟踪，阐述了网络舆情热点检测与跟踪技术的研究背景和意义、基本原理、发展现状等；在广泛阅读相关舆情热点检测与跟踪技术文献的基础上，分析前人的工作和当前舆情热点检测与跟踪技术的发展状况，比较和借鉴现有的算法，分析其优缺点，对相应的网络舆情热点检测与跟踪技术进行改进和完善，提出自己的算法，取得良好效果。文章本文的主要贡献如下：

1、本文对网络舆情热点检测与跟踪技术的现状和发展进行了简要的回顾。文中分别对网络舆情信息挖掘的相关技术和热点检测与跟踪算法作了分析。网络舆情信息挖掘主要包括网络舆情信息采集、网络舆情信息预处理、网络舆情信息分析等部分，并且介绍了现有的国内外网络舆情系统，对网络舆情热点检测与跟踪技术的研究主要包括话题检测与跟踪任务和话题检测与跟踪的关键技术等。在数据的采集上，主要是基于 Larbin(一种开源的网络爬行器)的一种应用，按照 Labin 自身的配置，运行速度比较快，占用的内存空间比较小，将主题网页中以文本为主的内容块重点抽取出来。

2、本文提出了基于共词分析的网络舆情热点检测方法。传统的共词分析方法一般运用在某一专业的学科领域中，通过判断学科领域中主题间的关系，进而展现该学科的研究结构。本文提出将共词分析运用到网络舆情热点检测方法中，而BBS是网络舆情的主要载体之一，主要是将设经过预处理之后带有年、月、日时间标签的论坛语料归入各个时间戳，经过分词、停用词过滤等预处理操作之后，统计出词语i的词频和文档频率，候选关键词提取主要是根据这两个参数得到的。然后通过词与词之间在文档中的共现程度构建出一个共词矩阵，并且根据词与词之间共现频数与彼此间距离成反比的思想，对构建出的共词矩阵进行标准化。这样不仅可以很直观的衡量各个词之间的距离，而且可以作为Bisecting K-means聚类算法的入口，从而进行分层聚类，通过大量的实验，发现在BBS论坛的环境下可以得到各个时间戳的BBS特定版块中的热点话题。实验证明本文提出的算法在BBS环境下的应用具有稳定性和高效性，并具有一定的可信度。

3、本文在总结了现有的主题关注度提取方法的基础上，分析了它的优缺点，并提出了一种基于关注度的热度提取方法，即综合考虑论坛帖子权重值和主题的媒体关注度对主题热度的影响。首先通过主题词回溯的方法提取关于相关主题的篇数，紧接着主要根据主题距离构建出主题进化图，将相对熵的概念引入到主题距离提取的方法上，并介绍了一些相对熵的应用。通过相对熵的阈值判断，从而

发现各个时间戳中主题的延续性。最后根据主题距离构建出主题进化图，通过主题进化图，可以直观的反映各个时间段主题之间的关系。

最后，分别使用大规模数据语料和真实论坛语料对本文提出的基于共词分析的网络舆情热点检测算法和基于热度分析的网络舆情热点跟踪算法进行了实验，并对测试结果进行了分析。实验结果表明，本文的算法对处理网络舆情热点检测与跟踪问题具有一定的可用性。

虽然，本文从对基于共词分析的网络舆情热点发现算法和基于热度分析的网络舆情热点跟踪算法的分析研究中取得了一定的成果，但同时也引出了许多新的问题，这些问题仍然需要进一步的研究。

- 本文仅从网络舆情热点检测与跟踪算法上作了较深入的研究，但是由于文本数据的特殊性，文本数据的预处理过程，包括分词、文本的计算机表示形式、如何确定候选关键词、更有效的排序方法等方面对热点主题的准确提取也有很大的影响，本文对此没有作更深入研究，所以，作为下一步工作，拟对此作深入研究。
- 本文提出的基于相对熵的主题进化距离算法，对于主题进化距离阈值的设置均需有一定的专业知识，如何设置更准确的阈值参数也是下一步工作的一个方向。
- 基于共词分析的网络舆情热点发现算法和基于热度分析的网络舆情热点跟踪算法还存在很多问题有待解决，例如，在候选关键词提取方面，有些词语可能在某一时间段内权重值趋于稳定，但是由于网络的不可预见性，这些词语也有可能会成为候选关键词。同时需要更好的去改进聚类算法，这样可以发现是否可以更好的发现热点话题。在将来的工作中，我们可以对上述问题进一步地研究。展望将来，网络舆情热点检测与跟踪技术将会得到广泛的应用，其技术的进步也会极大地提高热点话题发现的效率和效果。

关键词：热点检测，跟踪技术，网络舆情，共词分析，中文信息处理