分类号：_____ 密级：_____

UDC：_____ 编号：_____

工学硕士学位论文

互联网话题演变与传播分析技术研究

硕士研究生 ：郑希文

指导老师 ：张志强 教授

学位级别 ：工学硕士

学科、专业 ：计算机软件与理论

所在单位 ：计算机科学与技术学院

论文提交日期 ：2009 年 1 月

论文答辩日期 ：2009 年 3 月

学位授予单位 ：哈尔滨工程大学

摘 要

随着互联网技术的迅猛发展，网络舆情监管工作的重要性逐渐被人们认同。目前，网络舆情分析技术已经成为国内外的研究热点，并取得了一定的研究成果，主要的研究领域包括：话题检测、话题跟踪、自动摘要、趋势分析、舆情预警等。本文在已有研究工作的基础上，针对互联网话题演变和传播问题，力图在网络舆情分析技术领域做更深层次的研究，为网络舆情监管工作提供有力的技术支持。本文的主要研究内容有以下两个方面：

提出基于多中心和向量分解的话题演变分析技术。为了解决话题漂移问题、呈现话题演变过程，采用多中心话题模型来描述话题的多个侧面，提出向量分解思想发现后续文档的新颖特征，从而实现对话题中心的建立和更新，最后结合增量聚类算法，提出了解决话题演变问题的完整方案。实验证明，该方案能够有效提高话题检测性能、清晰呈现话题演变过程。

提出基于传播图和多元线性回归的话题传播分析技术。在论坛间，提出基于相似度比较和关键词匹配的转帖关系发现技术，并结合初始传播论坛发掘以及传播周期的计算，建立论坛话题传播图；在论坛内，提炼影响传播行为的指标体系，并结合多元线性回归理论，实现了对传播趋势的预测。实验证明，上述方案能够发现话题传播路径、准确预测话题传播趋势。

综上所述，本文在研究和总结现有舆情分析技术的基础上，重点针对话题内容和行为特征，对话题演变和传播分析技术进行了研究，并通过实验验证方案的可行性和实际效果，为网络舆情监管工作的进步做出了贡献。最后，本文还展望了该领域的发展趋势。

关键词：话题演变；话题传播；向量分解；传播图；多元线性回归

Abstract

With the rapid development of Internet technology, the importance of the supervision of network public opinion has been gradually recognized. At present, the analysis technology on network public opinion has become a hot topic at home and abroad, and has achieved some results including topic detection, topic tracking, automatic summary, trend analysis and early warning of public opinion. In order to solve the problem on topic evolvement and diffusion in network, in this thesis we tried to provide strong technical support for the supervision of network public opinion. The main content of this thesis are as follows.

The analysis technology of the topic evolvement has been studied based on multi-center and vector decomposition. To address the topic drift problem and show the process of the topic evolvement, we use the multi-center topic model to describe the aspects of a topic. In order to achieve the establishment and updating of the topic center we apply the vector decomposition method to find the new features of follow-up documents. Then, we put forward the project to solve the topic evolvement problem combined with incremental clustering. Experimental results show that the project can effectively improve the performance of the topic detection, and show the process of the topic evolvement clearly.

The topic diffusion analysis technology based on diffusion map and multivariate linear regression has been studied. Among Bulletin Board System(BBS), we study the technology to discover transmitting relations based on comparing similarity and matching words. We establishes the diffusion map on BBS to explore the initial diffusion BBS and computing diffusion period. In BBS, we study the index system affecting topic diffusion to achieve the diffusion trend forecast based on multivariate linear regression theory. Experiments show that the project can detect the topic diffusion path, and predict the topic diffusion trend accurately.

To sum up, based on researching and summarizing the analysis technology on network public opinion, in this paper we focus on the content and behavior characteristics of the topic and analyze topic evolvement and diffusion. In experiments we validate feasibility and actual results of the project. It contributes to the advancement of the supervision of network public opinion. Finally, this paper also analyze the trend of development in the field.

Key words: Topic Evolvement;Topic Diffusion; Vector Decomposition; Diffusion Map; Multivariate Linear Regression

哈尔滨工程大学

学位论文原创性声明

本人郑重声明：本论文的所有工作，是在导师的指导下，由作者本人独立完成的。有关观点、方法、数据和文献的引用已在文中指出，并与参考文献相对应。除文中已注明引用的内容外，本论文不包含任何其他个人或集体已经公开发表的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者（签字）：郑希文

日期：2009 年 3 月 17 日

哈尔滨工程大学

学位论文授权使用声明

本人完全了解学校保护知识产权的有关规定，即研究生在校攻读学位期间论文工作的知识产权属于哈尔滨工程大学。哈尔滨工程大学有权保留并向国家有关部门或机构送交论文的复印件。本人允许哈尔滨工程大学将论文的部分或全部内容编入有关数据库进行检索，可采用影印、缩印或扫描等复制手段保存和汇编本学位论文，可以公布论文的全部内容。同时本人保证毕业后结合学位论文研究课题再撰写的论文一律注明作者第一署名单位为哈尔滨工程大学。涉密学位论文待解密后适用本声明。

本论文（□在授予学位后即可　□在授予学位 12 个月后　□解密后）由哈尔滨工程大学送交有关部门进行保存、汇编等。

作者（签字）：郑希文　　导师（签字）：张志强

日期：2009 年 3 月 17 日　　2009 年 3 月 17 日

第 1 章 绪论

1.1 课题研究背景

随着互联网技术的迅猛发展，网络在传达社情民意方面的优势逐步显现出来。继传统的报纸、广播、电视之后，互联网已经确立了自己第四媒体的主导地位，并在表达民众心声、反映社会舆论方面发挥极其重要的作用。网络舆情是通过互联网传播的公众对现实生活中某些问题所持的有较强影响力、倾向性的观点和言论，是网民关注的热点，是民众讨论的焦点，集中反映一个时期网络舆论的中心。当今，信息交互与舆论传播空前迅捷，网络舆情的表达形式也日趋多元化。如果管理不善，负面的网络舆情将对社会公共安全造成极其恶劣的影响。如何对网络舆情加以有效的监督和引导，使和谐的互联网环境为构建社会主义和谐社会发挥重要作用，已成为网络舆情工作面临的一个重要课题。

目前，话题检测与跟踪技术（Topic Detection and Tracking，简写为 TDT）由于其在自然语言处理领域的优势，已经成为网络舆情分析工作主要的研究方向。TDT 起源于早期的面向事件的检测与跟踪。现在的 TDT 主要是面向语言文本和语音形式的新闻报道，负责自动整理报道边界、发掘突发性新闻话题、跟踪话题后续发展，以及跨语言检测与跟踪等相关任务。也就是说，TDT 检测与跟踪的对象除了特定时间和地点发生的事件，还扩展为定义更为宽泛的话题，相应的理论研究也从传统对于事件的检测发展到对突发事件及其后续相关报道的发掘等方面[1]。TDT 主要包括五项基础研究：面向新闻广播报道的切分任务；面向未知话题的检测任务；面向已知话题的跟踪任务；对未知话题首次报道的检测任务和报道间相关性的检测任务。应该说，本文就是在话题检测与跟踪任务的指导下，结合当前网络舆情工作的应用背景，重点在话题演变和传播方面进行的研究和探索[2]。

本文的题目是互联网话题演变与传播分析技术研究，显然互联网话题是主要的分析对象。TDT 对话题的准确定义经历了一个过程：最初的 TDT 研

究将话题定义为事件。事件是发生在特定时间和地点的事情。比如“2008 年 8 月 8 日北京奥运会开幕”就是一个事件。此外，事件还包括可预期事件和突发事件。从 TDT2 开始，话题有了更加广泛的含义，不仅包含了由最初事件引发的后续事件，同时还包含了与其直接相关的其他事件或活动。直到 TDT5，话题都一直沿用如下定义：一个话题由一个种子事件或活动以及与其直接相关的事件或活动组成。根据上述定义，一篇报道只要论述的事件或活动与一个话题的种子事件存在某种联系，那么这篇报道就与该话题相关，它们共同属于一个“话题”。结合网络舆情的特点，话题不仅包括对已经发生事件的记录和描述，更包含民众对该事件的讨论和观点。根据目前互联网服务的发展状况，本文对网络话题定义如下：“一个话题能够叙述新闻事件、反映社情民意，其具体表现形式为一系列新闻报道、网站评论、论坛帖子、博客日志等的集合。”

当前，基于话题的网络舆情分析技术已经成为国内外的研究热点，并取得了一定的研究成果，主要的研究领域包括：热点话题检测、敏感话题识别、话题后续跟踪、倾向性分析、自动摘要等。然而，针对日益严峻的网络舆情态势，目前以检测和跟踪为基础的分析技术还不能完全满足舆情监管工作者的需求。他们迫切希望在发掘网络话题的基础上，能够对这些话题做更深层次的分析和处理。本文正是在这样的应用背景下，结合已有的研究成果，重点从演变和传播两个角度分析网络话题，力图发现话题演变过程，研究话题传播机制，为网络舆情监管工作提供有力的技术支持。

1.2 课题研究内容

针对网络舆情工作的现状，本文的研究内容包括话题演变分析技术和话题传播分析技术，重点在演变和传播两个方面对网络话题做深层次的分析，并通过实验对本文提出的理论进行验证。研究工作的具体内容如下：

话题演变的定义是：一个话题在动态发展过程中，话题讨论的重点也随着时间相应的变化，表现为内容上重心的转移和分化。针对上述话题演变的定义，本部分的研究工作主要有以下两个方面：第一，提高已有话题检测与跟踪的召回率：应该说，绝大多数话题都是动态发展的，最初建立的话题模型仅由少数该话题的初始报道组成，报道数量少，话题模型简单。后续相关

报道讨论的内容可能是该话题的其他侧面，即话题的中心会发生迁移和分化，导致最初建立的话题模型不能跟踪到该话题的后续报道。因此，如何建立和更新话题模型，使其能够跟踪到话题的后续报道是话题演变要解决的问题之一；第二，呈现话题演变过程：在浏览某一话题的相关报道时，传统方法是将报道按其发布时间顺序简单罗列，用户不清楚该话题是怎样动态演化的。因此，使用什么样的话题模型使其在准确识别话题的基础上，能够清晰展现话题动态发展过程是话题演变分析技术要解决的另一个主要问题。

如果说话题演变是基于内容的研究，那么话题传播就是基于行为的分析。话题传播的定义是：一个话题在内容上没有大的演变的前提下，在时间空间上的发展和扩散，它主要是一种基于行为的分析与处理。当前，网上信息传播的方式多种多样，包括网络新闻和专题、网站评论专栏、虚拟论坛言论、博客帖吧、点击排行和在线调查等。其中，论坛主要是以发帖和跟帖的形式实现话题的发布和交流。由于目前论坛仍采用匿名登陆制，随意性较强，约束力不够，并且近几年来网上论坛迅猛发展，不同规模和主题的论坛层出不穷，许多敏感话题、虚假消息甚至反动言论都是通过论坛散播开来，对社会危害极大，为网络舆情监管工作带了新的挑战和难题。在这样的现实情况下，话题传播分析技术重点分析论坛上话题传播的趋势和动态，尤其关注敏感话题以及可能造成大规模传播的话题，力争对可能造成恶劣影响的话题提早进行防范和控制。以这样的应用需求为背景，本文力图发掘话题在论坛间的传播路径，分析话题在论坛内部的传播行为，客观描述话题传播状态、准确预测话题传播趋势。

1.3 论文组织结构

本文以日益严峻的网络舆情为应用背景，详述了网络舆情分析工作的研究现状，重点针对话题演变和传播技术进行了深入的研究和分析，结合已有的研究成果，提出了基于多中心与向量分解的话题演变分析技术、以及基于传播图与多元线性回归的话题传播分析技术，并通过实验对上述技术和理论进行验证。本文具体组织结构安排如下：

第二章是话题演变与传播分析技术研究，在已有舆情分析技术的基础上，重点介绍了网络话题演变和传播技术的研究现状，并且总结了该领域未来的

发展方向。

第三章提出了基于多中心与向量分解的话题演变分析技术，重点论述了话题多中心模型和向量分解的思想，并且提出了以增量聚类算法为基础的、解决话题演变问题的整体方案和流程，最后通过实验验证该方案在话题检测跟踪和发掘话题演变过程方面的效果。

第四章是基于传播图与多元线性回归的话题传播分析技术，首先阐述了论坛间话题传播图的构建方法，然后提出一套影响论坛内话题传播的指标体系，并结合多元线性回归方法对传播行为进行描述和预测，最后通过实验对上述理论加以验证。

结论部分总结了本文研究的主要内容和创新点，分析了当前研究工作的不足之处，最后提出本领域未来的发展趋势。

第 2 章 话题演变与传播分析技术研究

2.1 引言

在网络舆情日益严峻的背景下，结合该领域已有研究成果，本文将互联网话题演变与传播分析技术作为主要的研究内容。目前，国内外的研究者在演变和传播方面已经开展了一系列的工作，并取得了一定的研究成果，为本文的顺利完成奠定了基础。本章主要介绍了话题演变与传播分析技术的国内外研究现状，重点论述了目前中文话题跟踪技术的研究现状、国内学者在话题演变方面的探索、在宏观微观两个层次解决话题传播问题的思想、将复杂网络模型引入话题传播研究中的策略等，最后分析了本领域的发展趋势。

2.2 话题演变分析技术研究现状

为了分析话题演变现象，首先要提高已有话题跟踪技术的召回率，解决话题漂移问题，然后才能发现话题内容重心的转移、呈现话题演变过程。下面，主要讨论了目前话题跟踪技术的研究现状，并介绍了研究者在话题演变方面的探索和实践。

2.2.1 话题跟踪技术

为了提高话题跟踪技术的准确率和召回率、全面获取话题的后续文档，这也是解决话题演变问题的前提条件，众多研究者将工作的重心放在设计和实现动态话题模型上，力图使模型随着后续数据的到来自适应的动态调整，相关研究工作如下：东北大学的王会珍提出了自适应的话题跟踪方案，并分别采用主动学习和反馈学习的思想。首先构建基于数据流的主动学习框架，通过特征权值调整和话题向量转移两种方法动态的更新话题模型，并将确定性和不确定性同时引入到样本选择的策略中来，提高了话题跟踪的性能；同时，王会珍还采用增量方式对话题追踪模型进行修正，为每次修正后的话题追踪模型构建一个弱话题追踪器，保留初始的话题追踪器以及每一次修正后

构建的弱话题追踪器，用于追踪后继文档的总话题追踪器是这些弱追踪器的线性组合，同时还将时间因素考虑进来。该方案实现了无监督的调整话题模型，优化了已有的跟踪算法[3-4]。复旦大学的黄萱菁用可以动态调整的模板来实现文本过滤，模板和阈值的调整受用户反馈信息的控制[5]。莫倩将有监督自适应机制引入到话题跟踪中来，通过给报道打分、增量学习和调整关键字权重的方法实现话题模型的自适应调整，同时还提出了动态调整阈值的策略，在一定程度上提高了话题跟踪的召回率[6-7]。哈尔滨工业大学的洪宇将自适应学习的策略应用到层次聚类算法中，在众多的反馈信息中自动提取最优信息作为调整话题模型的依据，改进了传统的反馈机制，提高了反馈学习的应用效果，完善了自适应学习策略[8]。上述研究成果主要体现在“动态”上，动态的模型、动态的阈值等。研究者力图使用动态调整的模型和阈值来适应话题内容的演变，从而提高跟踪的性能。然而，上述研究并没有提出如何展现话题演变过程的策略。

由于目前文本挖掘大多采用传统的基于统计学的方法，一些研究者便将工作的重心放在 VSM 向量空间模型上，试图通过改进 VSM 向量来提高话题跟踪的性能：Juha Makkonen 基于新闻事件的四大要素：时间、地点、人物、事件内容，把传统的单一向量按照不同的词义划分为四个子向量：为普通词组成向量、为地点名词组成向量、为人名组成向量、为时间名词组成向量。分别计算四个子向量的相似度，最后统一为一个相似度。该方案是针对新闻要素进行向量空间划分，对新闻分类效果较好；但并不适用于其他网络文本，如论坛的帖子[9]。国内学者在改进向量空间模型方面也进行了一定的研究：大连理工大学的宋丹提出了一种改进的向量空间模型，按照语义将特征词分为四组，分别是时间、地点、任务和内容，并分别生成四个子向量。该方案与 Juha Makkonen 的思路一致，在对标准新闻文本的处理中获得了比较好的效果[10]。南京大学的李昕使用多维文档模型来处理论坛消息，所谓的多维具体包括关键词、用户、时间和话题线索四个维度。同时还引入加窗分析技术解决语义漂移问题。该方案与上述文献的策略相似，也是根据语义不同对话题特征分别处理，从而发现论坛中完整的语义信息[11]。哈尔滨工业大学的赵华用两个中心向量来表示一个英文话题：标识中心向量和内容中心向量。用报道中出现的、但不是位于句首的首字母大写的词生成标识中心向量；用报道

中出现但不是标识词的词生成内容中心向量。同时，结合 Single-Pass 聚类算法，使用双向量模型实现对新闻报道的精确区分。不足之处是，该方案仅适用于英文文档，无法对中文文档进行处理[12]。东北大学的王会珍同样将向量分成两个子向量，但分组的策略是根据命名实体。结合命名实体识别技术，用命名实体特征生成一个子向量，用其余特征生成另一个子向量。在话题跟踪的过程中，首先分别计算两个子向量的相似度，再将两个相似度值加权求和，最终得到一个总的相似度，作为两个文档内容上的距离。由于命名实体的引入，命名实体库质量的好坏、命名实体识别算法的效果将直接影响话题跟踪的性能[13-14]。上述研究工作力图在传统向量空间模型上有所突破，研究者改进了已有的 VSM 向量空间模型，根据语义等特征将向量分组，先组内计算，再组间合并，在一定程度上提高了话题跟踪的性能。但在实际应用中仍具有局限性，并且没有对话题演变现象进行分析和处理。

2.2.2 话题演变分析技术

在传统话题跟踪任务的基础上，一些学者已经开始研究话题演变分析技术，并取得了一定的研究成果：哈尔滨工业大学的赵华提出了一种面向动态演化的双质心话题模型。在双质心话题模型中，话题由初始质心和当前质心表示，以分界点为界。初始质心代表了话题在分界点之前关注的内容，当前质心表示话题从分界点到当前时刻所关注的内容。分界点、初始质心和当前质心随着文档的到来而不断更新，体现了话题动态演化的思想。如图 2.1 所示，实线椭圆代表一个话题，虚线圆形代表该话题的当前质心，虚线椭圆代表该话题的初始质心，小圆圈代表该话题的一篇报道，箭头代表质心之间的分界点。图 a 表示话题刚刚建立，只包含一篇文档，该文档属于当前质心；图 b 表示第二和第三篇文档被陆续跟踪到，加入当前质心；图 c 表示，第四篇文档被跟踪到，但和前三篇文档在内容上有所不同，建立分界点，原来的当前质心转化为初始质心，用第四篇文档建立新的当前质心；图 d 表示，随着后续文档陆续到来，初始质心不断扩充，当前质心也在不断的建立和更新。这样，话题的当前质心始终代表话题最新的讨论重心，而初始质心则是对话题以前讨论内容的概括，双质心模型就这样建立起来。该模型对于各个讨论重心文档顺序出现的话题，能够展现话题动态发展的过程；但不同重心的文

档有可能交叉出现，双质心模型无法准确识别，限制了双质心模型的应用[15]。

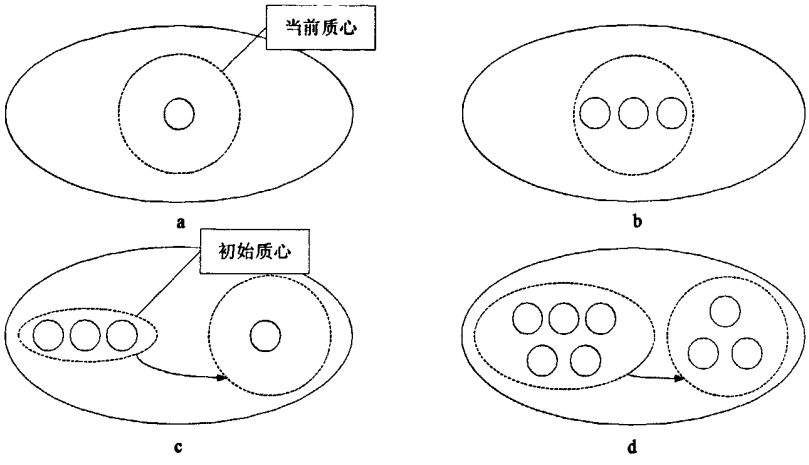

图 2.1 双质心话题模型示意图

中科院贾自艳的贡献是提出了事件“来龙去脉”的概念，实际上就是对话题演变问题的另一种表述。首先为文档生成摘要，然后使用段落主题相关性判别模型来维护摘要与话题的相关性，使用新颖性计算模型祛除内容基本相同的摘要，只保留内容新颖的摘要，最后还定义了一种事件来龙去脉评测模型对分析结果进行检验。实验验证，该方案提出的模型基本可行，能够生成一个话题的来龙去脉，在一定程度上呈现了话题演变的过程。贾自艳还提出了基于时间距离的相似度计算模型、事件模板进化策略、以及动态阈值设置思想，并借鉴 Single-Pass 聚类算法对话题检测与跟踪技术进行了改进。基于时间距离相似度计算模型通过引入时间因素区分不同的话题，如果两个事件之间的公共特征很多，并且发生时间不同，仅从内容上计算相似度容易误判为同一个话题，考虑时间因素就很容易区分开[16-17]。

在国外，Ramesh Nallapati 提出新的话题模型来分析话题演变现象。话题是由其内部的子话题及这些子话题之间的联系组成的。针对话题这一特性，首先要把属于话题的文档聚类到不同的子话题中，然后建立这些子话题之间的依赖关系。这里话题模型用一个有向图表示，图的节点代表子话题。如果两个子话题的相似度超过阈值，则建立一条有向边，由发生较早的话题指向

发生较晚的话题。有向图方法描述了如何在话题内部的各个子话题之间建立演变关系，对属于某一话题的报道进行更清晰的组织，方便用户的浏览。但文献中未解决话题演变根本的问题，即如何跟踪到内容上已经发生演变的报道[18]。

大连理工大学的金珠针对话题发生迁移和分化的特点，提出了事件框架的思想。通过对事件的不同侧面抽取敏感词，构成一种分类体系，在建立完整的事件框架的基础上，对事件的相关文档进行归类和整理，从而提炼出该事件的演变过程。在生成事件框架时，必须首先对事件样本文档按照不同的侧面进行聚类，然后从各个侧面中抽取框架的敏感词，最后人工排除一些明显的干扰。文献还提出了判别文档倾向性的方案，主要是基于知网的支持，收集正反双方立场的若干报道，提取敏感要素，在报道中找出包含敏感要素的关键句，进行报道立场的分析和计算[19-20]。清华大学的吴平博同样在事件框架方面进行了研究，提出了一种基于事件框架和事件主体信息的文档检索方法。该方法首先从事件样本中提取事件的框架要素，从事件文档中提炼事件的主体信息，然后将框架要素和主体信息转化成向量，从而优化了事件相关度评价函数。实验证明该方案提高了文档检索的性能[21]。在上述事件框架思想的基础上，大连理工大学的林鸿飞提出了一种基于语义框架的话题跟踪方案。将话题各个内容侧面定义为“槽”。在进行事件跟踪时，独立计算各个槽的相似度，并通过内容槽扩展的方法解决话题漂移问题。实验证明这种方法是有效的。上述方案都是基于一种“框架”的思想，力图用事先训练好的框架还原话题演变过程。该策略在训练的时候，需要收集大量的样本语料，样本中必须包含事件各个侧面的文档和报道，因此训练语料收集的全面与否、事件框架生成的客观与否都将直接影响算法的性能[22]。

2.3 话题传播分析技术研究现状

通过查阅国内外相关文献，发现目前话题传播领域的研究主要集中在以下几个方面：网站间宏观层次的研究、网站内微观层次的研究、结合复杂网络模型进行仿真实验的思路等。下面具体介绍各个研究方向的工作进展，并进行对比分析。

2.3.1 宏观层次的研究

在宏观层次上，研究者主要关注信息在网站间的传播行为。这里，把网站看作是信息传播的基本单位，并将其抽象成具有多个属性值的点，属性值用来描述该网站在信息传播中所做的贡献，点与点之间通过有向边连接，该有向边即网站间信息传播的路径，从而构造出网站间的信息传播图。通过对传播图的构造和维护，可以在宏观层次上获得网站间信息传播的概况。宏观层次的管理者主要是国家相关机构和部门。

万小军将一个新闻事件所有的文档都获取下来，然后基于元数据特征、关键词特征和文本相似度特征，采用 SVM-Light 二值分类算法，得到每一篇文档的“源文档”，即新闻转载的出处，进而可以得到该新闻事件在各大新闻网站传播的路径图，如图 2.2 所示。根据该路径图，能够得到传播起始网站——该话题文章最先出现的网站，以及传播中心网站——向其他网站转载文章总数最多的网站，从而获得一个新闻事件通过各个网站传播的整体状况。不足之处是网站间关联方式单一，就是转载文章，并且在论坛上缺乏可行性[23]。Avaré Stewart 的主要工作是分析博客之间信息交互的行为特征。首先需要建立了一个博客序列数据库，存储若干博客序列。所谓博客序列就是一个话题文档在各个博客上的传递次序。通过对该数据库的挖掘，得到信息传播路径，即对信息传播贡献最大的博客序列，从而可以对一些敏感话题的传播做出有效的防范和控制[24]。

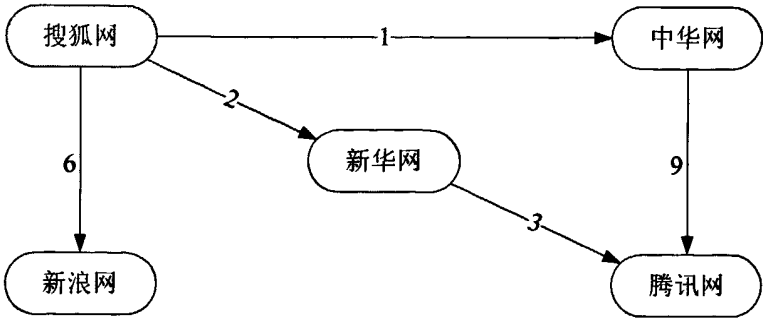

图 2.2 网站新闻传播路径图

对于宏观层次上的工作，理论上应该覆盖全网所有知名论坛，构建一张

“论坛话题传播图”，达到监控全网论坛的目的。该策略能够获得一个话题在论坛间的传播路径，以及每个论坛对传播行为的贡献。优点是监控范围广、处理数据全面，有助于总体的分析和决策；缺点是将论坛抽象成孤立的点，忽略了对论坛内部话题传播机制的挖掘和研究。并且，被监控论坛的多少，将直接影响话题传播图的客观性和真实性。同时，由于监控范围扩大到全网各大论坛，海量数据的获取和更新也提高了实现上的难度。

2.3.2 微观层次的研究

微观层次上，研究者主要关注网站内部信息传播的机制和特征。这里，研究的对象是人、以及人与人之间的信息交互关系。具体的说，把人看作是信息传播的基本单位，每个人都有发布信息的权利；同时，对于已经发布的信息，每个人都可以选择接受或者拒绝，可以选择讨论、传播或者沉默。出于每个人的兴趣和背景不同，他们的行为也一定不同。正是这种群体行为的多样性，以及对这种多样性的挖掘，为我们分析话题传播过程奠定了基础。微观层次的管理者是由站长、版主等组成的网站内部的管理团队。目前主要的研究成果如下：

西安交通大学的宫辉利用社会网络矩阵分析法对网络虚拟社区中信息传播模式进行分析，概括出网络虚拟社区群体的基本特征，同时利用计量分析模型对这些特征如何影响虚拟社区的信息传递进行验证和分析，并提出了控制信息传播的对策。宫辉最大的贡献是提出了利用社会网络矩阵来描述论坛信息传播的方法。不足之处是抛开了具体的话题，只是对整个论坛上的信息交互网络进行分析和挖掘，重点是分析人与人之间的关系，而对某一特定话题的传播趋势关注不够[25]。白淑英重点研究论坛互动的结构性要素有哪些，其互动过程的内在机制是什么，存在哪些一般性的理论问题。研究者以哈工大紫丁香论坛为样本空间，使用社会关系矩阵法来测量论坛讨论关系的互动特征，进而对帖子进行量化处理，得出如下结论：论坛互动关系的建构至少需要四个结构性要素，即电子空间、话题、角色、帖子；论坛互动可分为焦点互动和非焦点互动两种基本类型；由帖子的性质所决定，论坛参与者具有多种角色关系，进而会达成多样化的互动模式，具体包括：单中心互动模式、多中心互动模式、跨网互动模式、两两互动模式和宣告阅读互动模式等[26]。

北京理工大学的于静提出了基于社会关系网络分析的论坛内容安全动态监测模型。综合了关键字匹配、统计模型和语义模型的特点和优势，提炼出一种综合的分析模型。该方案引入社会学和侦察学的知识，将论坛信息安全技术与社会关系网络分析、侦察学现场勘察技术相结合，成功发现论坛中传播信息的中心人物和中心群落，挖掘信息传播渠道和行为特征，为监控论坛内容提供了有效的技术支持[27]。

相对于宏观层次，微观层次的工作把注意力放在了论坛内部，重点研究论坛内人与人之间信息交互的行为特征。优点是：把研究的重心集中到传播的最基本单位——人，力图挖掘互联网信息传播的基本原理，研究工作更加细致入微，触及传播最底层的概念和机制；缺点也是显而易见的，分析范围小，仅仅关注一个论坛。更重要的是，由于论坛使用匿名制，每个人可以使用完全随意的用户名登陆论坛，而同一个人完全可能使用不同的用户名登陆不同的论坛，这就使把微观层次的研究扩展到宏观层次的想法变的困难重重。这里可以考虑使用 IP 地址识别论坛用户的真实身份。

2.3.3 结合复杂网络模型的研究

主要思路是应用复杂网络理论解决话题传播问题。其中，小世界网络和无尺度网络是目前两类最典型的复杂网络模型。研究者结合小世界和无尺度特征，通过仿真平台建立信息传播模型，并实现模型的动态演化。将现实网络环境数据与仿真实验数据进行对比，结果基本吻合，从而验证了网络信息传播符合复杂网络模型中的某些特性，为后续的工作奠定了理论基础。

厦门大学的张嘉龄提出了博客网络信息传播的博弈演化模型，利用无尺度网络的健壮性对模型进行了简化，研究了在网络高效率的传播机制下信息的扩张或湮灭。将现实中博客网络的数据与仿真实验的数据进行对比，发现实验结果和真实话题的传播过程基本吻合，最后展望了这类信息传播模型的推广方向。该文献的思路是利用模型仿真的方法证明博客信息传播网络具有无尺度网络和小世界网络的某些特征[28]。国防大学的刘常昱利用小世界模型构建人际关系网络拓扑，并以此为基础，通过设计个体的局部相互作用规则，引入个体心理因素和外界媒体影响，提出了利用计算机仿真建立舆论传播演化模型的基本思路。初步实现了我国某特定地区舆论传播模型，并对该模型

的构建过程进行了分析，为量化研究舆论传播这一复杂社会现象提供了有益的探索和思路。该文献同样是应用模型仿真的策略，证明舆论传播具有小世界网络的某些特性[29]。

对于结合复杂网络的思路，目前仍然处于实验阶段。优点是理论基础雄厚、前景广阔；然而，如何将小世界和无尺度特性应用到实际网络环境中解决实际问题，目前还没有可行的方案，需要进一步的攻关和研究。

2.3.4 其他有价值的研究

日本研究者 Naohiro Matsumura 提出了一种“影响力传播模型”的概念。主要是针对网络留言版上人与人之间的信息交互行为。该模型大致的思想是：用两个人发布文档共有信息的多少来衡量两个文档之间的影响力，进而通过文档影响力得到两个信息发布人的影响力，然后得出在这个信息交互网络上人与人之间的“距离”。最终根据影响力、距离等因素，对该信息交互网络做出整体的分析和评价。影响力传播模型最大的贡献是把内容引入到对行为的研究中；缺点是策略比较简单，可以考虑加入文本相似度计算对已有的影响力传播模型进行改进[30-34]。图 2.3 是影响力传播模型的示意图。

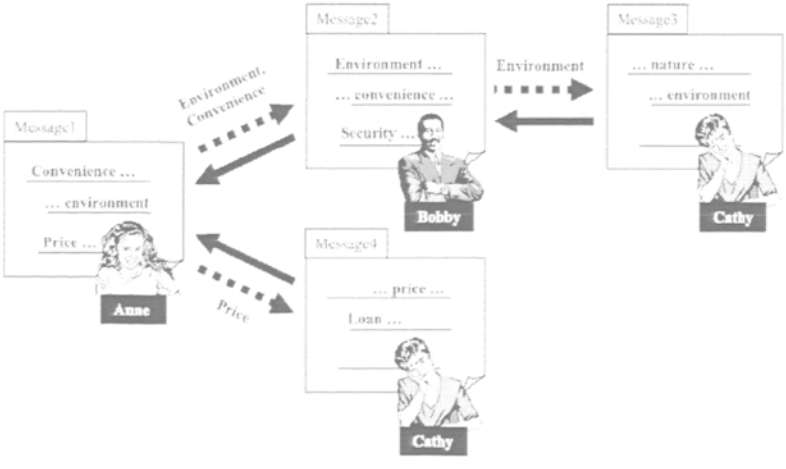

图 2.3 影响力传播模型示意图

浙江大学钱斌最大的贡献是研究了网络口碑再传播意愿的影响因素。将因变量——消费者网络口碑再传播意愿分为人际再传播意愿和群体再传播意

愿两类，将研究对象按照有无满意度感知体验分为两类，分别对其进行对比研究。在构建实证概念模型时，从口碑来源、口碑内容、接受者自身特征以及发送者和接受者关系强度这四个角度，梳理了各种因素，纳入到概念模型中去。应该说，话题传播已经涉及到社会学等其他学科，浙江大学钱斌所做的工作最具代表性，这就是一篇社会学论文。其中，提出假设、实验验证、总结影响因素等研究方法对研究话题传播现象很有价值[35]。

2.4 话题演变与传播分析技术发展趋势

2.4.1 话题演变分析技术发展趋势

目前，话题检测和跟踪大多是应用传统的文本分类聚类技术。在话题演变方面，主要有以下几种解决方案：首先，研究者提出了话题双质心模型，用于及时捕捉话题新出现的内容。双质心模型能够及时捕捉到话题新侧面的出现，但不适用于话题各个侧面报道乱序到达的情况；同时，话题演变是随着时间进行的，因此有些文献将时间因素引入话题相似度的计算，认为话题或者报道时间差越小，相似度越大，这种计算模型对于区分不同的话题有一定的效果，但不适用于对同一个话题不同侧面的区分；还有一种思路是建立完整的话题框架模型，需要事先收集话题不同侧面的报道，并提取侧面关键词。这种模型需要较多的人工干预，而且各个侧面信息收集是否全面对话题检测和跟踪的性能影响很大。通过总结分析以上研究成果，话题演变分析技术的发展趋势如下：

（1）建立完善的话题模型：话题模型从研究者给出其定义开始，经历了从简单到复杂、从静态到动态的过程。建立话题模型的目的是用较少的数据通过一定的组织和结构来展现话题特征、辅助实现话题检测与跟踪，而不必保存全部数据，从而减少了算法的开销。传统的话题模型只是由话题初始的几篇文档构成，在话题检测与跟踪的过程中保持不变。这样的模型不但召回率低，不能跟踪到话题后续的文档，更无法展现该话题在内容上重心的转移，不能呈现话题演变过程。在今后的工作中，力图设计这样一种模型：它能够随着话题演变实现动态调整，不但能够获取该话题所有的后续文档，更可以展现一个话题发展的来龙去脉。

（2）结合语义分析：目前，无论是对话题检测与跟踪任务的研究，还是对话题演变问题的探索，国内研究者的工作重心仍然是基于统计学的方法，如传统的 VSM 向量空间模型。该方法由于其实现简单、开销小、效率高等优点，被广大科研机构认同和应用。在语义分析方面，国内研究者也进行了一定的尝试和探索，但实现效果一般，相关的研究成果也比较少。本文认为，要在中文文本挖掘领域做进一步的突破，仅仅依靠 VSM 向量空间模型显然是不够的。例如，基于统计学的方法就不能解决同义词问题，语义分析的引入事在必行。因此，语义分析理论与统计学原理相结合的研究思路势必成为话题演变、乃至文本挖掘领域未来的发展趋势。

2.4.2 话题传播分析技术发展趋势

目前，针对论坛的话题传播分析技术方面的研究并不多，主要集中在以下几个研究方向：在论坛间发掘信息传播路径；分析论坛内的人际信息交互关系；应用复杂网络理论对话题传播行为进行仿真；社会学学者也在进行相关理论的研究。通过分析总结，话题传播分析技术的发展趋势主要有以下几个方面：

（1）多层次分析的融合：前文已经提到，可以粗略的将论坛话题传播分析技术分为宏观和微观两个层次，宏观层次分析覆盖面广、结果全面，但忽略了传播细节；微观层次分析更加细致入微，但仅仅是对一个论坛的监控，分析结果不够全面。如何将两个层次的研究与分析结合起来，得到一个多层次、多角度更加全面的话题传播分析结果是研究者未来努力的方向。

（2）引入复杂网络理论：目前该研究方向仍然处于仿真实验阶段。许多研究者提出并验证了复杂网络对解决现实问题的重大意义，例如病毒在计算机网络上的蔓延、传染病在人群中的流行、谣言在社会中的扩散等，都可以看作是服从某种规律的网络传播行为。所以，应用小世界、无尺度等网络模型分析话题传播行为仍然具有广阔的发展前景。

（3）跨专业、跨学科的合作：话题传播涉及到人与人之间沟通和交流的行为特性，或者说话题传播是社会关系在论坛这种特定交流平台上的具体体现。那么，要分析话题传播行为就必然涉及社会学、心理学等其他学科，学科间的合作也许会为该研究领域开辟一条全新的道路。

2.5 本章小结

本章首先介绍了话题演变分析技术的研究现状。对目前国内外话题跟踪领域的工作进行了总结，特别对构建动态话题模型和优化向量空间模型的思路进行了详细的论述；同时，介绍了目前话题演变领域的工作进展，重点论述了双质心模型、获得话题来龙去脉的策略以及构造事件框架的思想，为后续研究工作奠定了理论基础。

同时，本章还介绍了话题传播分析技术的研究现状。分别从宏观、微观、复杂网络等角度对现有的研究进行概括和总结，分析了各个研究方向的优势和不足。其中，建立论坛间信息传播路径的方法和梳理口碑传播影响因素的思想对本文的帮助很大。

最后，总结了话题演变和传播分析技术领域的发展趋势，为以后的工作指明了方向。

第 3 章 基于多中心与向量分解的话题演变分析技术

3.1 引言

为了提高话题跟踪算法的召回率、客观展现话题内容演变的过程，本文提出了基于多中心与向量分解的话题演变分析技术。应用多中心结构建立话题模型，每个中心代表话题的一个侧面。采用向量分解方法提取文档中的新特征，并根据新特征的比重维护话题中心。同时，结合增量聚类算法，设计并实现了解决话题演变问题的整体方案。最后，通过实验验证，本方案能够有效提高话题跟踪的性能，在一定程度上展现了话题内容演变的全过程，解决了话题演变问题。

3.2 话题多中心模型

为了实现话题跟踪、清晰呈现话题演变过程，需要建立一个合适的话题模型。该模型必须具备以下两个特性：首先，随着后续文档的加入，模型必须能够动态调整；第二，该模型必须支持话题多侧面的描述，能够动态维护已有的各个侧面，还能够建立新的侧面。在话题间实现分类的基础上，在话题内部还要进行二次的聚类和分类，从而有效解决话题演变问题，展现话题演变过程。基于上述两个特性，本文提出了话题多中心模型，通过多中心的结构描述话题的多个侧面，通过话题中心的更新实现模型的动态调整，通过向量分解思想建立新的话题中心。

3.2.1 话题多中心结构

要建立一个适应话题演变的模型，首先需要探讨话题演变的原理。谈到话题，就不得不引入“事件”这个概念。众所周知，事件是动态的，一个事件的中心是随着事件的发展而不断转换的。同时，话题和事件是不可分割的，没有事件就谈不上话题。话题是人们对某一事件的观点和言论。如上文所述，本文对话题定义如下：一个话题能够叙述新闻事件、反映社情民意，其具体

表现形式为一系列新闻报道、网站评论、论坛帖子和博客日志等的集合。这再一次证明了，话题和事件一样，是动态变化的，话题的中心是随着讨论的深入而不断演变的。这种变化可能是由于所讨论的事件出现新的进展，也可能是由于新观点的加入。正是由于这些主观客观因素的影响，使话题呈现出明显的动态演变的特征。下面具体研究话题演变的原理。

首先，在话题刚刚建立的时候，就无形中形成了一个中心，即初始中心，它相当于话题演变的“种子”。有些话题在种子阶段就逐渐消亡了，只形成了一个初始中心，没有演化出新的中心。而大多数话题都要发生演变，话题中心都要随着事件的发展而不断变化，由一个初始中心开始，不断生成新的中心，每个中心代表某个阶段话题讨论的侧重点。以一次沉船事件为例：最初对于事件的报道主要是关于时间、地点、人员伤亡情况等；随着对沉船事件的深入调查，讨论的核心内容集中于调查事故发生的原因；接下来，报道内容侧重于政府对遇难者家属的赔偿、善后措施，以及后续的相关工作等。在上述例子中，话题重点讨论了三个中心，即海难概况、原因调查和善后事宜。每个中心代表本次事故的一个内容重心，它们都是组成这次事故的重要方面。因此，在本例中采用三个中心的话题模型可以全面表示这次沉船事故。综上所述，本文使用多中心结构建立话题模型，中心的个数没有限定，完全根据话题演变的情况动态调整，话题演变出几个内容侧面，模型就建立几个中心，在结构上同话题演变的结果保持一致，如图 3.1 所示。

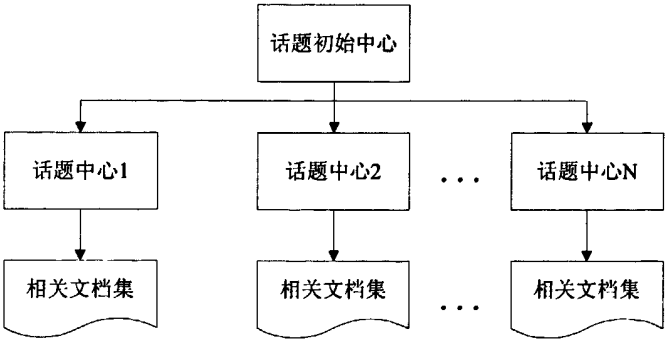

图 3.1 话题多中心结构示意图

该话题模型最大的特点是，用多个中心完全还原了话题多侧面的逻辑结构。更重要的是，它能够并行维护和更新话题的每一个中心。这里“并行”

这个概念十分重要，也就是说，不管后续文档到来的次序如何，本模型都能够一一对它们进行分析，确定它们属于哪个中心，或是用该文档建立新的中心。并不是所有的模型都具有“并行”这个特性。比如第二章中提到的“双质心”话题模型，它实际上是一种“双中心”结构，模型可以动态维护两个中心，具体实现的策略如下：初始质心代表话题在分界点之前关注的内容，当前质心表示话题从分界点到当前时刻所关注的内容。当新的分界点出现时，已有的初始质心和当前质心合并成新的初始质心，促使分界点建立的文档形成新的当前质心。如图 3.2 所示，文档 6 到来后需要建立分界点，原来的两个中心合并，即文档 1 到 5 合并成一个初始质心，而文档 6 形成新的当前质心。该模型可以在一定程度上再现话题演变的过程，但最大的问题就是不支持“并行”的概念。当前质心是话题最新讨论的内容，而初始质心中混杂了所有的历史数据，也就是说初始质心是由历史上多个当前质心合并得到的。当文档到来的次序和中心建立次序完全一致时，该模型没有问题；而当文档“乱序”到来时，由于历史数据被混杂进初始质心，模型不能对文档归属进行正确的分析和判断。后面的实验也验证了这一点。综上所述，多中心结构是对话题多侧面特征最大限度的重现，建立多中心模型是解决话题演变问题的有效方法。

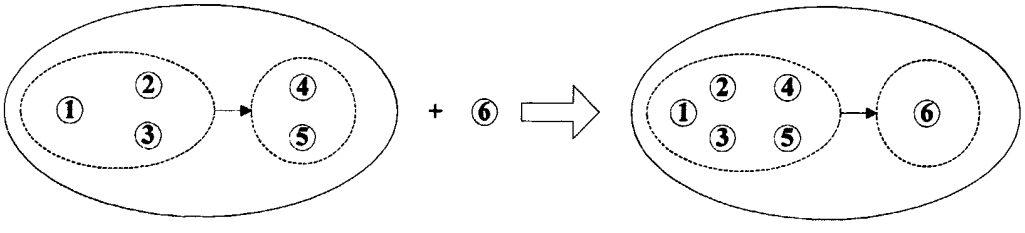

图 3.2 双质心话题模型示意图

3.2.2 话题中心的表示和更新

首先需要明确“中心”的概念。中心是对若干文本的概括和总结。为文本集提炼中心的目的是降低算法的开销。由于中心可以近似代替多个文本，在计算一个文本和一个文本集的相似度的时候，只需要计算该文本与文本集

中心的相似度即可，实现简单又不影响效果，降低了算法的时间复杂度。同时，在算法运行期间，只需要保存中心数据，而不必保存所有的文本，从而减少了程序对内存的占用，降低了算法的空间复杂度。可见，提取中心是十分必要的。对中心的应用也比较灵活，可以提炼话题的中心，也可以提炼话题中各个侧面的中心。本文采用的是第二个策略。下面详细论述如何提炼中心，以及如何动态更新中心。

曾经有文献提出，提取文本集中最具代表性的一篇文本作为中心，策略如下：首先，需要对文本集进行去重，对于那些内容基本相同的文本，只保留其中的一篇，其余的全部删除；然后，在剩下的文本集中，选取和其他文本距离最为平均的一篇文本作为中心，这里的距离也就是内容上的相似度。该方案实现简单，支持中心的动态调整；但用一篇原始文本表示一个文本集的内容大意，准确性有待考证。本文采取的方案是，使用文本集中的每一篇文本通过计算得出中心。该中心不是任何一篇原始文本，但又与每一篇文本相关，这才是真正意义上的中心。如图 3.3 所示：大圆表示一个文本集，小圆表示文本集中的文本，灰色小圆是提取出的中心，小圆之间的距离代表它们内容上的相似度，距离越小表示文本内容越接近。左图是第一种方案，文本 4 和其他文本距离最为平均，选取它为文本集的中心；右图是本文采取的策略，通过对七篇文本的计算得出中心，即右图中灰色的小圆。下面具体介绍中心的表示和更新方法。

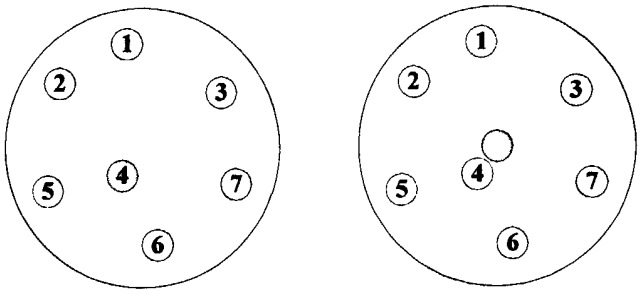

图 3.3 两种中心表示方法示意图

在本文中，中心的建立和更新过程是融合在一起的。当话题的一个侧面刚刚建立时，很显然只有一篇文档，它就是该侧面的第一个中心。第二篇文

档到来后，与第一个中心进行计算，更新原有中心，建立新的中心，后续过程以此类推。这里需要说明的是，文档和中心都使用传统的 VSM 向量空间模型表示，权值计算方法采用传统的 TF-IDF 公式，如公式（3-1）所示。

$$W_i=\frac{TF_i(t,d)\log(\frac{N}{DF(t)}+0.01)}{\sqrt{\sum_k TF_i^2(t,d)\log^2(\frac{N}{DF(t)}+0.01)}} \tag{3-1}$$

更新中心的方法是计算当前文档向量与中心向量的和，具体策略如下：假设 v_{di} 和 w_{di} 分别表示当前文档向量的特征项及相应权值，则当前文档向量 V_d 可表示为 $(v_{d1}, w_{d1}; v_{d2}, w_{d2}; \ldots; v_{dn}, w_{dn})$，同理中心向量 V_c 可表示为 $(v_{c1}, w_{c1}; v_{c2}, w_{c2}; \ldots; v_{cn}, w_{cn})$，则它们的和 SUM $(V_d, V_c) = (v_{s1}, w_{s1}; v_{s2}, w_{s2}; \ldots; v_{sl}, w_{sl})$ 的计算方法如下所示[15]。

$$w_{sr}=\begin{cases} w(v_{sr},V_d)+w(v_{sr},V_c), 若 v_{sr}\in F(V_d)\cap F(V_c) \\ w(v_{sr},V_d), 若 v_{sr}\in F(V_d)-F(V_d)\cap F(V_c) \\ w(v_{sr},V_c), 若 v_{sr}\in F(V_c)-F(V_d)\cap F(V_c) \end{cases} \tag{3-2}$$

其中 l 为相加后的特征数目。$F(V)$ 表示 V 的特征项集合，$w(v_{sr},V)$ 表示 v_{sr} 在 V 中的权值。如公式(3-2)所示，计算两个向量和的方法是：对于文档向量与中心向量的共有特征，取两个特征权值的和作为新特征的权值；对于只在文档向量或者中心向量中出现的向量特征，保留该特征的权值。该方案综合了已有中心和当前文档的文本特征，是一种有效的表示和更新话题中心的方法。

3.2.3 向量分解思想

本文力图通过呈现话题内容中心转换的过程来解决话题演变问题。首要的问题就是研究和分析话题内容发生演变的原理。如上文所述，文档和中心都使用传统的 VSM 向量空间模型表示，其中向量的每一维就是文档特征，也就是词。本文分词算法采用的是中科院的中文分词系统 ICTCLAS，是一种基于 N-最短路径方法的中文词语粗分模型。所以，要弄清话题演变的原理，就必须从一篇文本最基本的特征——“词”来入手。以“空难”这一话题为例：该话题生成后，可能会在以下几个方向上发生演变，“现场状况”、“搜救

过程”、“事故调查”、“善后事宜”等。但是，不管该话题演变出多少个侧面，各个侧面一定会保留若干共有的特征以标识这些侧面是属于同一话题的。如表 3.1 所示。

表 3.1 空难话题演变概况

序号	话题侧面	共有特征	相异特征
1	现场状况	空难、飞机失事…	目击、组图、现场、证人…
2	搜救过程	空难、飞机失事…	救援、搜救、打捞、进展…
3	事故调查	空难、飞机失事…	分析、调查、查找、排除…
4	善后事宜	空难、飞机失事…	赔偿、保险、遗体、善后…

如上表所示，话题演变的本质是新特征的出现。也就是说，有些特征在话题开始阶段并未出现，而是经过一段时间后才出现，那么这些特征很可能代表新侧面的出现，即话题演变的发生。本文对“共有特征”和“相异特征”定义如下：

定义 3.1 话题演变过程中，在话题各个侧面中都出现、用来标识该话题基本语义信息的特征称为共有特征。

定义 3.2 在一个话题中，除去共有特征，用来标识各个侧面新颖性的特征称为相异特征。

然而，仅仅根据少数几个新特征的出现还不能判定话题发生了演化，只有当出现的新特征数量达到一定规模时，才有可能发生演化。基于这种思想，本文提出采用向量分解法建立话题多中心结构模型。向量分解方法采用 VSM 模型表示报道和话题，采用夹角余弦公式计算相似度，如公式（3-3）所示。并且，计算报道与话题各个中心相异特征数量，相异特征是一个相对的概念，指报道相对于某个中心的新特征，与话题不同的中心计算出的相异特征可能不相同。然后，需要计算相异特征百分比，目的是衡量文档中出现相异特征的规模，具体的计算方法是：计算报道中新特征的数量与该报道特征总数的比值，如公式（3-4）所示。需要说明的是，计算报道与话题相似度用于判断报道是否属于话题类，而相异特征百分比用于判断报道讨论的话题中心。

$$Sim(D,T)=\frac{\sum_{i\in H} q_i d_i}{\sqrt{(\sum_{i\in H} q_i^2)(\sum_{i\in H} d_i^2)}} \tag{3-3}$$

$$相异特征百分比=\frac{文档D新特征的数量}{文档D特征总数} \tag{3-4}$$

综上所述，在话题多中心结构中，仅仅判断出文档所属的话题类是不完整的，还需要判断报道讨论的中心是哪个。基于上面的向量分解思想，本文提出了计算文档所属中心的方法：在判断文档讨论的中心时，相异特征数量越少，则文档越可能在讨论该中心；反之，相异特征数量越多，则越有可能在讨论不同的中心。判断文档所属中心算法如图 3.4 所示。

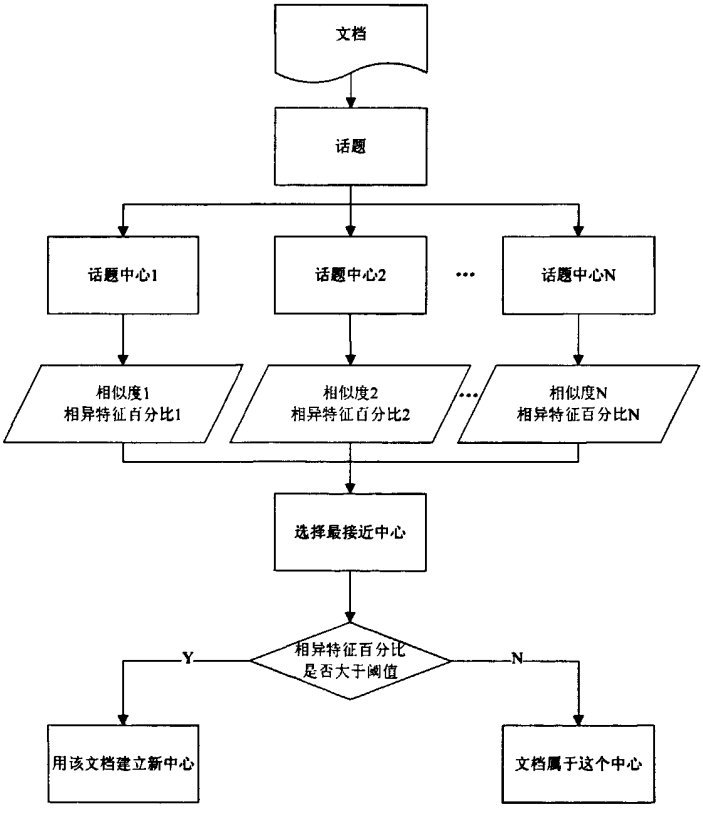

图 3.4 计算文档所属中心算法流程示意图

判断文档所属中心算法具体流程如下：

（1）计算文档与该话题所有中心的相似度和相异特征百分比。

（2）从话题各个中心中选择相似度最大的作为与文档最接近的中心。

（3）如果文档与最接近中心的相异特征百分比小于阈值，则将文档归入该中心，算法结束；否则，转向（4）。

（4）以该文档建立话题新的中心，算法结束。

3.3 话题演变分析算法

3.3.1 增量聚类算法

聚类算法是数据挖掘领域最常见的技术之一，用于发现数据集中未知的类对象。通过聚类形成的每一个组称为一个类。一般来说，聚类算法大致可以分为以下几种：分割聚类算法、层次聚类算法、基于密度的聚类算法、基于网络的聚类算法和基于模型的聚类算法等。这些聚类方法都具有各自的优点和不足，适用于不同的应用背景。本文基于网络舆情数据，力图通过一种有效的数据挖掘算法将杂乱的文档集归并成若干个话题。显而易见，这里的数据集不是静态的，而是动态激增的，实际上是一种数据流处理的概念。那么，选择什么样的聚类算法来实现对动态数据集的分析和处理呢？目前主要的策略有两种，一种是在新的数据集上重新计算的方法，另一种是针对更新数据采用增量式的聚类算法，如表 3.2 所示。

表 3.2 针对动态数据集的聚类算法比较

聚类策略	优点	缺点
重新聚类	实现简单，重复计算即可。	开销大，随着数据量的增加，算法的时间和空间复杂度都相应提高；同时，重复执行也是计算上的浪费。
增量聚类	开销相对较小，由于利用了已有的计算结果，算法的时间和空间复杂度不随数据量的增加而线性提高。	由于是一种类似于数据流的操作，所以数据输入的顺序必将影响算法最终的结果。

上表论述了重新聚类和增量聚类两种策略的优点和不足。显然，增量聚类算法更适用于流数据的分析和处理。上文提到，本文主要是基于网络上的

实际数据，众所周知网站的网页都是实时更新的，源数据的特性要求所采用的算法也具有“实时更新”的特征；同时，从算法开销的角度考虑，也迫切需要一种复杂度与数据量不直接相关的数据挖掘算法。综上所述，本文采用传统的 Single-pass 增量聚类算法来实现话题检测和跟踪，并在此基础上解决话题演变问题。该算法大致思想如下：算法顺序处理输入的每一篇文档，初始以第一篇文档为种子创建第一个类簇，对于每一篇新输入的文档，与前面生成的所有类簇进行相似度比较，如果该文档与某个类簇的相似度大于聚类阈值，则将该文档归入该类簇；否则，以该文档为种子创建一个新的主题类簇。

3.3.2 话题演变分析技术整体流程

为了清晰呈现话题演变过程，本文提出了基于多中心和向量分解的话题演变分析技术。首先，通过构建多中心话题模型准确描述话题多侧面的特征，通过向量分解的思想获取文档新颖特征、判断文档所属中心；最后，结合 Single-pass 增量聚类算法提出一套完整的解决话题演变问题的方案。以下是算法的整体流程：

（1）若当前文档是数据流中的第一个，则建立一个以该文档为初始中心的类，作为第一个话题类。继续处理下一篇文档。

（2）计算新文档与已有各个话题的相似度，并记录最大相似度及其对应话题类。计算文档和话题相似度时，应计算文档与话题每个中心的相似度，并将最大的相似度作为文档与话题最终的相似度。

（3）若最大的相似度小于创新阈值，则建立一个新话题类，同时该文档为话题初始中心，继续处理下一篇文档，转向（1）。

（4）若最大的相似度大于创新阈值，文档归入最大相似度对应的话题类。应用上文提到的算法判断文档所属话题中心。如果建立新的中心，则直接转向（1）；如果文档属于已有的中心，则转向（5）。

（5）更新中心向量：每当有新文档加入话题中心时更新相应的中心向量。继续处理下一篇文档，转向（1）。

图 3.5 是具体的算法流程图。该方案实现了基本的话题检测与跟踪功能，能够动态维护一系列话题类，每个类都采用多中心话题模型，通过建立和更

新话题中心的过程来描述话题内容重心的转换，进而在一定程度上呈现了话题演变的过程。

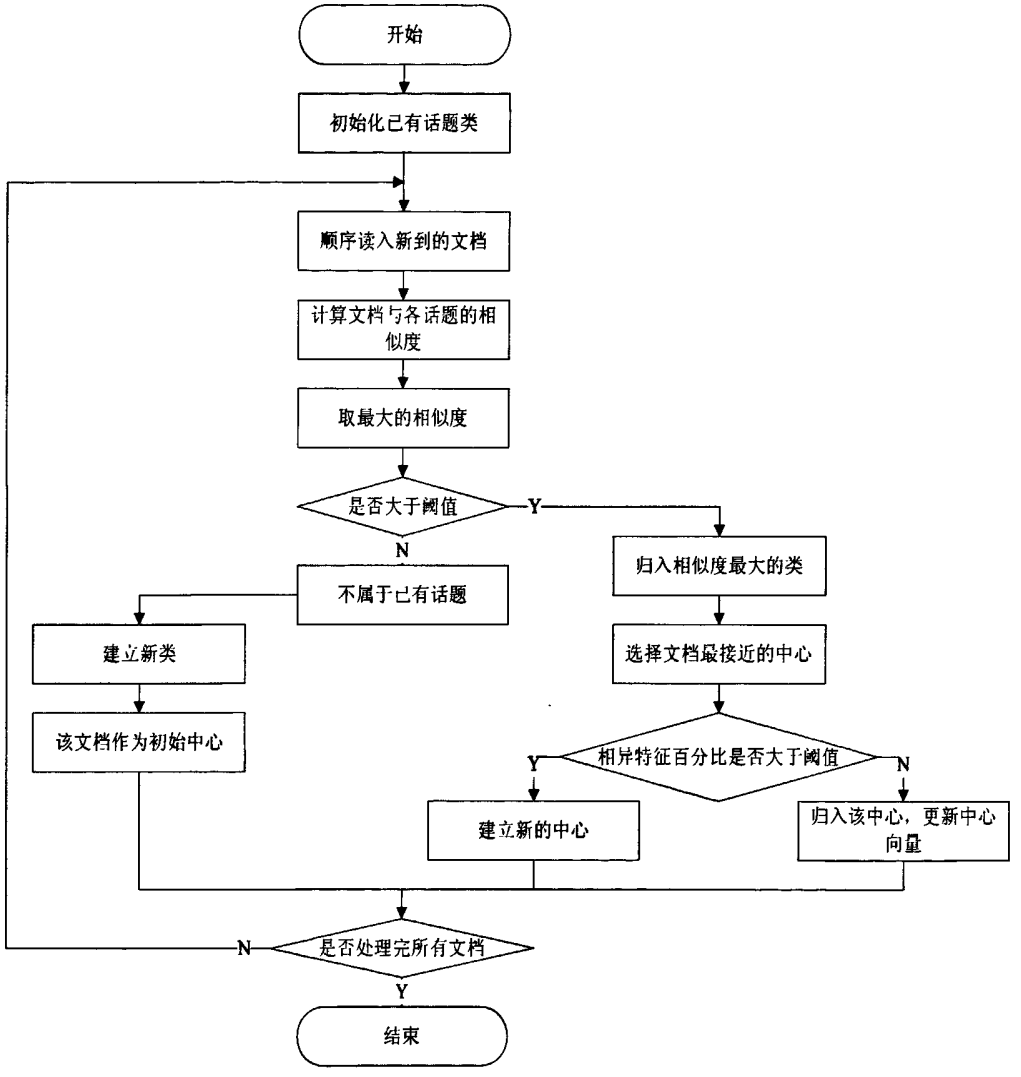

图 3.5 话题演变分析算法流程图

在传统的话题跟踪任务中，通常是将一个话题的文档归并为一类，所有属于该话题的文档按获取时间顺序组织。随着话题规模的不断扩大，这种组织方式不方便用户浏览，更无法展现话题演变的过程。本文基于上述话题演变分析方法，提出用“话题—中心—文档”的形式呈现多层次的话题特征。

其中话题中心层次的变化体现了话题内容上的演变，为浏览者全面了解话题的来龙去脉提供了便利的条件。

3.4 实验结果与分析

本实验的目的是验证基于多中心和向量分解的话题演变分析技术的可行性和实际效果，重点验证多中心模型能否准确刻画话题的多个侧面，以及向量分解方法能否准确地判断文档所属中心。实验软件环境包括：C、C++语言编写的实验程序、Linux 操作系统以及 Mysql 数据库系统；实验硬件设备为高性能服务器一台。下面介绍具体的实验过程。

首先，选择网络上的五个热点话题，分别是“西藏当雄地震”、“杭州地铁工地坍塌”、“乌鲁木齐商厦大火”、“三聚氰胺问题鸡蛋”、“山西黑砖厂虐待工人事件”。然后，分别获取各个话题对应的文档集。在网页预处理后，得到文档与 vsm_id 的对应关系，后续实验结果均用该 vsm_id 标注各篇文档，如表 3.3 所示。

表 3.3 话题演变实验源数据表

话题序号	话题内容	文档个数	文档 vsm_id
1	西藏当雄地震	26	1—27
2	杭州地铁工地坍塌	32	28—59
3	乌鲁木齐商厦大火	36	60—95
4	三聚氰胺问题鸡蛋	41	96—137
5	山西黑砖厂虐待工人事件	44	138—181

然后，基于上述五个话题的文档数据，应用本文提出的基于多中心和向量分解的话题演变解决方案进行分析和处理。为了便于比较算法性能和效果，本实验还实现了“双质心”话题模型，分别比较了两个算法的准确率、召回率以及 F1 值，具体计算方法如下面公式所示。

$$准确率=\frac{算法发现的相关文档数}{算法获取的文档总数}\times 100\% \qquad (3\text{-}5)$$

$$召回率=\frac{算法发现的相关文档数}{相关文档总数}\times 100\% \tag{3-6}$$

$$F1值=\frac{准确率\times 召回率\times 2}{准确率+召回率} \tag{3-7}$$

实验一：算法性能比较

首先，应用多中心模型进行话题检测跟踪。该实验有两个参数需要调整：相似度阈值和相异特征百分比阈值。相似度阈值用来判断一篇文档属于哪个话题，相异特征百分比用来决定文档所属中心。通过实验验证，当相似度阈值取 0.4、相异特征百分比取 0.6 时，多中心算法效果较好，如表 3.4 所示。

表 3.4 多中心算法性能

话题序号	文档总数	检测到的文档数	正确文档数	召回率	准确率	F1 值
1	26	24	24	0.923	1	0.960
2	32	31	31	0.969	1	0.984
3	36	34	33	0.944	0.971	0.957
4	41	36	36	0.878	1	0.935
5	44	23	18	0.522	0.783	0.626

表 3.5 双质心算法性能

话题序号	文档总数	检测到的文档数	正确文档数	召回率	准确率	F1 值
1	26	26	26	1	1	1
2	32	29	29	0.906	1	0.951
3	36	33	33	0.917	1	0.956
4	41	37	37	0.902	1	0.949
5	44	19	19	0.432	1	0.603

然后，应用双质心模型进行话题检测。该实验同样需要调整相似度阈值

这个参数，分别选取相似度阈值为 0.4、0.5 和 0.6。从实验结果可知，在相似度阈值为 0.4 时，话题检测性能指标较高。随着阈值的增加，性能指标有所下降，当阈值为 0.6 时，准确率、召回率很低。阈值取 0.4 时双质心算法的实验结果如表 3.5 所示。最后，分别比较多中心算法和双质心算法的准确率、召回率以及 F1 值，如表 3.6 所示。

表 3.6 多中心算法和双质心算法性能比较结果

话题序号	多中心算法			双质心算法		
	召回率	准确率	F1 值	召回率	准确率	F1 值
1	0.923	1	0.960	1	1	1
2	0.969	1	0.984	0.906	1	0.951
3	0.944	0.971	0.957	0.917	1	0.956
4	0.878	1	0.935	0.902	1	0.949
5	0.522	0.783	0.626	0.432	1	0.603

如上表所示，两种话题检测与跟踪算法，对多数话题检测的准确率都为 100%；而在召回率和 F1 值上存在一定的差距：在检测第一类话题和第四类话题时，双质心算法性能优于多中心算法；而在检测第二类、第三类和第五类话题时，多中心算法性能优于双质心算法。综合分析，多中心算法在准确率上略逊于双质心算法，而在召回率和 F1 值上都优于双质心算法。综上，本文提出的多中心模型在完成话题检测与跟踪任务上取得了一定的成果。

实验二：多中心算法呈现话题演变过程

本实验的目的是验证多中心模型在呈现话题多侧面方面的效果。如果相异特征百分比阈值设置过低，会导致建立过多的中心，每篇报道自成一个中心；设置过高，会使话题类的中心较少，可能整个话题只有一个中心。通过实验调整，阈值设置为 0.6 时效果较好。下面是多中心话题检测结果，每个话题内部建立了多个中心，属于同一中心的文章所讨论的内容相关性较大；而不同中心之间存在一定差别，如下面各表所示。

表 3.7 西藏当雄地震话题演变分析结果

中心序号	中心文档	中心内容摘要
1	1、13、15	中国地震台网地震通报
2	2	人员伤亡和财产损失情况
3	3、5	地震对青藏铁路以及拉萨机场的影响
4	4	地震对各个文物点的影响
5	7-9、10、12、14、17-20、21、22、24、26	启动三级应急响应，成立恢复重建工作领导小组
6	11	做好市场供应，确保物资供应充足
7	25	截至日前，已紧急转移受灾群众 4828 人
8	27	灾区急需物资源源不断地运往灾区

表 3.8 杭州地铁工地塌陷话题演变分析结果

中心序号	中心文档	中心内容摘要
1	28	今天下午 14：30 左右事故发生
2	29	现场情况复杂，救援工作一度被延迟
3	30、32、33、42、44-46、49、53、55、57	目前已造成多人伤亡，救援工作进行中
4	31	杭州市委书记王国平、杭州市市长蔡奇指挥救援
5	34、38-41、43、50、51、54、56、58、59	要求加强抢救工作，同时查明事故原因，落实责任
6	35、36	救援工作
7	37	救援工作
8	48	救援工作
9	52	事故原因分析

表 3.9 乌鲁木齐商厦大火话题演变分析结果

中心序号	中心文档	中心内容摘要
1	60	乌鲁木齐市区一栋 12 层商铺二日晚间突发火灾

（续表）

2	62	消防队员正在火灾现场救火
3	63、90	火灾目前仍未扑灭
4	64、86、88、89、91-93	3 名消防官兵在搜救被困群众过程中不幸牺牲
5	65、80、87、94、95	火灾事故调查、善后处理工作有序进行
6	66	消防队员火灾现场救火
7	67	消防队员火灾现场救火
8	68-70，73，75-79，81-85	火灾原因查明，火灾现场情况复杂，易燃可燃物多，有毒气体弥漫
9	71	乌鲁木齐市区又发生了两起小火灾
10	74	成立火灾调查和善后处理工作领导小组

表 3.10 三聚氰胺鸡蛋话题演变分析结果

中心序号	中心文档	中心内容摘要
1	96	香港检出含三聚氰胺鸡蛋
2	98	北京暂未发现含三聚氰胺鸡蛋
3	99、119、125	市场反应及部门说法
4	102、107、109	香港检出产自湖北的鸡蛋含三聚氰胺超标
5	103	重庆暂未发现三聚氰胺鸡蛋
6	105、110、113-116、120、123、124、126、128、135	湖北调查组赴京查三聚氰胺鸡蛋事件
7	106	杭州检出慈云祥牌鸡蛋含三聚氰胺
8	108	海口市场未发现问题鸡蛋
9	111、127、129	韩伟集团的鸡蛋被检测出含三聚氰胺
10	112、117、118、121、122、130-133、137	事件调查：饲料中添加三聚氰胺

表 3.11 山西黑砖窑事件话题演变分析结果

中心序号	中心文档	中心内容摘要
1	138、150-163、165、167、169、173、176、177、180	中央对黑砖窑事件进行调查
2	141	非法拘禁 32 名农民工，并强迫其从事无偿劳动

如上表所示，每个话题都被分为多个中心，各个中心在不偏离话题主题的前提下，在内容上都有所侧重，各个侧面具有较为明显的区分度，从一定程度上呈现了话题多侧面的特征，描述了话题演变的过程。

实验三：算法对文档输入顺序依赖性的比较

多中心话题模型最大的特点是能够并行维护多个中心，应用该模型进行话题检测，结果基本不受文档输入顺序的影响。这里所谓的顺序，是指文档输入顺序与话题中心生成顺序保持一致。同时，该特点正是双质心模型的不足之处。因此，本文设计如下实验：对于多中心模型和双质心模型，分别使用顺序输入和乱序输入的文档集进行话题检测和跟踪，比较两个算法对文档输入顺序的依赖性。

首先，从“乌鲁木齐商厦大火”话题类中选取 23 篇文档，并事先人为将其分为三个内容中心，各个中心的摘要如下：“三名消防员遇难”、“火灾事故调查和善后事宜”以及“分析总结事故原因”，这是对该话题进行话题演变分析的默认正确结果，如表 3.12 所示。然后，分别采用“顺序”和“乱序”两种输入策略进行话题检测，比较两个算法的性能。顺序输入是按照文档序号从 1 到 23 逐个输入；乱序输入次序为：1 到 10、13 到 16、11、12、17 到 23。最后，将多中心和双质心算法两种输入次序的分析结果与事先分类的结果进行对比，如表 3.13 所示。

表 3.12 话题乌鲁木齐商厦大火人工处理结果

中心序号	中心文档	中心内容摘要
1	1-7	三名消防员遇难

（续表）

2	8-12	火灾事故调查和善后事宜
3	13-23	分析总结事故原因

表 3.13 算法对文档输入顺序依赖性比较结果

中心序号	顺序输入文档处理结果		乱序输入文档处理结果	
	多中心算法	双质心算法	多中心算法	双质心算法
1	4-7	1-7	4-7	1-7
2	8-12	8-13	8-13	8-10、13-16
3	13-23	14-19	14、16-23	11、12、17-19

如上表所示：在处理顺序到来的文档时，多中心和双质心两种算法的效果没有明显差异；而在处理乱序到来的文档时，两种算法表现出性能上的差距。多中心算法基本不受文档输入顺序的影响，只有文档 13 的归类结果发生了变化；而观察双质心算法的分析结果，序号 8 到 12 的文档本属于同一个中心，应归入中心 2，双质心模型未能正确归类；文档 13 到 16 本属于中心 3，而双质心模型将其归入中心 2。综上所述，本文提出的基于多中心和向量分解的话题演变分析技术对文档输入顺序不敏感，能够同时支持多个中心的建立和更新，呈现话题内容演变的全过程。

最后，对上述实验结果进行分析和总结。通过准确率和召回率等评价指标比较多中心与双质心话题检测方案的性能，两种算法基本相当。只是在召回率和 F1 值上，多中心算法略优于双质心算法。可见，本文提出的多中心模型能够完成话题检测任务。同时，在呈现话题演变过程方面，多中心话题检测算法明显优于双质心算法。该算法不但能发现话题新出现的内容，还能对旧的内容进行二次归类。原因是多中心模型始终保存话题所有的中心，并及时对各个中心进行更新。而双质心模型仅依据新词的出现建立分界点，只适用于话题各个侧面顺序出现的情况。主要是由于双质心模型只关注话题当前内容的焦点，对历史话题焦点仅做简单的合并。双质心模型的这个特点决定了其对文档输入顺序的依赖性。

3.5 本章小结

本章提出基于多中心和向量分解的话题演变分析技术，力图在实现网络话题检测与跟踪的基础上，呈现话题演变过程。

提出多中心话题模型。通过多中心结构描述话题多侧面的特征，详细论述中心表示和更新的方法，重点解决了应用向量分解思想判定文档所属中心的问题。同时，结合增量聚类算法，提出了一套完整的分析话题演变过程的解决方案。

最后，通过实验与双质心算法进行对比，重点在话题检测跟踪性能和话题演变过程呈现两个方面对上述方案进行了验证。

第 4 章 基于传播图与多元线性回归的话题传播分析技术

4.1 引言

针对目前日益严峻的网络舆情态势，本文提出了基于传播图与多元线性回归的话题传播分析技术，力图在宏观和微观两个层次分析和解决话题传播问题。其中，宏观层次是指论坛间的分析，微观层次是指论坛内部的分析。在论坛间，试图挖掘话题的传播路径，建立论坛话题传播图，并量化每个论坛对该话题传播所做的贡献；在论坛内部，研究话题传播的行为和机制，建立一套影响话题传播的指标体系，并结合多元线性回归理论客观描述传播行为、准确预测传播趋势。

4.2 论坛间话题传播分析

对于论坛间的话题传播行为，本文提出了基于传播图的分析技术。本节详细论述了建立论坛话题传播图的方法和策略：通过比较发布时间发掘初始传播论坛，基于关键词匹配和相似度比较的转帖关系发现策略，基于帖子发布时间和更新时间的论坛传播生命期计算方法，以及基于传播图出度的论坛传播权值计算方案。需要说明的是，本文提出的传播图是根据一个话题在论坛间的传播行为构建出来的。

4.2.1 论坛转帖关系挖掘

要建立论坛话题传播图，首先需要发掘初始传播论坛。所谓初始传播论坛，就是最先发布一个话题的论坛，即该话题发起传播的起点。显然这是与时间相关的概念。统计该话题下所有帖子的发布时间，选择发布时间最早的帖子所在的论坛，就是初始传播论坛。这里需要说明的是，本文分析的数据来源于网络爬虫下载的论坛数据，将各大论坛的网址作为爬虫下载的种子。显然，爬虫下载的论坛数量是有限的，不可能实现对全网所有论坛的监控。

那么，初始传播论坛发现的准确性就必然依赖于爬虫的监控范围。本文发现的初始传播论坛是监控范围内各个论坛的传播起点，可能并不是真实网络环境下的传播起点。

论坛话题传播图将各个论坛抽象成图的结点，而连接结点的边就是话题在论坛间的转帖关系。通过合理组织论坛数据可以很容易地得到图的结点，那么要构建论坛话题传播图，对转帖关系的挖掘就变得尤为重要了。本文对转帖的定义如下。

定义 4.1 内容大致相同、在不同时间发布在不同论坛上的一系列帖子，称为转帖。

本文提出了基于关键词匹配和相似度比较的转帖关系发现策略：首先，统计该话题下所有帖子的主帖文本，计算任意两篇文本的相似度，如果帖子 A 和帖子 B 的相似度值超过一个很大的阈值（如 0.9）。这里的相似度计算是应用传统的基于 VSM 向量空间模型的夹角余弦公式。再判断这两篇帖子的发布论坛地址，如果它们属于同一论坛，那一定不是转帖关系；如果它们不属于同一个论坛，则很可能是一组转帖关系。假设帖子 A 的发布时间早于帖子 B，那么帖子 A 就是帖子 B 的“候选源帖”。这里需要解决的问题是：如果单纯根据文本相似度和发布时间发掘转帖关系，一个帖子很可能有多个“候选源帖”，既不符合现实情况，又产生了大量的冗余数据。需要对多个转帖关系进行筛选，在一个帖子的多个候选源帖中选择一个可能性最大的，然后删除其余转帖关系。这里使用的策略如下：假设帖子 A_1、A_2…A_n 是帖子 B 的多个候选源帖，现在需要从中选出帖子 B 真正的源帖。本文提出的转帖关系挖掘算法流程如下。

（1）基于关键词匹配的策略：根据候选源帖发布论坛的不同，每个候选源帖都对应着一个关键词库。例如，帖子 A_1 发布在天涯社区，其关键词就有“天涯社区”、“tianya.cn”等；帖子 A_2 发布在网易论坛，其关键词就有“网易”、“163.com”等。在帖子 B 中逐个匹配各个候选源帖的关键词，如果匹配成功，则确定该候选源帖就是帖子 B 真正的源帖，算法结束。如果没有匹配到关键词，或者有多篇匹配成功，则转向（2）。

（2）基于相似度比较的策略：对于各个候选原贴，比较它们与帖子 B 的文本相似度。取相似度最大的帖子作为帖子 B 真正的源帖，算法结束。图

4.1 为该方案的流程图。

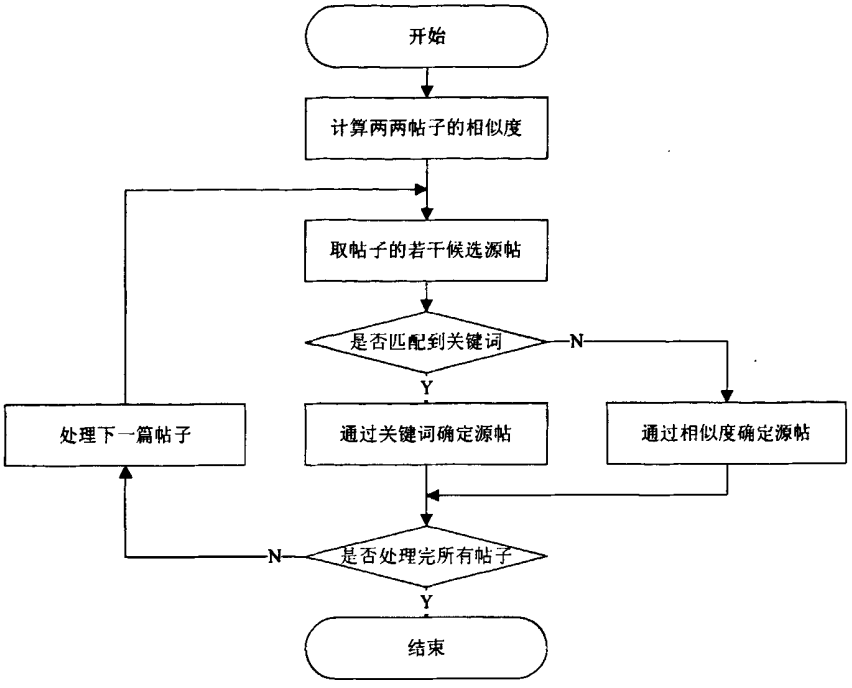

图 4.1 转帖关系挖掘算法流程图

4.2.2 论坛话题传播图

本文提出了基于传播图的论坛间话题传播分析技术，该传播图在描述话题传播行为的基础上，试图在传播时间和传播贡献上对各个论坛进行分析和标注。上文对转帖关系的发掘实际上就是对话题传播行为的分析，下面将着重论述本文在论坛传播时间和贡献上的分析方法。

在传播时间方面，需要计算某特定话题在一个论坛上的传播生命期。针对一个话题，将发布在同一论坛上的帖子搜集起来，可以得到一个文档集。在该文档集中，最早发布帖子的发布时间称为论坛的起始传播时间；最靠近现在的更新时间称为论坛的最近传播时间，二者之间的时间间隔就是该话题在一个论坛上的传播生命期。

同时，由于每个论坛对话题传播所做的贡献不同，本文用“传播权值”来描述这种贡献的大小。由于传播是一种扩散的行为，在方向上是向外的。所以，针对传播图本文认为：对于每一个论坛来说，在传播图中的出度越大，

其传播权值就越高。在图论中，结点 P 的出度是以 P 为前驱的结点数。具体的计算方法是：统计传播图中每个论坛结点的出度，然后对各个出度值进行归一化处理，最终得到各个论坛的传播权值。该传播权值都介于 0 到 1 之间，便于数据的比较和统计。

这样，建立论坛话题传播图所需要的基本要素已经齐全，具体包括：行为信息——话题转帖关系、时间信息——话题论坛传播生命期，以及地域信息——话题发布论坛地址。本文构造论坛话题传播图的策略是：基于传统的二维坐标模式，用横轴代表时间信息、纵轴代表地域信息，用论坛间的有向边代表话题传播路径，并且为每个论坛标注传播权值。图 4.2 是某一话题在六个知名论坛上的传播图。

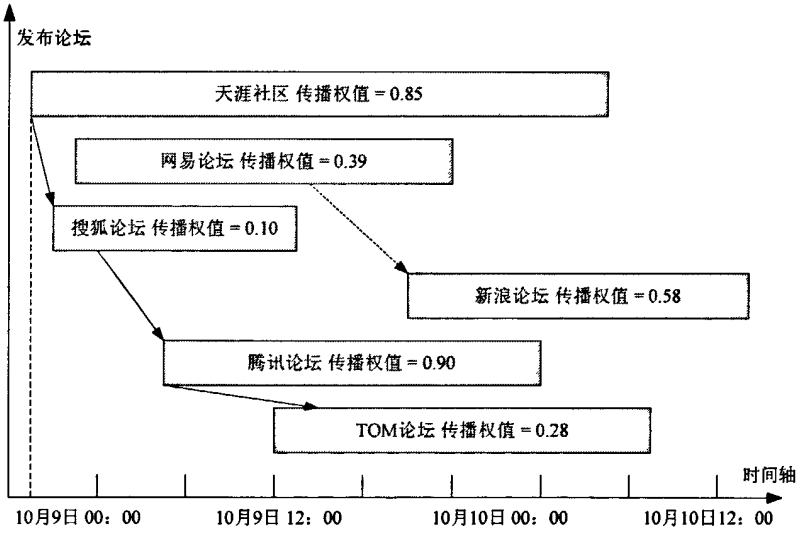

图 4.2 论坛话题传播示意图

由上图可以得到以下信息：该话题的初始传播论坛是天涯社区；同样的箭头代表同一组转帖引发的传播行为，发布在天涯社区的第一篇帖子被依次转发给搜狐、腾讯和 TOM 论坛，引起上述三个论坛对该话题的讨论；在网易论坛的传播生命期期间，一篇帖子被转发到新浪论坛，并引发了该论坛的讨论；方块的长度代表论坛的传播生命期，传播权值描述该论坛对话题传播所做的贡献。综上所述，该图分别从行为、时间、地域和贡献四个角度综合

描述了一个话题在多个论坛间的传播特征。

4.3 论坛内话题传播分析

本文提出了基于多元线性回归的论坛内话题传播分析技术。首先，分析了论坛内部话题传播的机制，提出一套影响传播行为的指标体系；同时，定义了一系列的变量和公式用来描述论坛内的传播行为；最后，结合多元线性回归理论，总结出预测论坛内话题传播趋势的完整方案。

4.3.1 传播指数定义

话题传播是一个话题在内容上没有大的演变的前提下，在时间空间上的发展和扩散，主要是一种基于行为的分析与处理。下面分析一个论坛内部话题传播的机制：话题定义为讨论或交谈的主题，它严格来说是一个抽象的概念，在论坛中它表现为一系列谈论该话题的帖子。一个话题就是一系列帖子的集合。那么，话题传播其实就是帖子传播。一篇帖子是如何把信息传播给其他人的呢？很显然，帖子通过吸引更多的人进行访问和回复来完成信息的交换和传递。所以说，在一个论坛中，话题传播实际上就是论坛用户访问和回复一篇或多篇帖子的具体行为。

下面设计一个虚拟的论坛，该论坛上新发布了一篇由用户 P 编辑的关于话题 A 的帖子 a，用户 Q 看到了这篇帖子在列表页的一系列信息，包括：主题、作者、访问数、回复数和更新时间等。Q 有几种可能的行为：访问帖子 a 不回复、访问帖子 a 并回复、在访问和回复帖子 a 后再写一篇新的帖子 b 继续该话题、对帖子 a 不予关注等。前三种行为都是对话题 A 或者说帖子 a 的传播，而第四种行为没有发生传播，至少对用户 Q 是这样的。那么，究竟是什么因素影响用户 Q 对帖子 a 的行为呢？可以通过简单的分析得到以下假设：

假设 1 Q 以前曾经访问或回复过 P 所写的帖子，那么这次重复相同行为的可能性较高。

假设 2 帖子 a 的敏感度和热度越高，Q 对其关注的可能性越大。

说明：敏感度和热度不是一个概念。敏感度是一个瞬间的概念，作用时间短，需要将数据在时间上加权，并且敏感度可能与某些特定的话题有关，

如“法轮功”、“藏独”等；热度是在一定时间段内积累的效果，作用时间长，是一种数据的累计。

假设 3 在时间上，发布时间越早的帖子，受关注的可能性越小；更新时间越靠近现在的帖子，受关注的可能性越大。

假设 4 对于曾经发生过大规模传播的话题、以及曾经发布过热门话题的用户，都会在一定程度上影响其他用户对该帖子的行为。

下面，将上述假设与论坛数据结合起来，得到影响一个人对一个帖子行为的传播指数。需要说明的是：这些影响因素并不是独立作用，而是综合起来一起影响话题传播的趋势，如表 4.1 所示。

表 4.1 话题传播指数

传播指数	指数含义	源数据
人际关系指数	主要是基于“人际关系矩阵”，论坛用户与该帖子作者和回复人的关系越紧密，该指数越大，这里的关系指对帖子访问和回复。	帖子作者 访问人列表 回复人列表
敏感度指数	短时间内（一小时）帖子访问数和回复数的累计结果，表示一个话题的敏感程度，实际上是一种单位时间内传播增量的概念。	访问数 回复数 帖子发布时间 当前时间
时间指数	发布时间距现在越久，受关注概率越低 更新时间距现在越近，受关注概率越高	帖子发布时间 帖子更新时间
传播历史指数	需要建立和维护一个“话题传播历史数据库”。存储内容包括：(1)曾经发生大规模传播的话题；（2）曾经发布热门话题用户的 ID；(3）一些人工设定的敏感话题。可以只保存关键字，新话题到来后与关键字进行比较。本数据库可以通过程序自动维护，也可加入人工的干预。	帖子标题 帖子作者 主帖文本 传播历史数据库

在计算人际关系指数的时候，涉及到“人际关系矩阵”的建立和维护。

所谓人际关系矩阵，实际上来源于社会学，是通过矩阵的形式描述群体中人与人之间的关系。本文结合社会学知识，引入人际关系矩阵的目的是要通过矩阵的形式描述论坛中用户之间的回复关系。具体方案如下：假设论坛中共用 100 名用户，为每个用户编制从 1 到 100 的编号，并建立一个横行为发帖者、纵列为回帖者的 100×100 的关系矩阵图。然后，将每一篇帖子中反映出的回复关系用上述编号进行矩阵的坐标表示，例如，1 号用户回复给 2 号的帖子表示为（1，2）。最后，将一个讨论区中所有帖子反映出的回复关系在矩阵图相应的位置标识出来，重复出现多次的坐标点按实际次数累计，没有发生回复关系的记为 0。以论坛中编号从 1 到 4 的四个用户为例，假设他们之间的回复关系为：用户 1 回复用户 3 两篇帖子，用户 2 回复用户 1 两篇帖子、回复用户 3 三篇帖子，用户 3 回复用户 2 一篇帖子、回复用户 4 五篇帖子，用户 4 回复用户 3 一篇帖子。根据上述关系建立的人际关系矩阵如表 4.2 所示。

表 4.2 人际关系矩阵

	1	2	3	4
1	0	0	2	0
2	2	0	3	0
3	0	1	0	5
4	0	0	1	0

4.3.2 变量定义

为了分析和研究论坛内的话题传播行为，本文定义了一系列的变量来描述话题传播行为、预测话题传播趋势。

定义 4.2 访问数（visit_num）和回复数（reply_num），表示特定时刻某一话题的访问数和回复数的累加值。

定义 4.3 传播程度值（diffuse_degree），表示当前时刻该话题在一个论坛内已经达到的传播规模和程度。话题在论坛内主要是通过用户访问和回复

帖子在实现传播的，传播程度值的计算方法见公式（4-1）。

$$diffuse\_degree(t) = visit\_num(t) + REPLY\_RIGHT * reply\_num(t) \quad (4-1)$$

定义 4.4 传播速度值（diffuse_ speed），表示一个话题在单位时间内的传播程度增量，这里的“单位时间”取决于数据采样周期。t 时刻的传播速度是从 t 时刻到下一采样时刻，该话题传播程度的增量值，其取值范围是 0 到正无穷。计算方法见公式（4-2）。

$$diffuse\_speed(t) = \frac{(diffuse\_degree(t+1) - diffuse\_degree(t))}{T} \quad (4-2)$$

定义 4.5 预测传播速度（pro_speed），表示通过预测得出的下一时间段的传播速度。一个话题在下一时间段内的传播速度与当前该话题的各个传播指数相关，也就是说各个传播指数综合影响一个话题未来的传播趋势。同时，同一话题各个传播指数的影响权重不同；不同话题相应传播指数的影响权重也不同。对于时间数据可以取不同的单位，如秒、分、时等。计算方法见公式（4-3）。

$$\begin{aligned} pro\_speed(t) = {} & a[0] * con\_index(t) + a[1] * sen\_index(t) \\ & + a[2] * time\_index(t) + a[3] \end{aligned} \quad (4-3)$$

定义 4.6 预测传播程度（pro_degree），通过当前传播程度和预测传播速度计算，得到下一时刻可能的传播程度值，见公式（4-4）。

$$pro\_degree(t+1) = pro\_degree(t) + pro\_speed(t) * T \quad (4-4)$$

定义 4.7 人际关系指数（con_index），表示该帖子在其讨论区内人际关系上的影响程度，计算方法是：与帖子作者有关的回复关系占整个讨论区回复关系总数的百分比。

定义 4.8 敏感度指数（sen_index），短时间内帖子的访问数和回复数的累计结果，表示一个话题的敏感程度。具体计算方法如下：首先，需要计算话题在某一个时刻的传播增量（increament），见公式（4-5）。

$$increament(t) = diffuse\_degree(t) - diffuse\_degree(t-1) \quad (4-5)$$

假设已知 t-1 时刻的敏感指数，需要计算 t 时刻的敏感指数。比较两个时刻的传播增量，以 t-1 时刻的传播增量为基准，将 t 时刻传播增量的变化程度与 t-1 时刻的敏感指数相加，即可得到 t 时刻的传播敏感指数。具体计算方法见公式（4-6）。

$$\text{sen\_index(t)}=\text{sen\_index(t-1)}+\frac{\text{increament(t)}-\text{increament(t-1)}}{\text{increament(t-1)}} \quad (4\text{-}6)$$

定义 4.9 时间指数（time_index），帖子更新时间与当前时间间隔的倒数。需要将时间统一转换成“年-月-日 时：分：秒”的形式，这样才能计算任意两个时刻间的时间间隔。同时，由于帖子更新时间与当前时间间隔越短，该帖子可能传播的范围越大，所以需要将计算出的时间间隔取倒数，作为时间指数。

4.3.3 多元线性回归分析

在许多实际问题中，常常遇到要研究一个随机变量与多个自变量之间的关系。例如，某公司的管理人员要预测来年该公司的销售额，研究认为影响销售额的因素不只是广告宣传费，还有个人可支配收入、产品价格、科研经费、各种投资和销售费用等。研究这种一个随机变量同其他多个变量之间的关系的主要方法是运用多元线性回归分析。下面对多元线性回归分析方法做简要的介绍。

设影响因变量 Y 的自变量个数为 P，并分别记为 $x_1,x_2,\cdots,x_{p,}$。多元线性模型是指这些自变量对 Y 的影响是线性的，即

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon,\ \varepsilon\sim N(0,\sigma^2) \quad (4\text{-}7)$$

其中 $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$，$\sigma^2$ 是与 $x_1,x_2,\cdots,x_p$ 无关的未知参数，称 Y 为对自变量 $x_1,x_2,\cdots,x_{p,}$ 的线性回归函数。

记 n 组样本分别是 $(x_{i1},x_{i2},\cdots,x_{ip},y_i)\ (i=1,2,\cdots,n)$，则有

$$\begin{cases} y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_px_{1p}+\varepsilon_1 \\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_px_{2p}+\varepsilon_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_px_{np}+\varepsilon_n \end{cases} \quad (4\text{-}8)$$

其中 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 相互独立，且 $\varepsilon_i\sim N(0,\sigma^2)$，$i=1,2,\cdots,n$，这个模型称为多元线性回归的数学模型。令

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\quad X=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},\quad \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},\quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{4-9}$$

则上述数学模型可用矩阵形式表示为

$$Y=X\beta+\varepsilon \tag{4-10}$$

其中ε是 n 维随机向量，它的分量相互独立。应用最小二乘法估计参数$\beta_0,\beta_1,\beta_2,\cdots,\beta_p$，解得$\hat{\beta}$就是$\beta$的最小二乘估计，即$\hat{\beta}$为回归方程。

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p \tag{4-11}$$

本文应用这种多元线性回归分析方法，通过对样本数据的回归分析，得到各个传播指数与传播速度之间的关系，并最终对传播速度和传播程度值进行了预测。

4.3.4 话题传播趋势预测

本文在提炼话题传播指数的基础上，结合上述多元线性回归分析方法，提出了一种预测话题传播速度和传播程度的实现方案。下面，针对一篇帖子，用 t1 到 t6 六个时刻的数据来预测 t7 时刻的传播趋势，通过这个过程来说明论坛内话题传播分析的整理流程。图 4.3 是算法整体流程图。

步骤一：录入样本数据：收集 t1 到 t6 的访问数和回复数，使用公式（4-1）计算 t1 到 t6 各个时刻的传播程度值(diffuse_degree)。根据公式（4-2）计算 t1 到 t5 各个时刻的传播速度值(diffuse_ speed)。根据传播指数的定义，分别计算 t1 到 t6 各个时刻的传播指数。

步骤二：将 t1 到 t5 各个时刻的传播指数和实际传播速度代入多元线性回归方程，得到公式（4-3）中的 a[0]、a[1]、a[2]和 a[3]，从而确定了由传播指数计算传播速度的公式。将 t6 时刻的各个传播指数代入刚刚确定的公式，求得 t6 时刻的传播速度 pro_speed(t6)。这个速度实际上是预测得出的结果。将 t6 时刻的预测传播速度代入公式（4-4），即可得到 t7 时刻的传播程度值 pre_degree(t7)，从而实现对传播趋势的预测，如表 4.3 所示。以此类推，可以预测出 t8、t9、t10 等未来时刻的传播程度值。

表 4.3 论坛内话题传播分析过程

T	人际指数	敏感指数	时间指数	传播速度	传播程度
t1	con(t1)	sen(t1)	time(t1)	diffuse_speed(t1)	diffuse_degree(t1)
t2	con(t2)	sen(t2)	time(t2)	diffuse_speed(t2)	diffuse_degree(t2)
t3	con(t3)	sen(t3)	time(t3)	diffuse_speed(t3)	diffuse_degree(t3)
t4	con(t4)	sen(t4)	time(t4)	diffuse_speed(t4)	diffuse_degree(t4)
t5	con(t5)	sen(t5)	time(t5)	diffuse_speed(t5)	diffuse_degree(t5)
t6	con(t6)	sen(t6)	time(t6)	pre_speed(t6)	diffuse_degree(t6)
t7	-	-	-	-	pre_dgree(t7)

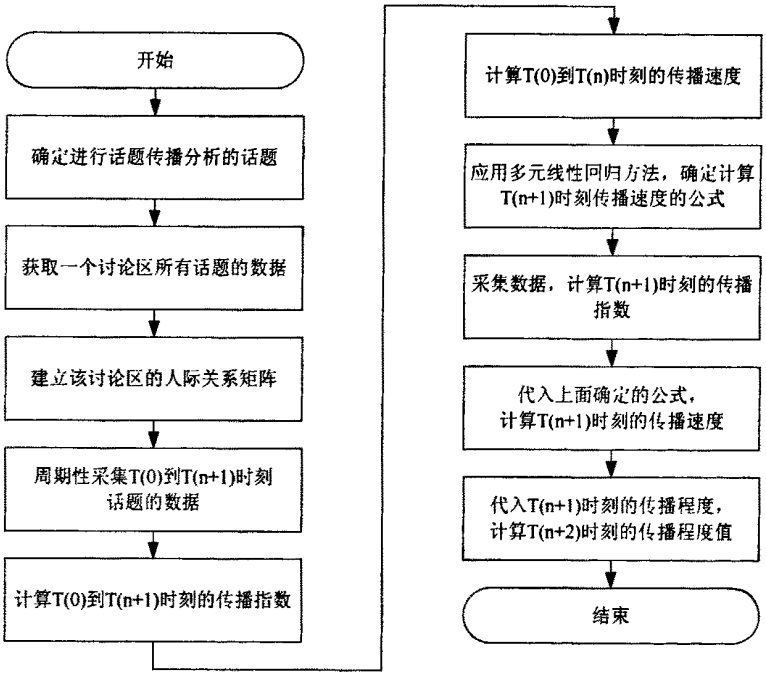

图 4.3 论坛内话传播分析算法流程图

本文提出的预测传播趋势的方案是一个逐步叠代的过程。在预测的初始阶段，样本数据较少；随着时间的推移，样本数据随之增加，预测结果将越来越准确。

4.4 实验结果与分析

本实验的目的是验证上述策略是否能够有效描述话题传播过程、解决话题传播问题。同时，根据实验结果，对已有的方案进行优化和调整，使其具备更大的理论研究和实际应用价值。本实验主要分为两大部分：实验一是基于传播图的论坛间话题传播分析技术实验，实验二是基于多元线性回归的论坛内话题传播分析技术实验。下面具体介绍两个实验的实验过程，以及对实验结果的分析和总结。

实验一：首先需要采集实验数据。实验程序从若干文件中读取一系列论坛帖子的文本，这些帖子都是关于一个特定话题的，并且属于不同的论坛。本实验将话题“中国政法大学男生课堂上砍死副教授”作为分析对象，抓取49篇关于该话题的帖子，这些帖子分别来自新浪、网易、天涯等7个国内知名论坛，每个论坛抓取7篇帖子。然后，根据发布论坛的不同将帖子分组。接着，计算每个论坛的传播生命期。同时，解析论坛内转帖关系，并发现初始传播论坛。这样，根据以上信息就可以构建该话题的传播图。最后，还要根据该图计算每个论坛的传播权值。

实验二：第一步也是采集实验数据。周期性采集帖子的数据写入数据库，数据采样周期取1小时。然后，使用公式（4-1）和（4-2）计算各个时间点的传播程度和传播速度。同时，计算该话题每个数据采样点的传播指数，具体包括：敏感度指数、人际关系指数和时间指数。将各个时刻的传播指数存入数据库中相应的数据项中。接着，应用多元线性回归理论，将若干样本数据输入到多元线性回归方程中。需要说明的是，由于本文定义了三个传播指数，那么至少需要三组样本数据才能得出回归结果。本实验在具体实现时，第一次回归分析使用了五组样本数据。然后，将下一时刻的传播指数代入多元线性回归模型确定的公式，得到传播速度的预测值。最后，将传播速度的预测值代入公式（4-4），得到传播程度的预测值。这样，就完成了一次对话题传播趋势的预测。

下面，对本实验需要获取的数据进行总结。一个完整的论坛根据关注内容不同分为多个讨论区，每个讨论区负责发布和管理网友编辑的文章——俗称“帖子”。应该说，帖子是论坛中话题传播的基本单位，是论坛运作的基本

要素。一篇帖子包含大量信息：帖子的主题和文本反映其关注的话题内容；帖子的作者和回复人体现论坛中用户之间的联系；发布时间和更新时间反映帖子在时序上的行为特征。通过总结帖子各项相关数据，本实验需要采集下列数据，如表 4.4 所示。

表 4.4 论坛内话题传播实验数据

数据类型	数据源
文本数据	标题文本(topic_text)
	主帖文本(main_text)
	跟帖文本(reply_text_list)
人际关系数据	作者(author_name)
	回复者队列(reply_name)
	访问者队列（一般论坛不提供）
受关注度数据	访问数（visit_num）
	回复数（reply_num）
时间数据	发布时间（issue_time）
	更新时间（update_time）
	当前时间（current_time）
地域属性数据	发布论坛站点关键字（BBS_key）

其中，时间数据用字符串格式存储，统一转化成“年-月-日 时-分-秒”的格式，如 2008-10-29 6:51:00。论坛关键字是用一个字符串来标识该帖子发布在哪个论坛上，如“tianya”表示天涯论坛。下面分别给出了两个实验的实验结果。

实验一：论坛间话题传播分析实验

（1）解析论坛站点信息：

用“帖子 ID”标识每一篇帖子，用“论坛关键词”标识不同的论坛。本实验总共收集了 7 个论坛的帖子，每个论坛分别收集 7 篇帖子，如表 4.5 所示。

表 4.5 论坛站点信息

论坛 ID	论坛名称	论坛关键词	帖子 ID	帖子个数
0	新浪论坛	sina	0~6	7
1	网易论坛	163	7~13	7
2	中华网论坛	zhonghua	14~20	7
3	天涯社区	tianya	21~27	7
4	新华网论坛	xinhua	28~34	7
5	人民网强国社区	renmin	35~41	7
6	雅虎论坛	yahoo	42~48	7

（2）解析论坛话题传播时间：

表 4.6 论坛时间信息

论坛 ID	论坛名称	话题初始讨论时间	话题最新更新时间
0	新浪论坛	2008-10-29 6:45:0	2008-11-1 18:23:2
1	网易论坛	2008-10-29 9:45:2	2008-11-1 19:05:2
2	中华网论坛	2008-10-29 10:18:2	2008-11-3 1:57:2
3	天涯社区	2008-10-29 9:45:2	2008-11-2 19:13:2
4	新华网论坛	2008-10-28 23:20:2	2008-11-2 18:5:2
5	人民网强国社区	2008-10-29 11:14:2	2008-11-2 5:14:2
6	雅虎论坛	2008-10-29 13:58:2	2008-11-2 21:43:2

（3）解析初始传播论坛：

初始传播论坛 ID 为 4，论坛关键词为 xinhua，即“新华网论坛”；初始讨论时间为 2008 年 10 月 28 日 23:20:2，最新更新时间为 2008 年 11 月 2 日 18:5:2。该帖子就是本话题传播的起点。

（4）解析转帖关系：

其中，帖子 2 是由帖子 1 转发而来。“sina-0”表示新浪论坛中帖子 ID 为 0 的帖子。如表 4.7 所示。

表 4.7 论坛间转帖关系

转帖关系 ID	帖子 1（论坛关键词-帖子 ID）	帖子 2（论坛关键词-帖子 ID）
0	sina-0	xinhua-31
1	163-7	zhonghua-17
2	tianya-21	163-12
3	xinhua-31	zhonghua-14
4	yahoo-45	zhonghua-16
5	xinhua-31	renmin-35
6	xinhua-31	yahoo-43

（5）计算每个论坛的传播权值，如表 4.8 所示。

表 4.8 论坛传播权值

论坛 ID	论坛名称	传播图中相应的出度值	归一化后的传播权值
0	新浪论坛	1	0.366667
1	网易论坛	1	0.366667
2	中华网论坛	0	0.100000
3	天涯社区	1	0.366667
4	新华网论坛	3	0.900000
5	人民网强国社区	0	0.100000
6	雅虎论坛	1	0.366667

本文采用的数据归一化方法如公式（4-12）所示。该方法与传统归一化方法相比最大的特点是：结果的最大值不为 1、最小值不为 0，从而便于将该归一化结果图形化表示。

$$X = 0.1 + 0.8(X - MIN)/(MAX - MIN) \qquad (4\text{-}12)$$

（6）建立论坛话题传播图：

通过该传播图，我们能够得到以下信息：首先，基于实验所采集的数据，话题“中国政法大学男生课堂上砍死副教授”最先出现在新华网论坛上；同时，传播图展现出的传播路径为：有一篇帖子由新浪网转发到新华网，然后

又由新华网论坛分别转发到中华网论坛、人民网强国论坛以及雅虎论坛，还有三篇不同的帖子完成了转发，分别是天涯到网易、网易到中华网、以及雅虎到中华网。综上，该传播图分别从行为、站点、时间等多角度描述了一个话题的传播过程，在一定程度上解决了话题传播问题。如图 4.4 示。

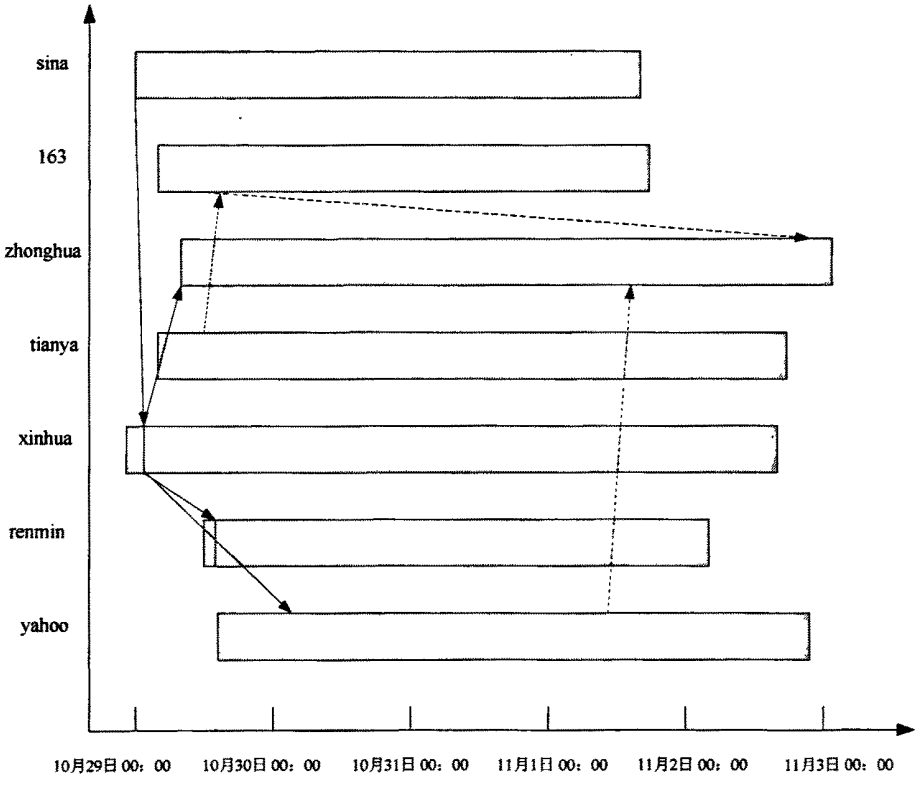

图 4.4 论坛话题传播图

这里传播权值的计算主要是基于上面生成的传播图。本文提出的策略是：在传播图中结点的出度越大，即向外转发的帖子越多，其对应的传播权重就越大。分析实验结果，新华网论坛在传播图中的出度最大，是 3，则其传播权重也最大；中华网和人民网的出度是 0，所以它们的传播权重最小。通过传播权重的计算，比较和评价了各个论坛对特定话题传播所做的贡献。特别是当某敏感甚至反动话题在论坛间传播的时候，为政府相关部门的监管工作提供了技术保障。

实验二：论坛内话题传播分析实验

（1）计算传播指数：

计算各个时刻的传播指数，具体包括人际关系指数、敏感度指数和时间指数，并录入到数据库中。根据敏感度的定义，第一个数据采样点的敏感度值无法计算。如表 4.4 所示。

表 4.4 话题传播指数

时序序号	人际关系指数	敏感度指数	时间指数
1	0.013222	-	0.333333
2	0.013454	0	0.5
3	0.013852	0.116	1
4	0.013514	0.055581	0.25
5	0.013552	-0.04814	0.25
6	0.013468	-0.013147	0.1
7	0.013465	-0.026315	0.0625
8	0.013545	-0.041015	0.066667
9	0.012158	0.136882	0.066667
10	0.013695	-0.048217	0.022222
11	0.011954	0.004107	0.022222
12	0.011853	-0.043852	0.013333
13	0.011712	0.030909	0.020408
14	0.011568	-0.038213	0.012658
15	0.011019	-0.007525	0.012987

（2）预测传播趋势：

计算实际的传播速度和传播程度值，并应用多元线性回归分析方法，对传播速度和传播程度进行预测。实验预测了 t7 到 t14 时刻的传播速度值，以及 t8 到 t15 时刻的传播程度值。如表 4.5 所示。

表 4.5 预测传播速度和传播程度值

时序序号	实际传播速度值	预测传播速度值	实际传播程度值	预测传播程度值
1	1.666667	-	2588	-
2	3.6	-	2688	-
3	5.183333	-	2904	-
4	2.4	-	3215	-
5	2.2	-	3359	-
6	1.65	-	3491	-
7	1.016667	1.51437	3590	-
8	2.45	1.135405	3651	3680.862213
9	0.933333	0.701263	3798	3719.124325
10	1.016667	1.603606	3854	3840.075809
11	0.566667	1.154615	3915	3950.216333
12	0.766667	0.6474	3949	3984.276929
13	0.45	0.618067	3995	3987.843999
14	0.433333	0.540814	4022	4032.084019
15	-	-	4048	4054.448854

（3）回归效果检验：

本文在分析论坛内话题传播机制的基础上，定义了一套传播指数，并应用多元线性回归方法，对一个话题未来的传播趋势进行预测。这里需要对多元线性回归的效果进行检验。本实验采用基本的拟合优度检验：拟合优度检验是检验回归方程对样本观测值的拟合程度，即检验所有解释变量与被解释变量之间的相关程度。检验的方法是构造一个可以表征拟合程度的指标，这个指标是通过对总离差的分解而得到。总离差平方和（TSS）是各个观察值与样本均值之差的平方和，反映了全部数据之间的差异；残差平方和（ESS）是总变差平方和中未被回归方程解释的部分，由解释变量 X_1 、X_2 、X_3 、X_4 … …X_k 中未包含的一切因素对被解释变量 Y 的影响而造成的。一个拟合较好的回归模型体现在总体平方和与回归平方和的接近程度，即 TSS 中 ESS

越小越好。所以定义复相关系数 R 来衡量拟合优度，检验回归的效果。具体计算公式如下[36]。

残差平方和：
$$ESS = \sum (Y_i - \hat{Y}_i)^2 \tag{4-13}$$

总离差平方和：
$$TSS = \sum (Y_i - \bar{Y}_i)^2 \tag{4-14}$$

复相关系数：
$$R = \sqrt{1 - ESS/TSS}\,(0 \le R \le 1) \tag{4-15}$$

表 4.6 列出了每次回归的残差平方和、总离差平方和以及复相关系数，每次回归的复相关系数都基本接近于 1，可见相对误差接近于 0，回归方程的拟合优度较高，回归效果较好。

表 4.6 回归效果检验

序号	ESS	TSS	R
1	3.834209e-03	7.949221e+00	9.997588e-01
2	1.677242e-01	1.124930e+01	9.925171e-01
3	1.109553e+00	1.129270e+01	9.496030e-01
4	1.110475e+00	1.384986e+01	9.590727e-01
5	1.349907e+00	1.562333e+01	9.558225e-01
6	1.425261e+00	1.824136e+01	9.601388e-01
7	1.431176e+00	1.986156e+01	9.632977e-01
8	1.450522e+00	2.200824e+01	9.664843e-01

（4）预测结果分析。

最后，分别针对传播速度和传播程度值，用曲线图的形式对比预测数据和实际数据。如图 4.5 和图 4.6 所示，图中实线是实际数据，虚线是预测数据。图 4.5 展现了本实验对话题传播速度的预测效果，数据采样点 8 和 10 处的误差较大，其余各点的预测较为准确，预测曲线在整体上基本符合真实数据的发展趋势。图 4.6 展示了对话题传播程度的预测效果，很明显预测数据曲线与实际数据曲线基本重合，预测结果比较准确。分析原因，由于预测传播程度值是通过上一时刻真实程度值和预测速度值计算得到的，这里真实值的引入必然会提高预测的准确性；而对传播速度的预测则要依赖于多元线性回归

的效果，以及当前时刻各个传播指数的计算，所以准确度相对会低一些。

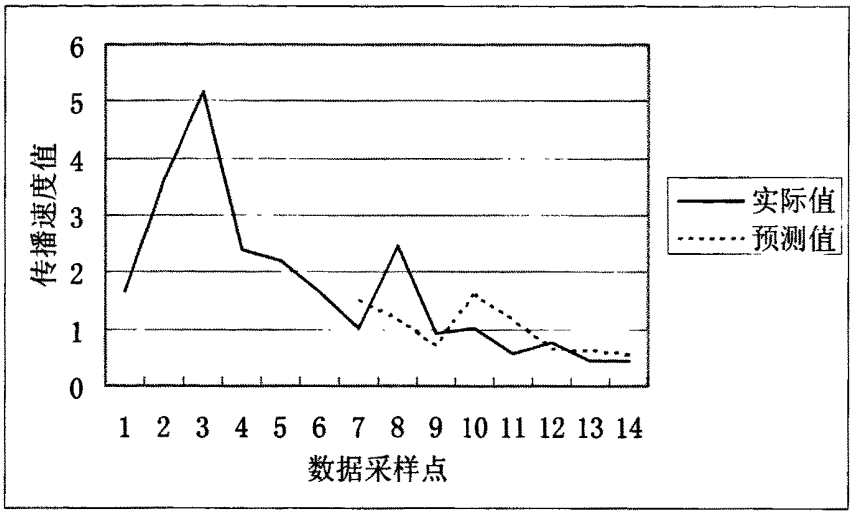

图 4.5 话题传播速度图

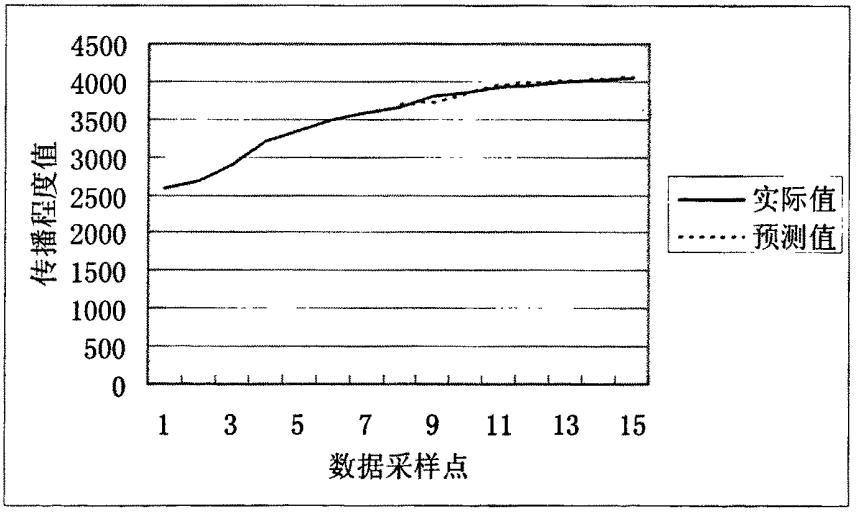

图 4.6 话题传播程度图

综上所述，本文提出了分析论坛内话题传播趋势的方法，基于实时获取的论坛数据，提出一套影响传播行为的指标体系，定义了一系列变量和公式，并应用多元线性回归分析方法，对话题传播速度和传播程度进行了分析和预

测。除个别数据误差较大外，预测结果基本能够反映话题未来的传播趋势，在一定程度上解决了论坛内的话题传播问题。

4.5 本章小结

本章提出基于传播图和多元线性回归的话题传播分析技术，力图发掘论坛间和论坛内的话题传播机制、分析话题传播行为。

在论坛间，提出基于相似度比较和关键词匹配的转帖关系发现技术，并结合传播权值的计算，构建论坛话题传播图；在论坛内，提炼影响传播行为的指标体系，并结合多元线性回归分析理论，提出一套完整的预测话题传播趋势的实施方案。

最后，通过实验对上述方案的可行性和实现效果进行检验。

结 论

本文在研究和总结现有网络舆情分析技术的基础上，重点针对话题演变和传播现象，提出基于多中心和向量分解的话题演变分析技术以及基于传播图和多元线性回归的话题传播分析技术。具体的工作成果和创新点有以下几个方面：

1、在话题多中心模型的基础上，提出向量分解思想。将文档向量分解成共有子向量和相异子向量，提取后续文档的新颖特征，并根据新特征的比重建立和更新话题中心，从而呈现话题演变的全过程。

2、为了发掘话题传播路径、建立论坛话题传播图，提出基于相似度比较和关键词匹配的转帖关系发现技术。该策略是用基于内容的分析方法解决基于行为的问题，能够有效发现帖子之间的转发关系，为构建论坛话题传播图奠定了基础。

3、在对论坛内话题传播的研究中，结合社会学知识，提炼出一套影响传播行为的指标体系，并结合多元线性回归分析理论，实现对话题传播趋势的预测。该方案着重提出了趋势预测的概念，为网络舆情监管工作提供了有力的技术支持。

在网络舆情日益严峻的背景下，本文重点对话题演变和传播分析技术进行了研究。该领域未来的发展趋势如下：在话题演变方面，可以考虑将统计学策略和语义分析结合起来；在话题传播方面，力图在完善现有传播指标体系的同时，通过大量实验工作的验证，寻找一种更为有效的方法来分析论坛内的话题传播行为。

参考文献

[1] 洪宇，张宇，刘挺，李生. 话题检测与跟踪的评测及研究综述. 中文信息学报. 2007，21(6):71-84 页

[2] Charles L. Wayne. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. In proceedings of the 2nd International Conference on Language Resources & Evaluation (LREC 2000)，2000:1487-1494P

[3] 王会珍，张希娟，朱靖波，张斌. 基于主动学习的自适应话题追踪 2006. 中国中文信息学会二十五周年学术会议论文集. 2006:373-381 页

[4] 王会珍，朱靖波，季铎，叶娜，张斌. 基于反馈学习自适应的中文话题追踪. 中文信息学报. 2006，20(3):92-98 页

[5] 黄萱菁，夏迎炬，吴立德. 基于向量空间的文本过滤. 软件学报. 2003，14(3):435-442 页

[6] 谭应伟，莫倩. 基于 Web 的有监督自适应话题追踪系统的设计与实现. 郑州大学学报. 2007，39(2):25-29 页

[7] 莫倩，刘书家，李凯. 主题追踪系统的研究与实现. 计算机工程与应用. 2006，02(179):179-181 页

[8] 洪宇，张宇，刘挺，郑伟，龚诚，李生. 基于层次聚类的自适应信息过滤学习算法. 中文信息学报. 2007，21(3):47-52 页

[9] Juha Makkonen. Investigations on event evolution in TDT. Proceedings of HLT-NAACL. 2003:43-48P

[10] 宋丹，王卫东，陈英. 基于改进向量空间模型的话题识别与跟踪. 计算机技术与发展. 2006，16(9):63-67 页

[11] 李昕，朱永盛，武港山. 论坛消息的语义漂移分析. 计算机工程. 2006，32(4):88-93 页

[12] 赵华，赵铁军，张姝，王浩畅. 基于内容分析的话题检测研究. 哈尔滨工业大学学报. 2006，38(10):1740-1743 页

[13] 王会珍，朱靖波，季铎，张斌. 基于多向量模型的中文话题追踪. 全国第八届计算语言学联合学术会议论文集. 2005:669-671 页

[14] 王会珍. 面向话题追踪的特征选取与文本表示技术的研究. 东北大学硕士学位论坛. 2004:39-44 页

[15] 赵华，赵铁军，于浩，张姝. 面向动态演化的话题检测研究. 高技术通信. 2006，16(12):1230-1235 页

[16] 贾自艳，何清，张海俊，李嘉佑，史忠植. 一种基于动态进化模型的事件探测和追踪算法. 计算机研究与发展. 2004，41(7):1273-1280 页

[17] 贾自艳. Web 信息智能获取若干关键问题研究. 中国科学院研究生院博士学位论文. 2004:83-97 页

[18] Ramesh Nallapati，Ao Feng，Fuchun Peng，James Allan. event threading within news topic. information retrieval and knowledge management. 2004:446-453P

[19] 金珠，林鸿飞，赵晶. 基于 HowNet 的话题跟踪及倾向性分类研究. 情报学报. 2005，24(5):555-561 页

[20] 金珠. 基于知网的话题跟踪和倾向性跟踪研究. 大连理工大学硕士学位论文. 2005:26-47 页

[21] 吴平博，陈群秀，马亮. 基于事件框架的事件相关文档的智能检索研究. 中文信息学报. 2003，17(6):25-31 页

[22] 林鸿飞，宋丹，杨志豪. 基于语义框架的话题跟踪方法. 中国中文信息学会二十五周年学术会议论文集. 2006:383-392 页

[23] Xiaojun Wan，Jianwu Yang . Learning Information Diffusion Process on the Web. WWW 2007 Poster Paper . 2007:1173-1174P

[24] Avaré Stewart，Ling Chen，Raluca Paiu，Wolfgang Nejdl. Discovering Information Diffusion Paths from Blogosphere for Online Advertising. ADKDD`07. 2007:46-53P

[25] 宫辉，徐渝. 高校 BBS 社群结构与信息传播的影响因素. 西安交通大学学报. 2007，21(1):93-96 页

[26] 白淑英，何明升. BBS 互动的结构与过程. 社会学研究. 2003，5:8-18 页

[27] 于静，赵燕平. 基于社会网络分析的 BBS 内容安全动态监测模型. 北京

理工大学学报. 2006，26(1):319-328 页

[28] 张嘉龄，李茂青. 博客信息传播的网络模型构建. 软件导刊. 2008，7(5):67-69 页

[29] 刘常昱，胡晓峰，司光亚，罗批. 基于小世界网络的舆论传播模型研究. 系统仿真学报. 2006，18(12):3608-3610 页

[30] Naohiro Matsumura, David E. Goldberg, Xavier Llor. Mining Social Networks in Message Boards. Illinois Genetic Algorithms Laboratory Department of General Engineering University of Illinois at Urbana-Champaign. 2005:1-12P

[31] Naohiro Matsumura, David E. Goldberg, Xavier Llor. Mining Directed Social Network from Message Board. International World Wide Web Conference. 2005: 1092-1093 P

[32] Naohiro Matsumura. Modeling Influence Diffusion in Human Society. Graduate School of Economics, Osaka University. 2006: 137-153P

[33] Naohiro Matsumura, Yukio Ohsawa, Mitsuru Ishizuka. Discovery of Emerging Topics between Communities on WWW. Proceedings of the First Asia-Pacific Conference on Web Intelligence. 2001:473-482P

[34] Naohiro Matsumura, Yukio Ohsawa, Mitsuru Ishizuka. Future Directions of Communities on the Web. Lecture notes in computer science. 2001: 435-443P

[35] 钱斌. 餐饮类论坛中口碑再传播现象的实证研究与仿真模拟. 浙江大学硕士学位论文. 2008:35-93 页

[36] 刘京娟. 多元线性回归模型检验方法. 湖南税务高等专科学校学报. 2005，5(18):48-59 页

攻读硕士学位期间发表的论文和取得的科研成果

[1] 郑希文，杨武，王巍. 基于传播图的论坛间话题传播分析技术研究. 已投稿

致 谢

在论文完成之际，衷心感谢所有关心、帮助过我的老师、同学、朋友和亲人。

首先，感谢我最尊敬的导师杨武教授、张志强教授。在课题研究和论文撰写阶段，两位老师给予我很大的帮助和指导。他们渊博的学识和敏捷的思维令人钦佩，严谨求实的治学态度和孜孜不倦的工作精神令我感动。两位老师对待工作、对待研究的态度将是我一生的榜样。

我还要特别感谢哈尔滨工程大学信息安全研究中心的王巍副教授、郑军博士和苘大鹏博士，论文自始至终都渗透着实验室各位老师的心血，感谢各位老师在工作和学习方面给予的巨大帮助和支持。实验室两年半的学习生活使我获益良多，在此衷心的表示感谢！

感谢舆情监管项目组的赵慧杰、康喜、吴昊、董红臣和严俊同学，还有已经毕业的齐海凤师姐，在大家的共同努力下，我的论文才得以顺利完成。

感谢信息安全研究中心的孙敏、于骁丹、王凯琢、韩利辉、敖日格勒、张斌、刘炜、张强和周洋同学，平日里与大家的交流使我受益颇多。

感谢我的父母！他们在毕业设计期间给予我无微不至的关怀和爱护，让我时刻信心百倍的迎接新的挑战，勤劳朴实的父亲母亲是我永远的骄傲！

最后，谨向百忙之中审阅本文的老师们表示衷心感谢！